

# Proceedings of the XVI EURALEX International Congress:

## The User in Focus

15-19 July 2014, Bolzano/Bozen



## **Acknowledgements**

We would like to thank all those who have made the XVI EURALEX International Congress possible, by contributing to the reviewing, to the logistics and by financially supporting the event. In particular, we would like to thank our sponsoring partners and patrons:

A.S. Hornby Educational Trust  
Accademia della Crusca (scientific patronage)  
Autonomous Province of Bolzano/Bozen (patronage)  
Consortium Südtiroler Speck Alto Adige  
Dr. Schär AG/SPA  
Istitut Ladin Micurà de Rù  
k Dictionaries  
Municipality of Bolzano/Bozen (patronage)  
Oxford University Press  
SDL  
Sketch Engine  
Sprachstelle im Südtiroler Kulturinstitut  
Südtirol Marketing (SMG)  
Verband der Südtiroler Obstgenossenschaften (VOG)

## **Programme Committee**

Andrea Abel (EURAC)  
Janet DeCesaris (Universitat Pompeu Fabra)  
Tinatin Margalidze (Ivane Javakishvili Tbilisi State University - Next EURALEX Congress Organizer)  
Natascia Ralli (EURAC)  
Marcello Soffritti (Università di Bologna/EURAC)  
Pius ten Hacken (Universität Innsbruck)  
Ruth Vatvedt Fjeld (Universitetet i Oslo)  
Chiara Vettori (EURAC)

## **Reviewers (with number of reviewed papers)**

Andrea Abel (EURAC, 18), Hauke Bartels (Serbski Institut, 5), Maria Paz Battaner Arias (Catedrática jubilada - Universitat Pompeu Fabra, 10), Patrizia CORDIN (Università degli Studi di Trento, 8), Janet DeCesaris (Universitat Pompeu Fabra, 19), Anne Dykstra (Fryske Akademy, 15), Ruth Vatvedt Fjeld (Universitetet i Oslo, 13), Thierry Fontenelle (Translation Centre for the Bodies of the European Union, 5), Alexander Geyken (Berlin-Brandenburgische Akademie der Wissenschaften, 7), Rufus Gouws (Universiteit Stellenbosch University, 7), Sylviane Granger (Université catholique de Louvain, 3), Reinhard Hartmann (University of Exeter, emeritus, 5), Ulrich Heid (Universität Hildesheim, 6), Adam Kilgariff (Lexical Computing Ltd, 7), Annette Klosa (Institut für Deutsche Sprache, Mannheim, 8), Simon Krek (Jozef Stefan Institute & Amebis, 7), Iztok Kosem (Trojina, Institute for Applied Slovene Studies, 14), Lothar Lemnitzer (Berlin-Brandenburgische Akademie der Wissenschaften, 8), Robert Lew (Uniwersytet im. Adama Mickiewicza w Poznaniu, 7), Carla Marelllo (Università di Torino, 8), Tinatin Margalidze (Ivane Javakishvili Tbilisi State University, 11), Carolin Müller-Spitzer (Institut für Deutsche Sprache, Mannheim, 8), Martina Nied Curcio (Università degli Studi Roma Tre, 7), Magali Paquot (Université catholique de Louvain, 8), Natascia Ralli (EURAC, 21), Michael Rundell (Lexicography MasterClass, 9), Felix San Vicente (Università di Bologna, 9), Marcello Soffritti (Università di Bologna/EURAC, 14), Joan Soler Bou (Institut d'Estudis Catalans, 9), Pius ten Hacken (Universität Innsbruck, 19), Carole Tiberius (Instituut voor Nederlandse Lexicologie, 7), Lars Trap Jensen (Det Danske Sprog- og Litteraturselskab, 16), Serge Verlinde (KU Leuven, 7), Chiara Vettori (EURAC, 22), Geoffrey Williams (Université de Bretagne Sud, 16).





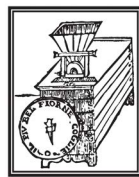
**Proceedings of the XVI EURALEX  
International Congress:  
The User in Focus**  
15-19 July 2014, Bolzano/Bozen

Edited by Andrea Abel, Chiara Vettori, Natascia Ralli

**EURAC**  
research







ACCADEMIA DELLA CRUSCA





Institute for Specialised Communication and Multilingualism  
Viale Druso/Drususallee 1, 39100 Bolzano/Bozen  
Tel. +39 0471 055100  
Fax +39 0471 055199  
[communication.multilingualism@eurac.edu](mailto:communication.multilingualism@eurac.edu)  
[www.eurac.edu](http://www.eurac.edu)

The digital copy of the Proceedings of the XVI EURALEX International Congress  
can be downloaded at the following URL  
<http://www.eurac.edu/en/research/institutes/multilingualism/Publications>

© Bolzano/Bozen, 2014

Director: Stephan Ortner  
Prepress: Pluristamp  
Press: EURAC research

ISBN: 978-88-88906-97-3



## Foreword

The XVI edition of the EURALEX International Congress took place in Italy at the European Academy of Bolzano/Bozen (EURAC) from 15 to 19 July 2014. On behalf of the organising committee, I am extremely pleased to present this volume of proceedings. A novelty of this latest EURALEX edition is our tribute to the digital age: the proceedings collect selected contributions presented at the congress and are published exclusively in digital format on the Internet.

EURALEX congresses take place every second year and usually attract a large international audience. This was true also for the XVI edition. You will find that the authors contained in this volume come from all over the world and have drafted their papers in several languages. Even though English remains the most popular, we have received contributions written in other major European languages, such as French, German, Italian and Spanish. The number of languages that the authors treat within their papers is even higher and goes well beyond the frontiers of the European continent, covering a wide range of languages.

Being a European organisation devoted to lexicography and related fields, EURALEX is very sensitive to issues of language diversity. Regional and minority languages on the one hand and large international languages on the other hand therefore receive equal consideration. In this respect, South Tyrol (called *Alto Adige* in Italian and *Südtirol* in German) represents the ideal location for an edition of the EURALEX congress. The region being officially trilingual, South Tyrol is a perfect example of language diversity. Italian and German are both official languages of the province of Bolzano/Bozen, but also the local minority language Ladin has gained official recognition. Moreover, South Tyrol is becoming growingly multilingual and multicultural, thus adding new languages from all over the world to an already quite colourful linguistic landscape.

EURAC's [Institute for Specialised Communication and Multilingualism](#) (ISCM) was particularly honoured to host the XVI EURALEX International Congress. The main aim of our institute's research is to provide scientific answers to current issues of language and education policy as well as to economic and social questions. Our activities at local and international level comprise applied research (research projects and networking), training and consulting (consulting services, monitoring, seminars) as well as dissemination (scientific publications, dictionaries, databases, corpora) in three main research areas: bilingualism and multilingualism, specialised communication and language technologies. Our interests range from general to special languages, from old to new minorities and from language policy to language planning. Our research activities also centre on the observation of language usage, language documentation, language consulting, multilingual knowledge management and the management of linguistic and cultural diversity. To support our research initiatives, we also create and develop dedicated research infrastructures.

Against this background, at the ISCM we have developed long-term experience and in-depth expertise into different types of lexicographic activities concerning both general language and special language. We particularly focus on terminology and electronic learner lexicography, but have also pro-

duced reference works for the Italian Sign Language and Ladin terminology. A pioneering feature of our work consists in applying new technologies to lexicography, as in the Information system for legal terminology [bistro](#), the electronic learner dictionary for German and Italian [eldit](#) and the first bilingual electronic dictionary Italian Sign Language - Italian [e-LIS](#). To this end, we have created and developed resources to support lexicography and terminography, such as online corpora (e.g. [Korpus Südtirol](#) and [PAISÀ](#)) and tools for visualising linguistic data ([LinfoVis](#)).

Our researchers at the ISCM are active in large international projects and networks, such as the academic DFG-network “[Internet Lexicography](#)” and the COST action “[European Network of e-Lexicography](#)”. They are members of various international organisations like the Council for German language terminology [RaDT](#) and the International Information Centre for Terminology [Infoterm](#). Given such fitting background, the EURAC Institute for Specialised Communication and Multilingualism seemed particularly appropriate as a host and organiser of a EURALEX International Congress.

The 2014 edition was driven by the motto “the user in focus”, since one of the challenges of lexicographic products is to respond to specific and diverse user needs. Therefore, we particularly welcomed submissions focusing on the user perspective at all levels of dictionary production and consultation (e.g. experts vs. laymen as users and/or as producers, professional lexicographers as users of dedicated lexicographic tools, user needs before and during dictionary consultation, considering user needs during the whole lexicographic process, groups with specific needs). Nevertheless, we did not wish to exclude any areas of potential interest for lexicography, intending to cover – as for all preceding editions of the EURALEX congress – a large and varied array of relevant and current themes, which you consequently find in the present proceedings.

This volume contains the four plenary lectures given at the congress. The plenaries represent significant contributions by well-known experts in their fields. Thierry Fontenelle of the Translation Centre for the Bodies of the European Union in Luxembourg argued for considering the space between lexicography and terminology as a continuum rather than a hard-and-fast dichotomy. Ulrich Heid of the Universität Hildesheim analysed natural language processing techniques with respect to their impact on the user friendliness of electronic dictionaries. Carla Marellò of the Università di Torino discussed whether the reference skills of native digital EFL students are developed well enough for them to take advantage of large bilingual dictionaries on their smartphones. Rosamund Moon of the University of Birmingham explored ideological meanings in learner dictionaries with a particular focus on age and ageism. As in past EURALEX editions, the Hornby Trust generously sponsored one of the plenary lectures (R. Moon) in honour of A.S. Hornby, a pioneering figure in learners’ dictionaries for non-native speakers. The organising committee would like to extend its sincere thanks to all plenary speakers for setting the tone of the congress and of this volume, which we believe represents a significant contribution to the literature of dictionary making theory and practice. Also, we were particularly glad that they all most excellently considered and discussed the motto of the XVI edition.



Submissions to EURALEX 2014 were filed under several categories: full papers, short papers, posters and software demonstrations. The contributions in these proceedings are grouped according to the following sections, which set the frame for the call-for-papers:

- The Dictionary-Making Process
- Research on Dictionary Use
- Lexicography and Language Technologies
- Lexicography and Corpus Linguistics
- Bi- and Multilingual Lexicography
- Lexicography for Specialised Languages, Terminology and Terminography
- Lexicography of Lesser Used languages
- Phraseology and Collocation
- Historical Lexicography and Etymology
- Lexicological Issues of Lexicographical Relevance
- Reports on Lexicographical and Lexicological Projects
- Others

Within each section, the papers are organised alphabetically by the surname of first author.

EURALEX Congress proceedings have become an important reference for dictionary research, based on the high quality of the papers selected. During the review process every submitted abstract was evaluated by two independent blind referees. In case of doubt, a third independent opinion was sought. This process led to rejecting about one third (35%) of the papers originally submitted.

On behalf of everyone associated with the organisation of EURALEX 2014 at EURAC, I would like to express our gratitude to all the contributors for submitting relevant and interesting work as well as for meeting the tight production schedule of this online publication. I personally would like to thank also all the colleagues who participated in the review process and those who joined me on the EURALEX 2014 programme committee to jointly prepare the final congress programme. The generous patrons and sponsors who supported us for this edition are all listed on a dedicated page within these proceedings. Last but not least, I am particularly indebted to the three members of the organising committee who joined efforts with me to make EURALEX 2014 a successful event: Chiara Vettori, Nascia Ralli and Daniela Gasser from EURAC. Their dedication and constant commitment deserve a special mention. Chiara Vettori is the main responsible for the timely production of these online proceedings. It is a pleasure to acknowledge their precious work here, but also that of all staff who contributed to a successful congress week and who cannot be named here.

**Andrea Abel**

Chair, XVI EURALEX International Congress

May, 2014



# Index

<b>Plenary Lectures</b> .....	<b>23</b>
From Lexicography to Terminology: a Cline, not a Dichotomy .....	25
<i>Thierry Fontenelle</i>	
Natural Language Processing Techniques for Improved User-friendliness of Electronic Dictionaries .....	47
<i>Ulrich Heid</i>	
Using Mobile Bilingual Dictionaries in an EFL Class .....	63
<i>Carla Marengo</i>	
Meanings, Ideologies, and Learners' Dictionaries .....	85
<i>Rosamund Moon</i>	
<b>The Dictionary-Making Process</b> .....	<b>107</b>
The Making of a Large English-Arabic/Arabic-English Dictionary: the Oxford Arabic Dictionary .....	109
<i>Tressy Arts</i>	
Simple and Effective User Interface for the Dictionary Writing System .....	125
<i>Kamil Barbierik, Zuzana Děngeová, Martina Holcová Habrová, Vladimír Jarý, Tomáš Liška, Michaela Lišková, Miroslav Virius</i>	
Totalitarian Dictionary of Czech .....	137
<i>František Čermák</i>	
Dictionary of Abbreviations in Linguistics: Towards Defining Cognitive Aspects as Structural Elements of the Entry .....	145
<i>Ivo Fabijancić</i>	
La definizione delle relazioni intra- e interlinguistiche nella costruzione dell'ontologia IMAGACT .....	159
<i>Gloria Gagliardi</i>	
A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition .....	169
<i>Ellert Thor Johannsson, Simonetta Battista</i>	
Making a Learner's Dictionary of Academic English .....	181
<i>Diana Lea</i>	
The Danish Thesaurus: Problems and Perspectives .....	191
<i>Sanni Nimb, Lars Trap-Jensen, Henrik Lorentzen</i>	
From a Dialect Dictionary to an Etymological One .....	201
<i>Vilja Oja, Iris Metsmägi</i>	
<b>Research on Dictionary Use</b> .....	<b>211</b>
Wörterbuchbenutzung: Ergebnisse einer Umfrage bei italienischen DaF-Lernern .....	213
<i>Carolina Flinz</i>	

Translation, Cultural Adaptation and Preliminary Psychometric Evaluation of the English Version of “Strategy Inventory for Dictionary Use” (S.I.D.U) .....	225
<i>Gavriilidou Zoe</i>	
The Authentic Voices of Dictionary Users – Viewing Comments on an Online Learner’s Dictionary Before and After Revision .....	237
<i>Ann-Kristin Hult</i>	
Mobile Lexicography: A Survey of the Mobile User Situation .....	249
<i>Henrik Køhler Simonsen</i>	
Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht .....	263
<i>Martina Nied Curcio</i>	
Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses .....	281
<i>Sascha Wolfer, Alexander Kopleinig, Peter Meyer, Carolin Müller-Spitzer</i>	
La performance dell’utente apprendente di italiano LS/L2 e la microstruttura dei dizionari: sussidi per lo sviluppo della Lessicografia Pedagogica .....	291
<i>Angela Maria Tenório Zucchi</i>	
<b>Lexicography and Language Technologies .....</b>	<b>303</b>
ALITUOT – Atlante della Lingua Italiana QUOTidiana .....	305
<i>Michele Castellarin, Fabio Tosques</i>	
Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples .....	319
<i>Paul Cook, Michael Rundell, Jey Han Lau, and Timothy Baldwin</i>	
Cross-linking Austrian dialectal Dictionaries through formalized Meanings .....	329
<i>Thierry Declerck, Eveline Wandl-Vogt</i>	
Nutzung des DWDS-Wortprofils beim Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen .....	345
<i>Jörg Didakowski, Nadja Radtke</i>	
Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies .....	355
<i>Iztok Kosem, Polona Gantar, Nataša Logar, Simon Kreh</i>	
A Corpus-based Dictionary of Polish Sign Language (PJM) .....	365
<i>Jadwiga Linde-Usiekiewicz, Małgorzata Czajkowska-Kisil, Joanna Łacheta, Paweł Rutkowski</i>	
Laying the Foundations for a Diachronic Dictionary of Tunis Arabic: a First Glance at an Evolving New Language Resource .....	377
<i>Karlheinz Mörth, Stephan Procházka, Ines Dallaji</i>	
BabelNet meets Lexicography: the Case of an Automatically-built Multilingual Encyclopedic Dictionary .....	389
<i>Roberto Navigli</i>	
A Simple Platform for Defining Idiom Variation Matching Rules .....	399
<i>Koichi Takeuchi, Ulrich Apel, Ray Miyata, Ryo Murayama, Ryoko Adachi, Wolfgang Fanderl, Iris Vogel, Kyo Kageura</i>	

Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon .....	405
<i>Jana Šindlerová, Kateřina Veselovská, Jan Hajič jr.</i>	
<b>Lexicography and Corpus Linguistics .....</b>	<b>415</b>
Compatible Sketch Grammars for Comparable Corpora .....	417
<i>Vladimír Benko</i>	
From GLÀFF to PsychoGLÀFF: a Large Psycholinguistics-oriented French Lexical Resource .....	431
<i>Basilio Calderone, Nabil Hathout, Franck Sajous</i>	
RIDIRE. Corpus and Tools for the Acquisition of Italian L2 .....	447
<i>Alessandro Panunzi, Emanuela Cresti, Lorenzo Gregori</i>	
Empirical Approaches to German Paronyms .....	463
<i>Petra Storjohann, Ulrich Schnörch</i>	
Pragmatic Meaning in Lexicographical Description: Semantic Prosody on the Go .....	477
<i>Mojca Šorli</i>	
<b>Bi- and Multilingual Lexicography .....</b>	<b>493</b>
Linking a Dictionary to Other Open Data – Better Access to More Specific Information for the Users .....	495
<i>Ulrich Apel</i>	
Bilingual Word Sketches: the <i>translate</i> Button .....	505
<i>Vít Baisa, Miloš Jakubiček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý</i>	
Creating a Bilingual Italian-English Dictionary of Collocations .....	515
<i>Barbara Berti, Laura Pinnavaia</i>	
La valencia del adjetivo en diccionarios bilingües alemán-español-alemán .....	525
<i>Andreu Castell, Natàlia Català, Maria Bargalló</i>	
Esame storico dei “realia” nei dizionari bilingui italiano/ungheresi .....	537
<i>Zsuzsanna Fábíán</i>	
Quello che i dizionari possono fare: l’esempio dei Dizionari di Tedesco (Giacoma/Kolb – Zanichelli/Klett) .....	551
<i>Luisa Giacoma</i>	
Bilingual Dictionary Drafting. The Example of German-Basque, a Medium-density Language Pair .....	563
<i>David Lindemann, Iker Manterola, Rogelio Nazar, Iñaki San Vicente, Xabier Saralegi</i>	
Illustrative Examples and the Aspect of Culture: The Perspective of a Tshivenda Bilingual Dictionary .....	577
<i>Munzhedzi James Mafela</i>	
Corpus, Parallélisme et Lexicographie Bilingue .....	587
<i>Adriana Zavaglia, Gisele Galafacci</i>	

<b>Lexicography for Specialised Languages, Technology and Terminography</b> .....	<b>599</b>
EcoLexicon .....	601
<i>Pamela Faber, Miriam Buendía Castro</i>	
Experts and Terminologists: Exchanging Roles in the Elaboration of the Terminological Dictionary of the Brenner Base Tunnel (BBT) .....	609
<i>Elena Chiocchetti, Natascia Ralli</i>	
Cloud Terminology Services Facilitate Specialised Lexicography Work .....	621
<i>Tatiana Gornostay, Andrejs Vasiljevs</i>	
Good Contexts for Translators—A First Account of the Cristal Project .....	631
<i>Amélie Josselin-Leray, Cécile Fabre, Josette Rebeyrolle, Aurélie Picton, Emmanuel Planas</i>	
Kontextbasierte lexikalische Kontrolle von Anforderungsdokumenten .....	647
<i>Jennifer Krisch</i>	
From Term Dynamics to Concept Dynamics: Term Variation and Multidimensionality in the Psychiatric Domain .....	657
<i>Pilar León-Araúz, Arianne Reimerink</i>	
<i>Bon usage</i> vs. Fachliches: Fachsprache in der Geschichte der französischen Sprachpflege und Lexikographie .....	669
<i>Martina Mayer</i>	
EU-Terminologie in den einsprachigen Wörterbüchern des Deutschen .....	685
<i>Diana Stantcheva</i>	
Text Boxes as Lexicographic Device in LSP Dictionaries .....	697
<i>Elsabé Taljard, Danie J. Prinsloo, Rufus H. Gouws</i>	
Station Sensunique: Architecture générale d’une plateforme web paramétrable, modulaire et évolutive d’acquisition assistée de ressources .....	707
<i>Izabella Thomas, Blandine Plaisantin Alecu, Bérenger Germain, Marie-Laure Betbeder</i>	
Station Sensunique: une plateforme Web modulaire, collaborative et évolutive d’acquisition assistée de ressources terminologiques et non terminologiques (orientée Langues Contrôlées) .....	727
<i>Izabella Thomas, Blandine Plaisantin Alecu, Bérenger Germain, Marie-Laure Betbeder</i>	
<b>Lexicography of Lesser Used Languages</b> .....	<b>737</b>
Towards an Integrated E-Dictionary Application – The Case of an English to Zulu Dictionary of Possessives .....	739
<i>Sonja Bosch, Gertrud Faaß</i>	
Zur (Vor-)Geschichte der saamischen Lexikografie: ein lateinisch-saamisches Wörterverzeichnis aus dem 17. Jahrhundert .....	749
<i>Eino Koponen</i>	
Compiling a Basic Vocabulary for German Sign Language (DGS) – lexicographic issues with a focus on word senses .....	767
<i>Gabriele Langer, Susanne König, Silke Matthes</i>	

Dearcadh na nDéise – Representations of Gaeltacht na nDéise in Dineen’s Bilingual Irish-English Dictionary (1927) .....	787
<i>Chris Mulhall, Seamus ó’Diollúin</i>	
The <i>eLexicon Mediae et Infimae Latinitatis Polonorum</i> . The Electronic Dictionary of Polish Medieval Latin .....	793
<i>Krzysztof Nowak</i>	
From DANTE to Dictionary: The New English-Irish Dictionary .....	807
<i>Pádraig Ó Mianáin, Cathal Convery</i>	
User Support in e-Dictionaries for Complex Grammatical Structures in the Bantu Languages .....	819
<i>Danie J. Prinsloo, Theo J.D. Bothma, Ulrich Heid</i>	
Concerning the Treatment of Co-existent Synonyms in Estonian Monolingual and Bilingual Dictionaries .....	829
<i>Enn Veldi</i>	
<b>Phraseology and Collocation</b> .....	<b>837</b>
Unusual Phrases in English MLDs: Increasing User Friendliness .....	839
<i>Stephen Coffey</i>	
Harvesting from One’s Own Field: A Study in Collocational Resonance .....	855
<i>Janet DeCesaris, Geoffrey Williams</i>	
The Use of Corpora in Bilingual Phraseography .....	867
<i>Dmitrij Dobrovol’skij</i>	
Comparing Phraseologisms: Building a Corpus-Based Lexicographic Resource for Translators .....	885
<i>Laura Giacomini</i>	
Lexical Variation within Phraseological Units .....	893
<i>Tarja Riitta Heinonen</i>	
<i>Prendere il toro per le corna o lasciare una bocca amara? – The Treatment of Tripartite Italian Idioms in Monolingual Italian and Bilingual Italian-English Dictionaries</i> .....	905
<i>Chris Mulhall</i>	
Especialización y Prototipicidad en Binomios N y N .....	915
<i>Ignacio Rodríguez Sánchez</i>	
Syntax and Semantics vs. Statistics for Italian Multiword Expressions: Empirical Prototypes and Extraction Strategies .....	927
<i>Luigi Squillante</i>	
<b>Historical Lexicography and Etymology</b> .....	<b>939</b>
Il DiVo (Dizionario dei Volgarizzamenti). Un archivio digitale integrato per lo studio del lessico di traduzione nell’italiano antico .....	941
<i>Diego Dotto</i>	

Informatiser le Französisches etymologisches Wörterbuch: la nécessaire prise en compte de l'utilisateur .....	955
<i>Pascale Renders, Esther Baiwir</i>	
A Morphological Historical Root Dictionary for Portuguese .....	967
<i>João Paulo Silvestre, Alina Villalva</i>	
<b>Lexicological Issues of Lexicographical Relevance .....</b>	<b>979</b>
What can Lexicography Gain from Studies of Loanword Perception and Adaptation? .....	981
<i>Miroslaw Bańko, Milena Hebal-Jeziarska</i>	
Pejorative Language Use in the Satirical Journal “Die Fackel” as documented in the “Dictionary of Insults and Invectives” .....	993
<i>Hanno Biber</i>	
The Presence of Gender Issues in Spanish Dictionaries .....	1001
<i>Ana Costa Pérez</i>	
Reflexive Verbs in a Valency Lexicon: The Case of Czech Reflexive Morphemes .....	1007
<i>Václava Kettnerová, Markéta Lopatková</i>	
Polysemous Models of Words and Their Representation in a Dictionary Entry .....	1025
<i>Tinatin Margalidze</i>	
One Lexicological Theory, two Lexicographical Models and the Pragmatemes .....	1039
<i>Lena Papadopoulou</i>	
Analyzing Specialized Verbs in a French-Italian-English Medical Corpus: A Frame-based Methodology .....	1049
<i>Anna Riccio</i>	
Neoclassical Formatives in Dictionaries .....	1059
<i>Pius ten Hacken, Renáta Panocová</i>	
<b>Reports on Lexicographical and Lexicological Projects .....</b>	<b>1073</b>
Revision and Digitisation of the Early Volumes of <i>Norsk Ordbok</i> : Lexicographical Challenges .....	1075
<i>Sturla Berg-Olsen, Åse Wetås</i>	
A Dictionary Guide for Web Users .....	1087
<i>Valeria Caruso, Anna De Meo</i>	
What a Multilingual Loanword Dictionary can be used for: Searching the <i>Dizionario di italianismi in francese, inglese, Tedesco</i> (DIFIT) .....	1099
<i>Matthias Heinz, Anne-Kathrin Gärtig</i>	
The Basic Estonian Dictionary: the first Monolingual L2 learner’s Dictionary of Estonian .....	1109
<i>Jelena Kallas, Maria Tuulik, Margit Langemets</i>	
Die fremdsprachige Produktionssituation im Fokus eines onomasiologisch konzeptuell orientierten, zweisprachig-bilateralen Wörterbuches für das Sprachenpaar Deutsch - Spanisch: Theoretische und methodologische Grundlagen von DICONALE .....	1119
<i>Meike Meliss</i>	



Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries .....	1135
<i>Peter Meyer</i>	
At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project) .....	1145
<i>Pavla Kochová, Zdeňka Opavská, Martina Holcová Habrová</i>	
Frame Semantics and Learner's Dictionaries: Frame Example Sections as a New Dictionary Feature .....	1153
<i>Carolin Ostermann</i>	
Translating Action Verbs using a Dictionary of Images: the IMAGACT Ontology .....	1163
<i>Alessandro Panunzi, Irene De Felice, Lorenzo Gregori, Stefano Jacoviello, Monica Monachini, Massimo Moneglia, Valeria Quochi, Irene Russo</i>	
Degrees of Synonymity as the Basis of a Network for German Communication Verbs in the Online Reference Work <i>Kommunikationsverben</i> in OWID .....	1171
<i>Kristel Proost, Carolin Müller-Spitzer</i>	
Job-hunting in Italy: Building a glossary of "English-inspired" job titles .....	1187
<i>Virginia Pulcini, Angela Andreani</i>	
A Small Dictionary of Life under Communist Totalitarian Rule (Czechoslovakia 1948-1989) .....	1203
<i>Věra Schmiedtová</i>	
A Frequency Dictionary of Dutch .....	1211
<i>Carole Tiberius, Tanneke Schoonheim, Adam Kilgarriff</i>	
The Corpus of the Croatian Church Slavonic Texts and the Current State of Affairs Concerning the Dictionary of the Croatian Redaction of Church Slavonic Compiling .....	1221
<i>Vida Vukoja</i>	
<b>Others</b> .....	<b>1237</b>
Considerations about Gender Symmetry in the Dictionary of Bavarian Dialects in Austria .....	1239
<i>Isabella Flucher, Eveline Wendl-Vogt, Thierry Declerck</i>	
Advancing Search in the Algemeen Nederlands Woordenboek .....	1247
<i>Carole Tiberius, Jan Niestadt, Lut Colman, Boudewijn van den Berg</i>	



# Plenary Lectures



# From Lexicography to Terminology: a Cline, not a Dichotomy

Thierry Fontenelle

Translation Centre for the Bodies of the European Union, Luxembourg

thierry.fontenelle@cdt.europa.eu

## Abstract

In a paper presented at the Euralex 2012 conference, ten Hacken (2012) discusses the OED's problematic claim to be the "definitive record of the English language". He argues that what distinguishes the OED from other dictionaries is the information it provides about English words and the range of problems this information can be used to solve. Dictionaries are not descriptions of a language, he claims, but tools with which users of the dictionary solve problems of a particular type. The nature of the dictionary therefore determines which types of problems it can solve.

In this paper, I would like to extend the parallel made by ten Hacken between general dictionaries, learners' dictionaries and historical dictionaries such as the OED to what is traditionally perceived as a dichotomy, namely the distinction between dictionaries and terminological databases. Instead of viewing term bases as a totally distinct type of linguistic product, I would like to argue that they should rather be seen as a specific kind of tool which provides information that specific users will use in order to solve specific linguistic problems, usually related to translation. In the course of their careers, translators will indeed need to make use of a whole range of dictionaries, starting from learners' dictionaries when they learn foreign languages, to monolingual dictionaries and bilingual dictionaries to learn translation techniques, to term bases as soon as they start translating highly specialized and technical texts. We will focus on terminology databases such as IATE, the European Union's interinstitutional term base, which is the natural tool to which they turn to obtain information about technical terms in the medical, legal, environmental, chemical fields, to cite only a few domains covered by this resource. With 8.7 million terms covering the 24 official languages of the European Union, including 1.4 million English terms and half a million abbreviations, this database is a highly popular tool in the translation world (44 million queries in 2013).

In addition to general term bases such as IATE, we will also discuss other specialized EU terminological databases such as ECHA-Term, a term base compiled for the European Chemicals Agency (ECHA) to help industry comply with the legal requirements of the REACH Directive and of the regulation on the classification, labeling and packaging of chemicals. We will show how user requirements have been taken into account to meet the needs of the users of the database, who resort to ECHA-Term to obtain reliable, coherent and up-to-date multilingual terminology in the chemicals field, a *sine qua non* for clear specialized communication. The description of these databases will make it clear that the distinction between 'traditional' lexicography and terminology is more a cline than a dichotomy,

insofar as the types of linguistic information included in the respective products created by both disciplines all correspond to the specific needs of their users.

**Keywords:** term banks; terminological database; translation; European Union; LSP dictionaries; IATE; ECHA-term

## 1 Introduction

In a paper presented at the Euralex 2012 conference, ten Hacken (2012) discusses the Oxford English Dictionary's problematic claim to be the "definitive record of the English language". He points out that the OED is often regarded as authoritative and that one of the aspects of authority is the comprehensive lexical coverage of the dictionary. Yet, he argues, even if lexicographers such as Simpson (2000:1) call the OED "the principal dictionary of record for the English language", there is in fact no empirical entity corresponding to "the English language" for which the OED could be taken as a description (ten Hacken 2012: 838). Simpson himself is aware of the impossibility of providing a comprehensive coverage in a dictionary. It is therefore a myth to assume that a dictionary should contain every word. A dictionary can therefore only be a partial record and it is unrealistic to assume that a dictionary can provide a full record of a language.

Ten Hacken argues that what distinguishes the OED from other dictionaries is the information it provides about English words and the range of problems this information can be used to solve. Dictionaries are not descriptions of a language, he claims, but tools with which users of the dictionary solve problems of a particular type. The nature of the dictionary therefore determines which types of problems it can solve.

It is interesting to note that the controversy around the comprehensive nature of a dictionary such as the OED usually involves a comparison with other general-purpose monolingual dictionaries, including learners' dictionaries. The inclusion of usage notes is also used as a criterion to distinguish descriptive and prescriptive dictionaries. Surprisingly, no mention is usually ever made of a different type of dictionary, namely terminological databases, which should also be seen as specific types of dictionaries designed to solve specific types of linguistic problems. Why is that nobody ever questions the comprehensiveness of a terminological database? Why would anybody expect the OED to include "all possible" words of the English language, while recognizing that the list of acronyms and abbreviations in a language is potentially infinite and that even a huge term base such as IATE, the interactive terminology database of the European Union described below, can only provide a partial record of specialized terminology, even with 1.4 million English terms and half a million abbreviations and acronyms?

## 2 From Learner's Dictionaries to Mono- and Bilingual Dictionaries to Terminological Bases

I would like to extend the parallel made by ten Hacken between general dictionaries, learners' dictionaries and historical dictionaries such as the OED to what is traditionally perceived as a dichotomy, namely the distinction between dictionaries and terminological databases. Instead of viewing term bases as a totally distinct type of linguistic product, I would like to argue that they should rather be seen as a specific kind of tool which provides information that specific users will use in order to solve specific linguistic problems, usually related to translation.

If, as proposed by Ten Hacken (2012, 2009), dictionaries are tools with which users solve problems of a particular type, the various types of dictionaries available on the market actually correspond to the range of problems users are faced with at different moments of their career. A teenager or a university student who learns a foreign language will most probably require a small bilingual dictionary at the beginning of the learning process because a learner's dictionary can only be used by someone who is not a total beginner in this foreign language. Once knowledge of the foreign language reaches a certain level, the student will be encouraged to make use of a learners' dictionary, which will provide useful information in a decoding and an encoding perspective, thanks to its simplified definitions, its system of grammar codes, its illustrative examples, etc. General-purpose monolingual dictionaries target a different kind of public, made up of advanced native speakers or of non-native speakers who have an in-depth knowledge of the language of the dictionary. Historical dictionaries such as the OED, with their focus on etymology and the evolution of words, are yet for other users who expect the dictionary to provide descriptive records of the development and use of words over time.

A parallel may be drawn with the tools used by translators. At the beginning of their career, students in translation will primarily make use of general-coverage bilingual dictionaries which will provide them with information about collocations, idioms, sense distinctions, etc. The role of translations in bilingual dictionaries is to provide target-language equivalents of the source-language headword (see also Fontenelle forthcoming). At a later stage, however, seasoned translators will tend to consult their bilingual dictionaries less and less, and will turn more frequently to terminological databases (a.k.a. term banks), which will enable them to translate highly-specialized texts and to make communication possible between specialists, or between specialists and the general public. Such tools have become a *sine qua non* in our multilingual world where access to technical information across multiple domains is a must.

## 3 Terminology and Term Banks

Understanding terminology, i.e. the specialized vocabulary which is used in a specific domain, is a key element in communication. This poses a number of challenges in the case of translation in a multi-

lingual context, since knowing the exact meaning of a technical term is necessary to understand a text, but also to reproduce the text as faithfully as possible in another language. It therefore no surprise that the various translation services of the European Union have traditionally dedicated significant resources to the compilation of terminological information in the official languages of the EU. In order to describe the vocabulary of special subject fields, terminologists create term banks, which are compilations of the collections of words associated with a given domain. A terminology database will then be seen as a repository of descriptions of concepts, which are seen as mental constructs which are distinct from the terms they correspond to in a given language (see also Fontenelle and Rummel, forthcoming). The traditional approach assumes that terms can be organised into networks of concepts to structure a given domain. This is indeed well-suited for normalization, but, as is pointed out by Jacquemin and Bourrigault (2003), there seems to be a flaw in this reasoning, because this approach is not really suitable for computational term analysis. A terminologist indeed bases his or her work upon the analysis of textual data (corpora) and the term base is actually the result of this analysis, and not the result of some introspection whereby abstract conceptual maps would be derived.

Even if terminological databases are traditionally seen as distinct from dictionaries, it cannot be denied that terminological entries have a lot in common with dictionary entries. Of course, at the macrostructural level, it is clear that some entries found in a traditional dictionary will not be found in a term base: some parts of speech will be absent from term banks. Prepositions or adverbs for instance will most probably not be found in term banks. Even verbs are traditionally underrepresented in such resources because the vast majority of terms are noun phrases. This is not enough to consider that a term base is not a dictionary, however. After all, rare scientific words will also be excluded from learners' dictionaries. At the microstructural level, definitions will be as essential in term bases as in a traditional monolingual dictionary and the NLP community has always been interested in how candidate terms could be extracted from corpora, together with possible definitions (see Person 1998, who proposed an analysis of the defining mechanisms signalling the presence of a term in a corpus, using linguistic patterns such as 'X is known as Y' or "X is called Y" to link a definiens and a definiendum).

## **4 IATE: The European Union Terminological Database**

### **4.1 An Interinstitutional Database**

IATE stands for 'Inter-Active Terminology for Europe' and is the term base of the language services of the European Union. This concept-oriented, large-scale multilingual database covers all fields of activity of the European Union. IATE was initially launched in 1999 by the Translation Centre for the Bodies of the European Union, located in Luxembourg. Today, the Translation Centre manages the technical aspects of the project on behalf of the project partners: the European Commission, the Eu-



ropean Parliament, the Council, the Court of Auditors, the Court of Justice of the European Union, the European Investment Bank, the European Central Bank, the Economic and Social Committee, the Committee of the Regions and the Translation Centre. Before the launch of the project and its opening to the public in 2007, nearly every institution had its own term base (Fontenelle and Mergen 1998; Reichling 1998), while, today, IATE can be seen as the shared terminology database of all existing terminology collections of the translation services of all EU institutions and bodies. It can be consulted free of charge at <http://iate.europa.eu>.

IATE can be searched for specific terms or abbreviations in a given source language and for its equivalent(s) in any of the 23 other languages (IATE contains mainly terminology in the 24 official languages of the EU, as well as some content in non-official languages). After the accession of Croatia on 1 July 2013, the 24 official languages are: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, Swedish.

The database is constantly updated by the terminologists and translators of the various participating institutions. In 2013, around 97 000 new terms were added and over 150 000 existing terms were modified. The contents of the database as of 1 January 2014 can be broken down as in Table 1.

Language	Number of terms	Language	Number of terms
en - English	1402006	ga - Irish	57879
fr - French	1339496	lt - Lithuanian	53802
de - German	1034088	hu - Hungarian	49858
it - Italian	701735	et - Estonian	41489
nl - Dutch	691801	sl - Slovenian	41337
es - Spanish	617124	cs - Czech	38043
da - Danish	603481	sk - Slovak	37219
pt - Portuguese	533623	mt - Maltese	35732
el - Greek	523086	ro - Romanian	35451
fi - Finnish	329491	bg - Bulgarian	34420
sv - Swedish	314220	lv - Latvian	28617
la - Latin	64159	hr - Croatian	8863
pl - Polish	59576		

**Table 1: IATE: Number of terms (1/1/2014).**

The table reflects the history of the European Union and its successive enlargements, of course. Altogether, the IATE database contained 8,705,334 terms on 1/1/2014. The languages with the most terms were the European Union's most often used working languages, viz. English, with more than 1.4 million terms in first place, followed by French (1.3 million terms) and then German (1 million terms). It is interesting to note that Latin is also fairly well-represented, with 64159 terms: taxonomies of animal

and vegetal species are crucial for translators who deal with the translation of texts related to the Common Fisheries Policy or the Common Agricultural Policy, as well as texts written by the European Environment Agency, the Community Plant Variety Office or the European Food Safety Agency.

The public version of IATE received 44 million queries in 2013 and the number of queries in the internal version of IATE (accessible only to EU staff) amounted to 14,206,137 (vs. 11,178,323 queries in 2012). Queries in the public version of IATE came from all over the world, from France to Somalia, Argentina to South Korea. The country where the most queries originated from in 2013 was Italy, followed by France, Spain, Germany, Belgium, Greece, Portugal, the United Kingdom, the Netherlands and Switzerland.

Terms in IATE have a fairly standard data structure. One of the challenges which the creators of the database faced was the mapping rules between the data structures of the existing databases and the new format of this interinstitutional database. A concept-oriented approach was adopted to express the various aspects of concepts via a series of three interrelated levels:

- (1) a language-independent top level, containing information pertaining to the whole concept. Information about domains is a case in point (i.e. the field of knowledge in which the concept is used). Other types of information can also be stored at that level, including pictures or images;
- (2) An intermediate 'language' level for definitions, explanations and comments, which can be stored for each of the languages of the terminological record;
- (3) The 'term' sub-level, at which several terms can be distinguished to store synonyms of a given concept or abbreviations. Reliability codes, term references and the date when a record was last edited will typically appear at that level.

For instance, *BSE*, *bovine spongiform encephalopathy* and *mad cow disease* are three distinct terms that are synonyms and refer to the same concept. The definitions would be coded at level 2 (the three terms will have the same definition in a given language), while the domain (Animal health) is at level 1, the conceptual level, as is illustrated in Figure 1 and in Figure 2 below.

English (en) Search Screen Help

BSE Search

en > de fr it nl (domain: Any domain, type of search: All)

Result 1- 10 of 32 for BSE

Animal health [COM]		Full entry
	mad cow disease	★★★★ *@
EN	bovine spongiform encephalopathy	★★★★ *@
	BSE	★★★★ *@
	bovine spongiforme Enzephalopathie	★★★★ *@
DE	Rinderwahnsinn	★★★★ *@
	spongiforme Rinderenzephalopathie	★★★★ *@
	BSE	★★★★ *@
	maladie de la vache folle	★★★★ *@
	maladie des vaches folles	★★★★ *@
FR	encéphalite spongiforme bovine	★★★★ *@
	encéphalite spongiforme bovine	★★★★ *@
	encéphalopathie spongiforme bovine	★★★★ *@
	ESB	★★★★ *@
	folia bovina	★★★★ *@
IT	encefalopatia spongiforme bovina	★★★★ *@
	BSE	★★★★ *@
	gekkekoeienziekte	★★★★ *@
NL	bovine spongiforme encefalopathie	★★★★ *@
	BSE	★★★★ *@

Figure 1: Query on BSE in IATE.

<b>Domain</b>	<b>Animal health</b>
<b>en</b>	
<b>Definition</b>	progressive, fatal, neurologic disease of adult domestic cattle that resembles <i>scrapie</i> ( IATE:1257587 ) of sheep and goats
<b>Definition Ref.</b>	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, <a href="http://www.merckvetmanual.co...">http://www.merckvetmanual.co...</a> [10.1.2012]
<b>Note</b>	<p>It was first diagnosed in the UK in 1986. Epidemiological studies conducted in the UK suggest that the source of BSE was cattle feed prepared from bovine tissues, such as brain and spinal cord, that was contaminated by the BSE agent. Speculation as to the cause of the appearance of the agent causing the disease has ranged from spontaneous occurrence in cattle, the carcasses of which then entered the cattle food chain, to entry into the cattle food chain from the carcasses of sheep with scrapie.</p> <p>For further information please refer to: WHO &gt; Programmes and projects &gt; Media centre &gt; Fact sheets &gt; Bovine spongiform encephalopathy, <a href="http://www.who.int/mediacent...">http://www.who.int/mediacent...</a> [10.1.2012]</p>
<b>Term</b>	<b>mad cow disease</b>
<b>Reliability</b>	3 (Reliable)
<b>Term Ref.</b>	Centers for Disease Control and Prevention > CDC A-Z Index > BSE (Bovine Spongiform Encephalopathy, or Mad Cow Disease), <a href="http://www.cdc.gov/ncidod/dv...">http://www.cdc.gov/ncidod/dv...</a> [10.1.2012]
<b>Date</b>	10/01/2012
<b>Term</b>	<b>bovine spongiform encephalopathy</b>
<b>Reliability</b>	3 (Reliable)
<b>Term Ref.</b>	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, <a href="http://www.merckvetmanual.co...">http://www.merckvetmanual.co...</a> [10.1.2012]
<b>Date</b>	10/01/2012
<b>Abbreviation</b>	<b>BSE</b>
<b>Reliability</b>	3 (Reliable)
<b>Term Ref.</b>	The Merck Veterinary Manual > Nervous System > Bovine Spongiform Encephalopathy > Introduction, <a href="http://www.merckvetmanual.co...">http://www.merckvetmanual.co...</a> [10.1.2012]
<b>Date</b>	10/01/2012

**Figure 2: set of synonyms for bovine spongiform encephalopathy.**

The system offers today the following features (see also Fontenelle and Rummel in press):

- One common database for all institutions and agencies containing all legacy data;
- Basic and advanced search features (including stemming and base character conversion);
- On-line access in read and write mode, i.e. the possibility for users to carry out modifications, to add entries directly to the central database and hence to allow their colleagues to benefit from this work in real time;
- A validation workflow that ensures that all newly added or modified terminology is reviewed;
- Role-based user management;
- Auditing features that keep track of all changes made to the terminology in the database;
- Features for the export and import of data;
- Statistics on the content of the database and user activity;
- A basic messaging system as communication mechanism between the actors in the terminology workflow.

In 2014, a new functionality will also be offered to allow users to download or copy the contents of the IATE database which is not protected by third-party copyrights, for research or for commercial purposes.

## 4.2 Language and Terminology: A Dynamic Organism

Languages are dynamic organisms. New terms are created every day (nobody was talking about 3D printers five years ago and the translation of selfie is a hot topic in many linguistic communities in 2014). It is therefore not surprising to see that IATE is not just concerned with terms proper, but also with abbreviations and acronyms, which are condensed versions of often long and complex terms. Table 2 shows the contents of the IATE database per term type.

Abbrev	519161
Formula	698
Phrase	142784
Short Form	20579
Term	8022112

**Table 2: Number of terms per type in IATE.**

With more than half a million abbreviations, one can immediately see that the question of the ‘completeness’ of a dictionary discussed in the context of the OED above is a myth, indeed, and a quick glance at the list of abbreviations included in such a database will convince anybody that there cannot be such a thing as a “definitive record” of any language.



DROIT [CdT]		Entrée entière	
EN	the Administrative Board shall adopt rules of procedure	★★★★ +@	
DE	der Verwaltungsrat gibt sich eine Geschäftsordnung	★★★★ +@	
ES	el Consejo de Administración adoptará su reglamento interno	★★★★ +@	
FR	le Conseil d'administration arrête son règlement intérieur	★★★★ +@	
IT	il Consiglio di amministrazione adotta il proprio regolamento interno	★★★★ +@	

DROIT [CdT]		Entrée entière	
EN	an Opposition Division shall consist of three members	★★★★ +@	
DE	eine Widerspruchsabteilung setzt sich aus drei Mitgliedern zusammen	★★★★ +@	
ES	una División de Oposición constará de tres miembros	★★★★ +@	
FR	une division d'opposition se compose de trois membres	★★★★ +@	
IT	una divisione di opposizione è formata da tre membri	★★★★ +@	

DROIT [CdT]		Entrée entière	
EN	the Secretariat for the Administrative Board shall be provided by the Office	★★★★ +@	
DE	die Sekretariatsgeschäfte des Verwaltungsrates werden vom Amt wahrgenommen	★★★★ +@	
ES	la Oficina se encargará de la secretaría del Consejo de Administración	★★★★ +@	
FR	le secrétariat du Conseil d'administration est assuré par l'Office	★★★★ +@	
IT	l'Ufficio svolge la funzione di segretariato del Consiglio di amministrazione	★★★★ +@	

DROIT [CdT]		Entrée entière	
EN	the Administrative Board shall take its decisions by a simple majority	★★★★ +@	
DE	der Verwaltungsrat faßt seine Beschlüsse mit der einfachen Mehrheit	★★★★ +@	
ES	el Consejo de Administración tomará sus acuerdos por mayoría simple	★★★★ +@	
FR	le Conseil d'administration prend ses décisions à la majorité simple	★★★★ +@	
IT	il Consiglio di amministrazione prende le sue decisioni a maggioranza semplice	★★★★ +@	

Figure 3: semi-fixed phrases including shall in IATE.

The inclusion of “formulae” and “phrases” in the term base is also a sign that terminologists are more and more concerned with phraseological units and various types of formulaic expressions. The need to standardize language (especially in legal texts, but also in technical fields such as aviation and aeronautics) encourages people to make use of ready-made phraseological patterns which sometimes go way beyond the traditional notion of terms viewed as a noun phrase. In this context, the new European financial supervision authorities which organize stress tests at the European level in coordination with the European Central Bank regularly issue Guidelines which must be translated in all the official EU languages. The translators working in the financial field to provide the various linguistic versions of these Guidelines have received strict recommendations concerning the translation of modal auxiliaries like “shall” and “should” (e.g. should needs to be translated by *devoir* in French in the specific context of Guidelines for financial supervision). These questions are debated by the terminologists of the various language teams of the European Union and IATE now includes a fair amount of

“fixed phrases” which show how these modal auxiliaries should be translated (e.g. “shall” in English most frequently translates as a present tense in French legal texts, when in normal, non-legal language, the default is usually a future tense in French). Figure 3 above illustrates some of these “semi-fixed phrases” included in IATE (see “shall be provided by the Office” à “est assuré par l’Office”). It is clear that there has been an evolution over the last few years with respect to the traditional distinction between lexical items included in dictionaries and terms included in terminology databases and what used to be a clear-cut distinction now increasingly appears as a cline between lexicography and terminology.

### 4.3 Metadata for User Preferences

Very much like a traditional dictionary which makes use of usage notes and a variety of labels aimed at capturing levels of formality (formal, informal, slang, taboo...), a terminological database such as IATE makes extensive use of metalinguistic labels which guide translators in their daily work. Official terminology in any field may indeed change rapidly and terms which are commonly used today may become deprecated tomorrow. It is therefore important to capture the preferences expressed by the main “consumers” of a translation (official organizations and administrations, public bodies and authorities, scientific communities, etc). Such preferences may result from simple stylistic or even sometimes arbitrary preferences and conventions. They may also reflect historical evolutions and a need to avoid geo-political problems. IATE will therefore make use of metalinguistic labels such as “preferred”, “obsolete” or “deprecated” to provide information about changes in the pragmatic use that is made of these terms.

The concept for the disease known as A(H1N1) is a case in point. IATE indicates that several “synonymous” terms can be used to refer to that disease, which was first observed in Mexico in 2009, including the term Mexican flu as well as swine flu. The term Mexican flu was used at the very beginning of what later became a worldwide epidemic, but international organisations such as the European Centre for Disease Prevention and Control (ECDC), one of the Translation Centre’s clients, expressed a strong preference in favour of the term A(H1N1)v, rather than Mexican flu or swine flu. Such preferences are captured through the use of labels like Preferred (appearing in green on the IATE web site) or Deprecated (in red on the IATE web site), as illustrated in Figure 4.

Medical science [COM]		<a href="#">Full entry</a>
	influenza A(H1N1)v virus <b>(Preferred)</b>	★★★★ +@
	novel flu virus	★★★★
EN	novel influenza virus	★★★★
	novel influenza virus A(H1N1)	★★★★ +@
	Mexican influenza virus <b>(Deprecated)</b>	★★★★
	swine influenza virus A(H1N1) <b>(Deprecated)</b>	★★★★

**Figure 4: Preferred terms vs Deprecated terms.**

The usage note included at the conceptual level of the terminological record reads as follows:

*ECDC prefers to use the term influenza A(H1N1)v (where v indicates variant), which has been chosen by WHO's Global Influenza Surveillance Network and helps distinguish the virus from seasonal influenza A(H1N1) viruses and A(H1N1) swine influenza viruses. A name for the disease caused by the virus has yet to be determined by WHO but the term 'swine flu' is inaccurate for what is now a human influenza. REF: European Centre for Infectious Disease Control ECDC Interim Risk Assessment > Health Topics > Documents > Human cases of influenza A(H1N1)v, <http://www.ecdc.europa.eu/en...> (27.8.2009)*

Metalinguistic labels such as Deprecated/Obsolete or Preferred are also complemented by the systematic use of reliability codes, which are captured in a different field in the database. Less reliable terms may indeed be included in the database in order to offer as much information as possible to the translators, but a low reliability code provides an indication that the term is based on information coming from less trusted sources, as is the case for the distinction between the deprecated term *dual fuel vehicle* (reliability =2) and the preferred term *bi-fuel vehicle* (reliability =3), as in Figure 5 below.



**Domain** ENERGY, Means of transport

<b>en</b>	
<b>Definition</b>	vehicle with two separate fuel storage systems that can run part-time on two different fuels and is designed to run on only one fuel at a time
<b>Definition Ref.</b>	Regulation (EC) No 692/2008 implementing and amending Regulation (EC) No 715/2007 on type-approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5 and Euro 6) and on access to vehicle repair and maintenance information <a href="#">32008R0692/EN</a>
<b>Note</b>	On internal combustion engines one fuel is petrol (gasoline) or diesel, and the other is an alternate fuel such as natural gas (CNG), LPG, or hydrogen. The two fuels are stored in separate tanks and the engine runs on one fuel at a time, unlike <i>flexible-fuel vehicles</i> ( <a href="#">IATE:210005</a> ), that store the two different fuels mixed together in the same tank, and the resulting blend is burned in the combustion chamber.  REF:Wikipedia > Bi-fuel vehicle, <a href="http://en.wikipedia.org/wiki...">http://en.wikipedia.org/wiki...</a> (28.8.2009)
<b>Term</b>	<b>bi-fuel vehicle (Preferred)</b>
<b>Reliability</b>	3 (Reliable)
<b>Term Ref.</b>	Regulation (EC) No 692/2008 implementing and amending Regulation (EC) No 715/2007 on type-approval of motor vehicles with respect to emissions from light passenger and commercial vehicles (Euro 5 and Euro 6) and on access to vehicle repair and maintenance information <a href="#">32008R0692/EN</a>
<b>Date</b>	06/12/2013
<b>Term</b>	<b>dual fuel vehicle (Deprecated)</b>
<b>Reliability</b>	2 (Minimum reliability)
<b>Term Ref.</b>	Wikipedia > Bi-fuel vehicle, <a href="http://en.wikipedia.org/wiki...">http://en.wikipedia.org/wiki...</a> [6.12.2013]
<b>Term Note</b>	'Dual fuel vehicle' may refer either to a 'bi-fuel vehicle' (as defined here) or to a 'flex fuel vehicle' (see <a href="#">IATE:210005</a> ).
<b>Date</b>	06/12/2013

Source: COM

IATE ID: 2242494

**Figure 5: Preference labels and reliability codes.**

A label such as Obsolete may be used to mark terms which are no longer in official use. Official denominations for countries or cities may indeed change, as was the case for Bombay, which was changed to Mumbai. Depending upon the context, translators may need to preserve the former name (in historical documents, for instance), which is why the two terms need to coexist in the database, with labels distinguishing them.

## 5 ECHA-term

IATE is a “generalist” database covering many subject fields. It is one of the key resources used by translators, who contribute to its enrichment alongside the terminologists who manage the database. The Translation Centre for the Bodies of the European Union also created much more specialized databases, such as the ECHA-term database, which was developed for the European Chemicals Agency (ECHA), located in Helsinki, Finland. ECHA was created in 2007 to implement the European Union’s

chemical legislation, and more particularly the REACH Directive which was adopted to improve the protection of human health and the environment from the risks that can be posed by chemicals, while enhancing the competitiveness of the EU chemicals industry. REACH also promotes alternative methods for the hazard assessment of substances in order to reduce the number of tests on animals. With the REACH regulation, companies are responsible for providing information on the hazards, the risks and the safe use of the chemical substances they manufacture, import and transport throughout the European Union. The Classification, Labelling and Packaging (CLP) Regulation introduced a globally harmonised system for classifying and labelling chemicals in the EU, thereby ensuring that the hazards presented by these chemicals are clearly communicated to workers and consumers.

The European Chemicals Agency therefore has very important multilingual communication tasks (the Translation Centre produced over 25,000 pages of translations for ECHA in 2013, mainly leaflets, technical guidance, web content, IT manuals, administrative documents and news items translated into all the official languages of the European Union). In 2009, the ECHA-term project was launched with the objective to provide ECHA and its stakeholders with a reliable, coherent, and up-to-date source of terminology to harmonise the use of terminology in the REACH and CLP context, to enhance clear communication and ultimately to reduce costs for the stakeholders. The more general aim was to help industry comply with the legal requirements, to support the national authorities in their work and to improve the quality of the translated material. The Translation Centre collaborated closely with ECHA to create a terminological database, which included the compilation of a multilingual database of over 1200 terms related to REACH, CLP and the biocides regulation, as well as “substances of very high concern” (over 50 terms) in 23 languages. The project also included the development of a platform for the dissemination of the contents.

## 5.1 Compilation of Contents for ECHA-term

The general process for the creation of the terminological contents of the database can be described as follows:

- Definition of a relevant corpus in the source language (usually English);
- Semi-automatic extraction of concepts and completion with definition, reference, context, note etc. by terminologists;
- Validation of the monolingual glossary by 2 or 3 translators to ensure that the data is relevant (are any key concepts missing)
- Formal revision by English terminologist;
- Validation of the monolingual glossary by ECHA’s experts;
- Multilingual phase: target equivalents and relevant information are completed by the Translation Centre’s terminologists;
- Ideally target language equivalents are validated by experts (ECHA);

- Import of data in ECHA-term;
- Maintenance of data following user feedback

### 5.3 Main Features

The database, which can be consulted free of charge at <http://echa.cdt.europa.eu>, offers a range of search options. Users can search by:

- Terms
- EC numbers
- CAS numbers (unique numeric identifier designating one substance in the Chemical Abstracts Service)
- GHS codes (Globally Harmonised System)
- Hazard
- Precautionary statements

In addition to the search criteria above, an alphabetical list of terms can also be browsed. A word cloud also appears to the left of news items on the home page, displaying the most-frequently searched terms, as illustrated in Figure 6. A lemmatization tool is also included to provide a match with respect to the contents of the database even if the query is inflected (as in a query on the plural form in substances of very high concern).

**ECHA-term** Multilingual Chemical Terminology by ECHA

Search criteria  
EN - English substances of very high concern Define  
Translate to FR - French

Show alphabetical list

11/02/2014  
**More terms in Croatian on ECHA-term: the biocides terminology is now online**  
The key terms related to biocides can now be found on ECHA-term in Croatian. In total, the database contains 810 entries in Croatian (including REACH and CLP terminology and the substances of very high concern).  
More terms related to biocides will be available in 23 EU languages later in March.  
→ News archive

article authorisation candidate list clp compliance check dnel downstream user endpoint exposure scenario guidance hazard information requirement inquiry intermediate mixture notification pbt reach registrant registration registration dossier restriction risk management measure safety data sheet sief stakeholder substance svhc testing proposal use

Figure 6: ECHA-term home page.

The user can choose to either define the monolingual term or to translate it into one of the 23 supported languages. The presentation of term entries is very similar to the layout offered by IATE, as one can see in Figure 7 below.

The screenshot shows the ECHA-term website interface. At the top, there is a search bar with 'EN - anglais' selected and 'SVHC' entered. Below the search bar, there are buttons for 'Définir', 'Traduire en', and a dropdown menu for 'FR - français'. A link 'Afficher la liste alphabétique' is also visible. The search results show 'Résultats 1-1 sur 1 pour SVHC'. The results table lists the term in English and French, with a definition popup for SVHC. The definition text is as follows:

Définition: Les SVHC dans le contexte de REACH sont:  
 1. Les CMR, catégorie 1 ou 2,  
 2. Les PBT et vPvB répondant aux critères de l'annexe XIII et  
 3. Les substances - telles que celles ayant des propriétés de modulation endocrinienne ou celles ayant des propriétés persistantes, bioaccumulables et toxiques ou très persistantes et très bioaccumulables, qui ne répondent pas aux critères de l'annexe III - pour lesquelles il existe des preuves scientifiques d'effets graves pour la santé humaine ou pour l'

**Figure 7: EN->FR translation of SVHC (“substance of very high concern” - “substance extrêmement préoccupante” in French).**

Figure 8 below illustrates the inclusion in ECHA-term of precautionary statements. Such statements are phrases that describe the recommended measures to minimise or prevent adverse effects resulting from exposure to a hazardous substance or mixture due to its use or disposal. It is easy to understand why standardisation is crucial in this context, since the manufacturers and the chemical industry need to make use of “pre-fabricated language” which is in a way similar to the use of controlled language and vocabulary in the aviation and aeronautic domain. The labelling and packaging of dangerous goods need to be clear and unambiguous, which leaves little room for creativity and stylistic variations. It is therefore not surprising to find fairly lengthy phrases and linguistic material in this terminology database which goes beyond the traditional notion of a term typically viewed as a noun phrase. Consider the following examples of precautionary statements whose translations are provided into all the EU official languages:

- Do not spray on an open flame or other ignition source. [see Figure 8]
- Avoid contact during pregnancy/while nursing.
- Keep cool. Protect from sunlight.
- Rinse cautiously with water for several minutes.

- Do not pierce or burn, even after use.
- Keep away from any possible contact with water, because of violent reaction and possible flash fire.

The database also includes hazard statements such as:

- Contains gas under pressure, may explode if heated
- In contact with water releases flammable gases which may ignite spontaneously.
- Harmful to aquatic life with long lasting effects.
- Toxic by eye contact.

The examples above make it abundantly clear that the traditional distinction between lexicography and terminology is more and more blurred. The phraseological patterns displayed in the precautionary statements above sometimes correspond to entire sentences (in the imperative form). In some cases, they correspond to what would be considered a collocation in a traditional dictionary. *Keep cool* is a case in point: absent any context, the phrase is highly ambiguous ('stay calm and relaxed', 'gardez votre calme' in French, is a possible meaning). In the chemical legislation covered by our specialized database, the advice is not ambiguous and indicates that a product should not be exposed to high temperatures ('tenir au frais' in French). One can imagine the possible disastrous consequences of mistranslations of such labels if the appropriate terminology is not respected.

### 5.3 Pictograms as Terms

Terms included in terminology databases are most often noun phrases, although, as we have seen above, verbal collocations such as *keep cool*, complex imperative sentences, other types of prefabricated phraseological patterns and even modal auxiliaries may be granted term status in such databases. A more recent trend appears to be the inclusion of items which are traditionally seen as non-translatable material, such as images or diagrams. ECHA-term has innovated in this context, with the inclusion of pictograms used in the chemical CLP legislation. Figure 8 illustrates the information displayed for a pictogram corresponding to corrosion, which refers to a type of physical or health hazard. Other pictograms such as skull and crossbones, exclamation marks, gas cylinders or exploding bombs will allow users to quickly find out the meaning of such graphical representations, a crucial element for the industry, but also for firemen and civilian protection specialists in emergency situations. Such CLP pictograms can be used when creating safety data sheets and training material in national languages in the various Member States of the European Union.

Back to search screen

**Alphabetical search**

EN - English    Precautionary statement    Do not spray on an open flame or other ignition source.

Feedback

Other languages: [BG](#) [CS](#) [DA](#) [DE](#) [EL](#) [ES](#) [ET](#) [FI](#) [FR](#) [HR](#) [HU](#) [IT](#) [LT](#) [LV](#) [MT](#) [NL](#) [PL](#) [PT](#) [RO](#) [SK](#) [SL](#) [SV](#) [All](#)

<b>Domain</b>	Chemical legislation, CLP
<b>Code</b>	P211
<b>FR</b>	
<b>Precautionary statement</b>	Ne pas vaporiser sur une flamme nue ou sur toute autre source d'ignition.
<b>Reliability</b>	★★★★
<b>Term Ref.</b>	Annexe IV, Règlement (CE) n o 1272/2008 du Parlement européen et du Conseil du 16 décembre 2008 relatif à la classification, à l'étiquetage et à l'emballage des substances et des mélanges, modifiant et abrogeant les directives 67/548/CEE et 1999/45/CE et modifiant le règlement (CE) n o 1907/2006, <a href="#">32008R1272/FR</a>
<b>Term Note</b>	Conseils de prudence — Prévention
<b>EN</b>	
<b>Precautionary statement</b>	Do not spray on an open flame or other ignition source.
<b>Reliability</b>	★★★★
<b>Term Ref.</b>	Annex IV, Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006, <a href="#">32008R1272/EN</a>
<b>Term Note</b>	Precautionary statements — Prevention

Figure 8: Precautionary statements.

**Alphabetical search**

EN - English    Pictogram    corrosion

Feedback

Other languages: [BG](#) [CS](#) [DA](#) [DE](#) [EL](#) [ES](#) [ET](#) [FI](#) [FR](#) [HU](#) [IT](#) [LT](#) [LV](#) [MT](#) [NL](#) [PL](#) [PT](#) [RO](#) [SK](#) [SL](#) [SV](#) [All](#)

<b>Domain</b>	Chemical legislation, CLP
<b>Code</b>	GHS05
<b>Pictogram</b>	
<b>Related</b>	<a href="#">acrylamide</a> , <a href="#">ammonium bichromate</a> , <a href="#">ammonium chromate</a> , <a href="#">ammonium dichromate</a> , <a href="#">chromic acid [H2Cr2O7] diammonium salt</a> , <a href="#">diammonium dichromate</a> , <a href="#">dichromic acid diammonium salt</a>
<b>EN</b>	
<b>Pictogram</b>	corrosion
<b>Reliability</b>	★★★★
<b>Term Ref.</b>	Annex V, Regulation (EC) No 1272/2008 of the European Parliament and of the Council of 16 December 2008, <a href="#">32008R1272/EN</a>
<b>Term Note</b>	Physical hazard; health hazard

Figure 9: pictograms in ECHA-term (corrosion).



## 5.4 A User Perspective

In 2012, the European Chemicals Agency conducted a survey in order to better understand to what extent the ECHA-term users were satisfied with the glossary and who these users were. With around 3200 visitors per month and about 300 search queries per day, ECHA-term is obviously a much more 'confidential' database than IATE. Yet, contrary to original expectations, only 22% of the users who responded to the survey are translators. The majority (56%) of the respondents indicated that they actually work in the chemical industry, the remainder working for EU Member States, for international organisations and for NGOs such as environmental protection agencies. The vast majority of the users indicated that the database helps them understand the REACH and CLP Regulations and stressed that the terms were relevant to them. The multilingual nature of the database is considered a key feature for the users, who mainly visit the ECHA-term web site in order to look up the translation of specialized terminology. A feedback mechanism also allows users to suggest new terms (around 100 terms are added every year) and to provide comments about existing entries. Users can also download the entire database.

Such results explain why this project has been considered a success. They show that a small, but well-maintained glossary can make a difference for the Agency's stakeholders, who use it on a daily basis. It contributes to the use of a unified and consistent terminology in all the translations related to the regulations in the chemical field, a *sine qua non* in multilingual communication.

## 6 Conclusion

I started this paper with a reference to ten Hacken's remark that "dictionaries are not descriptions of a language, but tools with which users of the dictionary solve problems of a particular type". The nature of the dictionary therefore determines which types of problems it can solve, he argued in his 2012 Euralex paper. The same is true of terminology databases, as we have seen. The use of such databases in today's globalized economy and multilingual world accounts for the nature of the linguistic information included in these electronic resources. Translations, definitions, acronyms, subject field labels, usage notes and examples are similar to what can be found in monolingual or bilingual dictionaries. Some other types of information are used somewhat differently or to a larger extent in terminology databases, however. References and reliability codes are crucial in term bases, although they are virtually absent in traditional monolingual dictionaries, even if historical dictionaries such as the OED do make use of reference information, an essential element when sketching the historical development of a given lexical item. Information about the author of a term entry is also important in term bases, given that terminology is frequently associated with standardized language used by specific communities of users.

Metalinguistic labels, which are not different from prescriptions, as is pointed out by ten Hacken (2012: 843), are found both in traditional dictionaries and in terminology databases. Labels such as *nonstandard*, *preferred* or *obsolete* reflect the lexicographer's or the terminologist's attempt to capture a judgment which will be exploited by the end user of the linguistic resource. This label will help the user decide whether it is pragmatically appropriate to use a given form (e.g. a translation), depending upon the context, the nature of the document produced, the client for whom the translation is made, etc.

The distinction between terminological items and lexical items is also more and more blurred. The nature of the linguistic items discussed by terminologists has undoubtedly evolved over the last 10-15 years. The inclusion in specialized electronic glossaries and term bases of items such as modal auxiliaries, complete sentences, collocational or phraseological patterns, images, diagrams and pictograms is driven by the needs of the target users and the requirements of modern multilingual communication. In this respect, the road from lexicography to terminology is more a continuum, a cline, rather than a hard-and-fast dichotomy.

Another issue which will also need to be addressed in the future is the level at which terminology should be managed (see also Fontenelle and Rummel, in press). Should terminology management be centralized or should it rather be done at the local level, down to the level of individual translators in big translation services? How then should the data be made available to its users? Clearly, web technologies have made it possible to disseminate terminological knowledge to millions of users (the publicly available version of IATE received 44 million queries in 2013). However, one of the major stumbling blocks in the dissemination process is that it is still up to the individual translator or user to 'suspect' that a term base such as IATE or ECHA-term is able to provide interesting and useful information about a given term. What is therefore needed is a mechanism which can alert a translator that a word or a sequence of words appearing in the source text she is dealing with corresponds to a term entry in a specialized database for which an equivalent exists in the target language. Such tools exist at the local level, but will need to be linked to huge databases like IATE, without forcing the translator to host a local copy of a 9-million-term database, which is not recommended for obvious performance reasons. A number of initiatives are currently under way to tackle this crucial issue. Another burning issue is also related to the use of term checkers which ensure that only recommended (read 'validated') terminology is used and that 'dispreferred', obsolete or deprecated terms are not used by the translator. Once again, such obstacles require some level of linguistic processing to match inflected forms in a text and the canonical forms recorded in the quality assurance mechanisms. Organization challenges are also at stake here, since it is crucial to determine who is doing what. Such challenges are different from the question revolving around the distinction between terminological and lexical items, but they are equally important from a user's perspective. Should translators themselves take care of the terminological work? Where should they capture the preferences expressed by the "clients"? Should it be done centrally or locally? How can we make sure these preferences are not one-off information, but can be recycled in future translations to avoid repeating the same mistakes?



These questions do not seem to have clear-cut answers: what is clear, however, is that the solutions can only be effective if they combine technological innovation, using the appropriate amount of linguistic processing, together with organizational changes to make the best use of what modern technology can offer to language workers.

## 7 References

- Fontenelle, Th. and Mergen, C. (1998): « Les interfaces terminologiques au Service de Traduction de la Commission européenne », *Terminologie et Traduction*, 1.1998, Commission des Communautés Européennes, Luxembourg, 210-221.
- Fontenelle, T. (in press) 'Bilingual Dictionaries'. In Durkin, P. (ed.) *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.
- Fontenelle T., Rummel, D. (in press) 'Term banks'. In Hanks, P., De Schrijver, G.-M. (eds) *International Handbook of Modern Lexis and Lexicography*. Springer.
- Jacquemin, C. and Bourrigault, D. (2003) 'Term extraction and automatic indexing', in Mitkov, R. (ed.) *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press. 599-615.
- Pearson, J. (1998) *Terms in Context*. Studies in Corpus Linguistics. John Benjamins Publishing Co. Amsterdam.
- Reichling, A. (1998) 'Gestion centrale de la terminologie, EURODICAUTOM et ses outils satellites'. *Terminologie et Traduction*, 1.1998, Commission des Communautés Européennes, Luxembourg, 172-201.
- Rey, A. (1992) *La terminologie - Noms et notions*. 2ème édition. Collection « Que sais-je ? ». Presses Universitaires de France, Paris.
- Sager, J. (1990) *A Practical Course in Terminology Processing*. John Benjamins Publishing Company. Amsterdam.
- Simpson, J. (2000) 'Preface to the third edition of the OED'. Accessed at: <http://www.oed.com/public/oed3preface/preface-to-the-third-edition-of-the-oed>. [05/04/2014]
- Ten Hacken, P. (2012) 'In what sense is the OED the definitive record of the English language?', Fjeld, R., Torjusen, J. M. (eds) *Proceedings of the 15th EURALEX International Congress*, University of Oslo, 834-845.
- Ten Hacken, P. (2009) 'What is a Dictionary? A View from Chomskyan Linguistics'. *International Journal of Lexicography*. 22.4: 399-421.
- Wright, S. E. and Budin, G. (2001) *Handbook of terminology management (Volume 2): Application-oriented terminology management*. John Benjamins Publishing Company.



# Natural Language Processing Techniques for Improved User-friendliness of Electronic Dictionaries

Ulrich Heid  
Universität Hildesheim, Germany  
heid@uni-hildesheim.de

## Abstract

We discuss Natural Language Processing (NLP) tools and techniques which may be used to enhance the user friendliness of electronic dictionaries. Intended properties of electronic dictionaries on which we focus are improved guidance for text production, as well as easy and efficient access to lexical data in text reception dictionaries. In this talk, we focus on those NLP techniques which are mostly available for the major European languages: morphological analysis of inflection and word formation, as well as syntactic analysis. We also address the relevance of a detailed classification and representation of lexical data categories within the dictionary: this is a central prerequisite for any integration of dictionaries and NLP tools. Our discussion is embedded in an interpretation of the claims of the lexicographic Function Theory with respect to user orientation in dictionaries.

**Keywords:** electronic dictionaries; NLP tools and techniques; user orientation

## 1 Introduction

In this article, we give a short overview of existing and possible applications of Natural Language Processing (NLP) that could be used to enhance the user friendliness and usability of electronic dictionaries. Enhancing user friendliness in this context means providing better guidance in text production dictionaries, as well as improving access to the data provided by reception dictionaries. In the long term, we envisage integrated lexical information systems that combine a dictionary with a number of Natural Language Processing components.

We first situate our discussion in the framework of our view on the lexicographic Function Theory (cf. e.g. Tarp 2008), and we summarize those aspects of the Function Theory that are directly relevant for the integration of language processing and dictionaries (section 2). As the internal representation of lexical and lexicographical data is a key element in the interaction between the lexicographic and the language processing components, we devote a short section to the issue of data categories and markup (section 3). In section 4, we then discuss existing and likely upcoming language processing devices for text production dictionaries (section 4.1) and for text reception dictionaries (section 4.2). Before we conclude, we address a few general design issues and questions related with the presentation of language processing results to the user (section 5).

## 2 User Orientation in Dictionaries

Requesting user friendliness of printed and in particular of electronic dictionaries has almost become a common place of lexicographic theory, but also of the advertisements for dictionaries. An example from metalexigraphy is the lexicographic Function Theory (cf. e.g. Tarp 2008) which places the user and his needs in the centre of its reasoning about dictionary design.

### 2.1 Metalexigraphic Viewpoint

As stated by Tarp (2008), dictionaries are to be seen as utility tools. The dictionary is a (possibly network-like) set of structured texts from which a user may be able to extract textual data that allow him, by means of an interactive interpretation, to derive information. The user will go through this interpretation process in response to a need that arises from a non-lexicographic situation. Tarp classifies these needs into several types; the needs immediately relevant to the discussion in this paper are either cognitive or communicative in nature. Cognitive needs arise in situations where the user wants to know about or to learn certain facts, be they about things, concepts or words. Communicative needs arise in (the preparation for) communication activities, i.e. text reception (reading or hearing – and understanding) or text production (writing or speaking – and lexical or grammatical choice). Translation, the revision of texts in a foreign language etc. are also communicative situations, and thus translation towards the mother tongue is a receptive activity, and translation to a foreign language is a production-oriented one.

Dictionaries are supposed to provide appropriate data for users to satisfy needs of the above types. An ideal dictionary, according to the lexicographic Function Theory (henceforth: FT), satisfies exactly one type of need. In terms of FT, the “dictionary function” is to satisfy such a need, and the optimal dictionary is monofunctional. While this ideal is hardly ever commercially viable in printed dictionaries, it can be approximated in electronic dictionaries (cf. e.g. Bergenholtz/Bergenholtz (2011), for an exemplification).

The development of dictionaries, be they printed or electronic, is typically governed by lexicographic processes. These have in the past been exclusively geared towards the production of paper dictionaries, but since the advent of electronic dictionaries (or electronic versions of print dictionaries) they may also be more general (cf. Gouws, to appear), aimed at setting up a repository of lexicographical data that can be used in both an electronic and a print dictionary. We situate the discussion of the use of Natural Language Processing (NLP) tools and techniques in a scenario which is aimed at such a possible double or even multiple use of lexicographic data.

---

1 For reasons of practicality, we use the masculine form throughout this paper, meant to denote both genders.

## 2.2 User Orientation in Electronic Dictionaries

In the above sense, we assume that lexicographers collect data to feed more than one dictionary; or, conversely, that not all collected data will show up in one given dictionary. Rather, most dictionary publishers create a broad repository of lexicographic data from which they will select appropriate items for individual dictionaries. This notion is close to the idea of a “mother dictionary” that feeds into several specific dictionaries, a concept introduced into the discussion by R. H. Gouws. Implicitly, the same idea is present in work by the proponents of the Function Theory: to avoid overloading the user, at query time, with (unnecessary) data, they require lexicographers to carefully select the data categories they want to present to the user for a given dictionary function.

While this requirement is very clear at an abstract, metalexigraphic level, the way in which it can be satisfied in practice is described in the Function Theory in much less detail (cf., however, the tables given by Tarp 2008: 75-77).

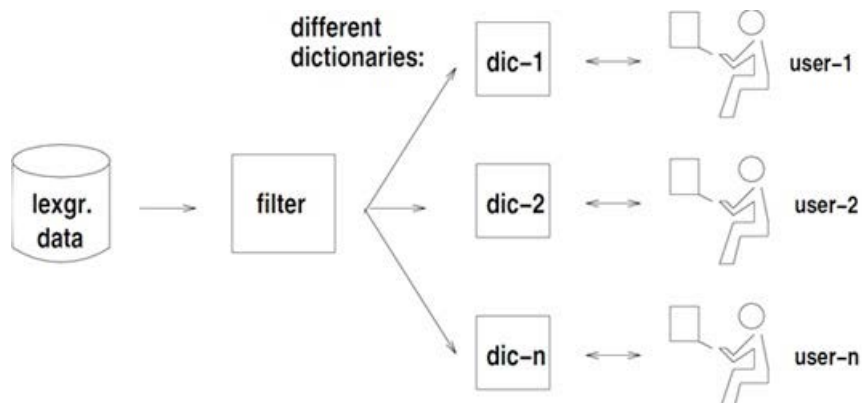
In a scenario where electronic dictionaries are to be produced from an electronic data collection, providing users with data appropriate for a given need involves filtering the contents of the data collection. In the spirit of the well-known distinction between lexicographic data description and lexicographic data presentation, filtering has the following aspects:

- (1) selection of data categories relevant for a given dictionary function;
- (2) selection of presentational properties appropriate for the targeted user public, in terms of the ordering of microstructural items, of their layout, metalanguage, presentational devices, and of the provision of appropriate access routes to the data.

Both of the above selections depend crucially on the following factors:

- (1) the dictionary function;
- (2) the pre-existing knowledge of users, in terms of the language (or languages) dealt with in the dictionary, as well as of general aspects of dictionary use (Tarp 2008) or of the use of online information tools;
- (3) possibly the complexity of the language phenomena described in the targeted dictionary article(s).

The illustration in figure 1 schematically summarizes the relationship between data repository, filtering and user-oriented dictionary versions.



**Figure 1: Scenario of the production of user oriented dictionaries: data repository – filtering – monofunctional dictionaries.**

The definition of the above mentioned filtering criteria, as well as of the selected presentational devices is in principle part of the dictionary concept (“Wörterbuchplan”, in Wiegand’s and Gouws’s terms); it is similar in nature to specifications of a piece of software; the selection of data categories is a contents-related specification, while the definition of presentational devices and of access routes is mainly a matter of the rendering of individual data categories for printed or on-screen presentation. The specification of access to the data is mainly conditioned by the lexicographer’s assumptions about the user’s pre-existing knowledge of the domain or language treated in the dictionary. For example, it is plausible to assume that users in need of collocational data, for text production, will know the base of the collocation (in the sense of Hausmann 2004), and will search for an appropriate collocate to express a given idea. Thus a sort of onomasiological access would be preferred: one that allows the user to search for expressions of ideas around a base concept, irrespective of whether these ideas are expressed by collocations, compounds or single words (cf. Giacomini 2012)

This difference in access is illustrated, in figure 2, below, for a printed dictionary (or a print-like electronic one) with sample data from the OCDSE (Oxford Collocations Dictionary for Students of English) for the lemma *advance*: for text production, the data are sorted according to OCDSE’s principles (right side of fig. 2), i.e. per reading of the base, with subdivisions per syntactic model and semantic groups (cf. also Heid/Zimmermann 2012); for text reception (left side), a semi-integrated microstructure is suggested, with a section on the readings of the base, followed by an alphabetical listing of collocational adjectives (and, later in the article, but not shown in figure 2, of collocational verbs and nouns); in an electronic dictionary, reception-oriented access would be more flexible, i.e. from both elements of the collocation, as well as from the collocation as a whole.

Reception	Production
<ul style="list-style-type: none"> <li>• Readings               <ul style="list-style-type: none"> <li>(1) [military] forward movement</li> <li>(2) development</li> <li>(3) amount of money</li> </ul> </li> <li>• Typical adjectives               <ul style="list-style-type: none"> <li>- Allied etc. (cf. German etc) (1)</li> <li>- big (= considerable) (2)</li> <li>- cash (3)</li> <li>- considerable (= big) (2)</li> <li>- dramatic (2)</li> <li>- German (cf. Allied, etc) (1)</li> <li>- great (2)</li> <li>- important (1)</li> <li>- large (3)</li> <li>- notable (2)</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>Reading 1: forward movement [military]               <ul style="list-style-type: none"> <li>• ADJ + advance                   <ul style="list-style-type: none"> <li>- [speed] rapid ~</li> <li>- [agent] German ~, Allied ~, etc.</li> </ul> </li> <li>• V + advance                   <ul style="list-style-type: none"> <li>- [make] make an ~ on X</li> <li>The regiment made an advance on the enemy lines</li> </ul> </li> </ul> </li> <li>Reading 2: development (often in the plural)               <ul style="list-style-type: none"> <li>• ADJ + advance                   <ul style="list-style-type: none"> <li>- [amount] considerable ~, big ~, substantial ~, dramatic ~, enormous ~, great ~, spectacular ~, tremendous</li> </ul> </li> <li>• V + Advance                   <ul style="list-style-type: none"> <li>- [make] make ~ es (in/on) [plural]</li> </ul> </li> </ul> </li> <li>Reading 3: amount of money               <ul style="list-style-type: none"> <li>• ADJ + advance                   <ul style="list-style-type: none"> <li>- [quantity] small ~, large ~ - [type] cash ~</li> </ul> </li> <li>• V + Advance                   <ul style="list-style-type: none"> <li>- [provide] give so, an ~, pay so, an ~</li> <li><i>The university pays me an advance for thir business trip.</i></li> </ul> </li> </ul> </li> </ul>

**Figure 2: Microstructures for easy access to collocations: for text reception (left) vs. text production (right).**

### 3 Data Representation and User Orientation

For an electronic dictionary scenario which involves a central data collection, monofunctional (or at least function-related) dictionaries and the appropriate filtering techniques, a cornerstone of successful implementation is a detailed classification of the available lexicographic data. If, for example, no distinction is made between collocations and idiomatic expressions, and if bases and collocates of collocations (in the above mentioned sense of Hausmann 2004) are not distinguished and marked up, it will be hard if not impossible to provide appropriately differentiated access to collocations vs. idioms. If, for example, both are classified as “multiword expressions”, it will be hard to decide which items will go into a text production dictionary and which ones into a text reception dictionary. Most lexicographers would however agree that collocations are relevant for text production (but not needed – with the exception perhaps of what Grossmann/Tutin 2003 would call “opaque collocations”, such as FR *peur bleue* – for text reception), while idiomatic expressions would need to be semantically explained in a text reception dictionary, but would rather not be described, for instance, in a learner-oriented text production dictionary.

If the classification of data categories is central, so is the functional markup of different data categories in the central repository used by a publisher. Some authors call this repository a “database” (e.g. Bergenholtz 2011). This may be technically adequate in the implementation described by Bergenholtz

(2011), but in the general case, this repository need not be a database in the technical sense; publishing houses may also use XML-based data models, or a representation within a content management system, or any other implementation: what counts is that different data categories are distinguished and identifiable. There are examples of publishing houses which use the fine-grained data classification of their data repository to “extract” dictionaries for certain functions and user-groups, without much need for adding new data. Such detailed data categorization and “markup” is also necessary if certain subsets of the lexicographical data available to a publisher are to be provided for the purposes of Natural Language Processing (NLP).

A note of caution may be in order here, with respect to the abovementioned example of collocations and idioms. Some lexicographers might rightly assume that users will not be able to distinguish between the two types of multiword expressions, and claim that they don’t need to (cf. e.g. Tarp 2008). This is certainly an appropriate viewpoint, but it does not distinguish the internal representation of lexic(ographic)al data from the presentation of such data to the user. If the abovementioned assumptions about the differences between idioms and collocations, in terms of data selection and access, are correct, an optimal presentation of such data to the user will clearly have to rely on the distinction between the two types of multiword expressions, and on an appropriate markup of each multiword item contained in the data repository. In other words: an optimal presentation of dictionary contents to the user is (trivially) dependent on an adequate internal classification and representation of this contents.

### **3 Natural Language Processing Tools in Support of User Orientation**

In our discussion, so far, Natural Language Processing (= NLP) tools have not been mentioned; a sensible level of user-friendliness can, as has been shown above, be reached without NLP technology, by adhering to good practice in data category classification and markup. A legitimate question is thus what the added value of computational linguistic technology is, in terms of a surplus of user-friendliness of the dictionary. Here, we understand user orientation in a wide sense; it is meant, here, to include aspects of enhanced usability of electronic dictionaries, such as improved access to lexicographic data, individualized support according to the user’s pre-existing knowledge, or the availability of information (on demand) which goes beyond the amount of material encoded in the data repository underlying the dictionary, e.g. by means of the presentation of corpus data.

We will address this question in the following by first analyzing monolingual dictionaries for text production, then dictionaries for text reception. We will not discuss the use of NLP techniques for the provision of corpus data to the lexicographer, i.e. corpus analysis and query tools, such as, for instance, those embodied in the “Sketch Engine” (Kilgarriff et al. 2004). Such tools are essentially aimed at the



lexicographer, and we will show in which way the end user may profit from other kinds of corpus analysis tools.

## 4.1 NLP Tools for Text Production Dictionaries

For text production, especially in a foreign language, a rich microstructure is necessary, for example one which explains to the user for each treatment unit which morphological forms it has, which syntactic patterns it follows, or which collocations it enters into. All these properties involve lists of options from which a user may want to select in a text production situation; while syntax and collocations are idiosyncratic (Hausmann 2004 talks of “coded combinatorics”) and need thus be listed individually, morphological forms are often more regular and may be provided to the user by means of a morphological generator or of a list of inflection forms. The latter is limited and may require regular updates by the lexicographer, whereas a morphological generator may provide the advantage of comparatively easy ways of extending its coverage. In any case, offering users of a production dictionary on-demand access to inflection forms is certainly very useful.

### 4.1.1 Access to Corpus Data

Another possible use of NLP techniques in text production dictionaries has to do with the much discussed facilities that give the user access to corpus data, from a dictionary entry, (cf. e.g. Asmussen 2013; Heid et al. 2012; Tarp 2012). It has been claimed that links from the dictionary to corpus data or to the internet provide users with data about language in use which can serve as a model for the users’ own text production. Such data thus serves the need for checking one’s own formulation hypotheses against putatively “standard” usage.

Asmussen (2013) discusses the issues related with the realization of such links; using the German DWDS dictionary as an example, he shows that the mere coexistence of dictionary and corpus data in one portal is not sufficient to provide adequate service to users. Asmussen has examples of lemmas present only in one of the two sources of information, and he discusses the impossibility of linking corpus data, given today’s technology, to readings of a dictionary entry, at least in a large-scale high-quality way. Asmussen’s best example of the linking of dictionary and corpus data is from the domain of collocations. In fact, the portal of [ordnet.dk](http://ordnet.dk) provides direct access to usage examples for collocations, with example sentences retrieved from Korpus 2000, the current Danish corpus underlying the DDO dictionary (Den Danske Ordbog, cf. [ordnet.dk](http://ordnet.dk)). To activate the link (in the sense of an information-on-demand offer), the user has to press a button next to the collocation item in the dictionary. Technically, this activates a query in the corpus which is created from the (text form of) the collocation item (cf. Heid et al. 2012).

This facility could be further enhanced if the corpus were preprocessed at a more advanced level of linguistic analysis (e.g. by means of (flat) syntactic analysis), and if the query were made more sensitive to potentially ambiguous corpus sentences. For example a search that starts from the collocation

give + resultater (EN “produce/lead to results”) provides many relevant examples, e.g. gav betydelige resultater (“gave important results”), gav de ønskede resultater (“gave the intended results”); but it also provides gav 10.3 km/l til resultat (“gave 10.3 km/l as a result”), which is not an example of the searched collocation.

Other, similar facilities involve the retrieval of contexts for specialized terms in specialized texts, e.g. in the intranet of a company, or lists of cooccurents of items (for the user to choose from) sorted by association strength, as they are provided within Verlinde’s ILT tool (URL: <https://idp.kuleuven.be/idp/view/login.htm>).

On the basis of syntactically analyzed and annotated texts, the same device could also be offered for syntactic subcategorization. Parsed corpus data tend to identify subjects, objects, prepositional complements, verb-dependent clauses or infinitivals in each analyzed sentence; this is true for dependency parsing, which has reached, at least for several European languages, a degree of maturity makes its use in the intended context possible (cf. e.g. Bohnet 2010). Syntactic valency is, as mentioned above, an important property of lexical predicates (verbs, adjectives and nationalizations) that must be learned by foreign language learners. Illustrating valency in full sentences has the advantage of showing the user not only an abstract indication, but also concrete instantiations of it. This principle has been followed, very successfully, in the ELDIT dictionary (Elektronisches Lernerwörterbuch Deutsch-Italienisch, <http://eldit.eurac.edu/>), where the authors have provided the user with four complementary types of indications for the syntactic construction of verbs (cf. Abel 2002):

- (I) a formula of the type “someone suggests something to someone”;
- (II) example sentences for each pattern;
- (III) on-demand indications of the involved grammatical functions (e.g. “object” for “something” in the above example);
- (IV) on-demand highlighting of the respective phrases in the example sentences, when the user points the mouse to an element of the formula (i): if, for example, the user points the mouse to “to someone” in the above example, not only the grammatical function (indirect object) is displayed, but also the respective stretch of the example sentence is highlighted.

In ELDIT, these devices are applied to prefabricated examples, i.e. a closed list of verbs and example sentences for these. By use of NLP tools, such a device could be made dynamic, i.e. provided on demand by the user, on the basis of a pre-analyzed (dependency-parsed) corpus and extraction tools for syntactic patterns. If the corpus is big enough and adequately annotated, also frequency data for individual valency constructions and lists of the most prominent fillers of valency relations could be provided. Finally, as Engelberg et al. (2012) have shown, the different possible valency constructions of verbs are used differently in different genres or text types: not all possible patterns are equally frequent in all kinds of texts; if notions like “genre” or “text type” are applied to the annotation of a corpus which is exploited to offer examples for valency patterns explained in a dictionary, such valency preferences by genre or text type can be made visible.

All this may appear futuristic to some readers; it is, however, only dependent on two conditions: an adequately detailed inventory of valency patterns in the dictionary, and good quality corpus parsing. The device would allow users to get real-text models of syntactic constructions, from which they could take inspiration for their own text production.

Such devices are in principle thinkable for all those linguistic properties of lexical items that can unambiguously be identified in a (syntactically parsed) corpus. With adjective+noun-collocations and, to a lesser extent, verb+object- and verb+subject-collocations, this is well possible; the same holds for syntactic valency, for the contextual use of terms from a specialized language; but it does not yet so for lexical semantic properties. Some partial results could be obtained if additional resources are used, e.g. WordNets that would support a search for word combinations and the pertaining sentences according to semantically defined sets of lexical items. Again, what is needed as a prerequisite, are appropriate classifications of the lexicographic data, mappable onto the classifications annotated in the corpus. In all cases, the combination of lexical items and targeted linguistic properties acts as search criteria for corpus data extraction.

Instead of lexical items and their linguistic properties, also pairs of translational equivalents from an electronic bilingual dictionary may be used as search constraints, in this case on parallel corpora; Verlinde's ILT tools provide access to the Europarl corpus (URL: <http://www.statmt.org/europarl/>), but they use only the source language item as a search criterion, in the hope of providing the user in this way a broad range of equivalence candidates; for very advanced users, and especially for those who are used to work with parallel corpora, this may provide indeed new insight. A more modest, though perhaps more focused (and thus easier to use) version of such search would be one that retrieves example sentence pairs for a given equivalent pair from the dictionary; it may then on demand also provide sentence pairs that do not contain the equivalents mentioned in the dictionary entry, as a complement of information.

#### **4.1.2 Lexico-grammatical Guidance**

With the above mentioned provision of example sentences for a given lexical property of an item from the dictionary, one problem mentioned by Asmussen (2013) still remains: for the dictionary user, the relationship between the text he is in the process of producing, and the sentences retrieved from the corpus, may still be rather indirect; the examples may illustrate the behaviour of the searched item, but they will still not necessarily provide an exact solution to the actual text production problem which the user is confronted with, as the other lexical items he intends to use are absent from the retrieved examples.

While the transfer between the corpus examples and the upcoming text of the user may be relatively simple for most European languages, it is much harder in languages with massive rule-based morphosyntactic variation, where lexical choice and grammatical choice interact in more complex ways. Examples of such situations are provided by the South African Sotho and Nguni languages. The morphosyntactic complexity of the noun class system of these languages, of their concordial and prono-

minal morphemes, as well as of their tense and mood systems interferes with issues of lexical choice that depend on semantic selection criteria. Examples are discussed, among others in this conference, by Bosch/Faaß (2014) and Prinsloo et al. (2014); they analyze possessive constructions in Zulu (type: the medicine of this doctor) and subjects, objects and relative clauses in Northern Sotho (type: the boy who helped the woman), respectively, from the viewpoint of English → Zulu or English → Sotho learner's dictionaries.

Both model the respective phenomena in an NLP tool that implements the morphosyntactic (agreement) rules of the language and interacts with a dictionary whose nominal entries are classified by noun classes and whose verbal entries can be inserted into the grammar of the constructions under analysis. Bosch/Faaß's (2014) system can operate in two modes: one that provides a translation from English to Zulu of the intended possessive construction, and one that in addition explains to the user which construction and agreement rules have been used. A next step could be a system that allows the user to produce his own solution and which then proposes modifications where necessary. Prinsloo's system is not yet implemented but intended to provide similar optional guidance: if, for example, the user plans to construct a Northern Sotho subject-verb-object sentence where the object is not expressed by a noun (phrase), but by a concord, the following situations may occur:

- (I) the user may know the appropriate concord: the system will check the appropriateness and confirm it;
- (II) the user may know the noun which he wants to express by the concord, but not the concord itself: the system will retrieve the noun class of the item (from its standard dictionary), identify the appropriate concord (from morphosyntactic tables) and propose the appropriate concord, possibly with additional information about the underlying grammatical facts;
- (III) the user may only know the English equivalent of the nominal he intends to construct as an object: the system then retrieves the appropriate Sotho noun from a bilingual dictionary and then proceeds as explained under (ii).

In both cases, the objective is text production guidance for learners of the foreign language; the proposed solutions combine some amount of NLP with a well-structured dictionary. Such combined systems may be counted among online language learning systems, or among e-dictionaries. Irrespective of how they are classified, they combine lexicon and (partial) grammar data.

## 4.2 NLP Tools for Text Reception Dictionaries

While the function of NLP tools in the context of text production goes from a source of inspiration (through the selection of appropriate example sentences from corpora) to guidance in issues of lexico-grammatical choice, NLP tools function mostly as access tools in text reception dictionaries. Text reception starts from an existing text and aims at detecting its meaning and possibly other properties. Access support tools ease the user's retrieval of the right dictionary entry and the right indications within this entry. This kind of support starts with inflectional morphology and may involve

word formation, syntax and possibly, at least to some extent, multi-word items and semantics. In all cases, the basic function of the tools is to analyse a word(form), possibly in the context of the sentence the user is reading, and to relate it to an entry or ideally even a reading in the dictionary.

For inflectional morphology, such devices work relatively well and are quite established: if the user enters a word form, the dictionary relates it to the appropriate base form and displays the entry of the pertaining lemma. Such interfaces exist in several online dictionaries, and they may either depend on large lists of inflected forms (related with the appropriate lemmas), or on morphological analysers.

However, similar devices may be used also for word formation: when Bergenholtz/Johnsen (2005) analyzed the log files of their Danish Internet Dictionary (Dansk Netordbog), they discovered that a considerable number of items searched by users, but not found in the dictionary, were word formation products. In Germanic languages, compounding is so productive that a standard monovolume dictionary cannot cover even a small portion of the items found with a non-trivial number of occurrences in a corpus. The same holds, at least to some extent, for derivation products: these also will likely not be covered in full in standard monovolume dictionaries. If the dictionary is meant for text reception, it may even reasonably adopt a policy of focussing on semantically non-transparent compounds and derivations, i.e. on those whose meaning needs to be explained beyond a simple recall of the morphological structure, e.g. because they have idiosyncratic meanings. In such a case, no space may be left for the treatment of transparent compounds.

A morphological word formation analyzer may be useful in such a situation, as it would be able to split a compound that is not part of the nomenclature of the dictionary into its components and guide the user to their entries in the dictionary. For derivation, even generic paraphrases may be given, as is the case in the DériF system for French (URL: <http://www.cnrtl.fr/outils/DeriF/>). Alternatively, a structural or morpheme decomposition hypothesis may be given, as in the canoo tools (URL: <http://www.canoo.net/>). In all cases, the user would get partial information in reply to his query, thus at least some guidance towards the analysis of the word formation product. Often, a recall of the morphological structure and links to the dictionary entries of the components may help users understand complex word.

While the above functions are in general seen independent from the context (see below for a discussion), any kind of tool intended for guidance of text readers towards an appropriate entry in the dictionary will inevitably be confronted with ambiguity and context-based disambiguation. This starts with categorial homographs (EN: can: modal verb or noun, cf. Bothma/Prinsloo 2013: 176) or forms which are homographous with items from other word classes (thought: participle or noun, *ibid.* 174) and goes all the way up to polysemy and reading distinctions. While the latter problems are in the general case still not solvable, categorial homographs of different types can be disambiguated fairly well on the basis of word class tagging and/or syntactic analysis. For the major European languages, tools for these functions are available, also as web services that would allow for an on-the-fly treatment, as suggested by Bothma/Prinsloo (2013: 187 ff.).

If syntactic (dependency) analysis is available, a variant of the search for examples discussed above, in section 3.1, could be applied; if a syntactic analysis of a sentence can be produced, at least the verb and its potential complements can be extracted from the analysis result and matched against the list of valency patterns offered by a dictionary. In many cases, this would reduce the number of readings which the user would need to consult in order to find the meaning of the verb used in the sentence he is reading. Obviously, not always all arguments of a predicate are explicitly mentioned in a sentence, which might increase the number of syntactic readings that have to be looked up by the user in such a situation.

Another type of contextual analysis has been available since the 1990s, already. It deals with (idiomatic) multiword expressions; the objective is to provide the user with the (part of a) dictionary entry that explains the multiword expression, when the user clicks on any of the words that make up the expression. Due to the high level of lexical specificity this device works quite well for idioms (cf. Sereitan/Wehli 2013). In combination with a syntactic analysis of the text which the user is working on, this facility could be extended to collocations as well.

## 5 NLP Tools as a Part of Lexical Information Systems

In the two preceding sections, we have listed a few simple devices based on NLP technology that could enhance the user friendliness of electronic dictionaries. A number of issues should however be discussed in this context which concern more general aspects of user interaction.

An important first aspect concerns the problem of the quality of the tool output; it can not be guaranteed that the linguistic analysis underlying the integrated tool is correct in all cases. The more NLP components are involved, the more possibilities are there for errors to occur in the processing chain. Thus a setup where, for example, the text being read by the user is syntactically analyzed and then searched, in order to find the appropriate dictionary entry for a given item, may provide fully correct results only in a certain percentage of cases (we would assume at least in three quarters of the cases).

The dictionary developer and the users should be aware of this situation, and such a device should be offered as experimental (cf. Tarp 2012 on this issue in the context of corpus data provision). It would need to be presented to the dictionary user as being fully automatic and not cross-checked by the lexicographer. A warning in the user interface would be appropriate in such a case (e.g. 'N.B.: the following results were automatically produced and have not been checked by lexicographers'). This type of warnings is regularly produced by the canoo morphology system when analyses are displayed which have not been cross-checked by canoo's lexicographers.

Furthermore, the results should be shown in a part of the interface that is clearly recognizable as a part of the dictionary or of the lexical information system. Early realizations, especially of tools that link dictionary and corpus, did not make clear enough to the user that the corpus data shown on

screen were meant as an additional service of the dictionary (cf. Bank 2012 on an early version of the Base lexicale du français).

Another, related issue concerns the status of NLP-based services within a dictionary system. In our view, they should always be information-on-demand services: the user should have the possibility to explicitly decide in favour (or against) the use of the NLP-based service. In dictionaries with (personalized) user profiles, this choice may be an element of the user profile.

Finally the design of the overall integrated system should also be governed by the principles of user-oriented dictionary design: for each NLP component to be integrated, the lexicographer should assess which user need (and thus: dictionary function) it satisfies.

## 6 Conclusion

We have shown a few devices for the enhancement of text production and text reception dictionaries that are based on Natural Language Processing tools. While morphology systems or morphological tables are almost a standard component of current electronic dictionaries, this is much less so with tools for syntactic analysis, and technologies for semantic processing are still in an experimental phase, although some are very promising (cf. e.g. Cook 2014).

We have tried to show that a detailed classification of the data categories contained in an electronic dictionary (or in the data repository underlying it) is a major requirement for any work with NLP tools; this is due to the fact that these data categories (and a common “understanding” about them, between NLP tools and lexical repository) are the interface between the two components.

In our view, syntactic dependency - parsing has reached a degree - of maturity, at least for some European languages, which should allow for its experimental - and perhaps also productive - integration into lexical information systems of the kind discussed here. Prototypes of such systems should be built and tested with users.

If lexicographers join forces with NLP experts, and if they jointly produce integrated systems that present an added value, chances are good that users will accept these systems. Since people are particularly demanding with respect to the quality of language tools (and since dictionaries have a high reputation in this respect), offering integrated services as information-on-demand seems to be an adequate solution.

## 7 References

- Abel, A. (2002). Darstellung der Verbvalenz in einem elektronischen Lernerwörterbuch Deutsch - Italienisch (ELDIT). Neue Medien, neue Ansätze. In: Braasch, A. et al. (eds.): EURALEX-2002 Proceedings, pp. 413-418.



- Asmussen, J. (2013). Combined products: Dictionary and corpus. In: Gouws, G.H., Heid, U., Schweickard, W., Wiegand, H.E. (eds.) (2013). *Dictionaries. An International Encyclopedia of Lexicography*. Volume 5.4. pp. 1081-1090.
- Bank, C. (2012). Die Usability von Online-Wörterbüchern und elektronischen Sprachportalen. *Information - Wissenschaft & Praxis*. Volume 63/ 6. pp. 345-360. Accessed at: <http://www.degruyter.com/view/j/iwp.2012.63.issue-6/iwp-2012-0069/iwp-2012-0069.xml?format=INT> [28/05/2014].
- Bergenholtz, H. (2011). Access to and Presentation of Needs-Adapted Data in Monofunctional Internet Dictionaries. In: Fuertes-Olivera, P. A., Bergenholtz, H. (eds.): *e-Lexicography: The Internet, Digital Initiatives and Lexicography*, pp. 30-53.
- Bergenholtz, H., Bergenholtz, I. (2011). A Dictionary Is a Tools, a Good Dictionary Is a Monofunctional Tool. In: Fuertes-Olivera, P. A., Bergenholtz, H. (eds.): *e-Lexicography: The Internet, Digital Initiatives and Lexicography*, pp. 187-207.
- Bergenholtz, H., Johnson, M. (2005). Log Files as a Tool for Improving Internet Dictionaries. In: *Hermes*, (34), pp. 117-141.
- Bohnet, B. (2010). Very High Accuracy and Fast Dependency Parsing is not a Contradiction. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, Beijing, China.
- Bothma, T. J. D., Prinsloo, D. J. (2013). Automated dictionary consultation for text reception: a critical evaluation of lexicographic guidance in Kindle e-dictionaries. *Lexicographica: International Annual for Lexicography*, Volume 29, pp. 165-198.
- Bosch, S., Faaß, G. (2014). Towards an integrated e-dictionary application - the case of an English to Zulu dictionary of possessives. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/Bozen (EURAC).
- Canoo. Accessed at: <http://www.canoo.net/> [27/05/2014].
- Cook, P., Rundell, M., Han Lau, J., Baldwin, T. (2014). Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/Bozen (EURAC).
- DériF. Accessed at: <http://www.cnrtl.fr/outils/DériF/> [27/05/2014].
- ELDIT - Elektronisches Wörterbuch Deutsch - Italienisch. Accessed at: <http://eldit.eurac.edu/> [27/05/2014].
- Engelberg, S., Koplenig, A., Proost, K., Winkler, E. (2012). Argument structure and text genre: cross-corpus evaluation of the distributional characteristics of argument structure realizations. *Lexicographica: International Annual for Lexicography*, Volume 28, pp. 13-48.
- Fuertes-Olivera, P. A., Bergenholtz, H., Nielsen, S., Niño Amo, M. (2012). Classification in lexicography. The concept of collocation in the Accounting-Dictionaries. *Lexicographica: International Annual for Lexicography*, Volume 28, pp. 293-307.
- Fuertes-Olivera, P. A., Bergenholtz, H. (eds.) (2011). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. London: Bloomsbury.
- Giacomini, L. (2012). An onomasiological dictionary of collocations: mediostuctural properties and search procedures. *Lexicographica: International Annual for Lexicography*, Volume 27, pp. 241-267.
- Gossmann, F., Tutin, A. (eds.) (2003). *Les collocations: analyse et traitement*. Amsterdam : De Werelt.
- Gouws, R. H. (2013). Aspekte des lexikographischen Prozesses in Print- und Onlinewörterbüchern. To appear. *OPAL (IdS Mannheim)*.
- Hausmann, F. J. (2004). Was sind eigentlich Kollokationen? In: Steyer, K. (ed.): *Wortverbindungen - mehr oder weniger fest*. Jahrbuch des Instituts für Deutsche Sprache 2003. Berlin/New York: De Gruyter, pp. 309-334.
- Heid, U., Prinsloo, D. J., Bothma, T. J. D. (2012). Dictionary and corpus data in a common portal: state of the art and requirements for the future. *Lexicographica: International Annual for Lexicography*, Volume 28, pp. 269-289.



- Heid, U., Zimmermann, J. T. (2012). Usability testing as a tool for e-dictionary design: collocations as a case in point. EURALEX-2012 Proceedings, Oslo, Norway, pp. 661-671.
- ILT Tools. Accessed at: <https://idp.kuleuven.be/idp/view/login.htm> [27/05/2014].
- Kilgarriff, P. R., Smrz, P., Tugwell, D. (2004). The Sketch Engine EURALEX-2004 Proceedings. Lorient, France, July: pp. 105-116.
- Prinsloo, D. J., Bothma, T. J. D., Heid, U. (2014). User support in e-dictionaries for complex grammatical structures in the Bantu languages. 16th EURALEX International congress, 15-19 July, 2014, Bolzano/Bozen (EURAC).
- Seretan, V., Wehrli, E. (eds.) (2013). Context-sensitive look-up in electronic dictionaries. In: Gouws, G.H., Heid, U., Schweickard, W., Wiegand, H.E. (2013). Dictionaries. An International Encyclopedia of Lexicography. Volume 5.4. pp. 1046-1053.
- Trap-Jensen, L. (2013). Researching lexicographical practice. In: Jackson, H. (ed.): Bloomsbury Companion to Lexicography, pp. 35-47. London: Bloomsbury Academic.
- Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography. Tübingen: Max Niemeyer.
- Tarp, S. (2012). Online dictionaries: today and tomorrow. Lexicographica: International Annual for Lexicography, Volume 28, pp. 253-267.



# Using Mobile Bilingual Dictionaries in an EFL Class

Carla Marello  
Università di Torino  
carla.marello@unito.it

## Abstract

Are reference skills of native digital EFL students developed enough to take advantage of having a big bilingual dictionary on their smartphones? To carry out this research a class from an Italian technical high school was observed. Students aged 17 were split into three groups of users and were given different versions of the same bilingual dictionary, the Ragazzini Italian and English dictionary (Zanichelli Bologna 2013 edition). The first group was allowed to use the Android app; the second was given access to the online version on the web portal [ubidictionary.zanichelli.it](http://ubidictionary.zanichelli.it); the third group received paper copies of the dictionary.

Students were asked to answer some questions about their (un)familiarity with Italian monolingual and bilingual dictionaries. Then the three groups carried out the same activities during a two hour-English-lesson.

The case study reports similarities and differences in their performances, showing that linguistic proficiency proved more determinant than access to digital versions. Students were also asked to report on the main difficulties they had to overcome when looking words up in the dictionaries and were invited to suggest possible improvements in the way pieces of information were displayed on the mobile digital version.

**Keywords:** bilingual dictionary use; mobile dictionary app; advanced CLIL prerequisite

## 1 Bilingual Dictionaries in EFL Classes

In recent years whether English- L1 bilingual dictionaries should be banned from the EFL classroom has become an issue of debate (cf. Butzkamm 2009). This is the natural consequence of the reassessment of students' use of L1 to check the final results of a reading comprehension dealing with progressively more complex concepts and reasoning. In fact, an increasing number of students with B1 levels of English or higher are being asked to read and understand such texts. It is also the result of a dispute over binding defined as the 'cognitive and affective mental process of linking a meaning to a form' (Terrell 1986: 214) and not to a translation. Studies on language acquisition in learning situations that are at least partially guided have shown that interference from L1 (whether it is the real mother tongue and language used at school) and interference of a L2 studied previously and typologically similar to the foreign language being learned, involves more than the insertion of isolated words

in a foreign language text. Above all, research on learning use of verbs in directed motion events in verb-framed languages such as Spanish or Italian by native speakers of satellite-framed languages such as English are revealing that meaning often has a form in L1 that students carry over into words from L2 (See Cadierno 2008). Quadripartition of a linguistic sign as formulated by Hjelmslev (1961) may be (and has been) neglected in studies on language teaching of beginners and intermediate level students, in which the semiotic triangle “meaning-form-reference” is conveniently considered sufficient. Its inadequacy for structures used to make complex references and needed even by beginners, for example the expression of modes, is bypassed by memorization of conversational routines. At advanced levels, however, learners deal with texts in which they are really confronted with the diverse mode with which natural languages form the substance of content, not only for isolated words, but for morphosyntactic structures as well. At that point, teaching English relies upon learner’s dictionaries and their increasing effectiveness, resulting from studies based on learner corpora, usage notes and, more recently, on even more refined and frequent-mistake-based collocation dictionaries for learners of English.

Studies on teaching/learning languages other than English have never shown a rigid refusal to use bilingual dictionaries as a reference tool. The explanations can be found in the typological characteristics of the languages<sup>1</sup> and, therefore, in a teaching tradition that focuses (focused?) more on morphology and less on rote learning of word sequences. Studying them at intermediate and advanced levels has, even recently, been largely based on increasingly more effective bilingual lexicographic tools to describe verb patterns and rich in derivatives and phraseology. They are languages that linguists have dealt with extensively in terms of their collocations later because they provide less fixed, more easily split and variable collocations. Linguists truly appreciated the restrictions of the collocations in Romance languages by examining the concordance of the corpora, which were developed after the corpora of English. Italian lexicographers, for instance, recorded collocations first in bilingual and then in monolingual dictionaries, where today they are still found mainly in the examples, and therefore, do not receive the attention they deserve. In addition to the systematic characteristics a natural language has and the traditions tied to its didactic standards, the position a language holds in the global language market needs to be considered. The English only EFL classroom derives mostly from economic and political reasoning (see Cook 2010): a world market stimulates the creation of reference tools which can be sold everywhere. Attention to the user of a specific L1 is often entrusted to local publishing houses which “bilingualize” products by Anglo-American publishers for foreign learners of any mother tongue.

---

1 We do not deal with this topic here because it would lead, among other things, to a discussion on macro- and microlexicographic structures in monolingual and bilingual dictionaries of languages morphologically poorer – like English – compared to the macro – and microlexicographic structures of morphologically richer languages, like the Romance and Slavic languages: richer due to their visible differentiation of parts of speech as well as word formation with prefixes and suffixes. See, however, below in § 4.1 the reference to the question in regards to the problems encountered by the students when using the dictionary.

If an English learner's dictionary is "bilingualized", however, the different ways the substance of content is organized is not always evident to the learner: the entry's organization into the various meanings and examples retains the style formulated in English by English speakers (see Marellò 1998). A bilingual dictionary, however, set up using good monolingual microstructures is more likely to highlight the differences. For languages used less internationally, often monolingual dictionaries for foreigners are not produced and bilingual dictionaries are recommendable tools, not only for translating, but also for understanding a text in L2. They help to appreciate the different structures and grasp connotations while giving comparative information that the monolingual dictionary, formulated with the native speaker in mind, deemed unnecessary to the user. In addition, bilingual dictionaries for languages with relatively developed markets and a consolidated lexicographical tradition have reached good levels in the last few decades, thanks to the progress of studies on meta-lexicography, the data available in large corpora, dedicated editing software that makes it possible to check word classes and to verify the good quality of reverse translation routes<sup>2</sup>.

It remains to be seen whether the moment has arrived to allow bilingual dictionaries into an EFL class, even though English has very well developed monolingual lexicographical tools.

## 2 Bilingual Glosses enter EFL Learning through Mobile Devices

In the context of the 'connected' classroom online dictionaries and translation tools are readily available. Even if bilingual paper dictionaries are banned from the classroom, and use of online dictionaries is not practiced, teachers and researchers would have to admit that tablets and smartphones provide access to translation tools everywhere else and young people, in particular, use them frequently. Augustyn (2013) allowed her USA university students engaged in a first-year text-based approach to German to use whatever device they preferred to work with the electronic materials that were made available to them. She noticed that some learners rely entirely on translation because they choose to type every utterance they do not understand in L2, or want to produce in the L2, into a translation tool such as Google Translate (2013: 367-368). Augustyn concludes her paper with the following consideration: "The groundwork for bilingual practice and contrastive analysis for vocabulary acquisition has already been done by lexicographers. Maybe the convergence of a reassessment of translation in the context of SLA theory and foreign language pedagogy, on the one hand, and an increasing dependence on online learning tools and digital media will introduce the language learner to the bilingual dictionary in digital format?" (Augustyn 2013:381). It is now time, then, for the foreign

---

2 In Ragazzini 2013 for Engl. *endeavour* we find It. *sforzo*, *tentativo*, but in the Italian®English section we do not find *endeavour* as a translation of It. *sforzo* or It. *tentativo*. Through this unsuccessful reverse translation the user might conclude that the two Italian words, though precious to roughly understand what *endeavour* means, perhaps should be avoided in a written translation of the passage where the English word appears.

language teacher, and especially for teachers of English which has the most information online, to decide whether the use of L1, at least for comprehension purposes, should continue to be avoided. With the forthcoming introduction of CLIL<sup>3</sup> in Italian secondary schools a decision should be made, because such a form of teaching implies extensive reading and the development of literacy skills in addition to communicative skills. The experiment described in this paper has been carried out to verify if the digital natives are able to use a digital bilingual dictionary covering a vast vocabulary and numerous translations of specialized meanings.

### 3 The Ideal Class for the Test

A 4<sup>th</sup> year class was chosen for the experiment from an Italian technical high school, specializing in electronics and made up of 17 year old boys, four of which were not native Italian speakers but had been living in Italy for many years. It was important that learning English had a practical value for the future (professional as well) of the students involved. It was also important that they had no experience consulting reference materials, but were able to judge the pros and cons of having a dictionary online or on their smartphone. The teacher said that she had never used bilingual or monolingual dictionaries for work in the classroom because she did not use translation to teach, but used reading comprehensions and the creation of concept maps in English. Her teaching method focuses on the overall content of the message more than on its form.

About two weeks before the test, the class was given a pre-test to verify the size of vocabulary in reception activities at the B1 level and the ability to produce a derivative with prefixes and suffixes<sup>4</sup>. The instructions for one exercise (see Appendix 1) said “Use the word in the box to form a word that fits in the text” and asked students to use *produce* or *equip* to form the words *product* and *equipment* needed in the context. Particularly noteworthy is the fact that no one in the class wrote *useless* in the 6<sup>th</sup> blank, whose context was defined over two sentences: “The instruction booklets are always (6). They never help me at all”. One of the answers given, the word *unuseful*, is particularly interesting because it is used much more rarely than *useless*.

The other exercise was a cloze test with eight blanks. The students could choose the correct word out of the four proposed. The three distractors given for each gap were semantically similar.

---

3 Content Language Integrated Learning: teaching a subject in L2, usually English or French, will be implemented in Italian schools in the last year of secondary schools as of 2015; in high schools specialized in foreign language learning it will begin in the third year. It may be noted, however, that experiments with what is also called *dual learning* are widely implemented, especially in private schools, beginning in primary school.

4 For the more interesting parts of the experiment for language learning purposes and reference skills see the contributions in progress by Elisa Corino and Elena Martra.

Based on the results of the pre-test, it was decided that on the day of the test rather simple tasks entailing reference skills would be given<sup>5</sup>. The teacher divided the class into three groups: 5 that had a smartphone which could download the bilingual Ragazzini dictionary application, 5 who would use a paper copy and the remaining 7 who would consult an online version<sup>6</sup>. Most of the students (12 of 17) stated they had studied English for over 8 years. Three of the 4 foreign mother tongue students said they had been studying English for less than 8 years. They all owned a dictionary of their mother tongue, except the Philippine student who said he owns a bilingual Philippine-Italian dictionary. From the data taken from the research, 15 students (of which 3 out of the 4 not native Italian speakers) admitted owning a bilingual dictionary; 9 said they used it only for school; 1 student (Italian student B), moreover, indirectly stated that he preferred the online dictionary since he admitted using a paper dictionary only when he did not have an internet connection.

Only 2 students (Philippine student C, Italian student Q) said they had no Italian-English bilingual dictionary; 2 of the 3 non native Italian students who said they owned one, said they did not use it.

#### 4. The Test: a Description

The main objective of the experiment was

- a) to verify if students today, digital natives, are able to use the electronic version of a bilingual dictionary for reception activities without being trained to do so and
- b) if the online version or app really help them more than a paper dictionary.

We prepared five English sentences taken and slightly modified from examples contained in the glosses of the entry **to break** in the Ragazzini dictionary. One sentence was constructed in the same way for to serve and another one for the word **fast**. The verb to **break** was chosen because in Italian it gives the form in - **si** for the corresponding use of the English intransitive and has a vast phraseology. The instructions asked students to do something completely new for them “Translate the following sentences.” It also asked them to indicate which part of the gloss they used to identify the correct translation (See Appendix 2)

“Translate the following sentences.”

“**Which entry and which part of the dictionary entry** did you look up to translate the words in bold?”

An example was provided of what had to be done.

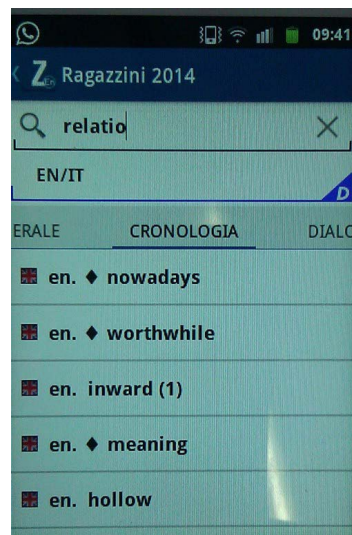
In any case, it was necessary to explain to the class what an entry was.

---

5 The students with the best results were C, L, N, O, Q. C and L are not native Italian speakers and most of their errors were spelling mistakes, for ex. the suffix *-ful* written with two ll.

6 The publishing house Zanichelli generously provided the class with 5 apps, access to the online dictionary and paper copies of the dictionary.

Despite the bold character and underlining, students often disregarded the second part of the request; however, behind each student was an observer that took notes of the ways they consulted the dictionary. Only the app version, and not the one online, allowed the researcher or teacher to access a type of log file using the “chronology” of the words searched (see fig. 1)



**Figure 1: Search Chronology in Ragazzini 2014 app.**

The observers were university and Ph. D students from the Torino University Department of Foreign Languages and Literature and Modern Culture trained to observe a series of behaviours (see the observation protocol given to the observers in Italian and reproduced in appendix 3 in an English translation). After the experiment the observers were invited to write a report on their experience and it became clear that many of the tested students were not really able to understand what a headword was and go straight to the entry with the right part of speech. The students' inability to choose the right English entry in the presence of homonymic headwords is widely reported in literature with students of different ages and mother tongues.

Here are two significant excerpts from the observers' reports:

The student was not able to lemmatize and looked up words as they were written in the text (for ex. He looked up “cheaper”, then “cheape”; “broke” etc).

What probably struck me most was that he checked the lexical entry of the noun “break” as well as the verb “to break” many many times, going back and forth from one to the other.



## 4.1 Two “Simple” Tasks

Two of the sentences to translate which contained different meanings of the verb *break* were:

*The engine broke when he tried to speed up and I need some small change, can you break a 5€ note?*

These examples were chosen because they are functionally closer to the type of English students use in a technical high school course and, therefore, easy to translate. This conviction was also based on the fact that the meaning of the verb in the sentences was indicated as a level A2 in the Dictionary of the English Vocabulary Profile.


### **To Break [omission]**

#### **NOT WORK**

A2 [I or T] If you break a machine, object, etc., or if it breaks, it stops working because it is damaged.

#### **Dictionary example:**

I think I've broken your camera.

In reality, the first problem was that students looked for **broke** in the search window and found broke /brəʊk/ 

A pass. di **to break**

B a. (fam.) senza soldi; in bolletta; spiantato; rovinato; fallito: **I'm broke**, non ho un soldo; sono in bolletta; **to go broke**, andare in rovina; fallire; (fam.) **flat broke**, completamente al verde

● (fam.) **to go for broke**, rischiare il tutto per tutto.

A second problem occurred when they were told to look for **to break**: 5 students out of 17 forgot that it was a past tense and translated it in the present. When indicating the part of the lexicographic entry in which they found the correct translation, student I says to broke v.t. 3, student B more correctly indicates v.i.<sup>7</sup>, while M and R indicate to break v.t.1.

---

7 In fact, in Italian the translation is *si rompe* or *si è rotto*, therefore, it is an intransitive pronominal verb form. Those students who indicated the intransitive verb 1 paid attention to the meaning *rompere*; *infrangere*; *spezzare* and then translated with the *si* needed.

Student N mistook *engine for engineer* and translated *engineer broke it when he tried to make it go faster*<sup>8</sup> : such a mistake might be explained by the fact that he was using the online version, where if you type *engine* in the search window some other words also appear listed below the word entered. They are words that follow in the alphabetical macrostructure: in this case the words that appear are *engined* and *engineer*.

For *I need some small change, can you **break** a 5€ note?* We were counting on the fact that in the first part of the sentence there was the very frequent *I need* ( A1 ) and the noun phrase *small change*, which we assumed the students would know: in the English Vocabulary Profile the meaning of the noun *change* is B1 and contains the same phrase *small change*

### Change [omission]

### COINS

B1 [U] money which is coins rather than notes

#### Dictionary examples:

She gave me £5 in change.

My dad always used to carry a lot of loose/small change in his pocket.

Unfortunately a good 12 out of 17 students took *small change* to mean ‘piccolo cambio’, mistaking it for an exchange of currencies; this did not invalidate the correct translation “puoi cambiarmi 5€?” in the second part of the sentence.

Only two translated it perfectly: student G and R who is Romanian.

Ho bisogno di spiccioli, puoi cambiarmi una banconota da 5€?

Mi servono delle monete, potresti cambiarmi una banconota da 5€?

And another two got very close (student O and Q who we had already noted as two of the best in the pre-test) translating with:

Ho bisogno di **pezzi piccoli** (literally. small pieces), puoi cambiare un 5€?

Ho bisogno di un **taglio più piccolo** (literally. lower denomination), potrebbe cambiare una banconota da 5€?

---

8 The various translations ‘accendere’, ‘velocizzare’, ‘andare avanti’ of *to speed up* will not be discussed here – the electronic versions often show the headword of phrasal verbs with labels of specialised fields such as *automobilismo*)

■ speed up

A v. t. + avv.1 (*autom., ecc.*) accelerare 2 sveltire; **velocizzare: to speed up production**, sveltire la produzione

B v. i. + avv.1 affrettarsi; affrettare il passo 2 (*autom., ecc.*) accelerare.

## Student H translated word for word

Io ho bisogno di una piccola variazione, potrei interrompere per 5€ facendo attenzione

Word for word: I need a small variation, I might interrupt for 5€ paying attention

This translation is strikingly similar to the one obtained using the site<sup>9</sup> <http://translate.google.it>; however, it is clear that a speck of human imagination played its part in constructing something different in the second part<sup>10</sup>.

*I need some small change, can you **break** a 5€ note?*

Ho bisogno di qualche piccola modifica, si può rompere una nota di 5 €?

Word for word I need some small adjustments, is it possible to tear a note of 5 €?

The two students that translated well used different formats, the Italian G the app and the Romanian R a paper dictionary and said, correctly, that they took information from meaning 5 of the transitive verb:

**to break** [omission]

5 cambiare (*una banconota, spec. pagando qc. e ricevendo un resto*); spicciolare: to break a £50 note, *cambiare un biglietto da 50 sterline*

The students that translated *piccolo cambio* managed, however, to get the translation of the second part of the sentence and they either did not indicate anything or they indicated coherently meaning 4, since they had translated *puoi dividere in valute da 5€? Puoi dividere una banconota da 5 €?*

**to break** [omission]

4 suddividere; dividere; frazionare: **to break a word into syllables**, dividere una parola in sillabe

Those who translated *piccolo cambio* used the bilingual dictionary the way you do when, you do not have an overall ideal of the meaning of the sentence so you form a plausible hypothesis and make the rest fit around it.

The Ragazzini bilingual dictionary gave every imaginable form of support possible: under *change* noun meaning 3 is dedicated to the equivalent ‘spiccioli’ with the expressions *loose* and *small change* translated as *spiccioli*; there is even a USAGE NOTE that, if read carefully, would have helped the student to avoid confusing *change* and *exchange*.

---

9 Consulted 19 April 2014

10 *You* is disregarded and continued as an *I*; *fare attenzione* is the first equivalent of the verb *to note* in Ragazzini 2013

#### NOTA D'USO

**change** o **exchange**? Quando si parla del tasso di cambio tra valute non si usa il sostantivo *change*, ma *exchange*: *What's the exchange rate of the euro against the dollar?* qual è il tasso di cambio tra euro e dollaro? (non *What's the change of euros and dollars?*). *Change* viene usato in relazione al denaro per indicare il “resto” o gli spiccioli in moneta: *Did the cashier give you the right change?* il cassiere ti ha dato il resto giusto?; *Do you have any change at all?* hai della moneta?

#### USAGE NOTE

**change** or **exchange**? When you talk about the exchange rate of currencies you do not use *change* but *exchange*: *What's the exchange rate of the euro against the dollar?* qual è il tasso di cambio tra euro e dollaro? (not *What's the change of euros and dollars?*). *Change* is used for money to indicate what “remains” or coins: *Did the cashier give you the right change?* il cassiere ti ha dato il resto giusto?; *Do you have any change at all?* hai della moneta?

Moreover, it should be noted that the instructions also contributed to the problem since they led students to start with the word **in bold**; we are convinced that had we put *small change* in bold the error would not have occurred, because the processing left to right would have led to the identification of the expression first.

On the other hand, it would be wrong to draw the conclusion from the results of our “easy” tasks that the bilingual dictionaries are not helpful or worse, lead to mistakes: users normally use them for texts on topics they are familiar with. We were interested in whether students could go from the example to the meaning and we wanted them to look up the word **break** many times, also demonstrating that they had identified the correct part of speech called for by the various sentences.

It has been correctly observed that the task of associating meanings in a dictionary to the contexts of use is a complex task and in some ways metalinguistically unnatural. Still more difficult and often impossible to decide is the task of matching concrete examples to the appropriate specific meanings given in a dictionary and this is not because a dictionary is badly arranged, but because it is possible for circumstances to exist in which a combination of families of meanings is applied collectively in a context of use and not necessarily just one isolated meaning. (Chiari 2012: 116)

To translate isolated sentences is unnatural, but the task was important to evidence the problems which occur when consulting a dictionary. The second part of the test, in fact, was based on a much more natural understanding of a rather easy text, and the same students carried out the task correctly, using the dictionary very little or not at all, as we expected, having encouraged them to use it only when needed.

## 4.2 A Difficult Task

Students were given a clearly more difficult task when asked to translate the sentence

*In a **fast break**, a team attempts to move the ball up court and into scoring position as quickly as possible*

because it was a matter of a) – understanding how to translate technical terminology that was not entered as a multi-word headword, and b) – clarifying that under **fast** and under **break** not much would be found. Even if the sentence was itself a definition of *fast break*, in the Ragazzini 2013 under **fast (2) adjective**<sup>11</sup> there is nothing very useful, under **break** noun you need to go to the end of the list of meanings to find something useful, but not decisive, because just the equivalents of *break* are given.

### **Break [omission]**

23 (*calcio, ecc.*) incursione; penetrazione; discesa

24 (*basket*) break ; sfondamento; vantaggio (o svantaggio) incolmabile .

Real answers are given by the *full text*<sup>12</sup> search mode of both words: when the expression is searched, two entries come up, in English **to stop** and in Italian **contropiede**. Clicking on **to stop** the app automatically brings it up exactly where it appears in the phraseology of the very long entry of the word,

(*basket, calcio, ecc.*) **to stop a fast break**, stoppare un contropiede

*fast break* is highlighted in yellow in the app version, framed in red online. Whoever knows something about sports understands that *contropiede* is the correct translation; clicking on **contropiede**, noun, masculine, the first meaning comes up:

(*sport:calcio*) counter-attack, fast break

---

11 The lexicographic norms always list the homonym that has part of speech noun first, followed by the adjective and by the verb.

In this case **fast (1)** is the noun that we translate as *digiuno*

12 It should be noted that the only suggestion given to students before the test was “Remember that in the app and online versions in addition to the word search there is a full text search”. In this specific case an observer gave the student misleading advice suggesting that he look for the “words in bold” separately, while the student had initially looked for the two words.

With the equivalent *fast break* highlighted in yellow in the app version and framed in red online, the observers recorded that

The exercise that wasted the most time was the one containing the expression “fast break”, as well as the exercise containing the idiom “break a leg” (I believe they were the exercises that took the student the longest time, about 16 and 34 minutes).

Only three students, F, L, Q out of 17 were able to translate *in un contropiede* and they all were using the online site. One did not indicate the part of the entry that helped him and the other two indicated the useful parts of the entry *fast break v.i. 1 fast break 4* and that, if referred to the verb **to break**, they did not make sense. As for the other students, three did not translate, six opted for ‘pausa veloce’ (“fast pause”), one for ‘azione veloce’ (“fast action”) and one for the most sensible ‘sfondamento veloce’ (“fast smashing”)<sup>13</sup>, indicating meaning 24 as his source of information. It can be concluded that a third of the class attributed the meaning ‘pausa’ to *break*, familiar to them because the word is used in Italian as a loan from English with this meaning and the meaning in tennis.

In this exercise having a paper dictionary was a handicap because only with the full text search can the equivalent be found and as De Schryver (2003: 146) observed only the implementation of fully integrated hypermedia access structures makes electronic dictionaries really different from their paper counterparts.<sup>14</sup>

## 5 Full Text Search is not like a “Google Search”

The full text search is not like a “google search”, as emerged from the students’ translations of the sentence

*It served him right to fail the exam: he had never studied hard*

When a few of the 7 that had translated well indicate “to serve 10” as part of the dictionary lexicographic article that helped them to translate, they really meant to indicate the phraseology which follows meaning 10.

**to serve** /sɜ:v/ 

v. t. e i. [omission]

10 (naut.: della marea) essere favorevole

<sup>13</sup> Also the translation of *to move the ball up court* caused significant problems, but we won’t discuss them here.

<sup>14</sup> In regards to hypermedia access structures, it should be noted that in the online version every word of the gloss can be clicked on, bringing up the entry that word belongs to, but in the app version this mode was not implemented, probably to use less memory space in the smartphone.

● (mil.) **to serve as an officer**, prestare servizio come ufficiale; **to serve as a reminder [as a spoon]**, servire da promemoria [da cucchiaino]; **to serve at table**, servire ai tavoli; **to serve behind the counter**, servire (o stare) al banco (in un negozio, ecc.); (mil.) **to serve a gun**, servire un pezzo; caricare un cannone; (fig. fam.) **to serve sb. hand and foot**, servire q. di barba e di capelli [...]; **to serve a purpose**, servire a uno scopo; **to serve sb.'s purpose**, servire a q.; andare bene (lo stesso): *I haven't got a screwdriver, but a knife will serve my purpose*, non ho un cacciavite, ma un coltello va bene lo stesso; **to serve sb. right**, trattare q. come si merita; (impers.) meritarsi: *It served him right to lose his job: he was always taking time off for no reason*, il licenziamento se l'è meritato: faceva sempre assenze ingiustificate (omissis)

What can be noticed is how they make the example fit: *il licenziamento becomes il fallimento dell'esame* in a good three students.

*Il fallimento dell'esame se l'è meritato: non ha mai studiato*

In this way the Italian in the translation is not wrong but rather unlikely. Another four, the same group L, N, O, Q as before, plus A, F, L, choose a more appropriate Italian translation *Si è meritato/ se lo è meritato di fallire l'esame; non ha mai studiato (duramente, bene, abbastanza)*

Student A translates *it served* in the present but is not the only one to forget it is in the past.

*Si merita di non passare l'esame: perché non ha mai studiato*

At the same time, however, he is the only one who does not fall into the trap *to fail the exam = fallire l'esame*, a translation that the students did not check in the Ragazzini that gives under **exam** “**to fail an exam**, non passare un esame” and under

**to fail** v.t.

[omissis]

2 non superare (un esame); essere respinto (o bocciato) in: *to fail one's driving test*, non superare l'esame di guida; *I failed maths*, sono stato bocciato (o mi hanno bocciato) in matematica

Reading quickly through the other ten translations of *to serve* and then right, it becomes clear that the translation task was not so banal. We cite them here:

*Esso ha suggerito correttamente a lui per fallire l'esame: lui non ha mai studiato duramente.*

*L'ha aiutato in modo corretto per non essere bocciato all'esame: non ha studiato molto.*

*Ha fatto il suo esame giusto ma lo ha sbagliato: non aveva studiato bene.*

*Ha servito di lei la cosa giusta per farlo fallire all'esame, non ha mai studiato tanto.*

*L'ha aiutato correttamente per fare l'esame, lui non ha mai studiato tanto.*

*Gli bastava a lui giusto a fallire l'esame: lui non aveva mai studiato intensamente*

*Gli sta bene di fallire l'esame: non ha mai studiato tanto.*

The last translation uses the final indication in the Ragazzini phraseology:

(fam.) **Serves you right!, ben ti sta!**

which is – very inappropriately – detached from **to serve sb. right**. And it is translated by a student using the paper version, acting almost as a confirmation of the experiment by Tono (2011), which concludes that we read the first and the last parts of a gloss.<sup>15</sup> The user of the app does not skim through other parts once he has found the example.

Student O used the paper version and the other six, the one online. In this case the app seems to penalize the students, if the search starts with the entry **to serve**, because the visualization on the limited display implies having enough patience to skim through the gloss many times to get to the phraseology; on the other hand, the entire entry fits on the computer screen (see fig.2).

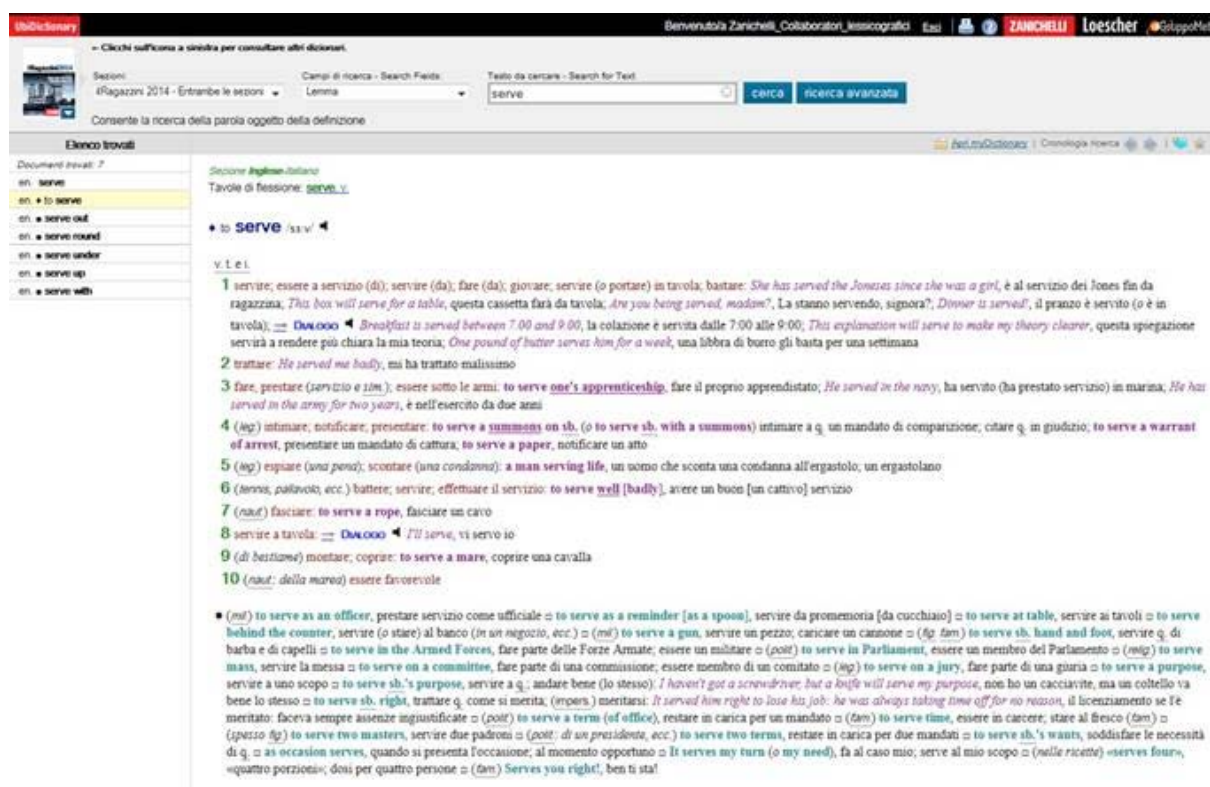


Figure 2: Entry to serve Ragazzini 2014 on line.

If the student had looked for the full text he wouldn't have found anything because it is one of the cases in which the user needs to substitute the variable part with placeholder pronouns. In other words, he needs to know how to go from *to serve him right* > **to serve sb. right**. Only this precise mode of set-

15 An observer reported that the his student began to realize that he had to read the whole lexicographic article and not just the beginning and the end of the big articles only after three exercises. Three weeks later, a post-test revealed that the lesson had been learned; students continued to read the entire entry even though it was a comprehension test and not translation.



ting up the full text search will bring up to serve, but it does not place it at exactly the right point of the article, nor is there the yellow highlighting or red framing that there were for **fast break**. The same problem emerged while trying to do the first exercise in the second part of the test reproduced below.<sup>16</sup>

Massive and unreliable, the first computers of thirty years ago are as dead as the dinosaur. **Today, computers which are 30,000 times smaller and 10,000 times cheaper can beat them hollow**

What does **the part in bold mean?** It means

- a) today computers have less hollow space in their cases
- b) modern computers can beat easily thirty year old computers
- c) thirty year old computers are as dead as the dinosaur.
- d) thirty year old computers can still beat modern computers

Was context enough to understand the meaning?  Yes  No

Which dictionary item helped you in answering the question?

In this exercise the bold print was much longer than the idiomatic expression, therefore the student should have identified the limits of the idiomatic expression and substitute them with the placeholder sb.; looking for **beat sb. hollow** in Full text brought up both **to beat** and **hollow** and it is found in their phraseology.

In the second part of the text, the task was not translating but choosing the right answer from four distractors; a clearly easier task and in fact a good 12 out of 17 gave answer b; three chose a) “today computers have less hollow space in their cases” and another two chose d). Only one student answered no to the first question, but identified the right answer. Seven said that consulting the entry **hollow** helped them, two found help in **beat**

Seven did not indicate anything, perhaps because they answered Yes to the question “Was context enough to understand the meaning?” and did not use their dictionaries.

The fact that in an electronic dictionary or apps typing in sequences of idioms with variables produces the response “no results found” rather upset the students who were expecting it to work the same way Google<sup>17</sup> does, where *them* automatically leads back to *somebody* and typing in *beat them hollow* brings up as the first site <http://idioms.thefreedictionary.com/beat+hollow>

**beat somebody hollow** (British & Australian)

16 The passage used in the first multiple choice exercise was taken from a B1 level English text formed by 432 words.

17 Typing *beat them hollow* in <http://translate.google.it/#en/it/beat%20them%20hollow> (19/04/2014) brought up the translation *batterli cava*, literally *beat them empty* adj. feminine singular or n. *quarry*

to defeat someone easily and by a large amount We played my brother's school at football and beat them hollow.

See also: beat, hollow

Cambridge Idioms Dictionary, 2nd ed. Copyright © Cambridge University Press 2006. Reproduced with permission.

Essentially students would like something more than an electronic dictionary; they would like what Tarp (2008:123) called a *leximat* “a lexicographical tool consisting of a search engine with access to a database and/or the internet, enabling users with a specific type of communicative or cognitive need to gain access via active or passive searching to lexicographical data”.

Here we have not dealt with tasks of consultation tied to production activities, but a *leximat* would be even more welcome when producing a text in L2.

## 6 Conclusions

The experiment has shown that without knowing how to consult a dictionary the user does not take full advantage of the electronic dictionary and, on the other hand, that a knowledgeable user is able to get the information needed from a paper dictionary just as well.

Compared to searching an electronic dictionary, consulting a paper version is slower, but penalizes the user only when looking up idioms or multiword units.

Regardless of the lexicographic tradition in lemmatization, the first 3000 most frequent English words are characterised by a peculiar morphological poverty in part of speech markers which makes homonymy abundant in the macrostructure. Students should at least be aware of this before using a handheld dictionary.

To deal with the polysemy, when opening long lexicographic article in the app an index, a menu of the microstructure -including signposts- should be adopted, similar to the ones already found in English learners dictionaries (see Medal), Latin and Greek dictionaries (see Montanari) and in the VanDale and Sansoni bilingual dictionaries published in the 1980s (see Marellò 1989: 77-98 )

Specialized field labels abound in Italian monolingual and bilingual lexicography but are not used enough by unskilled users, whether they are in paper or electronic versions.

Searches using jolly characters were not tested. Jolly characters are available only in the electronic dictionaries for purchase and not in the ones free on the Internet. Our observers, however, reported that students used a similar type of search, not based on morpheme boundaries and without jolly characters, taking advantage of the list of alphabetically near headwords which comes up on the electronic dictionary when the first letter are typed.<sup>18</sup>

---

18 For example, when typing *common* commonable, *commonage*, *commonality*, *commonalty*, *commoner*, *commonhold*, *commonly*, *commonness*, *commonplace*, *commonplaceness*, *commonweal*, *commonwealth* come up on the screen.

Observers reported that students gradually understand how an electronic dictionary works and they do not need a lot of training: they need practice, as noted previously, with the homonymy in the core English lexicon, also because in the so-called collaborative bottom-up dictionaries, produced by volunteers, like Wictionary and Wikizionario, homographs are not separated. As Chiari observes (2012: 108-109) this difference significantly changes the very idea that Wikizionario gives of a headword, which ends up matching the form keyed in and is strongly influenced by the user's need for a keypad search, so that it even includes some inflected forms. For now dictionaries are not considering the option "perhaps you were looking for..." but only give cross references similar to the one seen for broke in § 4.1. This is another "defect" our students accuse the dictionaries of, whether, apps or even more rightly so, online on computers, because it represents a valid aid especially for languages with complex spelling like English.

If students had been trained beforehand and given a common basis of elementary knowledge of dictionary use, the difference in the time of execution for online users and app users might have emerged. The difference between users of the paper version and of electronic versions are only indirectly revealing in our experiment, because 2 out of 5 students with paper dictionaries were non native speakers of Italian and another 2 were among the best students in the class. The latter two finished before the others both the test section requiring dictionary use and the comprehension in which they used the dictionary considerably less than the others while achieving a very good score. Students' proficiency influenced both score and speed in completing the test, more than the lexicographic tool.<sup>19</sup>

The convenience of having a dictionary on a smartphone is sufficient to justify allowing its use in the classroom: an expert user practices his/her ability to use a dictionary and a weaker one carries out a (meta)linguistic activity that, in any case, has important repercussions on his/her knowledge and ability to search more complex databases. However, interviewed on whether or not they would purchase or suggest purchasing the app for a bilingual dictionary to use on their own smartphones, the students in the class seemed perplexed since they were basically convinced that the translation programs and dictionaries free online were good enough for their extra-scholastic needs. As for their scholastic needs, so far they have been limited, but that could change with the introduction of teaching methods based on Content Language Integrated Learning. We hope to repeat the experiment in types of secondary schools where teachers are preparing students of the same age for university studies, and we dare say that having a good bilingual dictionary at hand - on a smartphone or tablet - and knowing how to use it might increase these students' understanding of important details in texts as well.

---

19 Their English teacher probably meant to help the two non native speakers giving them the paper dictionary and knew that for the best students a paper version would not be a handicap. See the contributions in progress by Elisa Corino and Elena Martra on these aspects. Of course, if we had performed the test in a lab we might have selected the coupling tool-testee differently, in order to obtain more telling results, but when an experiment is performed in a true class, the research has to respect the teacher's educational concern.

## 7 References

- Augustyn P. (2013). No dictionaries in the classroom: translation equivalents and vocabulary acquisition. In *International Journal of Lexicography*, 26 (3), pp. 362-385
- Bergenholtz, H. Johnsen, M. (2007). Log files can and should be prepared for a functionalistic approach. In *Lexikos*, No.17, pp. 1-21.
- Butzkamm, W. (2009). *The Bilingual Reform: A Paradigm Shift in Foreign Language Teaching*. Tübingen: Gunter Narr Verlag.
- Cadierno, T. (2008). Learning to talk about motion in a second language, in P. Robinson, N. Ellis, *Handbook of Cognitive Linguistics and Second Language Acquisition*, London: Routledge, pp. 239-275.
- Chiari I. (2012). Il dato empirico in lessicografia: dizionari tradizionali e collaborativi a confronto. In *Bollettino di Italianistica. Per Tullio De Mauro*, II, pp. 94-125.
- Cook, G. (2010). *Translation in Language Teaching: An Argument for Reassessment*. Oxford: Oxford University Press.
- de Schryver, G.-M. (2003). Lexicographers' Dreams in the Electronic-dictionary Age. In *International Journal of Lexicography*, 16 (3), pp. 143-199.
- Hjelmslev L. (1961<sup>2</sup>). *Prolegomena to a Theory of Language*. Madison: The University of Wisconsin Press
- Laufer, B. and N. Girsai. (2008). Form-focused Instruction in Second Language Vocabulary Acquisition: A Case for Contrastive Analysis and Translation. In *Applied Linguistics*, 39 (4), pp. 694-716.
- Macmillan English Dictionary Online*. Accessed at: <http://www.macmillandictionary.com> [19/04/2014].
- Marello C. (1989). *Dizionari bilingui con schede sui dizionari italiani per francese, inglese, spagnolo, tedesco*. Bologna: Zanichelli
- Marello C. (1998). Hornby's Bilingualised Dictionaries. In *International Journal of Lexicography*, Special Issue 1998, pp. 292-314
- Montanari F. (1995, 2013<sup>3</sup>) *GI - Vocabolario della lingua greca*. Torino: Loescher
- Il Ragazzini (2013, 2014) *Dizionario Inglese-Italiano Italian-English Dictionary* Bologna: Zanichelli; 2014 online version and app for iPhone e iPad
- Slobin, D. (1997). *The crosslinguistic study of Language Acquisition*, vol.5: Expanding the contexts, cap.1.3: The cross-Typological Approach. London: Erlbaum, pp. 11-34.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge*. Tübingen: Niemeyer
- Tono, Y. (2011). Application Of Eye-Tracking In EFL Learners' Dictionary Look-Up Process Research. In *International Journal of Lexicography* 24 (1), pp. 124-153.

### Acknowledgements

I would like to express our appreciation to the class IV E of the Avogadro technical high school, their English teacher, Elena Martra and the principal Tommaso Deluca, as well as to the publishing house, Zanichelli. We would also like to thank Elisa Corino for her assistance in preparing the tests and conducting the project, Andrea Belluccia for the tabulation, the observers for their observations and Kathy Metzger for the translation.

## Appendix 1

**Use the word in the box to form a word that fits in the text**

**How does the video work?**

When I was young, I always dreamed of becoming a famous (1) When I was at school I decided to study engineering, and then become a millionaire by inventing a wonderful new (2) which would make the world a better place. Unfortunately, I wasn't very good at technical subjects. Any time I operate any kind of (3), something terrible happens. Machines which use (4), such as computers or televisions, always seem to give me a (5) shock. The instruction booklets are always (6) They never help me at all. Nowadays you need to have specialised knowledge just to turn on the video. To my great embarrassment it is always a child of six who helps me out of my (7).

- 1 SCIENCE .....
- 2 PRODUCE .....
- 3 EQUIP .....
- 4 ELECTRIC .....
- 5 POWER .....
- 6 USE .....
- 7 DIFFICULT .....

From: Michael Vince, First Certificate Language Practice, Heinemann, Oxford, 1996, pag. 235

## Appendix 2

Translate the following sentences. **Which entry and which part of the dictionary entry** did you look up to translate the words in bold?

Ex. Her bones **break** easily

Le sue ossa si rompono facilmente

**Entry and part of the dictionary entry**      **to break B v.i. 1**

The engine **broke** when he tried to speed up

**Entry and part of the dictionary entry**

The news **broke** and everybody knew the truth

**Entry and part of the dictionary entry**

Well-wishers typically say “**Break a leg**” to actors and musicians before they go on stage

---

**Entry and part of the dictionary entry**

I need some small change, can you **break** a 5€ note?

---

**Entry and part of the dictionary entry**

In a **fast break**, a team attempts to move the ball up court and into scoring position as quickly as possible

---

**Entry and part of the dictionary entry**

It **served** him **right** to fail the exam: he had never studied hard

---

**Entry and part of the dictionary entry**

You can put it in the washing machine, it's a **fast**-colour T-shirt

---

**Entry and part of the dictionary entry**

### **Appendix 3**

Observation Protocol for 5 March 2014

Objectives of the Experiment

- (1) Establish the class's ability to use dictionaries
- (2) Establish whether students are aware of the pros and cons of a paper dictionary compared to the electronic versions
- (3) Establish whether students are aware of the pros and cons of an online dictionary compared to an app(lication).
- (4) Exercises were prepared which explicitly required the use of a dictionary to verify if students were able to trace a statement back to the schematization of microstructures in the dictionary (profile

of the gloss in a bilingual dictionary) or to use the examples to find the correct translations (semantic equivalents).

(5) Reading comprehension exercises were prepared which could be done without a dictionary, if students believed they knew how to answer.

Your job is to observe how students work

(1) The time they spend on each exercise

(2) If they look up words other than the ones in bold in the exercise.

Ex. in *We have to accept the microchip, or face the alternative of leaving off the free world market* they look up **face** as well as **leave off**

(3) If they check the translations (equivalents) by looking them up in the section in which they are entries.

N.B. online all the words can be clicked on the smartphone no.

(4) If the students with electronic versions use the **full text search**

(5) If they know how to lemmatize ( broke → break) or they get help from the online references

(6) If they go to the entry with **the right part of speech** without hesitating

(7) If, for idioms, they realize they are idiomatic or go to the phraseology section to look for it.

You need pen and paper to write down what they look up and the order they look things up for those that have paper copies. **Stop those with online access and smartphones BEFORE they close the session.** In this way we can take a picture of the list of entries looked up.





# Meanings, Ideologies, and Learners' Dictionaries

Rosamund Moon  
University of Birmingham  
R.E.Moon@bham.ac.uk

## Abstract

This paper looks at the treatment of ideologically loaded items in monolingual learners' dictionaries of English, and issues in the lexicographical description of their meanings. It begins by considering non-denotative meaning and the question of evidence; then considers a selection of entries relating to ethnocentricity, gender and sexuality, and age. Entries are drawn mainly from the standard British big five; the 1948 edition of Hornby; and two crowd-sourced dictionaries.

**Keywords:** ageism; critical lexicography; culture; ethnocentricity; gender; ideology; learners' dictionaries; meaning; sexism

## 1 Introduction

My talk at Euralex 2014 is concerned with the presentation of meanings in learners' dictionaries of English: in particular, the meanings of words which denote, represent, or reflect politicized concepts and phenomena – ideologically loaded items, totemic and socioculturally significant. Such words have been the frequent focus of linguistic investigations more widely, for example in corpus-led studies from a discourse analytic perspective, or sociological and cultural studies (Raymond Williams' discussions of “keywords” are a case in point). In relation to lexicography, ideology is where dictionaries collide with the social world: it brings in impolite and polite aspects of language, taboo items, evaluative orientation, connotation, and cultural allusion; the sublexicons, of course, of semantic fields such as politics, religion, ethnicity, sexuality, and so on; and above all the role of lexis, an unstable and mutable role, in naming and othering.

Ideologically positioned meaning is central to the concerns of critical lexicography, and particularly important with respect to learners' dictionaries because of their positioning as global texts for a pluralist multicultural usership. It is a topic that regularly surfaces at Euralex congresses<sup>1</sup> – though historically, perhaps not as often as might be expected – and in lexicographical journals; it is covered, at least tangentially, in lexicographical manuals and metalexicographical monographs (Svensén 1993;

---

1 Euralex papers include those by Coffey 2010; Hartevelde & van Niekirk 1996; Iversen 2012; Schutz 2002; Swanepoel 2010; van der Meer 2008; Veisbergs 2000, 2002, 2004; Whitcut 1983; and several due at Euralex 2014: see the Euralex database of past proceedings at <http://www.euralex.org/publications/> for these and other presentations on the topic.

Atkins & Rundell 2008: 422ff and passim; Landau 2001: 228ff and passim; Béjoint 2010), and more directly as the subject of a 1995 festschrift for Zgusta (eds. Kachru and Kahane). In returning to the topic today, I do not offer solutions, but there are, I think, a number of points that are still worth making in a 2014 context.

his written version of my talk provides an overview of topics to be discussed. In section 2, I look at non-denotative aspects of word meaning, the contribution of corpus evidence, and methodological issues; sections 3, 4, and 5 present and review a sample of words and dictionary explanations relevant to discussion of, respectively, ethnocentricity, gender and sexuality, and age/ageism. As a basis for observations, I draw largely on four current British monolingual learners' dictionaries of English (from Cambridge, Longman, Oxford, and Macmillan), specifically their free online versions as accessed in April 2014: by default, these four are the online learners' dictionaries referred to below. The fifth British learners' dictionary, *Cobuild*, is represented instead by its second print edition of 1995: partly because the current online version has restricted accessibility; partly because the 1995 text was based on an early version of the Bank of English corpus (BoE),<sup>2</sup> which I draw on below; and partly because of my own editorial involvement with the 1995 text (and as an editor of the first edition of 1987). Other dictionaries cited include the 1948 version of Hornby's first dictionary (ALD1); and the collaborative or crowd-sourced online texts *Simple English Wiktionary*<sup>3</sup> and *Urban Dictionary*.<sup>4</sup> It goes without saying that online dictionaries are complex multidimensional, multimodal texts, and although I focus mainly on linguistic explanations, examples, and labelling in my discussion, there are other parts of their entries where ideology is displayed and where othering may be performed.

## 2 Meanings: Culture, Connotation, Evaluation – and Evidence

In later sections, I will look at some words that are obvious sites for projection of contested and ideologically-bound attitudes. But many of the issues of lexicographical description overlap with more general issues of how and whether to represent non-denotative aspects of meaning, including connotation, evaluation, and culture. To demonstrate this, I want to look at the word *cardigan*, an item with a very clear concrete meaning and reference in the real world. It can be defined or explained straightforwardly, as in the current online entry in Oxford:<sup>5</sup>

- 
- 2 BoE is a 450-million word corpus, created by COBUILD at the University of Birmingham. 71% of its texts are British English, 21% North American, and 8% Australian, mainly drawn from the period 1990-2003; 86% of its texts are written, and the remaining 14% consists of transcribed spoken interaction and radio broadcasts.
  - 3 A simplified text, affiliated with Wiktionary, constructed with something of a controlled defining vocabulary, and claiming almost 22,000 definitions or "entries" in May 2014.
  - 4 In Fuertes Olivera's terminology, these are classifiable as "collective free multiple-language Internet reference works", in distinction to "institutional Internet reference works", which would include publisher-produced texts (2009: 103).
  - 5 To facilitate comparisons, here and below, I have standardized typography and layout irrespective of the original: headwords and phrases are in bold; where included, examples are italicized and set out on new lines; usage labels are italicized in parentheses.

(1) a knitted jacket made of wool, usually with no collar and fastened with buttons at the front where it is accompanied by a photographic illustration (and adverts from clothing companies). Compare Macmillan's online entry, which combines a broadly comparable descriptive and denotative explanation with a "cultural note":

(2) a jacket knitted from wool, that you fasten at the front with buttons or a zip  
Cultural note: cardigan

Cardigans are usually thought of as an old-fashioned, rather boring piece of clothing, worn mainly by older people.

We might dispute the explanations of cardigan (they can be made of fibres other than wool; not all cardigans have fastenings), as well as Macmillan's stereotyping comment (fashions change; babies, toddlers, schoolchildren wear cardigans), but there is an important point that even innocuous and simple items generate associations which reflect attitude and, in this case, ageism. Evidence for these associations can be detected in corpus data, as in this small sample from BoE, in particular the fifth, sixth, and seventh tokens:

this short green knitted dress and cardigan yesterday to promote the Irish  
Isabella was wearing a bulky cardigan with horizontal orange and red.  
Next. Not long ago a pure cashmere cardigan alone would have cost about £300.  
cardigan Invest in our classic cardigan and you'll wonder what you ever  
maths teacher who favours comfy cardigans and looks like the perfect grandad  
it. There are plenty of ladies in cardigans and old gentlemen in ties. It  
teeth; it meant shuffling around in cardigans and ranting about the youth of  
up the sleeves of her grey knitted cardigan and got to work. <p> Robina went  
for the day, or under a long cardigan for evening. This month we are  
was in the 80s, she wore a woollen cardigan buttoned to the neck and

Compare, too, two of the examples given in OED for the abbreviated form *cardy* (and discussion of middle-aged below):

(3) 1969 *Guardian* 3 Nov. 7/2 Grey gentlemen in shrunken cardies.

981 *Daily Telegraph* 29 Aug. 11/2 A flock of over-50s wearing pastel cardis and floppy hats.

Words that are more obviously politicized offer particular scope for critical lexicography: examination of differences between explanations in dictionaries, unpacking of attitudes projected towards the concept under definition, and dominant ethos of the lexicographers – or publishing/cultural context. For example, the four current online learners' dictionaries broadly agree on what materialism is (in the non-philosophical sense):<sup>6</sup>

(Macmillan)

the belief that money and possessions are the most important aspects of human existence

6 Cf. Williams 1988: 197-201; Bennett et al. 2005: 209-211.

(5) (Oxford)

(*usually disapproving*) the belief that money, possessions and physical comforts are more important than spiritual values *the greed and materialism of modern society*

but while Oxford adds a usage label to reflect sociocultural attitudes towards materialism and an example that indirectly condemns by collocation, Macmillan offers no evaluation at all of materialism as a *modus vivendi*.

Distinctions in attitude towards a concept can be subtle, as in these entries for *equality*<sup>7</sup> in the online dictionaries:

(6) (Cambridge)

the right of different groups of people to have a similar social position and receive the same treatment:

equality between the sexes

racial equality

the government department responsible for equalities

(7) (Longman)

a situation in which people have the same rights, advantages etc

**equality of** All people have the right to equality of opportunity.

**equality with** Women have yet to achieve full equality with men in the workplace.

**quality between** equality between men and women

**racial/sexual equality** The government must promote racial equality.

(8) (Oxford)

the fact of being equal in rights, status, advantages, etc:

racial/social/sexual equality

equality of opportunity

the principle of equality before the law (= the law treats everyone the same)

Don't you believe in equality between men and women?

The entries are broadly similar, all using examples to indicate arenas in which equality is an issue, particularly gender and race. But comparison of the genus words in explanations raises the question of what equality actually is – *fact* implies that it is a principle achieved; *situation* is non-committal; only Cambridge's *right* suggests that it is more of an ideal than real or implemented in practice, something hinted at in the following small sample from BoE:

---

7 Cf. Williams 1988: 117-119; Bennett et al. 2005: 109-111.

in an atmosphere of ease, equality and immense opportunities for the  
the Commission for Racial Equality and other ethnic minority and  
is further evident that political equality, co-existing with an increasing  
relationship characterized by equality, disagreement and conflict are  
aim of that document was to achieve equality for women by 2000. <p> The  
as a means of achieving human equality have been taken up by most of the  
movement to greater fairness and equality in schools seemed to go naturally  
to ensure that the new regime makes equality of opportunity, in terms of access  
I wish to take part. He believed in equality then and I believe in it now.  
in line with feminist notions of equality. Yet feminist therapists, most of

Top collocates of *equality* in BoE provide further data concerning its textual and contextual environments (see groupings below), and further hints of the elusive nature of equality:<sup>8</sup>

- (9) (arenas) gay, gender, homosexual, lesbian, men, race, racial, sexes, sexual, women  
(ideals and concepts) democracy, fairness, freedom, justice, liberty, rights  
(pressure groups and committees) campaign, commission, council, struggle  
(general items) commitment, economic, issues, law, legal, opportunity, political, principle\*, social, society, treatment

Similarly, to return to *materialism*, collocational evidence helps identify both semantic/philosophical contexts of occurrence and negative attitudes towards it, as in this from BoE:

- (10) (nouns) society, idealism, science, spirituality, greed, philosophy, hedonism, marxism, values, atheism, theory, feminism, selfishness, corruption, consumerism, ideology, capitalism, utilitarianism, pragmatism  
(prenominal adjectives) historical, dialectical, scientific, cultural, Western, crass, rampant, modern, new, gross, Marxist, atheistic, vulgar, subversive, secular

(further supported by lower-frequency collocates such as *self-interest*, *self-centredness*, *emptiness*, *godless*, *mindless*). At the same time, such evidence is a reflex of the discourse world of the corpus: a normative view that perhaps reveals more about prevailing attitudes in the language-owning culture at the time of data capture than about materialism itself. So there is a particular dilemma facing lexicographers attempting to deal with contested and ideologically loaded words: to balance a description of what data suggests about meaning with how in a postmodern inclusive society, the relevant concept “ought” to be regarded and represented. The English word *civilized* is another case in point: cf. Moon 1989: 88-90, and see discussion below.

---

8 For discussion of the role of collocation and co-text in relation to meaning, see for example Hanks 2013; Leech 1981: 16ff; Louw 1993; Sinclair 1991, 2004.

### 3 Ethnocentricity

It goes without saying that racist terms have to be labelled clearly as offensive in learners' dictionaries. However, the problem of ethnocentricity extends much further than labelling, something that has been explored at length by Benson (2001) and in Ogilvie (2013), as well as in multiple critical linguistic/lexicographical papers, including Krishnamurthy's examination of the words *ethnic*, *tribal*, *racial* in corpus and dictionary data (1996) and, for example, Hornscheidt's with respect to colonialist words in Danish dictionaries. In my own 1989 paper on ideology and lexicography, I was particularly concerned with the word *civilized*, its meanings, and dictionary representations of meaning. Here is the relevant entry from *ALD1*:

(11) *civilize*

1 bring out from a savage and ignorant state; give teaching in art, science, culture, good government, good customs and manners.

2 improve and educate.

Many a rough man has been civilized by his wife.

A product of its time and purpose, this now raises all sorts of questions: the paternalism of *bring out*; the subtext of *savage*; the meaning of *good*; and so on (similarly in the second extended sense, the use of *improve* and the sexism of its example). Fifty years later, and drawing on BoE corpus evidence, this is *Cobuild2*'s explanation for **civilized**:

(12) 1 If you describe a society as **civilized**, you mean that it is advanced and has sensible laws and customs.

I believed that in civilized countries, torture had ended long ago.

≠ barbaric

2 If you describe a person or their behaviour as **civilized**, you mean that they are polite and reasonable.

I wrote to my ex-wife. She was very civilized about it.

*Advanced? sensible? is barbaric* really an appropriate antonym? and with the example in the second sense, do the implicatures make it seem as sexist as *ALD1*'s? The next entries are from two online learners' dictionaries (extended senses have been omitted here):

(13) (Macmillan)

1 a civilized country, society etc has developed an advanced culture and institutions

A civilized society does not solve conflicts in a way that causes so much suffering.

(14) (Oxford)

- 1 well-organized socially with a very developed culture and way of life  
the civilized world  
rising crime in our so-called civilized societies  
civilized peoples
  - 2 having laws and customs that are fair and morally acceptable
- No civilized country should allow such terrible injustices.

While examples demonstrate something of what's implied by *advanced*, *developed* and indeed *civilized* itself, the overall meaning is unclear, almost insiderist (only someone from a "civilized" society would identify with the description - a circularity of meaning that is as problematic as the circularity of inspecting entries for *advanced*, *barbaric*, *cultured*, *developed*, *primitive*, *savage*, *uncivilized* etc. in order to locate what *civilized* might mean). Meanwhile, corpus data reflects, inevitably, an anglocentric view of the world:

to think of human beings as all civilised but they're not. Some remain terrorism. We British, and all civilised countries, should back America a world where a man who, in any civilised country, would - even though his fortunate brethren the benefits of civilized culture. Though more successful ask itself the big question: If the civilised European can allow the to a world seeking reassurance that civilised governments and legislatures that no country could call itself civilised if the sick are refused medical bullet, prohibited for use between civilised nations but sanctioned for big-seeds and causes of conflict among civilised nations will speedily appear. Of them, they too, can usher in a more civilised order. The Chinese took on idealism can cause seemingly civilised people to misuse, even destroy, human rights violations in a civilised secular democracy. These are not that an important hallmark of civilised societies is the extent to which by such horror close at hand, a civilised society has a choice. It can act, taxation system. <p> Any civilised society should provide education one every civilised nation, every civilised state, including the great European Communism to be open and civilised. The phenomenon of Eurocommunism syndrome is one of the <f> civilized world's most common diseases. It too close. `We warn you that the civilised world objects to your aggressive I think that the free world, the civilised world, understands that this

This presents a near-insoluble problem: should an explanation in a learners' dictionary present this kind of traditionalist anglocentric monocultural world-view evident in corpus data, thus promoting a particular ideological stance? Should an apologist usage note be added? Should there be instead a broader, multicultural, universalist, non-elitist explanation, even at the expense of misrepresenting of what the English word is actually used to mean, what mindset it reflects?

With *ethnocentric* itself, Macmillan is clearest in indicating that it is a derogatory label for inegalitarian perspectives:

(15) showing a failure to recognize that other people's cultures are also important and valuable while Longman's usage comment could be misinterpreted (exactly what is being disapproved of?):

- (16) based on the idea that your own race, nation, group etc is better than any other – used in order to show disapproval:

and Oxford fails to indicate that ethnocentricity is evaluated negatively at all:

(17) based on the ideas and beliefs of one particular culture and using these to judge other cultures  
Of the many items in the English lexicon which have potential for ethnocentric and racist usage, an important subset includes *foreign(er)* and other words which contribute to the othering of non-natives and non-nationals – cf. a critical linguistic study by Gabrielatos and Baker (2008), who are concerned with the discursive representation of refugees and asylum seekers in news media (and incidentally in definitions). The following brief discussion looks at *alien*, *asylum seeker*, *illegal immigrant/alien*, *migrant*, *refugee* in recent/current learners' dictionaries. Explanations are broadly comparable, but what's especially interesting is the selection of examples, particularly where these, in the decontextualized world of dictionary text, seem hortatory or imply moralistic value judgements. For example, these are *Co-build2*'s entries:

- (18) migrant

1 A **migrant** is a person who moves from one place to another, especially in order to find work. The government divides asylum-seekers into economic migrants and genuine refugees.  
... migrant workers following harvest northwards.

- (19) s.v. **illegal** 2

**Illegal** immigrants or workers have travelled into a country or are working without official permission. > Illegal immigrants or workers are sometimes referred to as **illegals**.  
... a clothing factory where many other illegals also worked.

Examples in the following entries for *refugee* show typical collocations with *flee*, *flow*, *stream* etc. – collocates which have been shown elsewhere to contribute to the negative discursive construction of refugees as a social group (Gabrielatos & Baker 2008: 22ff; Semino 2008: 87ff):

- (20) (Longman)

someone who has been forced to leave their country, especially during a war, or for political or religious reasons:  
Refugees were streaming across the border.  
refugee camps

- (21) (Oxford)

a person who has been forced to leave their country or home, because there is a war or for political, religious or social reasons  
a steady flow of refugees from the war zone  
political/economic refugees  
a refugee camp



(22) (Cambridge)

a person who has escaped from their own country for political, religious, or economic reasons or because of a war:

Thousands of refugees fled across the border.

Compare too the subtexts of examples in these:

(23) **asylum seeker** (Cambridge)

someone who leaves their own country for their safety, often for political reasons or because of war, and who travels to another country hoping that the government will protect them and allow them to live there:

genuine/bogus asylum seekers

(24) **alien** (Longman)

1 someone who is not a legal citizen of the country they are living or working in:

illegal aliens entering the country.

(25) **illegal** (Longman)

(*American English spoken*) an illegal immigrant:

Illegals are still slipping through in unacceptable numbers.

Also relevant, though this cannot be discussed in detail here, are the subtexts, the intertextual implicatures, which are created through the links to thesaurus entries or lists of semantically related words that are triggered automatically by searches for specific words. For example, Cambridge's entry for *refugee* displays a set "Runaways and refugees", listing items *boat people*, *deserter*, *displaced person*, *escapee*, *evacuee*, *fugitive*, *political asylum*, *refugee camp*, *transit camp*. Of course, these features are intended for vocabulary extension work, but many such examples of "interesting" juxtapositions can be found in online learners' dictionaries: these, by association, reinforce both othering and sociocultural evaluations.

## 4 Gender and Sexuality

The asymmetries of gendered nouns in English, together with the gendered collocational/semantic preferences of adjectives, have been widely discussed. With respect to dictionaries, discussion has tended to focus on sexism in general, asymmetric definitions of paired male/female terms, and representation of men and women in examples: see, for example, papers by Graham (1975), Whitcut (1984), Landau (1985), Barnickel (1999), Connor Martin (2005), etc., particularly with reference to orthodox (= androcentric) dictionaries; Russell (2012) examines feminist dictionaries which provide something of a counterdiscourse.

Where pairs of English words for (human) males and females are concerned, the lexicographical challenge is to balance two conflicting ideas. First, men and women in the UK, as in so many other

nations, now have equal status legally and legislation protecting their rights. Second, the continuing disparities in practice between the lives of men and women, along with biological/physiological distinctions, are reflected in lexis and language use and in attitudes communicated through language. Thus decisions taken when designing and constructing entries for paired terms cannot just be linguistic decisions: they must inevitably be ideological as well.

A pair such as *boy* and *girl* demonstrate the issues. Their primary and simplest senses – non-adult male/female, son/daughter – are clear counterparts; however, their symmetry changes when they are used to refer to adults (cf. discussion by Caldas-Coulthard & Moon 2010; Holmes & Sigley 2001; Sigley & Holmes 2002). In particular, *girl* continues to be applied to young women, especially in their late teens and twenties, whereas *boy* is more likely to be replaced by another term: *young man* or informal *lad*, *guy* etc. Both *boys* and *girls* are used informally of groups of adult male/female friends, and groups of male workers (soldiers, police, fire fighters, sometimes factory operatives, etc.) or female workers (typically in low-status occupations). While these are infantilizing usages, they are also ambivalent: showing affection and solidarity if used by speakers who are part of the group concerned, or who position themselves as part of the group; but often paternalistic, condescending, or demeaning if used by outsiders or those with higher status.

It seems reasonable now to expect that dictionary entries for *boy* and *girl* would have parallel explanations for primary senses, then present information about the various usages that relate to adults, including register and potential for offence. However, there are some surprising asymmetries and inconsistencies. Those in *ALD1* in 1948 could be predicted:<sup>9</sup>

(26) **boy**

- 1 a male child up to the age of 17 or 18
- 2 a son (of any age)3 a male servant

(27) **girl**

- 1 a female child of any age; a daughter
- 2 a female child not yet grown up; one who is not yet married3 a maidservant
- 4 a girl or woman working in a shop, office, etc.5 (*colloq., vulg.*) a sweetheart

But there are also curious asymmetries in *Cobuild2*, written in the 1990s by a strongly pro-feminist team (as was the 1987 first edition):

(28) **boy**

- 1 A **boy** is a child who will grow up to be a man.
- 2 You can refer to a young man as a **boy**, especially when talking about relationships between boys and girls.
- 3 Someone's **boy** is their son; an informal use4 You can refer to a man as a **boy**, especially when you are talking about him in an affectionate way.

<sup>9</sup> Here and below, for reasons of space I have mostly omitted examples given in dictionary entries for *boy* and *girl*, though these too are interesting and very relevant to examinations of ideological stance and sexism.

5 You can use **boy** when giving instructions to a horse or dog.

(29) **girl**

1 A **girl** is a female child.

2 You can refer to someone's daughter as a **girl**.

3 Young women are often referred to as **girls**. Some people find this use offensive.

4 Some people refer to a man's girlfriend as his **girl**; an informal use.

Missing altogether is the use of *boys/girls* to refer to friendship groups, and the offensive, mainly old-fashioned American, use of *boy* to address an inferior. There seems no clear reason for the different wordings of *boy* 1, 3, and *girl* 1, 2; nor for the inclusion of sense 5 of *boy* (or conversely exclusion of a parallel vocative for female horses and dogs).

While current online versions of Longman, Macmillan, and Oxford have more symmetrical entries and explanations for at least the primary senses of *boy* and *girl*, Cambridge does not:

(30) **boy**

a male child or, more generally, a male of any age

Their little boy (= their young son) is very sick.

**the boys** a group of male friends(also **our boys**) an approving way of speaking about your country's soldiers

(31) **girl**

a female child or young woman, especially one still at school:

a daughter:

a woman worker, especially when seen as one of a group:

a group of female friends

Macmillan is representative of the other three in its parity and careful labelling (though it is still partially asymmetric); it also adds a usage note at **girl**:

(32) **boy**

1 a male child

a. a son

2 a young man

3 a man of any age, especially when you are talking about where he comes from

a. (*American, offensive*) an extremely offensive word used for talking to a black man, especially in the past

4 **the boys** (*informal*) a group of men who are friends

(*British*) the members of a sports team

5 used when speaking to a male dog or horse

6 a boy or man of any age who has a particular job

(33) **girl**

1 a female child

a. a daughter

- 2 a female adult, especially a young one. This use is considered offensive by many women
- a. **girls** used for talking to or about a group of women, especially by women who are the same age or older. This is often considered offensive when used by men
- b. (*old-fashioned*) a young woman who works as a servant or in a shop, office etc
- 3 a female animal, especially a pet

#### **PHRASES**

**my girl** (*British spoken*) used by some people when talking to a girl or woman who is younger than they are, especially to show that they are angry. This is usually considered offensive

**someone's girl** (*old-fashioned*) someone's girlfriend

#### **Words that may cause offence: girl**

People sometimes say **girl** to refer to a young adult woman, but this use may cause offence. Avoid using **girl** if it would seem wrong to use **boy** about a young man of the same age. Do not use **girl** about an adult woman.

Though such uses of *girl* to refer to adult women have been problematized and contested – as has *lady*<sup>10</sup> – not all anglophone adult women feel so strongly about the words, and may even prefer to be called a girl, or lady, rather than woman, according to situational context.

It is interesting to compare entries for **boy** and **girl** in these mediated publishers' texts with those in crowd-sourced *Simple Wiktionary*, which are indeed simple and mainly asymmetric. Here, examples are included as reminders of the stereotyping potential with such words:

#### (34) **boy**

1 (countable) A **boy** is a male child.

He had a pretty wife and two little ones: a boy and a girl.

My oldest son was a Boy Scout in England.

The boys basketball team won five games in a row.

Two teenage boys died in the crash.

A 12-year-old boy stands at the window and watches two men outside.

#### (35) **girl**

1 A **girl** is a female child.

Many girls like to play with dolls.

I have two children: a boy and a girl.

2 (informal) A female person of any age (even a woman).

The girls are going out tonight, do you want to come?

I really love that girl.

3 (informal) A female animal.

My cat is a girl.

She is a girl cat.

---

10 See discussion by Lakoff (1975) and subsequent feminist linguists, who point out that *lady* trivializes even while apparently showing politeness and respect.

Where items referring to sexual behaviour and sexual preferences are concerned, comparisons between historical and current dictionaries show the extent of social change. For example, in all of the big five British learners' dictionaries, entries for *gay* give priority to the sense "homosexual", labelling as old-fashioned its sense "happy, cheerful"; the derogatory use of *gay* "stupid, absurd, inadequate", if included at all, is labelled as offensive. *Bisexual, heterosexual, homosexual, lesbian, same-sex, transgender* etc. are all routinely covered. Particularly interesting in April 2014 – at the time of writing, less than a month after same-sex marriage was legalized in the UK – is how far online versions of learners' dictionaries reflect this in entries for *husband, wife, marriage, married, marry*. For example, Longman and Oxford have primarily heteronormative, though symmetrical, explanations for husband and wife, while Cambridge (British English version only) has non-specific explanations but heteronormative examples:

(36) **husband**

the man that you are married to:

*I've never met Fiona's husband.*

(37) **wife**

the woman that you are married to:

*I met Greg's wife for the first time.*

*She's his third wife* (= she is the third woman he has been married to).

Macmillan's entries are also symmetrical and non-specific and include non-heteronormative examples (a change from its print edition of 2007, *MED2*):

(38) **husband**

a male partner in a marriage

Carole's husband died last year.

She isn't looking for a husband.

He may be separated from his husband and deported back to Venezuela.

(39) **wife**

a female partner in a marriage

I'd better phone my wife and tell her I'll be late.

a reception for the wives of the ambassadors

In April she became the proud parent of twins with her wife Alex.

The four online learners' dictionaries provide mostly non-specific explanations for *married, marry* and *marriage*, though often imply heteronormativity through choices of examples which indicate mixed-sex couples. However, Macmillan is explicit in extending its explanation (another change from *MED2*):

(40) **marriage**

the relationship between two people who are husband and wife, or a similar relationship between people of the same sex

*a long and happy marriage*

*Too many marriages end in divorce.*

**by marriage:** *I'm related to Bill by marriage* (= he is a relative of my husband or wife).

Compare the cultural information in entries and examples in *Simple Wiktionary*, which are almost entirely heteronormative:

(41) **married**

1. A man and a woman are **married** if they are husband and wife to each other. Usually when two people are **married** they live in the same house and they often have children. Two people have a special day to become **married**.

*I don't need to meet more young men – I'm already married.*

(42) **marry**

1 When two people **marry** they become husband and wife; that is, they become married. In many countries this is a legal agreement. In some cultures **marrying** is a part of the religion. **Marrying** is often done with a wedding (a special day for those people to marry).

*I cannot believe he **married** her when there are nicer girls out there.*

There are many other items which could be used to test how far dictionary texts represent attitudes towards sexuality and acknowledgement of changing paradigms, as realized in lexis and therefore in need of definition – not least *gender* itself, where all of the big four online learners' dictionaries currently offer purely male-female binary explanations of *gender*, though *gender* is now widely considered a social and cultural construction with non-binary variations. Oxford's explanation mentions the first of these points; of the dictionaries examined, but only *Simple Wiktionary* mentions both:

(43) (Longman)

the fact of being male or female

(44) (Oxford)

the fact of being male or female, especially when considered with reference to social and cultural differences, not differences in biology

(45) (*Simple Wiktionary*)

1 A living thing's **gender** is its sex: male (man, boy), female (woman, girl), both, or neither.

3 (*psychology*) Someone's gender is whether they behave like a boy or girl. This is called masculine or feminine, and not the same as male or female. [*sic*]

We can expect dictionaries eventually to adapt to new norms more fully; at the same time, we have to acknowledge that some of these new norms are not universally accepted by any means, and may seem abnormal, even abhorrent, to some sectors of the global usership. Thus the changing mores and attitudes of the culture within which dictionaries are written – specifically here the community of native-speakers of British English – may be at odds with the mores and attitudes of the markets and readerships to which the texts are presented: an interesting tension between language, lexicography, and receiver. Is it possible that culture-specific splinter dictionaries, with different ideological per-

spectives, may develop, for example for/in cultures where homosexuality is illegal or stigmatized, or where women do not have equal rights? Should this matter? and who has the right to say it matters?

## 5 Representing Age

The last area I want to consider is age and ageism, as represented in dictionaries: the subject of a case study and part of an ongoing research collaboration into discourses of ageing. Particular sites for potential ageism are those adjectives and nouns which reference age directly, such as *young*, *old*, *teenager*, *codger*, though many other items embed or entail notions of age indirectly, including adjectives with age-related semantic preferences: see discussion in Moon 2014.

This written version of my talk looks at just two items to demonstrate the issues in an area that has been, metalexically, underexplored. The first is *middle-aged* which, like *young*, *elderly*, *old*, is a generalized indicator of life stage. Since age labels carry evaluations – young is good, old is bad – any discussion of when a label begins or ceases to be appropriate is evaluatively loaded and ideologically weighted. Some dictionaries set time parameters for *middle age*, others are more vague; *Cobuild2* offers both strategies:

- (46) **Middle age** is the period in your life when you are no longer young but have not yet become old. Middle age is usually considered to take place between the ages of 40 and 60.

*Men tend to put on weight in middle age.*

These next entries and explanations are from current online dictionaries:

- (47) **middle age** (Cambridge)

the period of your life, usually considered to be from about 45 to 60 years old, when you are no longer young, but are not yet old:

*Once you reach middle age, you have to be sensible with your health.*

- (48) **middle age** (Longman)

the period of your life between the ages of about 40 and 60, when you are no longer young but are not yet old:

*Men who smoke are more likely to have heart attacks in middle age.*

- (49) **middle-aged** (Longman)

1 between the ages of about 40 and 60:

*a middle-aged businessman.*

- (50) **middle-aged** (Macmillan)

1 no longer young but not yet old:

*He seems prematurely middle-aged*

Oxford agrees with Cambridge as to age range, but while Longman agrees with *Cobuild2*, it does not agree with itself, since its online word focus feature for **old**, which appears automatically at **midd-**

**le-aged**, explains it as “aged between about 50 and 60 years old”. Examples, where included, sometimes have a hortatory flavour, or reference dullness and decline.

Dullness is more directly represented in subsidiary senses for *middle-aged* in learners’ dictionaries. The connotations of *middle-aged*, and its overall negative evaluation, in these senses are reflected in the following small selection of BoE corpus lines:

n’roll even, start to sound so middle-aged? <p> We’ll ignore the occasional  
with a son’s girlfriend. A middle-aged Conservative MP, the essence of  
tartly. Exactly: the balding, middle-aged fanclub is here early, packing DAT  
at his side, was a plump middle-aged lady in a brown jumper and navy-blue  
half the time!” <p> Meanwhile, middle-aged locals shuffle through the scene,  
<p> Wilson was a middle-sized, middle-aged man in a grey, herringbone suit. His  
a conversation between a middle-aged man and his wife about insurance.  
baldheads or of selfish middle-aged people.” Thomas Wentworth Higginson  
climbed out. An overweight middle-aged salesman trying to look slimmer in a  
into an outer office where a middle-aged woman sat at a secretarial desk

For example:

(51) (*Cobuild2*)

2 If you describe someone’s activities or interests as **middle-aged**, you are critical of them because you think they are typical of a middle-aged person, for example by being conventional or old-fashioned.

*Her novels are middle-aged and boring.*

(52) (*Cambridge*)

(*disapproving*) too careful and not showing the enthusiasm, energy, or style of someone young:  
*What a conventional, middle-aged attitude he has to life!*

(53) (*Longman*)

someone who seems middle-aged seems rather dull and does not do exciting or dangerous things:

*Living with Henry had made her feel middle-aged.*



(54) (Macmillan)

2 used for suggesting that someone's behaviour, clothes, etc. are boring and typical of middle-aged people:

*They are in their twenties, but have very middle-aged views.*

(55) (Oxford)

3 (disapproving) (of a person's attitudes or behaviour) rather boring and old-fashioned.

*He has a very middle-aged attitude to life.*

Perhaps Macmillan's example for its first sense seems to be semantically closer to its second sense; Longman's example provides a context of use, but nothing to distinguish *middle-aged* from *depressed*, *fulfilled*, *secure*, *young*, *happy*..., unless dullness is to be inferred from the name Henry (an unreasonable expectation). Compare too an entry in crowd-sourced *Urban Dictionary*, which in explaining its age reference also rationalizes its connotations:

(56) **middle aged**

- i. a period between early adulthood and old age, anywhere from 30 to 65 years old.
- ii. Something most people will not admit to being. (It sucks to be older than 29...)

Many other items, used to identify whole age groups or individuals within age groups, are also evaluatively charged and communicate attitude. *Youth* itself is ambivalent: sometimes a focus for nostalgia, the ideal, a life force; sometimes a focus for disapproval. Both evaluation and youth culture are bound up in its respelling *yoof*, my final example here, as in these BoE lines from journalistic media:

the computer games beloved of modern yoof. At the end we see him walking hand-competition striking a chord with yoof audiences who make the politically another attempt to hijack British yoof culture by taking over Arcadia, the have proved that, in an era where yoof is supposed to be paramount, age is me most by today's emphasis on Yoof is that when I was one of them it adduced, for not being attractive to yoof". Just how in touch with most yoof of a recent edition of the late night yoof prog will have spotted King (who West and amphetamine abuse. A great `yoof" read. <p> Bookshop To order these who are trying to get apolitical yoof to join the electoral register and having it better than the dole-bound yoof who came after them. They have a

The big four online learners' dictionaries all label *yoof* as informal and humorous:<sup>11</sup>

<sup>11</sup> Cf. discussion of youth/youth in Bennett et al. 2005: 380-382; Thorne 2009: 343-5.

(57) (Macmillan)

(*British, very informal, humorous*) young people. This word is used especially on television, in the newspapers etc, as a humorous way of spelling the word 'youth'

(58) (Oxford)

(*British English, informal, humorous*) a non-standard spelling of 'youth', used to refer to young people as a group, especially as the group that particular types of entertainment, magazines, etc. are designed for

But humour is only part of the pragmatics of its usage, which also seems to involve contempt and trivialization – condescension is evident in the corpus lines above. The explanation in *Urban Dictionary* is more expansive and explicit:

(59) **yoof**

2 Cynical description for a style of marketing or programming created by establishment or corporate interests that seeks to identify with the under-21's and thereby sucker them into parting with their cash or individuality with its promise of street credibility or non-conformity. Media vehicles or brands that tell kids what to do by creating an ersatz peer group for them which they then feel they have to conform to.

As with *middle-aged*, this alternative lexicography affords insights into attitudes and connotation that are beyond the mandates and controls of conventional dictionaries: see Smith (2011) for discussion of *Urban Dictionary* and its significance as a lexicographical text.

## 6 To Conclude

I have looked at only a small selection of entries and, as I warned at the outset, I have offered no solutions. My intention was instead to emphasize the problems that persist, are perhaps insoluble, in the lexicographical treatment of a disparate range of items where ideology and institutionalized attitudes come into play. I have focused almost entirely on monolingual learners' dictionaries, but definitions and analyses in inventory/concise dictionaries, bilinguals, dictionaries for children or school students, are no less problematic, as critical lexicography has repeatedly found. There is massive potential in online dictionaries, including collaborative crowd-sourced dictionaries, for radicalism and inventiveness in entry design and for more effective representations of meaning – including the meanings of ideologically loaded words; but there is also potential for the filtering of world views, re-presentation rather than representation, in ways that may not seem desirable to us, here with our western perspectives and our own filtered views.

In his seminal paper on dictionary definitions, or explanations, Hanks says:

In the last resort, perhaps, all meanings are displaced, since all meanings rely on constructive interpretation by the hearer/reader, as well as by the utterer. If this is true, there is no such thing as literal

meaning, and a dictionary explanation is no more than a compromise with the impossible, a desperate attempt to state the unstateable. (1987: 135)

Yet desperate attempts go on, and meaning remains at the heart of any dictionary, of overwhelming importance. Over thirty years ago, Béjoint drew attention to his survey finding that “87% of the students [advanced learners of English] placed meaning among the three most often sought-after pieces of information” in a dictionary (1981: 215), substantially more than any other information type; at the same time, the highest-ranked cause of failed look-ups (29%) was “unsatisfactory definitions” (1981: 217), almost one in three.<sup>12</sup> Has the situation changed that much? Certainly with respect to ideologically loaded words, the difficulties of producing satisfactory entries are compounded by the complexities and instability of their meanings, the questions of stance and audience, and the balancing of what words mean with what they can be said to mean or be allowed to mean: moreover, the very process of composing entries for such words is essentially an ideological act. I may not have offered solutions, but I hope that at least I have demonstrated something of the nature, and seriousness, of the issues.

## 7 References

### 7.1 Dictionaries Cited

- (ALD1) Hornby, A.S., Gatenby, E.V., Wakefield, H. (1948). *A Learner's Dictionary of Current English*. Oxford: Oxford University Press. (Previously published as *Idiomatic and Syntactic English Dictionary*, Tokyo: Kaitakusha, 1942).
- (Cambridge) *Cambridge Learner's Dictionary Online*. Accessed at: <http://dictionary.cambridge.org/> [April 2014].
- (Cobuild2) Sinclair, J., Bullon, S. (eds.) *Collins Cobuild English Dictionary* (1995, 2nd edition). London and Glasgow: HarperCollins.
- (Longman) *Longman Dictionary of Contemporary English Online*. Accessed at: <http://www.ldoceonline.com/> [April 2014].
- (Macmillan) *Macmillan English Dictionary Online*. Accessed at: <http://www.macmillandictionary.com/> [April 2014].
- (MED2) Rundell, M. (ed.) (2007, 2nd edition). *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.
- (OED) *Oxford English Dictionary* Accessed at: <http://www.oed.com> [April 2014].
- (Oxford) *Oxford Advanced Learner's Dictionary Online*. Accessed at: <http://www.oxfordlearnersdictionaries.com/> [April 2014].
- (Simple Wiktionary) *Simple English Wiktionary*. Accessed at: [http://simple.wiktionary.org/wiki/Main\\_Page](http://simple.wiktionary.org/wiki/Main_Page) [April-May 2014].
- (Urbandictionary) *urban dictionary* Accessed at: <http://www.urbandictionary.com/> [April 2014].

---

<sup>12</sup> Other later surveys have still higher figures.

## 7.2 Other Literature

- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Barnickel, K.-D. (1999). Political correctness in learners' dictionaries. In Th. Herbst, K. Popp (eds.) *The Perfect Learners' Dictionary?*. Tübingen: Niemeyer, pp. 161-174.
- Béjoint, H. (1981). The foreign student's use of monolingual English dictionaries: a study of language needs and reference skills. In *Applied Linguistics*, 2(3), pp. 207-222.
- Béjoint, H. (2010). *The Lexicography of English*. Oxford: Oxford University Press.
- Bennett, T., Grossberg, L., Morris, M. (2005). *New Keywords: A Revised Vocabulary of Culture and Society*. Oxford: Blackwell.
- Benson, P. (2001). *Ethnocentrism and the English Dictionary*. London: Routledge.
- Caldas-Coulthard, C.R., Moon, R. (2010). Curvy, hunky, kinky: using corpora as tools in critical analysis. In *Discourse and Society*, 21(2), pp. 1-35.
- Connor Martin, K. (2005). Gendered aspects of lexicographic labeling. In *Dictionaries*, 26, pp. 160-173.
- Fuertes Olivera, P.A. (2009). The function theory of lexicography and electronic dictionaries: Wiktionary as a prototype of collective free multiple-language internet dictionary. In H. Bergenholtz, S. Nielsen, S. Tarp (eds.) *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang, pp. 99-134.
- Gabrielatos, C., Baker, P. (2008). Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996-2005. In *Journal of English Linguistics*, 36(1), pp. 5-38.
- Graham, A. (1975). The making of a nonsexist dictionary. In B. Thorne, N. Henley (eds.) *Language and Sex: Difference and Dominance*. Rowley, Massachusetts: Newbury House, pp. 57-63.
- Hanks, P. (1987). Definitions and explanations. In J. Sinclair (ed.) *Looking Up*. London and Glasgow: Collins, pp. 116-136.
- Hanks, P. (2013) *Lexical Analysis: Norms and Exploitations*. Cambridge, Mass.: MIT Press.
- Holmes, J., Sigley, R. (2001). What's a word like *girl* doing in a place like this?. In P. Peters, P. Collins, A. Smith (eds.) *New Frontiers of Corpus Linguistics*. Amsterdam: Rodopi, pp. 247-263.
- Hornscheidt, A.L. (2011). Postcolonial continuities in Danish monolingual dictionaries: Towards a critical postcolonial linguistics. In E.A Anchimbe, S.A. Mforteh (eds.) *Postcolonial Linguistic Voices: Identity Choices and Representations*. Berlin: De Gruyter Mouton, pp. 265-298.
- Kachru B.B., Kahane, H. (eds.) (1995). *Cultures, Ideologies, and the Dictionary: Studies in Honor of Ladislav Zgusta*. Tübingen: Niemeyer.
- Krishnamurthy, R. (1996). Ethnic, racial and tribal: the language of racism?. In C. Caldas-Coulthard, M. Coulthard (eds.) *Texts and Practices: Readings in Critical Discourse Analysis*. London: Routledge, pp. 129-149.
- Lakoff, Robin (1975). *Language and Woman's Place*. New York: Harper Colophon Books.
- Landau, S.I. (1985). The expression of changing social values in dictionaries. In *Dictionaries*, 5, pp. 261-269.
- Landau, S.I. (2001, 2nd edition). *Dictionaries: The Art and Craft of Lexicography*.
- Leech, G. (1981, 2nd edition). *Semantics*. Harmondsworth: Penguin.
- Louw, B. (1993). Irony in the text or insincerity in the writer? - the diagnostic potential of semantic prosodies. In M. Baker, G. Francis, E. Tognini-Bonelli (eds.) *Text and Technology: in Honour of John Sinclair*. Amsterdam: John Benjamins, pp. 157-176.
- Moon, R. (1989). Objective or objectionable? Ideological aspects of dictionaries. In M. Knowles, K. Malmkjær (eds.) *Language and Ideology (ELR Journal)*, 3. Birmingham: English Language Research, University of Birmingham, pp. 59-94.
- Moon, R. (2014). From gorgeous to grumpy: adjectives, age and gender. In *Gender and Language*, 8(1), pp. 5-41.
- Ogilvie, S. (2013). *Words of the World*. Cambridge: Cambridge University Press.
- Russell, L.R. (2012). This is what a dictionary looks like: the lexicographical contribution of feminist dictionaries. In *International Journal of Lexicography*, 25(1), pp. 1-29.

- Semino, E. (2008). *Metaphor in Discourse*. Cambridge: Cambridge University Press.
- Sigley, R., Holmes, J. (2002). Looking at *girls* in corpora of English. In *Journal of English Linguistics*, 30(2), pp. 138-157.
- Sinclair, J.M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J.M. (2004). *Trust the Text*. London: Routledge.
- Smith, R.E. (2011). Urban dictionary: youth slang and the redefining of definition. In *English Today*, 27(4), pp. 43-48.
- Svensén, B. (1993) *Practical Lexicography: Principles and Methods of Dictionary-making*. Oxford: Oxford University Press.
- Thorne, T. (2009). *The 100 Words that Made the English*. London: Abacus.
- Whitcut, J. (1984) Sexism in dictionaries. In R.R.K. Hartmann (ed.) *LEXeter '83 Proceedings*. Tübingen: Niemeyer, pp. 141-144.
- Williams, R. (1988, expanded edition). *Keywords: A Vocabulary of Culture and Society*. London: Fontana.



# **The Dictionary-Making Process**





# The Making of a Large English-Arabic/Arabic-English Dictionary: the Oxford Arabic Dictionary

Tressy Arts  
Oxford Arabic Dictionary  
tressy.arts@gmail.com

## Abstract

In this presentation, we will illustrate the process of making our brand-new Arabic-English/English-Arabic dictionary, which is due for publication in print and online in August 2014. It is intended for speakers of both English and Arabic. It contains over 26,000 entries on each side, including many up-to-the-minute words and expressions. Collocations and examples are an important feature. The dictionary has been compiled using dictionary writing software that enables editors to work and communicate with one another regardless of their location. It will be available in both print and online.

We show the entire process of making an Arabic dictionary, from finding a reliable framework in both languages, to developing a unique online functionality. We show the difficulties lexicographers face when compiling an Arabic dictionary, and the ways in which we dealt with those. In addition, the Oxford Arabic Dictionary has quite a few features that are entirely new to Arabic dictionaries, and we illustrate how we went about developing those.

**Keywords:** Arabic; bilingual; dictionary-making process

## 1 Arabic Lexicography

Arabic lexicography has a centuries-old tradition, starting with the *Kitāb al-'Ayn*, a large monolingual dictionary, in the eighth century. This impressive legacy is both a blessing and a curse, since conservatism rules, and the mediaeval tomes like the 13<sup>th</sup> century *Lisān al-'Arab* still count as *the* standard in lexicography. Modern monolingual dictionaries often do little more than repeat what has been said before, regardless of whether the senses, examples, or even the headwords mentioned are still in actual use. Many Arabic lexicographers, linguists and laymen see Classical Arabic, the language of the Koran and the pre-Islamic poetry, as the perfect standard, and any deviations from that are seen as corrupting the language. Therefore many loanwords and words derived from colloquial Arabic are not included in modern dictionaries, despite them being commonplace in Standard Arabic texts. Haywood says in his "Arabic Lexicography": "The lexicographers helped to keep the written language static, and to aid the understanding of it, as the spoken dialects diverged more and more from it."

(Haywood 1960: 116). There are several Arabic Language Academies which aim to find proper Arabic words (that is, words which conform to the Arabic root-and-pattern system for word forming, rather than words which are simply English words written in Arabic letters) for new concepts, but they are slow-moving and don't always communicate clearly, so that loanwords for new concepts often are commonplace long before an approved version is released. Sometimes the two continue to exist side by side (*tilifūn* and *hātif* (telephone)), sometimes the loanword remains strong (*tilifizyūn* (television)), sometimes the approved version gains prominence (*hāsūb* (computer) is more common than *kumbyūtīr* these days).

Still, it is rare to find modern words in monolingual dictionaries, either loanwords, or ones formed according to Arabic morphology. One will be hard put to find *hāsūb* (computer) in most, and a common modern word like *nazzala* (to download) I haven't been able to find in any, not even online.

Bilingual lexicography with Arabic as source or target language also has a long history. Already in the 11<sup>th</sup> century al-Zamakṣarī published an Arabic-Persian dictionary, and al-Kāṣḡarī a Turkish-Arabic one. For Europeans, the first bilingual dictionary was Golius' *Lexicon Arabico-Latinum*, published in Leiden in 1653. The first Arabic-English one was produced by Edward William Lane in the late 19<sup>th</sup> century.

Looking at current-day Arabic-English/English-Arabic lexicography, however, one cannot help but be a bit disappointed with what's there, especially considering the status of both languages as major world languages. The standard Arabic-English work, Hans Wehr's *Dictionary of Modern Written Arabic*, was translated from German and not much updated since it was first published in 1961. Though groundbreaking at the time, it is nowhere near comparable to modern bilingual dictionaries for most other languages. Obviously the word list is outdated, but also the presentation of the entries is not really what one would expect from a modern dictionary: long lists of possible English translations are given for most Arabic words, without any guidance for the user on which translation to choose in which context; no word senses are distinguished. No examples are given and only very few collocations, many of which don't actually exist in modern Arabic. Yet, this is still the dictionary that everyone translating from Arabic into English will use.

The situation for English-Arabic is even direr. Arguably the best option is still Oxford's English-Arabic (hand-written!) dictionary from 1972 (cf. Benzehra 2012); the most popular work is *al-Mawrid* (Baalbaki & Baalbaki 2013), which is updated regularly, but has certain major flaws. The English used seems mostly derived from very old texts, and its policy for including new words in updated editions is not clear: the most recent edition, from 2013, has "fuscous" and "silvern", featured new entries "naughties" [sic] and "smirt", but not "blog" or "text message". Examples given seem to be taken from very old English dictionaries/texts ("swallows that affect chimneys", "the hes would quarrel and fight with the females"). As to the Arabic translations, many of the senses given cannot be attested to exist in modern English, senses are ordered from oldest to most modern, leading to the most common sense often being the last one, and the editors even create new Arabic "words" to serve as translations:

compilers of bilingual dictionaries are not only entitled to coin words which may or may not gain currency, but that coinage becomes an essential duty of theirs, especially with a language like Arabic, where a huge number of terms, particularly scientific ones, are lacking. (Baalbaki 2004: 68)

This statement is disputable: Arabic does not lack terms for scientific or any other areas, but some of the terms will be loanwords or multi-word constructions, so it depends on what is understood by “Arabic terms”. We believe that it’s more useful for an English-Arabic dictionary to give those loanwords or multi-word phrases, than to coin new words which are useful for neither decoding or encoding users.

Other European languages fare slightly better, but not much, with the exception of Dutch, which has some good and modern dictionaries, published at the beginning of this century (*Woordenboek Nederlands-Arabisch/Arabisch-Nederlands* (Hoogland et al. 2003) and *Leerwoordenboek Arabisch* (Van Mol & Bergman 2001)).

There are plenty of online Arabic dictionaries, but the accuracy there usually leaves much to be desired.

So in developing a large bilingual Arabic-English/English-Arabic dictionary that was corpus-based and could live up to modern expectations, our team was truly taking on a unique endeavour. Not being able to rely on any previous works, either bilingual or monolingual, all progress had to be made by painstaking research.

## 2 The Arabic Language

The lack of reliable predecessors is far from the only complicating factor in Arabic lexicography. To highlight a few more:

### 2.1 Diglossia

In the Arabic world a diglossia exists, with the “high” variant, Modern Standard Arabic, being the accepted language for any written and official spoken discourse. Meanwhile, a multitude of “dialects” or “colloquials” are the languages people actually speak. These are officially known as dialects, but are often mutually unintelligible, and can be considered different languages on purely linguistic grounds. It is hard to say how many exist, as there is a dialect continuum, but in many countries the dialect of the capital has the major status (so if you pick up a text book or travel guide in “Egyptian Arabic”, this will be the dialect of Cairo, etc.). The “dialects” have no agreed orthography, though they are used more and more in written communication, mostly on the internet, and even a few novels have appeared.

Since we wanted to make a broadly useful dictionary, not focusing on one or a few dialects, and since it is hard to write dialects because of the lack of standard orthography, we chose to use (almost) exclu-

sively Standard Arabic for the Arabic in the dictionary. Even for English expressions which are very informal or chatty, we have tried to give Standard Arabic equivalents, or if needed descriptions. Both languages have level markers so the user is alerted if there is a difference in level.

Then we had to devise a strategy for deciding which words are indeed Modern Standard Arabic, and should be included in the dictionary. We could not simply include all written words in general texts, since these days many dialect words are written. We could also not rely on dictionaries to verify existence of words, since many common modern words are not listed in monolingual or bilingual dictionaries. So the criterion we decided on was that words which are used without quotation marks in a reasonable number of non-specialist, otherwise Standard Arabic texts, could themselves be considered Standard Arabic, and therefore deserved a place in the dictionary. On the other hand, if a word is only found in dictionaries, we did not include it.

## 2.2 No Native Speakers

It is standard practice these days in writing bilingual dictionaries to have them compiled by native speakers of the target language. However, there are no native speakers of Modern Standard Arabic; children in Arabic countries grow up with one of the dialects or even a non-Arabic language (Berber, Kurdish) as their mother tongue. Standard Arabic is learnt, like a foreign language, at school and in the mosque. Research has shown that indeed Arabic people process Standard Arabic in the brain in the same way as foreign languages (Ibrahim & Aharon-Peretz 2005). So although the editors hired were highly educated linguists, they lacked that native speaker sense that lexicographers for most other languages have.

## 2.3 Geographical Spread

In addition, the Arab world has a very large geographical spread, and regional language preferences exist even in Standard Arabic, so editors need to constantly be wary of using words and phrases that are limited to their home country. Very little research has been done in this area, partly because of the prevailing ideal of Arabic being one uniform language.

If we take all these factors together (no reliable antecedents, no native speaker sense, uncharted geographical differences), we can understand that it's very hard for the "native speaker"<sup>1</sup> editors to feel secure in their translation of a word or phrase, necessitating elaborate checking in a corpus and on the internet, as well as discussion with native speakers from other locations, to find the right terms. This meant that the project took much longer, and was more expensive, than originally envisaged.

---

1 We will use this to refer to native speakers of one of the Arabic dialects with good knowledge of Standard Arabic, for want of a better term.

### 3 The Bases

For the Arabic-English side, the data of Hoogland's Arabic-Dutch *Woordenboek* mentioned above, which had got very good press, was licensed.

For the English-Arabic side, we used an English framework that had been developed for use as a basis for Oxford unabridged bilingual dictionaries, expanded with words that are especially relevant for Arabic, like the English names of the Islamic months.

### 4 Vocalization

The Arabic script is a consonant script. Vowels are indicated by diacritic signs over and under the letter (see figure 1), but are not commonly written, so a word like *al-maḡrib* (Morocco) will be written as *al-mḡrb*. In addition, the three cases Arabic has are most often only indicated by vowels, and hence not visible in most texts.



Figure 1: Vocalized (left) and unvocalized (right) Arabic.

For learners of Arabic it is useful to be able to find all vowels both in the Arabic-English side (if one doesn't know a word, one probably doesn't know how to pronounce it), and in the English-Arabic side (when one finds a new Arabic word, it's useful to find the vowels). Similarly, case endings are useful to understand the syntax and word combinations. So in order to be most useful for non-Arab users, as well as clearest for Arab users, we wanted all Arabic on both sides of the dictionary, headwords, translations, examples and descriptions, to have full vocalization, including the case endings. This placed an extra burden on our editors, since most people are not used to writing vocalized Arabic, and there is no standard vocalization system, so we had to devise our own rules (most were taken over from the system used in the Hoogland dictionaries). It also made the use of an adapted font necessary, since most Arabic fonts are not designed with the vowels in mind, and they can "disappear" in certain letter combinations.

For the English-Arabic side, not only did we want the translations to be written in the above-mentioned vocalization system, we also wanted to provide the unpredictable grammatical information for single-word Arabic translations (the plurals of nouns and adjectives, and conjugational information and infinitives of many verbs are unpredictable in Arabic, so it's useful when this is provided with a word functioning as a translation). This information was already present for all the headwords in the Arabic framework, so a "translation picker" tool was developed within Oxford University Press: the

English-Arabic translator could enter an unvowelled word in this tool in the dictionary database, and was presented with all possible vocalized headwords in the Arabic data, with their grammatical information (see figure 2).

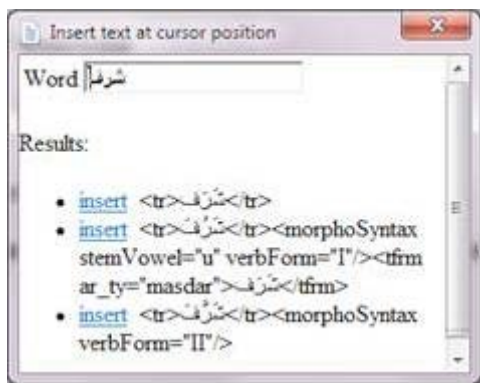


Figure 2: The Translation Picker in use.

By simply selecting the one they wanted, a correctly vocalized word with all relevant grammatical information was entered as translation.

This had an added advantage: if the translator could not find the word they wanted as translation in the Translation Picker, it meant that that word was not in the Arabic lemma list for the Arabic-English dictionary. The translator then made a note for the chief editor, who would decide if the Arabic word in question was a valuable addition to the Arabic-English lemma list, and add it to the latter if it was.

## 5 The Database Used

The dictionary writing software used was DPS (Digital Publishing System) produced by French company IDM, a database system specialized in creating and developing dictionaries, which is used by Oxford University Press and many other publishers. Its application, the Entry Editor, lets users connect to the central database in Oxford and download and upload entries to work on. It lets users edit entries uploaded by others, keeps track of who made which changes, lets you revert to any earlier uploaded version, and allows users to communicate with one another through a system of searchable annotations that can be made on every level (entry, sense, translation, etc.), a very valuable tool for a dictionary where much discussion of terms needs to take place. Via its search engine, editors could make highly specific searches, for example every translation that has a “region: Egypt” attached, and for the project managers its workflow manager allowed us to distribute the workload and keep track of progress.

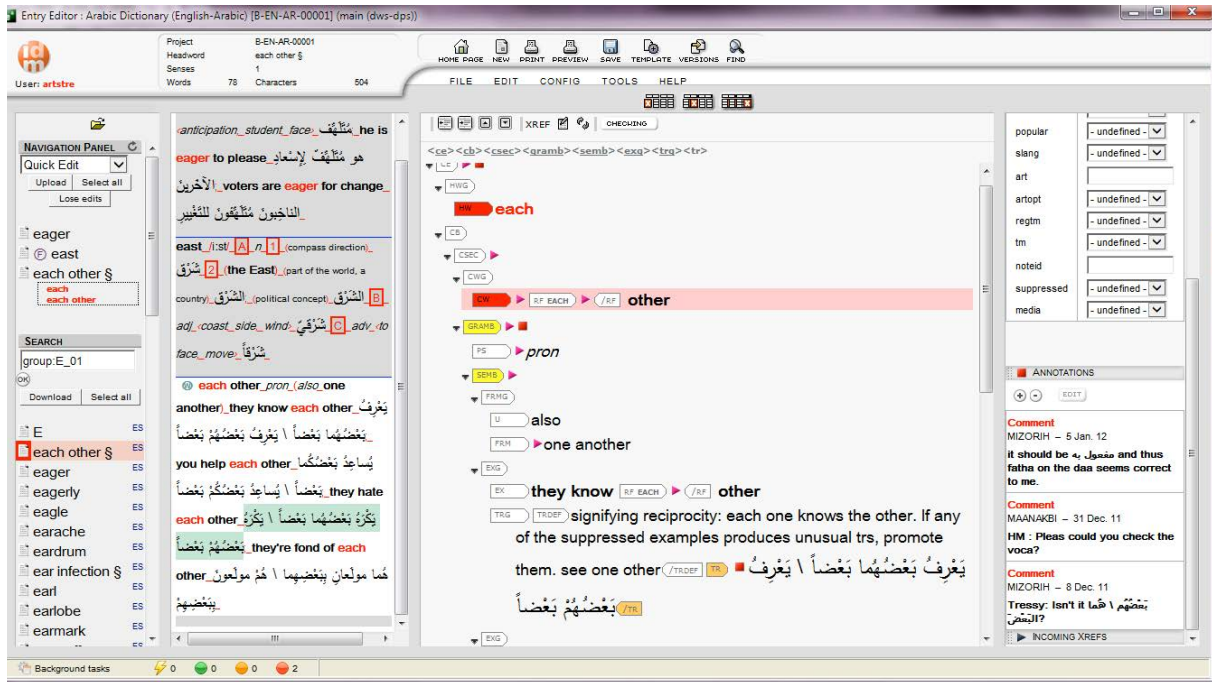


Figure 3: The Entry Editor showing entries in a group, preview, tag tree, and attributes and annotations.

## 6 The Corpus Used

The Oxford English Corpus contains 2.5 billion words of modern English from all over the globe, and as such was a very valuable resource for our editors. A unique new resource, however, was the Oxford Arabic Corpus, made searchable with the Sketch Engine software, developed by Lexical Computing Limited of Brighton. This enabled us to do truly unprecedented work.

The Oxford Arabic Corpus comprises the Arabic Gigaword Corpus Fourth Edition from Linguistic Data Consortium: 840 million words of news text from nine publications covering the period 1996-2008, plus 10 million words of fiction from the Arabic Writers Union of Damascus, and 30 million words from Arabic Wikipedia.

After the corpus was assembled, the raw data was processed using MADA (Habash, Rambow & Roth 2009). MADA uses the Buckwalter morphological analyzer (by Linguistic Data Consortium) to provide alternative analyses (part-of-speech, vowelised form, and lemma) of each input token, after which a Support Vector Machine classifier ranks the competing analyses in context. The highest ranked analysis is loaded into the Sketch Engine corpus software (Kilgariff et al. 2004), allowing items to be searched and concordanced by word form or lemma, and collocate phrases to be displayed by part of speech structure and grammatical class, as illustrated in Figure 4.

Here we see the Word Sketch for the word *tifl* (child). The abbreviation with the asterisk is the search term, and the sequence is from right to left, so the top left column is verbs (V) followed by the search



term (N for noun) in the accusative (a). The top collocate here, with 730 results, is *anjaba* (to give birth to). The results give us collocates in the form of verbs, nouns in so-called genitive constructs, adjectives, and prepositions with other nouns.

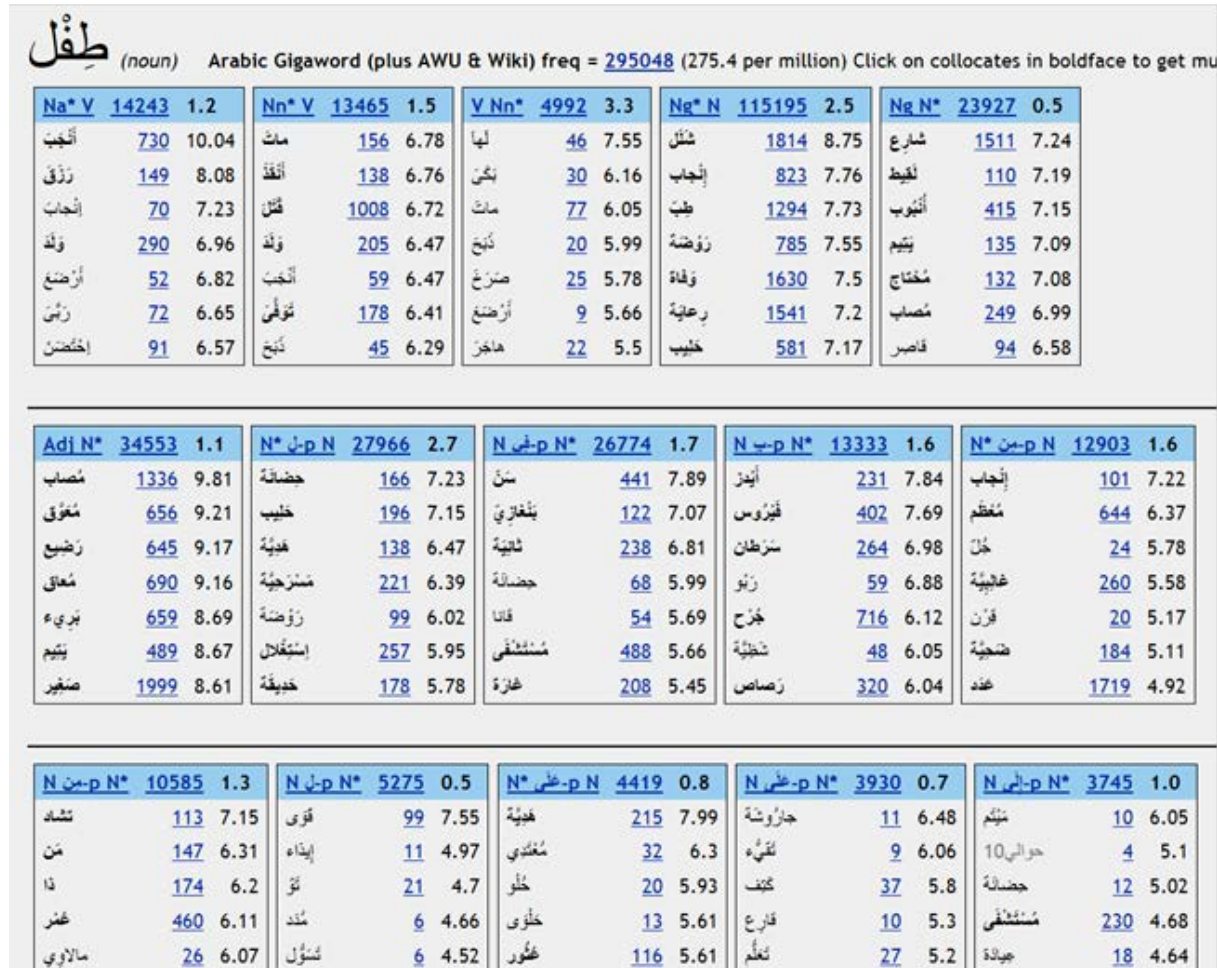


Figure 4: Word Sketch for ṭifl (child).

This allowed us to fine-tune translations, add relevant collocations and examples, and discover new words and senses.

Using the corpus wasn't without its pitfalls though. Correctly tagging Arabic is much more complex than English or Italian. Due to the complexity of Arabic's morphology and the surface-form ambiguity, mistagging in the corpus is inevitable - *wilz* (Wales) is consistently analysed as a form of *lazza* (to be compressed), because *wilz* is not in the lemma list and *wa-yaluzzu*, a conjugated form of *lazza*, has the same surface form: WiYLZ and WaYaLuZzu (the capitals are the letters visible in the surface form). Also, because the case endings aren't usually visible, and Arabic syntax has a VSO structure, differentiation between verb + nominative noun and verb + accusative noun is nearly impossible. In Figure 4, the biggest result in the verb + nominative noun table is *qatala* (to kill). We can be reasonably certain that in most of those cases the child is the unfortunate object, rather than the subject, and so is in fact an accusative noun.

Editors therefore needed a sharp critical eye when analysing the results.



## 7 Finding New Words and Examples in the Corpus

The English framework was developed by a skilled team of Oxford lexicographers, and continued to be expanded, so we could reliably assume that all relevant words and phrases were included. However, for the Arabic, this was a different story: the Arabic-Dutch dictionary was developed in the nineties, when Arabic corpus linguistics was still very primitive. The editors could only search the corpus on surface forms, which is remarkably difficult in a language where a simple verb like “to eat” can have no fewer than 46 surface forms, most of which also can mean “to feed”. They found ways to work around those problems though (Hoogland 2004), and the chief editor calculated that the dictionary covered 99.95% of (non-lemmatized) word forms in Modern Arabic texts (Hoogland 2003). However, with the new technology at our disposal in the form of the Oxford Arabic Corpus and Google, we could certainly see many possibilities for improving and expanding the Arabic framework.

We started on the level of the vocabulary, comparing the lemma list of the corpus to the Arabic-Dutch lemma list, and added any lemmas we thought warranted inclusion, like *rawwasa* (to sharpen; to supply with a header). However, this only discovered words that were already listed in the Buckwalter lexicon. Because of Arabic’s complex morphology, it’s often not possible to automatically distil a lemma from a string if the lemma is unknown, so many surface forms were still unidentified. Mohammed Attia, one of our translators and a computational linguist, had developed a system to distil potential lemmas from the corpus by checking frequency and compliance with Arabic morphological patterns (Attia et al. 2014). These potential lemmas were again compared to our lemma list, and thus we managed to find genuine new words, many of which not only weren’t listed in the Arabic-Dutch data, but had not been previously listed in any Arabic dictionaries, like *tamāhā* (to be congruent) and *ṭawā’ifi* (sectarian).

So altogether we had four ways to expand the Arabic lemma list to contain more, currently relevant, entries: listing Arabic translations for English words that weren’t in the Arabic data, comparing the lemma list of the Gigaword corpus with the Hoogland lemma list, using Attia’s potential lemma extractor, and old-fashioned handwork: critical reading of web sites and being alert to any newly developing words and collocations, like *al-rabi’ al-‘arabi* (Arab spring) and *taḡrīda* (tweet). In these ways we managed to expand the lemma list by over 2,000 entries. At the same time, we pruned entries that were deemed obsolete or archaic, thus removing about four hundred entries from the original Arabic data. This latter was not a priority however, since, from a user’s point of view, obsolete entries are not in the way: if a user doesn’t encounter the word, they are not going to look it up; so we focused on improving the relevant entries rather than on pruning the less relevant ones. All in all we raised the number of Arabic entries from 24,682 to 26,316.

On the level of individual entries, the editors used the corpus, as well as Google, to check the existing examples for relevance, to find new or better examples and collocations, and on occasion find entirely new senses for existing words (Arts & McNeil 2013). Especially on this microstructural level, the Arabic-English dictionary has been greatly expanded compared to its Arabic-Dutch predecessor.

## 8 Microstructure and Translations

On both sides of the dictionary, entries are divided into one or more senses, which have disambiguators in the source language indicating the meanings. Many senses have examples to illustrate typical uses of the headword in that sense. Idioms are given separately, outside the scope of the senses.

Several types of translations are given:

- the direct translation, which is an equivalent of the headword in that sense and can be used as its translation in most contexts, or which is the equivalent of the example given, e.g. “computer” and *ḥāsūb*.
- the approximate translation, which is a nearly equivalent translation, or a translation in certain contexts (which is then specified), e.g. *Allāhu yuḥallika*, literally “God bless you”, which is used to thank someone, gets the approximate translation “thank you”. Approximate translations are indicated with an approximation sign (≈).
- the translation or approximate translation followed by an explanation in brackets. Often the translation is useful for the encoding user, and the explanation for the decoding user, with the explanation further specifying the translation, e.g. “muezzin (*mosque official who recites the call to prayer*)”. The explanation is in brackets in the English-Arabic and in brackets and italics in the Arabic-English.
- the definition: where a headword or example doesn’t have an equivalent in the target language (for English e.g. “haggis”, for Arabic for example many Islamic terms), a description of what it means is given in the target language. In the English-Arabic these definitions are in square brackets, in the Arabic-English in italics.

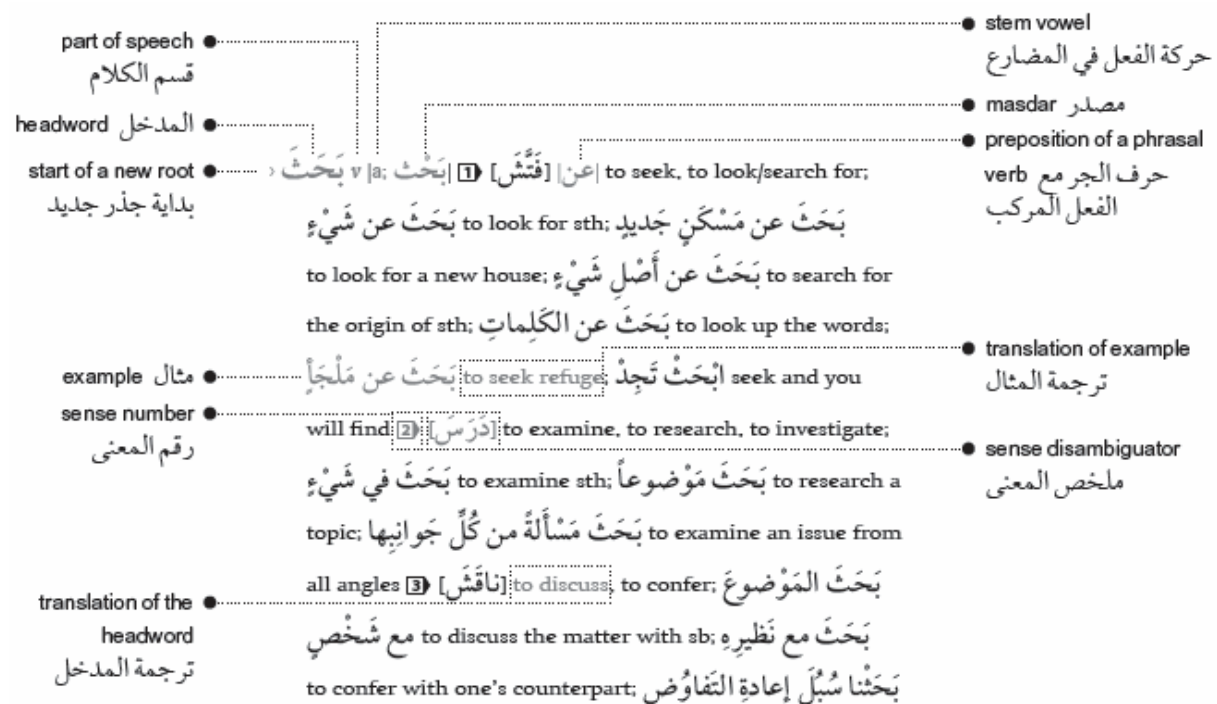


Figure 5: An Arabic-English entry with some aspects of its structure illustrated.

## 8.1 Arabic-English Translations

The existing Hoogland Arabic-Dutch dictionary was translated by Dutch Arabists with a good command of English. They made sure to keep the translation relation between the Arabic and the English in mind, rather than translating from Dutch to English, but the existing Dutch translations were too valuable a tool to disregard. Also, most of the editors had previously worked on the Arabic-Dutch dictionary, so they were familiar with Arabic lexicography and with this specific dictionary.

Then all entries were reviewed by native English-speaking Arabists, correcting the English where necessary and further improving the entries.

Since the original was made in a time when corpus research was hardly possible, the existing entries and examples were checked in the corpus and on Google: are the senses and examples representative of current Arabic, or are they so-called “dictionary words/examples”, copied from monolingual dictionaries that are not in actual use any more? The latter were weeded out and where needed we replaced the examples by new ones distilled from modern language found in the corpus and Google. We checked if all senses were attested, and checked senses we couldn't find evidence for with native speakers. Sometimes the corpus showed senses that were not yet listed, and we added those. During the entire process, we could ask native speakers' or other editors' opinion via annotations.

See figure 5 for an example of an Arabic entry and its structure.

## 8.2 Arabic-English Translations

The English-Arabic side of the dictionary was formed by having the English framework translated into Arabic by translators and lexicographers from several Arabic countries: Algeria, Tunisia, Egypt, Palestine, Lebanon, and Iraq. For certain fields, namely Medicine, Technology, and Law and Business, bilingual experts were found to advise on the terminology. The English framework contains field markers for many entries, so exporting all entries in a specific field was easily done.

Since, as I stated above, no one is a true native speaker of Arabic, it was important for the translators and reviewers to be able to check the translations not just against their own language sense, but also against a corpus. Even a 900-million-word corpus is not quite reliable enough to state that a certain word sense/construction is never used, so for verification of usage of Arabic translations of English terms, Google proved invaluable. Using quotation marks makes it possible to search for exact constructions, enabling translators and reviewers to verify that the translations are actually in use in modern Arabic.

Annotations were used to communicate with other native speakers verifying translations or asking for suggestions (“in Iraq we say this, but do you also use this in North Africa?”).

All entries were, after translation, reviewed one or more times by revising editors, at least one of which, for every entry, was a native speaker. Often the reviewer would discuss with the original translator to find the best solution for tricky aspects.

See figure 6 for an example of an English-Arabic entry and its structure.

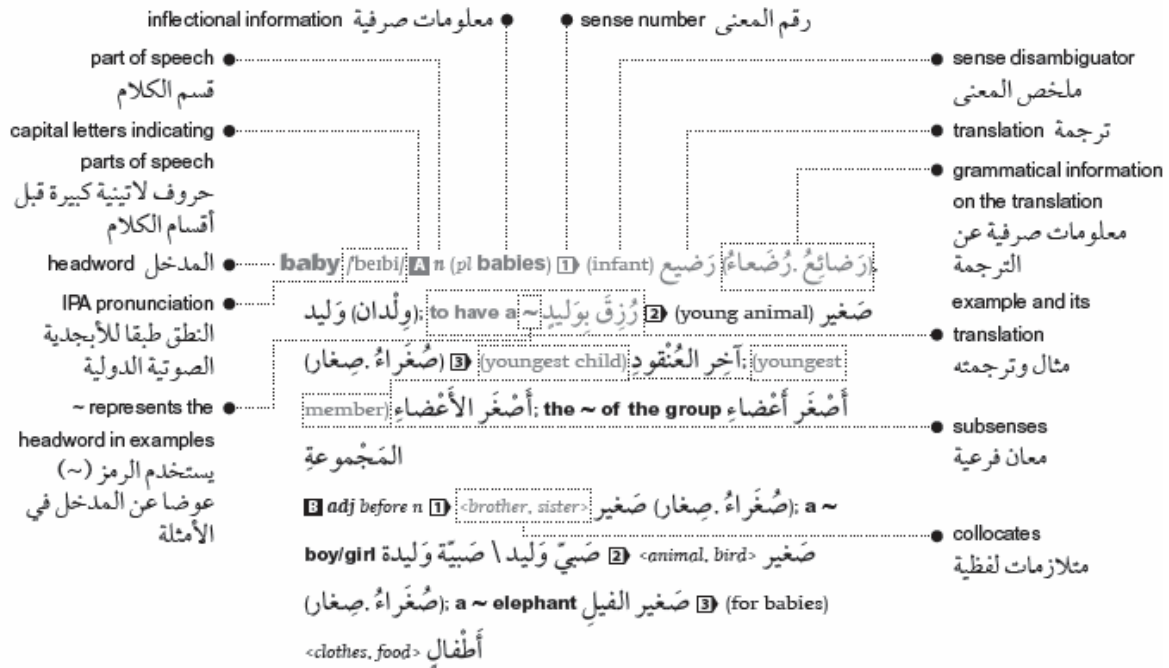


Figure 6: An English-Arabic entry with some of its structure illustrated.

## 9 Technical Challenges

A number of technical challenges had to be overcome in getting the dictionary writing software to work in harmony with the Arabic. The first challenge was the simple fact that Arabic is written from right to left and the letters within a word are joined up. This causes problems whenever Arabic and computers come into contact with each other, and made the news when Arabic and Persian visitors to the London Olympics were told the equivalent of N O D N O L O T E M O C L E W.

We had to make sure that the software could deal with English and Arabic, and even with English words inside Arabic tags, and vice versa, which gave the OUP dictionary technology staff no end of work. One of the largest problems turned out to be any numbers in examples, which in Arabic go in the same direction as in English, but which consistently were turned back to front in Arabic tags, since we had given the instruction that Arabic tags needed to output from right to left! The dictionary technologists then developed “bidirectional-override” tags, in which data that needed to output in a different direction than the default direction could be entered. This worked inasmuch as the numbers came out correctly... however, the rest of the data switched place, so the start of the sentence jumped to after the number and the back to before it. After months of tinkering we eventually had to instruct the typesetters to set all numerals back to front.

An additional complication is that our text is not just Arabic, but half Arabic and half English. Due to the large volume of text, we had to have English and Arabic constantly interplay, rather than having them in separate columns (see Figure 7).

<p>بَحْث n   بُحُوثُ, بُحُوثَاتُ, بُحُوثٌ [تَفْتِيشٌ] search, quest;          مُحَرِّكٌ بَحْثٍ search engine; البَحْثُ عَنْ شَيْءٍ the search for          sth [دِرَاسَةٌ] research, study, investigation; inquiry; بَحْثٌ          شَامِلٌ an in-depth investigation; a thorough examination;          عِلْمِيٌّ scientific research; مَيْدَانِيٌّ fieldwork;          تَجْرِبِيٌّ empirical research; مُكثَّفٌ intensive          research; نَظَرِيٌّ   تَطْبِيقِيٌّ theoretical/applied research.</p>	<p><b>balance</b> /bal(ə)ns/ n [تَوَازُنٌ] (equilibrium) <b>تَوَازُنٌ</b>; <b>to keep/lose</b>          one's ~ فَقَدَ تَوَازُنَهُ عَلَى \ حَافِظٌ عَلَى \ <b>aim for the right ~ between</b>  <b>quality and price</b> أَهْدَفَ إِلَى التَّوَازُنِ السَّلِيمِ بَيْنَ الْجُودَةِ وَالسَّعْرِ          أَفْقَدَ شَخْصًا تَوَازُنَهُ ~ (physically) <b>to throw sb off</b>; وَالسَّعْرِ          (to disconcert) أَرَبَكَ شَخْصًا [مِيزَانٌ] (scales) <b>on</b>          ~ <b>we had a good year</b> كَانَتِ السَّنَةُ فِي نَهَايَةِ الْأَمْرِ سَنَةً جَيِّدَةً          [رَاصِدَةٌ] (amount due) <b>المَبْلَغُ</b></p>
--	---

Figure 7: Both sides of the dictionary have text running on.

The direction of the dictionary is left-to-right, so fitting right-to-left phrases in there was a major challenge, not in the least because some information that is to the right (e.g. the start of an Arabic example sentence) needs to go on the top line in the case of a line break, whereas other information on the right (e.g. the grammatical information with an Arabic translation) needs to go on the bottom line in case of a line break. It took many tries and elaborate diagrams until we had all the kinks sorted out.

For some users it may seem a bit odd at first to have to “jump” when reading a translation, but we have found that one gets used to it very quickly, and indeed this is the way the entries are presented in most dictionaries. The alternative, writing the English on the left and the Arabic on the right, leaves too much white space and isn't feasible for a print dictionary of this size.

## 10 Finding an Arabic Word in the Print Dictionary

An important factor of Arabic is that it is root-based, that is, every word has a root of (usually three) consonants carrying the basic meaning of a word (e.g. *ktb* with basic meaning “writing”, which is modified by adding vowels and affixes, making *kātib* (writer), *kitāb* (book), *kataba* (he wrote), *kitāba* (writing), *maktaba* (library; bookshop), etc. Arabic dictionaries, with the exception of learners' dictionaries, are usually ordered by these roots, rather than in “proper” alphabetical order. We chose to do this as well, since the advantage of having all words of one root in one outweighs the difficulty a beginner may have in finding the root of a word. This means the Arabic-English side has a kind of double ordering, first the roots in alphabetical order, then a logical order for the words within one root.

Loanwords do not have an Arabic root. For them we listed each written letter as a root letter, and fitted them in into the root system like that.

## 11 Finding an Arabic Word in the Electronic Dictionary

Though all words are fully vocalized, we cannot expect the user of the electronic dictionary to enter a word with all its vowels, especially not if it's a word they don't know – not knowing a word means you could at best make an educated guess at its vowels, and there is no standard system of vocalization. So one of the challenges for the electronic dictionary was to make it so that the user is able to enter the unvowelled or partially vowelled form of a word, and be redirected to the entry or entries that correspond to that form.

But that is not the biggest challenge. I mentioned before the many possible surface forms of one lemma. Also, words are often joined together: the article *al-* is prefixed to the noun or adjective, object and possessive pronouns are suffixed, and some grammatical words like *wa-* (and) and *li-* (for) are joined to the word following them as well. All this can be combined, so for example *wali'uxtihi* (and for his sister) is one string. This can make strings of letters highly ambiguous – a common string like *l'nh* can be interpreted in at least seven different ways! We want the user to be able to enter any string they don't understand, without having to distinguish the different morphemes themselves, which can be a challenge for even quite advanced Arabic learners. For this, we again use the Buckwalter Arabic Morphological Analyzer integrated with our own headword list. Thus the strings are analysed into the appropriate morpheme(s) and the user will be redirected to the relevant entry. For example if a user enters the string *وكتبه* (and his books), they will be redirected to *كتاب* (book), with the information that it's preceded by *وَ* (and) and followed by *هُ* (his). In case of multiple possible analyses, like *كتب*, which can mean “he wrote”, “writing”, or “books”, it gives the possible entries with a summary of the part of speech and meaning, and allows the user to choose which entry to display fully.

## 12 Conclusion

The world of Arabic lexicography is a very challenging field, where little can be relied on, and expected and unexpected pitfalls abound. Despite, or maybe even because of this, it is a fascinating area, where truly significant achievements can be made. With English and Arabic being world languages, and being two of the six official languages of the UN, it is amazing that so few resources exist for translation between the two, with no reasonably modern ones.

We feel that this dictionary fills that hole, and with the opportunity to constantly update that online dictionaries offer, will continue to fill the gap for many years to come.

We hope this look behind the scenes has been interesting for lexicographers and other linguists.



## 13 References

- Arts, T. & McNeil, K. (2013). Corpus-based lexicography in a language with a long lexicographical tradition: The case of Arabic. In *Proceedings of WACL'2, Second Workshop on Arabic Corpus Linguistics, 22 July 2013*. Lancaster University, UK.
- Arts, T. et al. (Forthcoming 2014). *Oxford Arabic Dictionary*. Oxford: Oxford University Press.
- Attia, M., Pecina, P., Toral, A. & Van Genabith, J. (2014). A corpus-based finite-state morphological toolkit for contemporary Arabic. In *Journal of Logic and Computation*, 24 (2), pp. 455-472.
- Baalbaki, R.M. (2004). Coinage in Modern English-Arabic Lexicography. In *Zeitschrift für Arabische Linguistik*, 43, pp. 67-71.
- Baalbaki, M. & Baalbaki, R.M. (2013) *Al-Mawrid Al-Hadeeth*. Beirut: Dar El-Ilm Lilmalayin.
- Benzehra, R. (2012). Modern English-Arabic Lexicography: Issues and Challenges. In *Dictionaries: Journal of the Dictionary Society of North America*, 33, pp. 83-102.
- Doniach, N.S. (ed.) (1972). *Oxford English-Arabic Dictionary*. Oxford: Oxford University Press.
- Habash, N., Rambow, O. & Roth, R. (2009). MADA + TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2<sup>nd</sup> International Conference on Arabic Language Resources and Tools (MEDAR)*. Cairo, Egypt.
- Haywood, J.A. (1965). *Arabic Lexicography*. Leiden: Brill.
- Hoogland, J. (2003). Coverage. In *The Nijmegen Arabic/Dutch Dictionary Project*. Accessed at: [http://wba.ruhosting.nl/Content1/1.4\\_Coverage.htm](http://wba.ruhosting.nl/Content1/1.4_Coverage.htm) [10/04/2014].
- Hoogland, J. (2004). Working Methods. In *The Nijmegen Arabic/Dutch Dictionary Project*. Accessed at: [http://wba.ruhosting.nl/Content1/1.4\\_Working\\_Methods.htm](http://wba.ruhosting.nl/Content1/1.4_Working_Methods.htm) [10/04/2014].
- Hoogland, J. et al. (2003). *Woordenboek Arabisch-Nederlands*. Amsterdam: Bulaaq.
- Ibrahim, R. & Aharon-Peretz J. (2005). Is Literary Arabic a Second Language for Native Arab Speakers?: Evidence from Semantic Priming Study. In *Journal of Psycholinguistic Research*, 34 (1), pp. 51-70.
- Kilgariff, A., Rychly, P., Smrz, P. & Tugwell, D.A. (2004). The Sketch Engine. In *EURALEX Lorient Proceedings*. Lorient, France.
- Van Mol, M. & Bergman, K. (2001). *Leerwoordenboek Arabisch*. Amsterdam: Bulaaq.
- Wehr, H. (1979) *A Dictionary of Modern Written Arabic*. Rev. ed. Urbana: Spoken Language Services.





# Simple and Effective User Interface for the Dictionary Writing System

Kamil Barbierik, Zuzana Děngeová, Martina Holcová Habrová, Vladimír Jarý,  
Tomáš Liška, Michaela Lišková, Miroslav Virius  
Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i.  
kamil.barbierik@fffi.cvut.cz, dengeova@ujc.cas.cz, holcova@ujc.cas.cz,  
vladimir.jary@foxcom.cz, tomas.liska@foxcom.cz,  
liskova@ujc.cas.cz, miroslav.virius@fffi.cvut.cz

## Abstract

A new monolingual dictionary of the contemporary Czech language is being prepared by the Institute of Czech language at the Academy of Sciences of the Czech Republic, v.v.i. A dictionary writing software is being developed as part of a grant supported by the Ministry of Culture of the Czech Republic within the National and Cultural Identity (NAKI) applied research program. We will present the overall architecture of the software and then focus on its user interface and two modules: the referencing system and a new module – the editorial tool (that was promised in (Barbierik 2013)).

**Keywords:** dictionary writing system; DWS; cross-reference module; editorial tool; lexicography

## 1 Introduction

Since 2012, the Department of Contemporary Lexicology and Lexicography within the Institute of Czech Language at the Academy of Sciences of the Czech Republic, v. v. i., has been preparing a new monolingual dictionary of contemporary Czech.

Its working title is *Akademický slovník současné češtiny* (The Academic Dictionary of Contemporary Czech). It is a medium-sized dictionary with an expected number of 120,000–150,000 lexical units.

To aid this project, a new Dictionary Writing System (DWS) is developed. More information about the project can be found in (Kochová 2014). A detailed specification of the requirements from the lexicographer's point of view can be found in the article *A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System* (Barbierik 2013).

We introduce basic functionality of our DWS with emphasis on the user interface in this paper. Further, we focus mainly on the editorial tool which will be described in more detail.

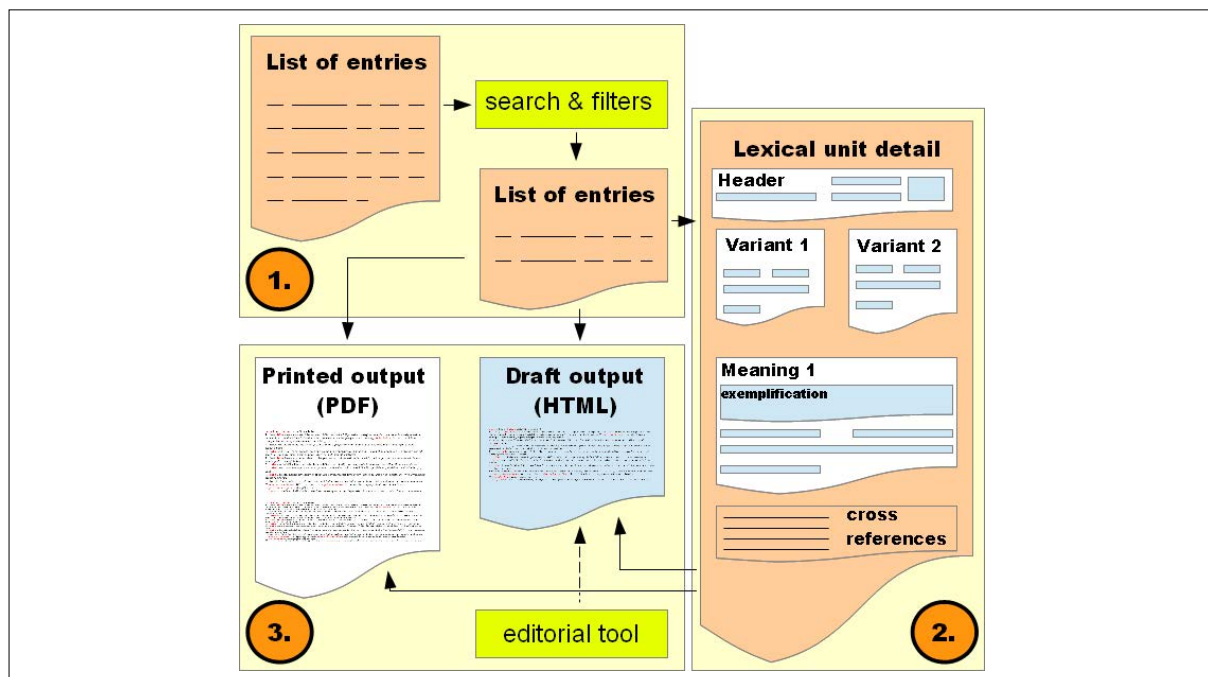


Figure 1: Basic scheme of our Dictionary Writing System.

## 2 Existing Dictionary Writing Systems

There are several commercial Dictionary Writing Systems (DWS) available (e.g. TshwaneLex (2013), IDM DPS (2013),

iLEX (2013)) as well as open-source systems (e.g. the Matapuna Dictionary Writing System (2013)). The DEB II (2013),

(DEBDict 2013), dictionary editor and browser, is available for the Czech language.

The lexicographic team at the Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i., considered three options: to buy an existing commercial DWS, to use one of the open-source systems or to develop their own system. One significant criterion used for the DWS selection was the amount of necessary adjustments due to specifics of the compilation of the dictionary and the time allocated to this task. Another criterion was the DWS price. After evaluation of the DWSs available, we decided to develop our own DWS that will fully respect the significant specifics of the compilation of the dictionary (Abel 2012: 1-23; Atkins 2008). The greatest advantage of this decision, which the lexicographic team benefits from, is the fact that any request for the user interface, some process modification or a new handy feature implementation etc., can be processed and implemented almost immediately.

### 3 Basic Functionality of Our DWS

From the common user point of view, the software is divided into three parts as it is shown on Fig. 1: the list of entries, the lexical unit detail and the output. The numbers of these main parts are used in the titles of the following subsections.

#### 3.1 List of Entries (part 1)

After successfully logging into the system, the user inputs the list of entries which mainly represents the macrostructure of the dictionary. It is a list of lexical units with some basic information that is important for linguists at this stage.

Some helpful functions are available for better orientation in the long list of entries. The most important one is probably the quick search engine together with a set of predefined filters. The quick search allows users to search for entries by selecting any field of the microstructure of lexical unit (through which the user intends to search) and entering a search query. The search query can contain wildcards granting users better control over the search results. In addition, the query field is automatically updating its mode according to the type of information the user is searching for to make the search process easier for the user. For instance, if the user is trying to find information in fields where only a few values are available (e.g. type of lemma), the query field updates itself to select box. Thus, the user does not have to guess what values are available within the selected field. To avoid typing errors, an auto-complete function is implemented when searching in fields that contain short texts (like lemma).



**Figure 2: The “Quick search” function with auto-complete (left) when searching for lemma and select box when searching lemmas of certain type (right).**

For example, the user is able to filter out, with combination of the quick search with predefined filters, the following entries:

- entries which I (the logged in user) founded and begin with the defined letter,
- entries of a specified word class created in some time interval,
- entries from manual selection containing some phrase in any of its exemplification, etc.

## 3.2 Detailed View of The Lexical Unit – Editing Module (part 2)

When the lexicographer finds the entry, he or she may continue to the detailed view of it. The detailed view of the entry contains all the information mainly from microstructure of the lexical unit that is available in a well-arranged and sophisticatedly structured way. Additionally, the user can edit any information required in this view. To make the editing process more effective, different input fields are used according to the type of information it needs to gather. The whole editing form is designed so that the linguists editing the lexical unit information do not need to learn any special markup language or have any advanced computer literacy skills.

This editing form is organized into 4 sections (see Fig. 1 – Part: Lexical unit detail):

- (1) Header
- (2) Section of variants
- (3) Meanings
- (4) Cross-references

**Header section.** General information about the lexical unit can be found in the header section. It contains the entry status indicator which shows the progress of the work on this entry. Also, the output status can be set in this section which indicates in which output (electronic or paper) the entry will be presented. Furthermore, it contains information about the time of creation and about the last editing of the entry. The header section also contains information on the responsible user as well as a field very where the lexicographers may leave a note concerning the entry.

**Section of variants.** The section of variants may contain one or more variants of the lemma with all required microstructure elements. The variants of variant lemmas are often equivalent in majority of values, thus the function “Add variant as a copy of the last one” was implemented. This function creates a new variant and copies all the values from the last existing variant to it. Thus, only a few values have to be edited in the new variant. Consequently, creating new similar variants is much more efficient.

**Section of meanings.** The section of meanings consists of one or more panels, where the meaning of the word is described together with other related information. The section is organized as a set of panels. Each panel contains a large form, where the information about the meaning can be edited. The user can change the order of the panels; this will affect the order in the dictionary printed or electronic output. Meanings are numbered and when reordered, the numbering of meanings is automatically updated. The panel containing the form, with information related to the meaning, can be minimized or maximized according to which panel the user intends to work with. It helps for better orientation, whereas some lemmas may have quite a lot of meanings. The quick navigation is also helpful when a word has a lot of meanings.

This allows navigating directly to the meaning with the certain number, without scrolling the page.

**Implementation of the editing form.** As we mentioned at the beginning of this section, the whole detailed view is basically a well arranged set of fields of different types. These field types were chosen according to the types of information contained in lexical unit microstructure. For the gathering of short textual information (mostly comments, but also pronunciation for instance, synonyms, etc.) we have used simple one row text input field. Multiline longer texts are collected using textboxes. Often it is necessary to format the input text in some way. Special textboxes with rich text functions were implemented for this purpose. Such fields are used to store the exemplifications or the meaning explanations. Probably the most complex input fields are administrated select boxes which provide lexicographers with finite number of options prepared by the administrative user. This prevents editors from committing typing errors and unifies the values in certain places in microstructure through the whole dictionary. But it does not limit them thanks to the option of adding their own entry if it is necessary. The statistics of these entries are collected, and if some value is used too frequently, the administrative user may integrate it to the select box and “standardize it” very easily. Due to limited range we cannot provide an adequately descriptive picture of the editing form. For more information about the editing form, please refer to our poster “Simple and effective user interface of our new DWS”.

### 3.3 The Output Module (part 3)

It is possible to evoke the output view of one or more lexical units from the detail of the lexical unit as well as from the list of entries. The output module takes the information collected using the editing form described above and utilizes some complex and very strict formatting rules on them to form the output. Thus, the user has a great possibility to preview the entry (or more entries at once) in its printed form and to see how it will exactly look like in the printed dictionary.

Two outputs are available in our DWS system: printed and electronic output. The printed output is not editable and it is presented in PDF format ready to be printed on the paper. The electronic or draft output is presented in HTML form. Even this output is not editable, but it is possible to implement additional interactive functions for it. Thanks to HTML format the user can interact with it using a web browser. One of the interesting functions we designed and implemented in this output is an editorial tool. It allows the lexicographers to fine tune the output or to correct mistakes or inconsistencies in cooperation with other lexicographers or editors.

## 4 Recently Implemented Features

### 4.1 Cross-References Module

A very necessary feature of the system, especially from the linguistic point of view, is the ability to define relations between entries. Relations are defined between two dictionary entries; one of the entries is considered to be the main or “master” entry, the second is the “slave”. The system always allows creating the connection from both sides. This means that the user may define the relation if he is editing the slave as well as the master entry.

There are different types of relations from the linguistic point of view: run-on entries, references between one-word and multi-word lexical units and linked entries. From the user point of view, each type of relation needs a slightly different approach, but from the system perspective it is always just a relation between two entries supplemented with some information that is important from the linguistic point of view. When the user is at the detailed view of the lexical unit, he can always define all available types of cross-references to another entry. The window of the referencing module is evoked by clicking on the buttons at certain sections of the editing form. These buttons are placed according to the element of the microstructure from which the user references the other entry. For instance, run-on entry may be referenced from the whole entry (the button is at the end of the form) or from any of its meanings or exemplifications (buttons are under the corresponding input fields). The referenced units are then displayed at correct places according to this information when the output is compiled.

Other type of referenced entry is the linked entry. It can be referenced from the whole lexical unit or from the particular meaning of the entry to other foreign entry or its meaning. Thus, the button for bringing up the referencing tool popup is always at the end of the editing form.

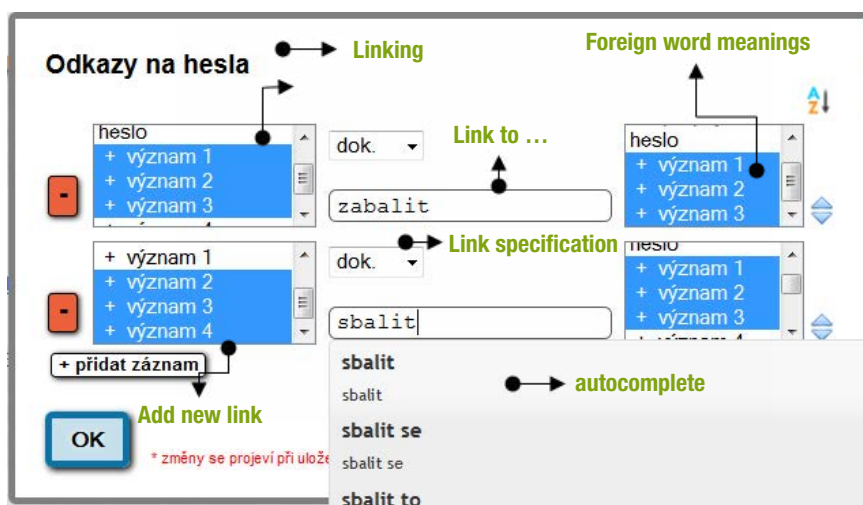


Figure 3: The cross-reference dialog box for linking words.

The interface for referencing is very simple and contains some clever functions to help the user to reference and to manage entries effectively.

Fig. 3 above is a snapshot of popup of referencing module evoked from lemma “balit” and it allows creating references (of type linked entry) to foreign entries and their meanings. As can be seen, references to multiple entries may be defined at once. The reference is created by putting the desired referenced lemma in the middle text input field. These inputs are provided by auto-complete functionality to make the process easier for the user. After writing in the foreign lemma, all its numbered meanings are loaded to the right select-box. Thus, the user knows how many meanings the foreign entry contains and he can comfortably choose the desired ones. Additional information to each reference can be added using a select box. When more references are defined (two in our case), it is reasonable to have an ability to sort the entries. It is possible to do it manually using little arrows next to each referenced entry, or to sort it alphabetically by the program using the AZ button in the top right corner. Links, if any, are according to the formatting rules for creating the monolingual dictionary attached to the end of each entry and our example will produce the output shown in Fig. 4 at the end of the word “balit” definition in the printed output.

► dok. k 1 - 3 → **zabalit**1 - 3, ► dok. k 2 - 4 → **sbalit**1 - 3

Figure 4: The printed output of linked words to the word “balit”.

## Editorial Tool

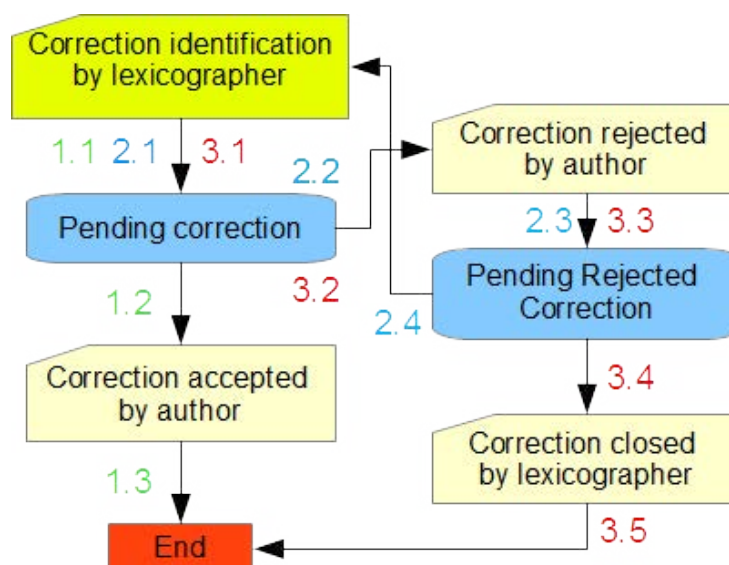


Figure 5: The editorial process.

In order to produce and maintain a high quality dictionary, our DWS system implements an editorial tool, which is, as mentioned above, connected to the electronic HTML output. This editorial tool has been projected to replace the standard editorial process, where some portion of the submitted entries is printed on the paper, sent to the other lexicographers or editors, and received back reviewed. By implementing this feature directly in the DWS system, we save a vast amount of paper, as well as the time and money for the transaction agenda. The whole dictionary, each and every entry of it, is always ready to be reviewed without printing or posting anything.

The editorial process on one entry is captured schematically on Fig. 5. When the author (lexicographer) of the entry submits it to the system, the reviewer is able to see it and to review it using the draft (electronic) output. This draft output is very similar to the final printed output, so he or she is revising the entry almost as it was printed on the paper.



Figure 6: The correction founding and sending it to editors.

By clicking on the information in the draft output, that the lexicographer wants to correct, he or she gets a popup window – see Fig. 6, where he or she can input his suggestion for correction, make a note about the correction for the author and send it to the author with a single click.

This is how the correction identification happens. There is a pending correction from this moment. This is indicated to the author of the entry (and not only him or her, but to other signed in users too) by highlighting the field that contains the corrected information – see Fig. 7.

By the click on the yellow “correction icon” nearby the highlighted field the popup will appear with the suggested correction from the lexicographer.



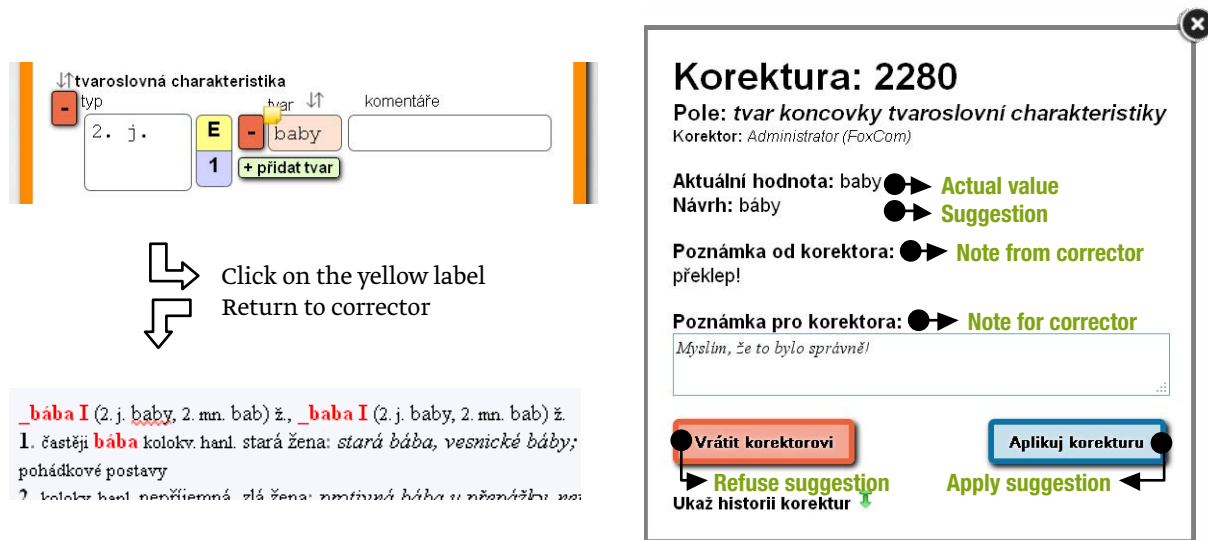


Figure 7: The correction from the author's point of view.

The author may now decide whether he accepts the correction (take the green path number 1 in scheme on Fig. 5), or reject it (the blue path number 2 or the red path number 3 on Fig. 5). In the case he accepts the pending suggested correction, the system automatically updates the entry and the process successfully ends.

If author rejects the pending correction suggested by the (other) lexicographer – the corrector, it will be indicated in the draft (electronic) output – see Fig. 8. By clicking on it, the (other) lexicographer may view it together with the author's note on why it was rejected. At this stage the lexicographer who suggested the correction has two options: to close the correction (following the red path number 3 on Fig. 5) or to suggest a new one (following the blue path number 2 on Fig. 5). If the correction is closed, no changes in the entry are made and the process ends. If a new correction is designed, the process is started again.



Figure 8: The correction refused by editor.

There is one more option for the lexicographer when he or she defines the correction that is not indicated in Fig. 5. He or she may immediately accept it – see Fig. 6 (the “Apply suggestion” button) - instead of creating a pending correction. Doing so, he or she directly updates the entry. This feature speeds up the process, when obvious typing errors are detected, because the pending correction does not have to wait for the author’s approval.

Every correction made using this editorial tool is recorded together with the old and new value of the time stamp, every action is signed by lexicographer, who changed the value or status, and their comments are also recorded. Thus, all the information that was ever corrected has an editorial history. It never gets lost and is always available in the editorial module pop-up. Thus, when the author is deciding whether to accept or reject some suggestion from another lexicographer, he can check the history of corrections made, find out who and when the suggestions were made. With the inclusion of a notes feature, he may even know why and under what circumstances they were made.

datum vzniku (korektor)	date and time corrector	heslo (pole) the lemma the field	zpracovatel poznámka zpracovatel editor note to editor
2014-04-14 10:28:28 (*ja*)	(druhá) pražská defenestrace (vznam-vyklad_vznamu)	*ja*	""
	vyhození dvou královských místodržících z oken Pražského hradu roku 1818		
	vyhození dvou královských místodržících z oken Pražského hradu roku 1818		
2014-04-14 10:01:12 (*ja*)	barzoj (vznam-vyklad_vznamu)	Pernicová	
	plemeno dlouhosrstých chrtů pocházejících z Ruska		"Nezačínat výklad slovem odrůda, druh, typ, forma, plemeno... - <a href="https://docs.google.com/document/d/1MZxKskjoCyIS2B1A_NNHSUIPsfxyGAgjXAFixKoP/edit">https://docs.google.com/document/d/1MZxKskjoCyIS2B1A_NNHSUIPsfxyGAgjXAFixKoP/edit</a> "
	plemeno dlouhosrstých chrtů pocházejících z Ruska		
2014-04-14 09:59:59 (*ja*)	step aerobik (vznam-vyklad_vznamu)	Pernicová	

Figure 9: The history of corrections.

## 5 Conclusion

Our DWS has been released and the lexicographic team uses it in their everyday work. Nevertheless, we are preparing additional modules for our DWS. This article was devoted to the editorial tool that has been deployed recently and we present it here for the first time.

We have strictly emphasized the quality of the user interface of our DWS. It must be designed according to the needs of the lexicographers that use it for the processing of large amount of lemmas.

Lexicographers are now processing lemmas using the described tools and preparing them to be published. Meanwhile we are preparing, except printed output, several applications, where published lemmas will be available for public. Using these applications like web pages, mobile applications for iOS or Android operating systems, users will be able to search and browse the dictionary on different devices.

Currently, we are preparing a very strong relation based search tool called xFilter which we will present sometime in the future.

## 6 References

- Barbierik, K. et al. A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System. In Proceedings of the 7th international conference Slovo, 13-15 November 2013. Slovenská akadémia vied, Jazykovedný ústav Ľudovíta Štúra, pp. 9-26.  
*DEB II*. Accessed at: <http://deb.fi.muni.cz/index-cs.php> [11/10/2013]  
*DEBDict*. Accessed at: <http://deb.fi.muni.cz/debdict/index-cs.php> [11/10/2013]  
*IDM DPS*. Accessed at: <http://www.idm.fr/products/dictionary-writing-system-dps/27/> [11/10/2013]  
*iLEX*. Accessed at: <http://www.emp.dk/ilexweb/index.jsp> [11/10/2013]  
Kochová, P., Opavská, Z., Holcová Habrová, M. (2014). At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project). *Poster presented on this conference. Matapuna*. Accessed at: <http://sourceforge.net/projects/matapuna/> [11/10/2013]  
*TshwaneLex*. Accessed at: <http://tshwanedje.com/tshwanelex/> [11/10/2013]  
Abel, A. and A. Klosa 2012. 'The lexicographic working environment in theory and practice.' In R. V. Fjeld and J. M. Torjusen (eds.), *Proceedings of the 15th EURALEX International Congress*. Oslo: University of Oslo, 1-23.  
Atkins, B. T. Sue and M. Rundell 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

### **Acknowledgement.**

This work has been supported by the grant project of the National and Cultural Identity (NAKI) applied research and development program A New Path to a Modern Monolingual Dictionary of Contemporary Czech (DF13P01OVV011).



# Totalitarian Dictionary of Czech

František Čermák

Institute of the Czech National Corpus, Charles University

frantisek.cermak@ff.cuni.cz

## Abstract

Due to a decades-long experience with the Communist totalitarian regime, an attempt was made to capture main features of its language, and thus the substance of the time, by a corpus-based dictionary. The corpus, based on large samples of Communist newspapers and booklets, which is now accessible on the web, served as basis for a frequency-based dictionary of the time; it has been provided with statistics according to time periods, with pertinent collocations, appearance in various genres and, most of all, with comparative figures with today's standard as it is to be found in a large contemporary corpus of Czech. This, by far the first systematic description of the period in any language, is complemented by a detailed study of the lexis of the time, many typical excerpts, etc.

**Keywords:** corpus; frequency dictionary; dictionary of a period; Communist totalitarian vocabulary

## 1 Goal and Historical Background

Twenty years after the decline of the Communist regime in Czechoslovakia, the first attempt (Čermák et al. 2010) has been made, on the basis of corpus-data, to map the language of the period of over four decades of the Communist rule (1948-1989) in the Czech-speaking part of the country, known now as Czechia. Whether too soon or too late, the decision was precipitated by realization how fast people tend to forget the totalitarian time, young generation not knowing it at all. However, it seems that it is through the vocabulary of a period that one gets to know it best. Needless to say that the objective of solving the situation is best served by corpus-based data offering an eminent source of information about ways of manipulation of the whole society through its rather effective propaganda. There seem to be very few similar dictionaries delving deep into the language of a period objectively and systematically, based on corpus data. Thus the book is both a lexicographic and linguistic record of that time as well as a valuable historical and sociological description of what had long been a taboo, namely propaganda, which, in many respects, drew amply on Hitler and Goebbels from an earlier period of fascism.

## 2 Data and Their Preparation

Since it was not possible to cover linguistically all texts from the period, the *Totalitarian corpus* (available on the web at korpus.cz) had to be limited to over 15 million positions made up of 422 762 word forms which equals to enormous 164 815 lemmas. The data have been drawn painstakingly from the major and all-embracing Communist daily *Rudé právo* (Red Right) from three historically critical but different periods of the second half of the year 1952 (6. 6. - 31. 12. 1952), i.e. from the peak of the most primitive communist propaganda, the second quarter of 1969, i.e. from after the Soviet and Russian occupation of the country in 1968 (1. 4. - 31. 4. 1969) and from the first quarter of 1977 (3. 1. - 31. 3. 1977), i.e. time of deep social recession and depression. The newspaper data amounting to some 10 million words have been complemented by a careful and typical selection of 91 propaganda books and booklets (about 5 million words) from the same period. All of this data have been manually scanned and corrected having to come to grips with several different spelling reforms and various standardisation processes. No attempt has been made to cover a later period when it was felt that the political regime was characterized by stagnation and inertia as well as gradual crumbling of its strongholds which culminated in 1989, i.e. its definite downfall. The data have been tagged and lemmatized which made it necessary to adapt the existing taggers and lemmatizers. Of course, the fact that the Totalitarian corpus is freely accessible on the Web now means that any further analysis is possible, although the primary goal has been to cover the period's typical vocabulary. It may be reasonably hoped that the core of the vocabulary is captured in dictionary form being accompanied by an extensive analysis, complements and comparison with today's language.

## 3 The Dictionary of the Communist Totality

The *Totalitarian Corpus*, based on a prior decision to choose data from three historical periods of eminent linguistic interest (and the available propagandist books and booklets), had to be further limited before dictionaries based on it have been compiled, both capturing only the top selection of the most important lemmas.

One of the main goals of the project was to point out to users what used to be typical in the vocabulary of the period (**A**). The tricky business of corpus comparison of the period with today's standard language, which is several decades remote from the time recorded in a large contemporary corpus of Czech (SYN 2005, see korpus.cz), has been projected into frequencies (**B**) in one of two dictionaries compiled on this basis, namely the **Specific Totalitarian Dictionary** (*Dictionary* in the following). Because this dictionary, based on such comparison, has been achieved due to use of statistical methods it is highly selective in its nature capturing only those lexemes whose frequency has had a high degree of deviation from today's use. The dictionary is alphabetical giving a summary of absolute (ABS) and relative (REL) frequency (in square brackets) and partial frequencies of each of the three

periods examined (1952, 1969, 1977, given in ppm, i.e. parts per million). Numerical differences between these three figures suggest a growing or falling use of the frequency of the words in the three periods. Obviously, many lexemes that have not seemed to be statistically different enough have been dropped and are not to be found here.

Where necessary, an additional brief explanatory note has been added identifying lexeme's meaning. Typical collocations, that offer a more direct information about the use of the word are offered following a full dot • Should the lemma be out of use in today's newspapers completely (i.e. not found in the contemporary 100-million corpus) it is preceded by a cross sign †, such as †**agrobiologický** (agrobiological). Positive, period-linked sign of a larger asterisk ★ standing before a collocation suggests a prominent use and specifically high frequency in the Communist era, while the sign † indicates, in contrast, a typical collocation of today, such as in **bezmezně** (boundlessly, infinitely, immensely) in „★ **b. oddáni** (SSSR) × † **b. věřit**“ (*immensely devoted* to USSR at the time, as against to *infinitely believe* in today's usage). The overwhelming devotedness to the Soviet Union, often found elsewhere, is captured in this example prominently, showing also how the usage of the verb *believe* (věřit) has eventually returned to normalcy now.

Collocations have been obtained by getting a combination of predetermined values of three association measures of lexemes having a frequency higher than 5, namely of MI-score, log-likelihood and Dice. It is specifically collocations that reveal the usage of heavily loaded words of the time such as fight/struggle:

**boj** (struggle, fight) • *boj dělníků* (workers' struggle); *boj lidu (za mír)* (people's fight for peace); *boj národů (za mír)* (fight of nations for peace); *boj proletariátu* (proletariat's struggle); *ideologický boj* (ideological struggle); *ideový boj* (struggle of ideas); *organisovat boj (pracujících)* (organise fight of the workers); *podporovat boj (národů)* (support fight of nations); *revoluční boj* (revolutionary fight); *rozvíjet boj (proti imperialismu)* (develop struggle against imperialism); *rozvinout boj (za mír)* (develop fight for peace); *stávkový boj* (fight of those on strike); *třídní boj* (class struggle); *řídít boj dělnické třídy* (direct the fight of the workers's class), etc.

On a closer inspection, it is evident that the society lived then in a state of frenzy feeling both endangered from outside and feeling that it must fight for almost everything, including even the most common, everyday things.

## 4 A Sample of the Specific Dictionary

Let us have a look at the Dictionary and some of its lemmas:

**Adenauer**                      ABS 348, REL 27 ppm  
[83 ppm | 1 ppm | 0 ppm] *Konrad Adenauer, kancléř NSR (1949-1963)*

**Adolf** ABS 92, REL 7 ppm  
[6 ppm | 12 ppm | 5 ppm] • Adolf Hitler

**agent** ABS 1292, REL 100 ppm  
[263 ppm | 22 ppm | 24 ppm] • agent buržoazie; agenti imperialismu; imperialistický agent;  
†titovský agent

**agitace** ABS 894, REL 69 ppm  
[104 ppm | 12 ppm | 85 ppm] • †agitace komunistů; názorná agitace; osobní agitace;  
politická agitace; účinnost ekonomické propagandy a agitace

**agitátor** ABS 811, REL 63 ppm  
[166 ppm | 2 ppm | 23 ppm] • †kolektiv agitátorů; †příprava agitátorů; †seminář agitátorů;  
†schůze agitátorů; stranický agitátor

**agrese** ABS 1213, REL 94 ppm  
[188 ppm | 50 ppm | 50 ppm] • hitlerovská agrese; imperialistická agrese;  
(definice) pojmu agrese; politika agrese

**americký** ABS 15089, REL 1170 ppm  
[2527 ppm | 601 ppm | 476 ppm] • americký agresor; američtí barbaři; americký businessman;  
americký imperialismus; americký katan; americký monopolista; americké okupační úřady;  
americký okupant; †američtí podněcovatelé války; †americká soldateska; američtí supermani;  
provokace amerických imperialistů

**angažovaný** ABS 275, REL 21 ppm  
[0 ppm | 38 ppm | 26 ppm] • angažované země; †schůzka neangažovaných zemí

**armáda** ABS 6993, REL 542 ppm  
[1067 ppm | 340 ppm | 260 ppm] • lidová armáda; (spartakiáda) spřátelených armád;  
západoevropská armáda; Svaz pro spolupráci s armádou (*viz Svazarm*)

**balistický** ABS 16, REL 1 ppm  
[0 ppm | 1 ppm | 2 ppm] ★ (mezikontinentální) b. raketa × ☞ b. expertíza

**bdělý** ABS 86, REL 7 ppm  
[18 ppm | 1 ppm | 2 ppm] ★ politický b. × ☞ b. stav



bezmezně

ABS 23, REL 2 ppm

[4 ppm | 0 ppm | 1 ppm] ★ b. oddáni (SSSR) × 🙅 b. věřit

## 5 Basic Dictionary of the Communist Totality

To balance this key, though highly (i.e. statistically) prominent, dictionary, a second dictionary, that of Basic Dictionary (Základní slovník) has been compiled that is complementary to the Specific Totalitarian Dictionary. **The Basic Dictionary** is represented by almost ten thousand most frequent lemmas from the Totalitarian Corpus (9994, i.e. those having more than 63 occurrences). Thus a comparison of the specific period-related vocabulary obtained statistically and that of the systematic full dictionary is enabled, the latter enabling to find there also those words that were used but were not statistically interesting enough. This dictionary, based on descending absolute frequency ordering, shows all lemmas provided within the given frequency limits. The sum-total of frequency for each lemma is broken into three different frequency figures for each of the three periods covered. For cross-reference purposes, those lemmas given also in the Specific Totalitarian Dictionary are marked by asterisk.

Lemma	Overall	1952	1969	1977
<b>a</b>	498450	166271	130175	202004
<b>v</b>	422528	129014	126641	166873
<b>být*</b>	339917	110927	101912	127078
<b>se</b>	225925	67849	72378	85698
<b>na</b>	214183	70624	66385	77174
<b>ten</b>	120081	36929	36986	46166
<b>s</b>	106963	31667	33946	41350
<b>který</b>	104580	36545	30776	37259
<b>že</b>	99118	34209	31456	33453
<b>z</b>	95681	27971	31637	36073
<b>strana*</b>	52942	22081	12821	18040
<b>všechen</b>	51970	22246	11180	18544
<b>rok*</b>	49349	13494	15128	20727
<b>práce*</b>	44491	17044	8300	19147
<b>jeho</b>	39773	11046	11304	17423
<b>aby</b>	37883	17291	8835	11757
<b>sovětský*</b>	36644	22690	4495	9459
<b>socialistický*</b>	35990	7450	5373	23167
<b>nový*</b>	34985	13992	8061	12932

Looking at the first 30 most frequent lemmas of the Basic Dictionary it is just appalling to see what kind of priority some of its words used to have in the Communist-ruled society at the time (although some words have been left out). Thus, the very first noun is *strana* (party, frequency 22) followed by *práce* (work) while the very first adjective used then was *socialistický* (socialist, fr. 29), followed by *nový* (new). The prominent presence of nouns does not need much comment telling what had been dominant at the time: Communist party and work. Likewise, the adjective socialist is to be expected here while the particular adjective new would deserve a special study, not possible here, since every second thing, notion, aspect of life coming from the past had to, on ideological grounds, be rejected and replaced by something new, i.e. *Communist*, most often *Soviet-inspired*, since Soviet Union had been presented as a new Promised Land serving as a model and exemplary goal to be followed by everyone and serving as an official inspiration for everyone. Some extensive collocations of this type showing this in detail are added in the Supplement to the Dictionary.

## 6 Additional Features of the Dictionary

Next to the two dictionaries, there is a large analytical study attached (*Slovník komunistické totality: lexémy, nominace a jejich užití, Dictionary of the Communist Totality: lexemes, nominations and their Use*) preceding both dictionaries, based on the data added, scrutinizing and researching typical words and collocations more deeply paying special attention to principles of the Communist propaganda. Some attention is paid to comparison of joint features of the fascist propaganda and the Communist one and a semiotic analysis of how certain lexemes were used for manipulation is shown.

Since all data have been published only after they underwent a severe censorship, an attempt is also made to counterbalance this *official* (official) language by a sample of the most important words of the *unofficial* language used by normal people and heard on the street that were not subject to the official parlance of the time and never printed. It is shown that beginnings of this second, unofficial language or rather vocabulary was born and used profusely wherever people were in private and not followed by the secret police; thus a parallel vocabulary of synonyms was born (e.g. *strana-partaj*, *komunista-komouš*, etc). The study is complemented by a number of typical samples of texts, vocabulary and collocations, showing, for example, the blind reliance on anything Soviet whose meaning has changed from a proper name so much as to just practically designate only what was „good, best, exemplary“.

## 7 Notes and Conclusions

**Slovník komunistické totality** (Dictionary of the Communist Totality) maps the gap between a general dictionary and one of an epoch or period enabling, in this case, comparison of two comparati-

vely close time periods, an attempt rather rare in lexicography, not to speak of its wide sociological and political impact.

The fact that there is a substantial specific corpus behind it (which is not the case of German, Polish or Russian where similar books have been published) is both a possibility and offer to anyone who wishes to study the period language more widely, and, should other dictionaries arise in other languages (from the period of Communism and Fascism primarily) it might become a contribution to a broader understanding of political and social movements spanning substantial parts of Europe not long ago - and not only that.

## 8 References

*Slovník komunistické totality* (Dictionary of the Communist Totality), eds. František Čermák, Václav Cvrček, Věra Schmiedtová, further coauthors Jan Koček, Dominika Kovářiková, Karel Kučera, Renata Novotná, NLN Praha 2010

- Arendtová, H. (2000). *Původ totalitarismu I a II*. Oikoymenh, Praha.
- Bartošek, K., J. L. Margolin (1999). *Černá kniha komunismu: Zločiny, teror, represe*; Paseka, Praha, Litomyšl.
- Bralczyk, J. (2001). *O języku polskiej propagandy politycznej lat siedemdziesiątych*. Wydawnictwo Trio, Warszawa.
- Brenner, Ch. (2008). *Znormalizovaný totalitarismus? Paradigmata výzkumu socialismu I*, Ročenka textů zahraničních profesorů; FFUK, Praha.
- Brousek, A. (1987). *Podivuhodní kouzelníci, čítanka českého stalinismu v řeči vázané z let 1945-1955*. Rozmluvy, Londýn.
- Cvek, B. (2005). *Proč a v čem je komunismus vlastně totéž co nacismus*; Britské listy 26.2.2005. Dostupné z WWW: <<http://www.blisty.cz/art/22182.html>>.
- Džilas, M. (1977). *Nová třída, kritika soudobého komunismu*. Demos, Curych.
- Fidelius, P. (1983). *Jazyk a moc*. Edice Arkýř, Mnichov.
- Fidelius, P. (2000). O totalitním myšlení, In *Kritické eseje*. Torst, Praha.
- Głowiński, M. (1991). *Nowomowa po polsku*. Wydawnictwo PEN, Warszawa.
- Hochel, B. (1991). Totalita v jazyku (a v nás). In *Kultúrny život*, 25, Bratislava.
- Jánský, P. (2004). *Totalita slovem, písní a obrazem (Dějiny hrůzovlády KSČ)*. Nakladatelství Music, Cheb.
- Klemperer, V. (2002). *Deníky 1933-1941, Chci vydat svědectví*. Paseka, Praha, Litomyšl (*Victor Klemperer: Die Tagebücher 1933-1945. Kritische Gesamtausgabe. CD-ROM. Berlin 2007*).
- Klemperer, V. (2003). *Jazyk Třetí Říše - LTI: poznámky filologovy*. Nakladatelství H&H, Jinočany.
- Kartous, B. (2004). *Může totalita zahltit celou civilizaci?* Britské listy 10.11.2004. Dostupné z WWW: <<http://www.blisty.cz/art/20532.html>>
- Macura, V. (1992). *Šťastný věk*. Pražská imaginace, Praha.
- Nowak, P. (2002). *Swoi i obcy w językowym obrazie świata: język publicystyki polskiej z pierwszej połowy lat pięćdziesiątych*. Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin.
- Mokijenko, V. M., T. G. Nikitina (1998). *Tolkovyy slovar jazyka sovdepii*. Sankt-Peterburg.
- Mokijeno, V. M. (2000). *Novaja ruskaja frazeologija*. Opole.
- Popper, K. (1994). *Otevřená společnost a její nepřátelé I. a II*. Oikoymenh, Praha.
- Pisarek, W. (1972). *Frekwencja wyrazów w prasie*. Ósrodek Badań Prasoznawczych, Kraków.
- Pisarek, W. (1976). *Język służy propagandzie*. Ósrodek Badań prasoznawczych, Kraków.
- Pisarek, W. (1983). *Analiza zawartości prasy*. Ósrodek Badań prasoznawczych, Kraków.

- Röhricht, A. (2008). *Ideologie, jazyky, texty: Analýza a interpretace textů Rudého práva z roku 1953 a 1975 a Práva z roku 1997*. Bor, Liberec.
- Šebesta, K. (2001). Studovat jazyk totality. In: *Institucionalizace (ne)odpovědnosti: Globální svět, evropská integrace a české zájmy*. Karolinum, Praha.
- Šlosar, D. (1995). Jazyk totality a jazyk dneška, In *Spisovná čeština a jazyková kultura 1993*. FF UK, Praha.
- Totalitarismus 3* (2007). Eds. I. Budil a T. Zíková. FF ZU, Plzeň.
- Totalitarismus 4* (2008). Eds. I. Budil a T. Zíková. FF ZU, Plzeň.
- Vítězové? Poražení? I. díl, Disent v období tzv. normalizace* (2005). Eds. M. Vaněk a P. Urbášek, Prostor, Praha.
- Vítězové? Poražení? II. díl, Politické elity v období tzv. normalizace* (2005). Eds. M. Vaněk a P. Urbášek, Prostor, Praha.
- Wierzbicki, P. (1986). *Struktura kłamstwa*, Glos, Warszawa.
- Włodek, M. (2002). *Sondy do jazyka totality*. diplomová práce, FFUK, Praha.

# Dictionary of Abbreviations in Linguistics: Towards Defining Cognitive Aspects as Structural Elements of the Entry

Ivo Fabijanić

English Department, University of Zadar, Obala k. Petra Krešimira IV., 2,

HR-23000 Zadar, Croatia

ifabijan@unizd.hr

## Abstract

The first part of the article deals with general information about the project of compiling a dictionary of abbreviations in linguistics. It also contains a short overview of past research together with their main results. So far, some specific theoretical and practical solutions were proposed. Theoretical solutions refer to the Multi-level approach in collecting data for submorphemic word-formations, which consists of three aspects: 1) *Structure and Modes of Production*, 2) *Cognitive Aspects*, and 3) *Functional Aspects*. Practical solutions for the structure and modes of production have already been recommended with the results properly substantiated by the examples of abbreviations. The second part presents results of the analysis for the cognitive aspect of the multi-level approach. The semantic part reveals the relationship between an abbreviation type and the semantic (sub-)field it might be assigned to. Semiotic part is achieved by designating a specific interpretation of a sign's meaning, while the analysis of lexicalization and institutionalization confirms their contingent addition to this scientific lexicon. The motivational aspect takes two conceptual perspectives in consideration: the narrower and broader senses of the formation, and various semantic relationship patterns which led to the classification of fully and partially motivated abbreviations.

**Keywords:** dictionary; abbreviations; linguistics; micro-structure; analysis; multi-level approach; cognitive aspect; semantics; semiotics; motivation; lexicalization; institutionalization

## 1 Introduction

The overall aim of this paper is to provide specific theoretical and methodological models in preparing both a macro- and micro-structural framework for compilation of the future bilingual and bidirectional (English – English – Croatian), specialized and explanatory dictionary of abbreviations in linguistics. The dictionary would cover the core areas of linguistics and its interdisciplinary areas as well. The main triggers which motivated us to study the lexicon of abbreviations are, by all means, the lack of consistent categorisation and typology, as well as fixed boundaries between the respective types of abbreviations, which are, unfortunately, their most distinctive characteristics. The classification of abbreviations used here largely relies on López Rúa's (2006) work. We find this taxonomy ap-

appropriate because it clearly distinguishes certain abbreviation types. *Abbreviations* are divided into *simple* and *complex abbreviations*. According to López Rúa, an *initialism* is made of initial letters or occasionally the first two letters of the words in a phrase and combined to form a new sequence (2006: 677). The term initialism denotes an abbreviation created through the usage of initial letters, which applies to both alphabetisms and acronyms. The term alphabetism denotes an abbreviation pronounced as a series of letters of the alphabet, while the acronym denotes abbreviations pronounced as whole words. *Clippings* are either shortened words or syllables without a change in meaning or functions (López Rúa 2006: 676). *Blends* are created by “[...] joining two or more word-forms through simple concatenation or overlap and then by shortening at least one of them” (López Rúa 2006: 677).

## 2 Previous Research in the Field

Previously explained and described in the past research (cf. Fabijanić 2014, in print), the main objectives in compiling the dictionary and its main characteristics refer to the following concepts: bilingual and bidirectional type of dictionary according to the number of languages, alphabetical order of headwords according to the order of presentation, appropriate use of the data which should be provided within the microstructural confines, and scientific, identified by the domain-specific collection of abbreviations in linguistics from specialized publications. So far, I have proposed some specific theoretical and practical solutions to be utilized in compilation of the entries (Fabijanić 2014, in print). These solutions refer to the concept of multi-level approach in collecting data for submorphemic word-formations (cf. Fandrych 2008a).

The triaspectual multi-level approach is comprised of the following stages: 1) *Structure and Modes of Production*, i.e. the structural aspects and word-formation potential, word class, medium and origin; 2) *Cognitive Aspects*, i.e. semantic, semiotic and motivational aspects, lexicalization and institutionalization, and 3) *Functional Aspects*, i.e. stylistic and sociolinguistic aspects, pragmatic and text-linguistic aspects. The application of this interdisciplinary approach will give a fuller and a more transparent picture of various orthographic, morphological, semantic, stylistic and functional processes involved in the production and uses of abbreviations.

As for the first aspect of *Structure and modes of production*, abbreviations have been classified according to two criteria – *narrower* and *broader sense* (cf. Fabijanić 2014, in print). The narrower sense refers to those formed by using initial letters of each element in the expansion (pertaining mostly to alphabetisms), and pronounced either by individual names of letters or as a word. The broader sense implies the ways and processes of formation, more or less different from the orthographic norms (pertaining mostly to hybrid forms, acronyms, blends and clippings featuring some orthographic changes), in consequence of which, one or more initials are used for various smaller elements of the expansion (smaller than words, yet bigger than initials). Due to this, initials for graphemes, compounds, and affixes, grammatical and lexical words found in the final form of an abbreviation, as well

as different orthographic changes, such as ellipsis, conversion, metathesis, addition, etc., were analysed and (sub-)classified.

For the purpose of their differentiation, a system of exclusive classification and subclassification of abbreviations was proposed (Fabijanić, Malenica 2013). Miscellaneous realisations of abbreviations are generally diversified into two main groups: those realised in the narrower sense and those in the broader sense. Abbreviations in the narrower sense are exclusively explained with an *LLL* descriptor for initials used in their formation. Abbreviations in the broader sense are represented with a whole set of additional different letters or initials (written either in capital or small letters) added to a three-letter descriptor: e.g. *l* for small letters, *P* for initial affixes, *N* for numerals, *S* for syllables, and *W* for a word. Further orthographic changes are explained by other descriptors, e.g. *E* for ellipsis, *C* - conversion, *M* - metathesis, and *A* for addition of a word or a diacritic sign not normally found in expansions. Comprehension and consequently classification of abbreviations depends on the degree of their (non-)coordination with the common orthographic norms.

The research in *Structure and modes of production* of abbreviations has proven that most of the alphabetisms are formed according to the criterion in the narrower sense, while the ratio of those formed in the narrower and the broader sense for acronyms (which were fewer in number than alphabetisms) was in favour of the broader-sense formations (cf. Fabijanić 2014 in print). As for the hybrid-form ratio, the broader-sense criterion is also more evident. The direct results of the analysis have attested the possibility of applying previously devised descriptors (Fabijanić, Malenica 2013), as well as some new descriptors, which have emerged in the analysis of abbreviations in linguistics (e.g. *P-LL* for alphabetisms, *FLL* for acronyms and some clippings in broader sense, *SFL* and *FFL* for blends, *Lll* and *lll* for clippings in broader sense, and *LLW*, *PLW*, *L/LW*, *FLW*, *F-LW*, *S-LW*, *SFL*, *WLL*, *W-LL* for hybrid formations). The Structure and modes of production aspect will be described by labels referring to form(s) of abbreviations, medium, word class, origin (cf. § 7).

### 3 Aims and Objectives of the Current Research

The immediate aim of this research is to bring forth the results of the analysis for the second aspect of the Multi-level approach, i.e. the cognitive aspect inherent to non-morphematic word-formations. The *Cognitive aspect* deals with semantic, semiotic and motivational aspects, as well as with the lexicalization and institutionalization of abbreviations. The semantic part gives answers to the relationship between a given word-formation of a specific abbreviation type and the semantic (sub-)field it may be assigned to. Semiotic part is achieved by designating a specific interpretation of a sign's meaning, while the analysis of lexicalization and institutionalization of abbreviations confirms their contingent addition to the lexicon of a language or to the specific scientific lexicon. As far as motivation is concerned, the relationship between the structural pattern of abbreviations, their meaning(s) and

phonemic, graphemic and sub-morphemic elements was analysed, which led to the classification of abbreviations into those fully or partially motivated.

## 4 The Corpus of Abbreviations

The abbreviations analysed in this article were taken from different dictionaries of general linguistics and dictionaries of various linguistic disciplines (e.g. phonetics and phonology, lexicography, etc.; cf. Sources). The corpus comprises 446 abbreviations, belonging either to the category of simple or complex abbreviations. There are 270 alphabetisms, 67 acronyms, 5 blends, 19 clippings, 22 simple abbreviations and 63 hybrid forms. Alphabetisms, simple abbreviations and clippings were mainly formed according to the criterion of narrower sense formations, while acronyms, blends and hybrid formations were mainly formed according to the criterion of broader sense (cf. Fabijanić 2014, in print).

The corpus provides additional information about each abbreviation in the following order: abbreviation, expansion, descriptor of abbreviation form, source, abbreviation type, and details of the analysis for lexicalization, institutionalization, semantics, semiotics, and motivation.

## 5 Research Methods

The explication of research methods in the article refers to the stages of analysis for the cognitive aspect of the Multi-level approach. They will be dealing with the description of methods applied for the analysis of semantics, semiotics, motivation, lexicalization, and institutionalization, i.e. the features of the mentioned subsidiary aspects which can be attributed to abbreviations.

The application of semantic aspect is understood through the possibility of assigning a specific abbreviation type to its semantic field, i.e. the (sub-)field of linguistics or the interdisciplinary disciplines. The practice of assigning a semantic field does not certainly mean that an abbreviation could have only been appointed to that specific field; on the contrary, each abbreviation can be assigned to other semantic fields as well, but what we wanted to point out by this practice is the immediate textual and contextual surrounding an abbreviation was found in. Semiotic aspect is realised by determining the relations between the elements of an abbreviation or a sign. Due to the fact that two elements of the sign – the sign itself and the object – are already present in the form of abbreviations and their expansions, I find the interpretant (the sense/meaning) to be the element which can and has to be analysed by the implementation of Peirce's three-graded diversification of sign's clarity or understanding, i.e. by implementing the grade of the immediate interpretant (sign's first meaning), the grade of the dynamic interpretant (the actual effect of the sign) or the grade of the final interpretant (the final interpretative result of the sign).

Motivation in the formation of abbreviations was analysed by taking two conceptual perspectives in consideration. The first concept is connected to the narrower and the broader senses of the formation,



while the second one is connected to the classification of abbreviations according to various motivational patterns (e.g. homophony, homography, homonymy, metaphor).

Lexicalization in this work is understood within the confines of the synchronic sense in which the lexicalization of abbreviations corresponds to *the process of listing* and the *listedness* (cf. Hohenhaus 2005: 356). For the purpose of the analysis, abbreviations are classified into those that were or were not listed or lexicalized. Being specific in their formation, listing/listedness of abbreviations will inevitably be sub-classified according to their specific structure and modes of production. Therefore, abbreviations like simple abbreviations and clippings, due to their graphic and spoken arbitrariness, will not be listed (or can be considered to be in the process of listing), while alphabetisms, acronyms, blends, and hybrid formations will be listed. Finally, institutionalization “[...] refers to the stage in the life of a word at (or form) [brackets in original] the transitional point between the status of ex-nonce-formation-turned-neologism and that of generally available vocabulary item, i.e. a formation that is listed but not (necessarily) [brackets in original] lexicalized in the diachronic sense yet [...]” (cf. Hohenhaus 2005: 359). Institutionalization, in terms of this part of the research, refers to the fact whether an abbreviation as such can be considered as institutionalized within the lexicon of linguistics or not.

## 6 Cognitive Aspects: The Analysis

### 6.1 Semantic and Semiotic Aspect

As described in the previous section, the semantic aspect of abbreviations was realised in accordance with the nearest corresponding semantic field. The analysed abbreviations were allocated to various sciences and disciplines, both linguistic or non-linguistic ones, according to the principle of immediate context within the entries. Here are the semantic fields with the information on total number of allocated abbreviations and an example with its expansion: *Applied linguistics* (39; ASTP - ‘Army Specialized Training Program’), *Cognitive linguistics* (1; ICM - ‘Idealized Cognitive Model’), *Computational linguistics* (38; COBOL - ‘Common Business-Oriented Language’), *Corpus linguistics* (9; BNC - ‘British National Corpus’), *Neurolinguistics* (6; TDH - ‘Trace-Deletion Hypothesis’), *Historical linguistics* (8; PIE - ‘Proto Indo-European’), *Linguistic anthropology* (2; LISA - ‘Language and Identity in Sociocultural Anthropology’), *Psycholinguistics* (12; PALPA - ‘Psycholinguistic Assessment of Language Processing in Aphasia’), *Sociolinguistics* (19; LSPT - ‘Language Status Politics’), *Pragmatics* (71; AP - ‘Applying Pragmatics’), *Theoretical approaches* (4; TG - ‘Transformational grammar’), *Phonetics and phonology* (16; VOT - ‘Voice Onset Time’), *Morphology* (27; SG - ‘Singular’), *Syntax* (56; NP - ‘Noun Phrase’), *Semantics* (13; SFH - ‘Semantic Feature Hypothesis’), *Dialectology* (1; ADS - ‘American Dialect Society’), *Stylistics* (1; DS - ‘Direct Speech’), *Lexicology* (1; ALLEX - ‘African Languages LEXical project’), *Lexicography* (25; OED - ‘Oxford English Dictionary’), *Linguistic typology* (2; ASL - ‘American Sign Language’), *Text analysis* (2; PISA - ‘Procedures for Incremental Structure Analysis’), *Discourse analysis* (1; SA - ‘Speech Acts’), *Literary*

*linguistic analysis* (1; NI – ‘Internal Narration’), *Translational studies* (1; TT-ST – ‘Target Text–Source Text’), *Philosophy of language* (1; CF – ‘Context-Free’), *Semiotics* (1; ISL – ‘Iconicity in Sign Language’), *Cognitive pragmatics* (1; ICM – ‘Idealized Cognitive Model’), *Cognitive technology* (1; CT/TC – ‘Cognitive Technology/ Technological Cognition’), *Speech recognition* (1; HMM – ‘Hidden Markov Model’), *Speech technology* (2; AVIOS – ‘American Voice Input and/Output Society’), *Media* (2; BBC English – ‘British Broadcasting Company English’), *Computational science* (9; WOZ – ‘Wizard-of-OZ simulation’), *Medicine* (1; TBI – ‘Traumatic Brain Injured patient’), *Literacy* (2; NLS – ‘New Literacy Study’), *Communicology* (6; SAT – ‘Speech Accommodation Theory’), *Education* (25; CABE – ‘Central Advisory Board of Education’), *Associations* (20; EURALEX – ‘European association for LEXicography’), *Organisations* (12; CALLSSA – ‘Center for Applied Language and Literacy Studies and Services in Africa’), *Conferences* (1; LTRC – ‘Language Testing Research Colloquium’), *Journals* (5; IJOAL – ‘International Journal Of Applied Linguistics’), and *Databases* (1; LAPTOC – ‘Latin American Periodical Table Of Contents’).

Most of the abbreviations were allocated for the following fields: *Pragmatics* (71), *Syntax* (56), *Applied Linguistics* (39), *Computational linguistics* (38), *Morphology* (27), *Education* (25), and *Lexicography* (25), while the least allocated, i.e. a field with one example, are: *Dialectology*, *Stylistics*, *Lexicology*, *Discourse analysis*, *Literary analysis*, *Semiotics*, *Cognitive Pragmatics*, *Cognitive technology*, *Speech recognition*, *Medicine*, *Conferences*, and *Databases*. A cross-sectional view of abbreviation types in some semantic fields will disclose the following data about the most frequent abbreviation type and its formation structure: the most frequent type of abbreviations in the fields of *Pragmatics*, *Applied linguistics*, *Syntax* and *Computational linguistics* is the type of alphabetisms with the narrower sense formation structure of *LLL*.

Semiotic aspect refers to the analysis of three grades of interpretants: immediate, dynamic and final. Most of the abbreviations were classified within the class of sign having an immediate interpretant (approx. 230 abbreviations), followed by the class of final interpretants (approx. 120), and the ones with the dynamic interpretant (approx. 90). I believe that in case of abbreviations, the classification of interpretants is firmly connected to the cognition of relationship between the abbreviation type(s), its/their expansion(s), variability of expansion, some inner features of different abbreviation types (e.g. those of acronyms’ when compared to alphabetisms), and frequency of use. The immediate interpretant indicates “[...] the effect the sign first produces or may produce upon a mind without any reflection upon it” (cf. *Semiotics and Significs* 1909: 110-1). From the previous quotation, it might be possible to assume that primary effects on our understanding of some abbreviations can be considered *sui generis*. Such is the case for some alphabetisms, clippings and hybrid forms whose understanding is conditioned by their immediate (con-)textual surrounding or expansion, e.g. primary understanding of the alphabetism AAAL (‘American Association for Applied Linguistics’) is conditioned by its expansion provided in the text. Furthermore, clippings like ACC (‘ACCusative’), ACT (‘ACTive’), COP (‘COPula’), or hybrid formations like, ALLEX (‘American Association for Applied Linguistics’), BSAfE (‘Black South African English’), CCSARP (‘Cross-Cultural Speech Act Realization Project’), LCPT (‘Language Corpus Politics’), LRs (‘Language Rights’), ATN grammar (‘Augmented Transition Network grammar’) is by all means conditioned by that primary effect of understanding. Direct evidence of this claim is

supported in the following example: the difference in understanding the ALLEX and EURALEX, although some extensional as well as structural and formational features are being shared, is evident in the finality of understanding of the latter hybrid (connotation for *EUR-* is more immediate than for *ALL-*).

With regard to the dynamic interpretant, understanding of abbreviations within this category is mostly conditioned by the variability of extensions or the form of an abbreviation. The dynamic interpretant “[...] is that which is experienced in each act of Interpretation and is different in each from that of the other [...]” (cf. Semiotics and Significs 1909: 110-1). I believe that this can be witnessed in the following examples of simple abbreviations, acronyms, alphabetisms and clippings: M for ‘Movement’ or ‘Metapragmatic joker’; ACE for ‘Automatic Content Extraction’ or ‘Australian Corpus of English’; CT for ‘Cognitive Technology’, ‘Conversational Theory’ or ‘Centering Theory’; AUX or Aux for ‘Auxiliary’.

The final interpretant is “[...] the one Interpretative result to which every Interpreter is destined to come if the Sign is sufficiently considered [...]” or “[...] the effect the Sign *would* [italics in original] produce upon any mind upon which the circumstances should permit it to work out its full effect [...]” (cf. Semiotics and Significs 1909: 110-1). Due to their completeness of graphic and phonetic forms, i.e. the possibility of being read as words, and their frequency of use, most of the analysed blends, acronyms, and some hybrids, which are very similar to acronyms, together with some infrequent alphabetisms, were classified into the class of abbreviations having the final interpretant. The sum of the meanings or the final interpretative result the signs would inevitably have, can be confirmed by the examples of: acronyms – COBUILD (‘Collins Birmingham University Information’), ECHO – (‘European Commission Host Organisation’); blends – (‘AFRiCAn association for LEXicography’), FORTRAN (‘FORMula TRANSlation’); hybrids – AUSTRALEX (‘AUSTRAliaN Association for LEXicography’), ITSPOKE (‘Intelligent Tutoring SPOKEn dialogue system’); alphabetisms – HTML (‘HyperText Mark-up Language’), L1 (‘First Language’).

## 6.2 Motivational Aspect

As it has already been explained (cf. § 5), motivational aspect was analysed through two conceptual perspectives. The first concept takes into consideration the formational difference between various types of abbreviations, previously classified according to the aspects of narrower and broader sense:

The narrower sense of their creation refers to those formed by using initial letters of each element in the expansion (mainly alphabetisms) [brackets in original], and pronounced either by individual names of letters or as a word. The broader sense implies the ways and processes of formation, more or less different from the orthographic norms (mainly hybrid forms, acronyms, blends and clippings featuring some orthographic changes) [brackets in original], in consequence of which, one or more initials are used for various smaller elements of the expansion (smaller than words, yet bigger than initials) [brackets in original]. (Fabijanić 2014; in print)

The motivational aspect for the abbreviations analysed in this work assumes that some of them are fully motivated, while the others are partially motivated. I find fully motivated abbreviations to be the ones which largely correspond to the norms of narrower sense creations, i.e. alphabetisms and acronyms, simple abbreviations, blends and clippings formed by the orthographic norm and in which every element of the expansion is traceable. Partially motivated are those that principally fall into the group of broader sense formations, i.e. alphabetisms, acronyms, simple abbreviations, clippings, blends, and hybrid formations which are not formed by the orthographic norms and in which more or less elements of expansions can be traced. The second conceptual perspective of the motivational aspect takes into consideration the specific patterns which motivated the emergence of abbreviations. These might be homophonous, homographic, homonymic, and metaphorical patterns. I shall provide some examples of different abbreviation types for each pattern. The homographic pattern, in which initials from extensions are repeated in abbreviations, is mostly evident in narrowly formed alphabetisms, e.g. IPA – ‘International Phonetic Association’, NLU – ‘National Lexicographic Unit’. The homophonic pattern, in which part(s) of extensions or initials (in acronyms) are either echoed in a resultant abbreviation or make a word having different meaning in the general lexicon, can be detected in the example of clippings ACT (‘ACTIVE’), INACT (‘INACTIVE’), in the acronym ACE (‘Automatic Content Extraction’), or the blend AFRILEX (‘AFRICan association for LEXicography). Sometimes the overlapping of patterns can be realized as in the example of BANA acronym (‘Britain, Australasia, and North America’) in which homographic and homophonic principles can be traced (the repetition of initials and homophony with other words and abbreviations like the surname *Bana*, a drink named *BANa* or *BANA* for ‘*British Acoustic Neuroma Association*’, ‘*Bulimia Anorexia Nervosa Association*’, ‘*Bath Area Network for Artists*’, etc.). The homonymic principle is realised in examples of hybrids and acronyms like BIT – ‘BINary digiT’ (homonymic with *bit* ‘amount of sth, part of sth’) and CAM – ‘Center of Auditory Memories’ (homonymic with *CAM* ‘Computer Aided Manufacturing’, *cam* ‘a wheel part which changes the movement of the wheel’, *cam* ‘a clipped form from *camera*’). *The homonymic principle can be disrupted by the addition or deletion of graphemes otherwise not found in original words, e.g. the acronym KWIC (‘KeyWord In Context’) in which <KWI> suggests the group of graphemes <qui>, while <C> suggests the group <ck> as in quick, or in case of the hybrid CHILDES database (‘CHILd Language Data Exchange System database’) in which its form might suggest the ungrammatical plural form of the noun child. Metaphorical principle of motivation is evident in subsequent forms of acronyms or hybrid forms: NORM – ‘Non-mobile, Older, Rural Male’; MACK – Multimodal Autonomous Conversational Kiosk’; CHAT – ‘Cultural-Historical Activity Theory’; BASIC English – ‘British, American, Scientific, International, Commercial English’, while humorous touch is felt in metaphorically motivated FUG (‘Functional Unification Grammar’), MUD (‘Multi-User Domain’), DARE (‘Dictionary of American Regional English’), LISA (‘Language and Identity in Sociocultural Anthropology’), ELI (‘English Language Institute’), PISA (‘Procedures for Incremental Structure Analysis’), etc.*

### 6.3 The Aspects of Lexicalization and Institutionalization

For the purpose of this research, Lipka's definitions on lexicalization and institutionalization of complex lexemes will be applied (2005: 4). According to Lipka, "[...] complex words [...] were coined according to productive morphological or semantic process [...], and they have [...] been affected – to a greater or lesser degree – by formal and/or semantic changes subsumed under the concepts of **lexicalization** and **institutionalization** [bold in original]." He defines lexicalization as: "[...] the process by which complex lexemes tend to become a single unit with a specific content, through frequent use. In this process, they lose their nature as a syntagma, or combination [of smaller units], to a greater or lesser extent" (1992: 107). In his later research (2005: 7), the definition of lexicalization was extended by the features of gradual, historical process which involve changes in phonology and semantics, as well as loss of motivation.

Since all the above definitions with some additional descriptions in Lipka's work (together with other work of specialists cited in his works), fit well to the topic of my research and since the lexicon to which the two processes can be applied is compatible with their patterns of verification, I shall propose a three-stage model of lexicalization for abbreviations, i.e. preliminary, primary and secondary stage. The *preliminary stage* is understood as a preparatory stage in the process of lexicalization. It refers to a small group of simple abbreviations whose final abbreviated form is too short or arbitrary to be considered either partially or fully lexicalized, e.g. A – 'Adjective', L – 'Location', M – 'Metapragmatic joker' or 'Movement', N – 'Noun'. The *primary stage*, with partially lexicalized abbreviations, is disclosed in formational incompleteness and variability of abbreviations. Thus we have alphabetisms with both small and capital letters (CmC – 'Computer mediated Communication'), clippings with small and capital letters, which, additionally, are not pronounced (Aux – 'Auxiliary', Utt – 'Utterance'), hybrid formations with small and capital letters (LHRs – 'Linguistic Human Rights', SaPs – 'Speech acts Projections'), and alphabetisms with various diacritics (R-A – 'Referentially Autonomous expression', CT/TC – 'Cognitive Technology/Technological Cognition'). The *secondary stage* with completely lexicalized abbreviations refers to those which are more easily recognized as lexical units, i.e. acronyms (LAD – 'Language Acquisition Device', LAPTOC – 'Latin American Periodical Table Of Contents'), hybrids in combinations with words (LISP language – 'LIST Processing language'), alphabetisms in narrower sense (MHG – 'Middle High German'), blends (AFRILEX – 'AFRICan association for LEXicography), some clippings with consistent and utterable orthography (COP – 'COPula', FEM – 'FEMinine').

Institutionalization, in Lipka's view, refers to "[...] the sociolinguistic aspect of this process and can be defined as the integration of a lexical item, with a particular form and meaning, into the existing stock of words as a generally acceptable and current lexeme" (2005: 8). Lipka further defines institutionalization as the process in which a specific speech community (e.g. doctors, medical people, linguists, etc.) accepts the specific lexemes into the lexicon. (2005:11). He also states that "[both lexicalized and] institutionalized words, ie item-familiar ones, are registered and listed in good dictionaries [...]"

(2005: 12). If we take into consideration all the above viewpoints and aspects of institutionalization, as well as the facts about our corpus confirmed and compiled from specialized (mostly encyclopaedic) dictionaries, than we can conclude that all the analysed examples of abbreviations have been fully institutionalized. For the purposes of defining the entry structure, institutionalization will be explained by two attributes – affirmed or not affirmed institutionalization.

## 7 The Entry and its Elements

As it had previously been suggested, after having completed the analysis of *Structure and modes of production aspect* (cf. Fabijanić 2014, in print), an entry would consist of the following elements: a headword, (a) variant(s), pronunciation, the type of word-formation, the information on the word class (where applicable), a descriptor which will inform users about the mode of production, expansion elements in English and their translation in Croatian, the information about the medium and the origin of abbreviations, both in English and Croatian.

The following examples of different abbreviation forms (an alphabetism, an acronym, a clipping, a blend and a hybrid form) present the micro-structure of the future dictionary entry with the addition of elements for the *Cognitive aspect*, i.e. information on the semantic field, interpretation of the sign's interpretant (immediate, dynamic or final), information on the motivational aspect (fully or partially motivated), the information on the lexicalization (preliminary, partial and complete lexicalization) and institutionalization (institutionalized or not). The subsequent (simple) abbreviations have been suggested for the use within the entry: *T* (type), *E* (expansion), *M* (medium), *D* (descriptor), *O* (origin), *SF* (semantic field), *SI* (sign's interpretant), *MT* (motivation), *L* (lexicalization), and *I* (institutionalization).

- **ADS** [ˈeɪˈdiːʔes] **T:** *alph.* | **E:** *American Dialect Society/Američko dijalektalno društvo* | **M:** *written/ pisani, spoken/govorni* | **D:** *LLL* | **O:** *ADS was founded in 1889 with the intention of creating a dictionary of American dialects. (ELL)/Američko dijalektalno društvo osnovano 1889. s namjerom stvaranja rječnika američkih narječja.* | **SF:** *dialectology/dijalektologija* | **SI:** *immediate/ neposredan* | **MT:** *full/potpuna* | **L:** *complete/dovršena* | **I:** *affirmed/potvrđena*
- **ACE** [eɪs] **T:** *acr.* | **E:** *1) Automatic Content Extraction/Automatsko ekstrahiranje sadržaja, 2) Australian Corpus of English/ Korpus australskoga engleskog* | **M:** *written/pisani, spoken/govorni* | **D:** *1) LLL, 2) LLL E = prep.;* | **O:** *1) The ACE program is a successor to ® MUC that has been running since a pilot study in 1999. (ELL)/ACE program provodi se još od pilot-projekta iz 1999., a naslijedio je MUC., 2) The corpus of Australian English compiled at Macquarie University using texts published in 1986. (HEL) / Korpus australskoga engleskog sačinjen na Macquarie sveučilištu iz tekstova objavljenih 1986.* | **SF:** *computational and corpus linguistics/ računalna i korpusna lingvistika* | **SI:** *dynamic/ dinamičan* | **MT:** *1) full/ potpuna, 2) partial/djelomična* | **L:** *complete/dovršena* | **I:** *affirmed/potvrđena*



- **AUX, Aux, aux** [-] **T:** *clip.* | **E:** *Auxiliary/pomoćni* | **M:** written/pisani | **D:** *Lll* | **O:** *Lat. auxiliaris* - 'giving aid' (RDLL)/*lat. pomoćni* - 'koji pomaže'. | **SF:** morphology/morfologija | **SI:** dynamic/dinamičan | **MT:** partial/djelomična | **L:** partial/ djelomična | **I:** affirmed/ potvrđena
- **AFRILEX** ['æfrɪ,leks] **T:** *blend* | **E:** *AFRICan association for LEXicography/Afričko leksikografsko udruženje* | **M:** written/ pisani, spoken/govorni | **D:** *LSF E= noun, prep.* | **O:** *AFRILEX* was founded in 1995 and strives to promote all aspects of lexicography on the African continent. (ELL)/Organizacija osnovana 1995. u svrhu promocije svih aspekata leksikografije na afričkom kontinentu. | **SF:** associations, lexicography/udruženja, leksikografija | **SI:** dynamic/dinamičan | **MT:** full/ potpuna | **L:** complete/dovršena | **I:** affirmed/ potvrđena
- **ALGOL** ['ælgɒl] **T:** *hybr. (syll+s.abb.)* | **E:** *ALGOritmic Language/algoritamski jezik* | **M:** written/pisani, spoken/ govorni | **D:** *SSL* | **O:** Programming computer language appeared in 1958. (RDLL)/ Programski računalni jezik nastao 1958. godine. **I:** affirmed/ potvrđena | **SF:** computer science/računalne znanosti | **SI:** final/konačan | **MT:** partial/djelomična | **L:** complete/ dovršena | **I:** affirmed/ potvrđena

## 8 Conclusion

In closing, I would like to stress again the overall and immediate aims of this research. The overall aim is to provide the basis for the future dictionary of abbreviations in linguistics, which would be bilingual, bidirectional, specialized (domain-specific, technical), synchronic, explanatory, alphabetically arranged dictionary, informative and encyclopaedic in content, and serve both non-specialized and specialized audience. The solution for the lexicographic presentation of abbreviations is based on Ingrid Fandrych's *Multi-level approach* which is comprised of three aspects: *Structure and Modes of Production*, *Cognitive Aspect*, and *Functional Aspect*. The immediate aim of the research is to bring forth the results of analysis for the second aspect, i.e. the Cognitive aspect, which deals with semantics, semiotics, motivation, lexicalization and institutionalization.

Summing up the results of this research, I would like to state that the sources used in compiling the corpus of abbreviations in linguistics, have proved to be trustworthy, valuable and fundamental for the cognitive aspect, just as they were for the structures and modes of production. Furthermore, the research has proved that the criteria of narrower and broader sense classification can be reiterated for the elements of the cognitive aspect too.

I hope to have shown that the stages of the cognitive aspect can be defined by the recommended methods in analysing semantic, semiotic, motivational features of abbreviations, as well as the facts about their lexicalization and institutionalization. The semantic aspect was realised in accordance with the nearest corresponding semantic field for a specific abbreviation. The abbreviations were allocated to various sciences and disciplines, such as *Pragmatics*, *Syntax*, *Applied Linguistics*, *Computational linguistics*, etc. The analysis of the semiotic aspect was realised by the application of three grades of interpretants: immediate (sign's first meaning), dynamic (the actual effect of the sign) and final (the fi-

nal interpretative result of the sign). Most of the abbreviations were classified within the class of sign having an immediate interpretant, followed by the class of final interpretants, and the class of the dynamic interpretant. The motivational aspect for the abbreviations analysed in this work assumes that some of them are fully motivated (those which largely correspond to the norms of narrower sense creations), while others are partially motivated (those that principally fall into the group of broader sense formations). The second conceptual perspective of the motivational aspect takes into consideration the homophonous, homographic, homonymic, and metaphorical patterns which motivated the emergence of abbreviations. As far as lexicalization is concerned, a three-stage model of lexicalization for abbreviations (preliminary, primary and secondary stage) is proposed. The preliminary stage is a preparatory stage in the process of lexicalization. The *primary stage*, with partially lexicalized abbreviations, is disclosed in formational incompleteness and variability of abbreviations. The *secondary stage* with completely lexicalized abbreviations refers to those which are more easily recognized as lexical units. In dealing with the aspect of institutionalization, the crucial moment for the application of dichotomous differentiation between affirmed and not affirmed institutionalization, is recognized through the fact that abbreviations have been lexicographically attested.

Finally, with regard to the micro-structure of the future entry, its second level will consist of five structural elements which will be represented by the following abbreviations/symbols: *SF* (semantic field), *SI* (sign's interpretant), *MT* (motivation), *L* (lexicalization), and *I* (institutionalization).

## 9 References

- Fabijanić, I., F. Malenica (2013). "Abbreviations in English medical terminology and their adaptation to Croatian". In *JAHN*, Vol. 4, No. 7., Faculty of Medicine, Rijeka.
- Fabijanić, I. (2014). "Dictionary of abbreviations in linguistics: towards a bilingual, specialized, single-field, explanatory dictionary". In Proceedings of the conference *Planning Non-existent dictionaries: Lexicographic Typologies*, Lisbon. In print.
- Fandrych, I. (2004). *Non-Morphematic Word-Formation Processes: A Multi-Level Approach to Acronyms, Blends, Clippings and Onomatopoeia*. Unpublished PhD Thesis. University of the Free State: Bloemfontein.
- Fandrych, I. (2008a). "Pagad, Chillax and Jozi: A Multi-Level Approach to Acronyms, Blends, and Clippings". In *Nawa Journal of Language and Communication*. Vol.2, No. 2.
- Fandrych, I. (2008b). "Submorphemic elements in the formation of acronyms, blends and clippings", In *Lexis 2: "Lexical Submorphemics / La submorphémique lexicale"*.
- Hohenhaus, P. (2005). "Lexicalization and Institutionalization". In *Handbook of Word-Formation* (Eds. Štekauer, P. and R. Lieber), Springer.
- Lipka, L. (1992). "Lexicalization and Institutionalization in English and German". In *Linguistica Pragensia*, Issue 1/ 1992, Faculty of Arts, Charles University of Prague, Prague.
- Lipka, L. (2002). *English lexicology, Lexical Structure, word semantics and word-formation*. Tuebingen: Gunter Narr.



- Lipka, L., S. Handl, W. Falkner (2004). "Lexicalization & Institutionalization: The State of the Art in 2004". In *SKASE Journal of Theoretical Linguistics*, Vol. 1/No. 1., University in Košice.
- Lipka, L. (2005). "Lexicalization and Institutionalization: revisited and extended". In *SKASE Journal of Theoretical Linguistics*, Vol. 2/No. 2, University in Košice.
- López Rúa, P. (2004). "Acronyms & Co.: A typology of typologies", In *Estudios Ingleses de la Universidad Complutense*. Vol. 12, pp. 109-129, Madrid.
- López Rúa, P. (2006). "Non-Morphological Word Formation". In *Encyclopedia of Language and Linguistics* (2nd Edition). Vol. 2., Elsevier: Oxford.
- Malenica, F., I. Fabijanić (2013). "Abbreviations in English Military Terminology". In *Brno Studies in English*. Vol. 39, No. 1., Faculty of Arts, Masaryk University, Brno.
- The Commens Dictionary of Peirce's Terms*. Accessed at <http://www.helsinki.fi/science/commens/terms> [30/03/2014].
- Semiotic and Significs: The Correspondence Between Charles S. Peirce and Victoria Lady Welby. Ed. by Charles S. Hardwick & J. Cook (1977). Bloomington: Indiana University Press. Accessed at <http://www.helsinki.fi/science/commens> [30/03/2014].

## 10 Sources

- Filipović, R. (1990). *Anglicizmi u hrvatskom ili srpskom jeziku: porijeklo – značenje – razvoj*. Školska knjiga: Zagreb.
- Anglicisms in European Languages*. Manfred Goerlach (ed.). Oxford University Press: Oxford. 2005.
- Concise Encyclopedia of Pragmatics*. Jacob L. Mey (ed.). Elsevier: Oxford. 2009.
- Hartmann, R. R. K., Gregory James (1998). *Dictionary of Lexicography*. Routledge: London
- Encyclopedia of Language and Linguistics*, 2<sup>nd</sup> edition. Keith Brown (ed.). Elsevier: Oxford. 2004.
- Roach, Peter (2009) *English Phonetics and Phonology, Glossary: A Little Encyclopaedia of Phonetics*. (<http://www.cambridge.org/elt/peterroach>)
- Kristal, Dejvid (1985). *Enciklopedijski rečnik moderne lingvistike*. Nolit: Beograd.
- The Handbook of English Linguistics*. Aarts, B., April McMahon (eds.). Blackwell Publishing: London. 2006.
- Bussman, Hadumod (1998). *Routledge Dictionary of Language and Linguistics*. Routledge: London.
- Trask, Robert Lawrence (2005). *Temeljni lingvistički pojmovi*. Školska knjiga: Zagreb.



# La definizione delle relazioni intra- e interlinguistiche nella costruzione dell'ontologia IMAGACT

Gloria Gagliardi  
Università degli Studi di Firenze  
gloria.gagliardi@unifi.it

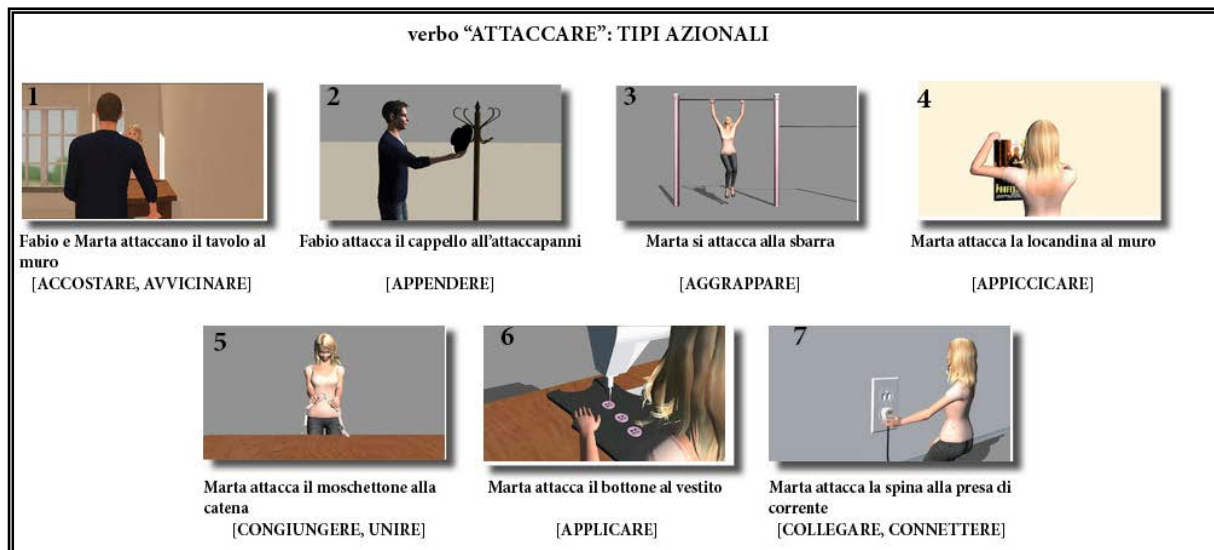
## Abstract

IMAGACT è un'ontologia interlinguistica che rende esplicito il *range* di variazione pragmatica associata ai predicati azionali a media ed alta frequenza in italiano ed inglese. Le classi di azione che rappresentano le entità di riferimento dei concetti linguistici, indotte da *corpora* di parlato da linguisti madrelingua, sono rappresentate in tale risorsa lessicale nella forma di scene prototipiche (Rosch 1978). Tale metodologia sfrutta la capacità dell'utente di trovare somiglianze tra immagini diverse indipendentemente dal linguaggio, sostituendo alla tradizionale definizione semantica, spesso sottodeterminata e linguo-specifica, il riconoscimento e l'identificazione dei tipi azionali. L'articolo illustra i criteri generali che hanno ispirato il *mapping* inter-/intra-linguistico dei dati derivati da *corpora* per la formazione dell'ontologia, le questioni di natura teorica e tecnica poste dalla costruzione della risorsa e le soluzioni adottate. Vengono descritte le tipologie e la natura delle relazioni tra le entità del database nella sua versione 1.0, e le modalità generali con cui i materiali linguistici annotati sono stati organizzati in una struttura dati coerente.

**Keywords:** ontologia; verbi di azione; relazioni interlinguistiche

## 1 Introduzione

I verbi d'azione veicolano informazioni essenziali per la corretta interpretazione delle frasi e quindi per la comprensione del linguaggio. Le relazioni che strutturano questa parte di lessico sono tuttavia molto complesse, in quanto non è possibile stabilire una corrispondenza biunivoca tra predicati ed eventi (Moneglia & Panunzi 2010). I verbi d'azione più frequenti nella comunicazione quotidiana sono infatti "generalisti", ovvero possono essere applicati in modo produttivo a classi di azioni pragmaticamente e cognitivamente diverse, come mostra la variazione pragmatica del verbo italiano.



**Fig. 1: Variazione pragmatica del lemma italiano *attaccare*.**

Tale variazione corrisponde alla competenza semantica referenziale dei parlanti, ed è quindi un dato essenziale per la modellizzazione dell'informazione lessicale. Tuttavia, essa è solo saltuariamente censita dai dizionari tradizionali e dalle più note ontologie e risorse computazionali, come ad esempio Wordnet (Fellbaum 1998) e Verbnets (Kipper-Schuler 2005).

Il progetto IMAGACT contribuisce a superare questa lacuna attraverso la realizzazione di un'ontologia interlinguistica dell'azione che esplicita lo spettro di variazione pragmatica associata ai predicati in italiano e in inglese (Moneglia *et al.* 2012). Le classi di azioni che identificano il riferimento di ogni verbo sono state individuate a partire dall'annotazione di grandi *corpora* rappresentativi dell'uso linguistico parlato spontaneo, e quindi associate attraverso una procedura di *mapping* inter-/intra-linguistico ad una serie di scene prototipiche, in grado di elicitare nell'utente la comprensione della classe di eventi rappresentata (Rosch 1978).

La metodologia di induzione delle classi di azioni e l'utilizzo di prototipi in sostituzione delle definizioni per la rappresentazione del riferimento, due tra gli aspetti più innovativi di IMAGACT, hanno però sollevato alcune questioni di strutturazione dell'informazione nella fase di formazione dell'ontologia (par. 3.1).

Verranno descritti, a partire da alcuni *case study*, i problemi e le soluzioni adottate per la costruzione della risorsa, ovvero le modalità secondo le quali il materiale annotato è stato organizzato all'interno di una struttura dati coerente, che, pur mantenendosi aderente all'intuizione dei parlanti madrelingua, risulti di facile consultazione per l'utente finale.

## 2 Induzione delle Classi Azionali da Corpus

Una strategia efficace per apprezzare l'uso dei verbi *action-oriented* è la diretta osservazione delle loro occorrenze nel parlato spontaneo, in cui il riferimento all'azione è decisivo. In IMAGACT è stata dunque adottata una procedura di tipo *bottom-up*; le classi di azioni sono state indotte da risorse linguistiche di parlato disponibili, su licenza, per scopi scientifici:

- *corpus* Inglese: una selezione del British National Corpus (BNC) di circa 2 milioni di parole;
- *corpus* Italiano: una collezione di risorse di parlato in lingua italiana (LABLITA corpus, LIP, CLIPS) per un totale di 1,6 milioni di parole (Moneglia *in press*; Gagliardi 2014).
- I materiali linguistici sono stati sottoposti ad una articolata procedura di annotazione (Moneglia *et al.* 2012; Frontini *et al.* 2012); i dati risultanti consistono di:
  - due elenchi di verbi, uno per l'italiano e uno per l'inglese;
  - per ogni verbo, una serie di "tipi" azionali, ovvero le classi di azioni fisiche tra loro tipologicamente e cognitivamente diverse che rientrano nell'estensione del predicato (fig.1);
  - per ogni tipo azionale, uno o più *Best Example*, cioè le istanze più rappresentative di tutte le strutture tematiche e delle proprietà aspettuative individuate;
  - per ogni *Best Example*, l'insieme delle occorrenze che costituiscono la variazione del verbo nel *corpus*, standardizzate da linguisti madrelingua in frasi semplici ed annotate a vari livelli.

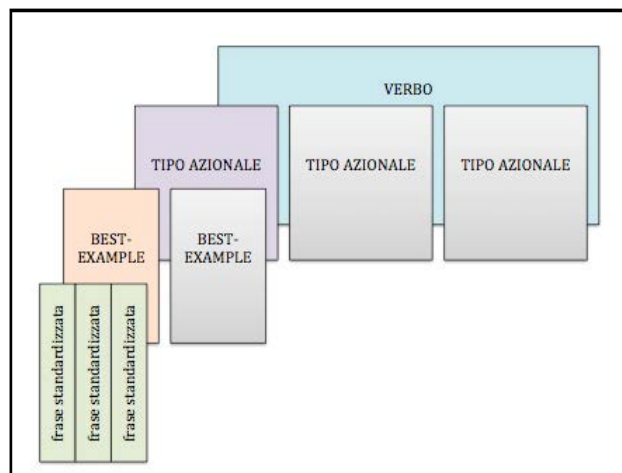


Fig. 2: Risultati della procedura di annotazione IMAGACT.

L'ontologia interlinguistica è stata costruita mediante il ricongiungimento dei tipi azionali in un'unica galleria di scene prototipali. Il requisito generale che ha ispirato la formazione della risorsa è che l'immagine standard associata a ciascun tipo sia facilmente riconoscibile e garantisca la corretta individuazione del concetto azionale, indipendentemente dalla lingua e dalla cultura di origine dell'utente.

### 3 Mapping

#### 3.1 Criteri generali

La costruzione di un'ontologia coerente dal punto di vista linguistico e formale a partire dai materiali estratti da *corpora* ha rappresentato una sfida molto impegnativa, sia a livello pratico, considerata l'enorme mole di dati da riconciliare in una struttura unitaria, che a livello teorico, data la novità della metodologia di induzione delle classi azionali.

I dati in *input* hanno influenzato fortemente la forma e la struttura del database e la concezione stessa della procedura di *mapping* inter-/intra- linguistico. La tipizzazione della variazione è infatti condizionata dalla semantica del verbo in oggetto: il senso del lemma, operando come “punto di vista” sulle categorie azionali, ha determinato la granularità dell'annotazione, anche a parità di eventi predicabili. Si considerino ad esempio le variazioni dei lemmi *attaccare* ed *appendere* riportate in fig. 3.

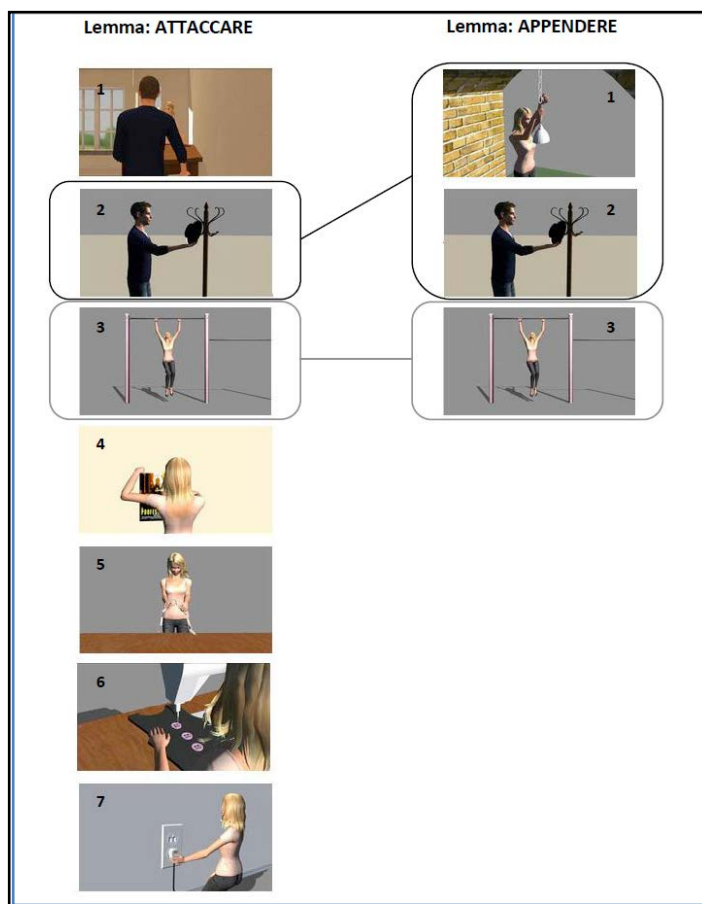


Fig. 3: Classi azionali dei lemmi attaccare e appendere a confronto.

I tipi azionali 1 e 2 del verbo *appendere* sono stati categorizzati come unico tipo (il 2) nel lemma *attaccare*: ciò significa che l'annotatore, tipizzando la variazione primaria del lemma *attaccare*, non ha ritenuto di dover distinguere gli eventi sulla base dello stato risultante del tema (il tema "pende dal riferimento" in 1, e non "pende" in 2); al contrario, il medesimo tratto è stato considerato rilevante per il lemma *appendere*. Eventi che costituiscono più classi azionali all'interno della variazione di un predicato più specifico sono stati insomma categorizzati come unica classe per predicati di maggiore generalità.

## 3.2 Ipotesi di lavoro

Per gestire casi come quello appena illustrato, nel corso della pianificazione della struttura del database sono state prese in considerazione due ipotesi di lavoro. Una prima soluzione prevede che la granularità dell'annotazione del lemma meno generale venga riprodotta nel lemma più generale. Riprendendo l'esempio in fig. 3, il tratto "sospensione" verrebbe considerato pertinente sia per il lemma *appendere* che per il lemma *attaccare*. Ciò avrebbe come conseguenza il fatto che il database ammetta un'unica tipologia di relazione, l'equivalenza. Ne risulterebbe un DB di notevole semplicità strutturale, in cui è sempre possibile stabilire relazioni 1:1 tra tipi azionali. Un *mapping* così concepito porterebbe però ad una anti-economica sovragerazione di tipi azionali per i verbi generali.

La seconda soluzione prevede che nella struttura dei dati vengano introdotte relazioni implicite di tipo IS\_A. La scelta avrebbe essenzialmente due conseguenze: il database creato conterrebbe gerarchie *implicite* di tipi, e uno stesso tipo potrebbe essere rappresentato nell'ontologia da più scene. A ciò corrisponderebbe un aumento della complessità delle relazioni nel DB; la soluzione, tuttavia, consentirebbe di mantenere contenuto il numero di tipi azionali e soprattutto di garantire l'aderenza della tipizzazione all'intuizione dei parlanti madrelingua.

In ragione della sua maggior coerenza rispetto ai requisiti progettuali, è stata scelta la seconda soluzione.

## 3.3 Relazioni del DB IMAGACT

Le entità del DB IMAGACT 1.0 sono organizzate mediante due tipologie di relazione:

- Relazione tipo-tipo;
- Relazione tipo-scena.

Nella prima categoria rientra la relazione L\_EQ, "*local equivalence*". Nel quadro teorico adottato in IMAGACT (Moneglia 1997) viene definita "equiestensionalità" o "equivalenza locale" la possibilità per due (o più) predicati di applicarsi allo stesso evento o insieme di eventi, sulla base di proprietà di senso. Tale proprietà è rappresentata *indirettamente* nel database dall'appartenenza di una stessa scena alla variazione primaria di due o più lemmi (fig. 4).

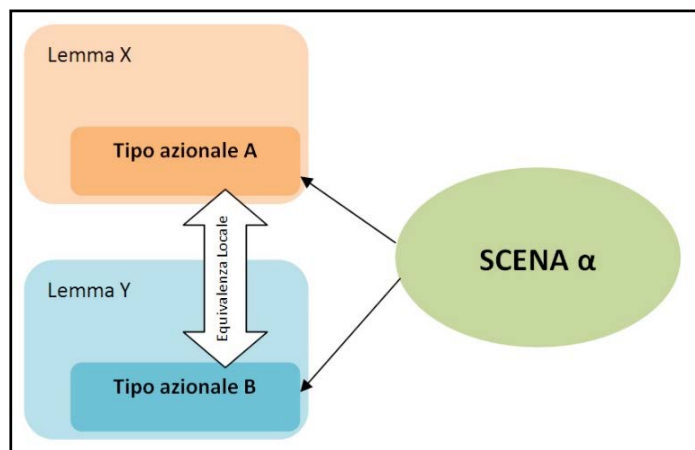


Fig. 4: Relazione di Equivalenza Locale (L\_EQ) nel DB IMAGACT.

Dati i tipi azionali a, b, c ed i lemmi X e Y, la relazione ha le seguenti caratteristiche:

- $a \in X, b \in X \Rightarrow \neg L\_EQ(a, b)$
- $a \mathcal{R} b \Rightarrow b \mathcal{R} a$  (simmetria)
- $a \mathcal{R} b, b \mathcal{R} c \Rightarrow a \mathcal{R} c$  (transitività)

Tipi azionali e scene vengono invece collegati in IMAGACT secondo due modalità (fig. 5):

- PRO, “prototipo”: la scena è un prototipo per il tipo;
- INST, “istanza”: la scena rappresenta una possibile realizzazione (non prototipica) della classe di eventi rappresentati in un tipo.

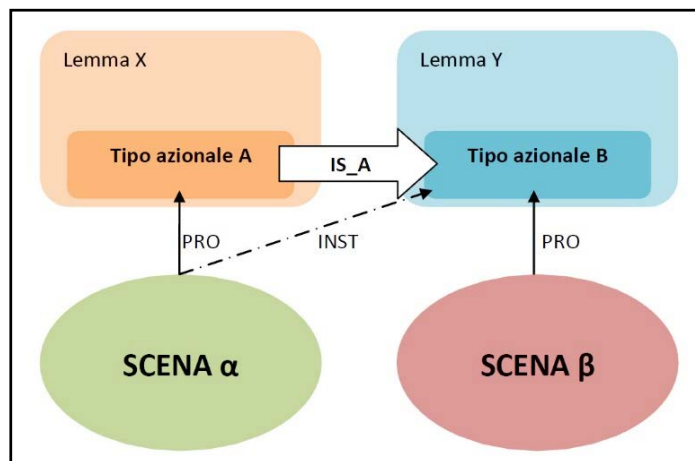


Fig. 5: Relazioni PRO e INST nel DB IMAGACT.

Ogni tipo azionale del database IMAGACT ha associata una, ed una sola, scena con relazione PRO; può invece avere, opzionalmente, più scene connesse con relazione INST.

Ciò corrisponde al fatto che un tipo di azione altamente prototipico per un verbo (es. “*attaccare/appendere* la lampada al soffitto” in relazione alla variazione primaria del lemma *appendere*), possa corrispondere ad una istanza periferica per un altro verbo (l’evento di “*attaccare/appendere* la lampada al soffitto”



è una fra le possibili istanze per il lemma *attaccare*, al pari di “*attaccare/appendere* il cappello all’attaccapanni”). Il fenomeno, che non ha natura logica, è connesso alla maggiore o minore marcatezza pragmatica dell’evento e a fattori semantici ancora da investigare.

### 3.4 Il concetto di “Famiglia di Prototipi”

La scelta di una strategia gerarchizzante ha avuto come effetto l’introduzione in IMAGACT del concetto di “famiglia di prototipi”: laddove siano presenti differenze di granularità di annotazione dovute al senso del lemma annotato e tali differenze appaiano consistenti e/o interessanti, classi di azioni distinte all’interno della variazione di un predicato specifico possono essere associate a costituire un unico *cluster* di prototipi in predicati più generali. Con la dicitura “famiglia di prototipi” si intende dunque in IMAGACT l’insieme delle scene connesse ad un predicato allo scopo di esplicitare differenziali linguistici.

In fig. 6 è mostrata la soluzione strutturale adottata per l’esempio discusso nei paragrafi precedenti. Dal punto di vista dell’architettura dell’ontologia, ciascuna scena corrisponde a un nodo della gerarchia.

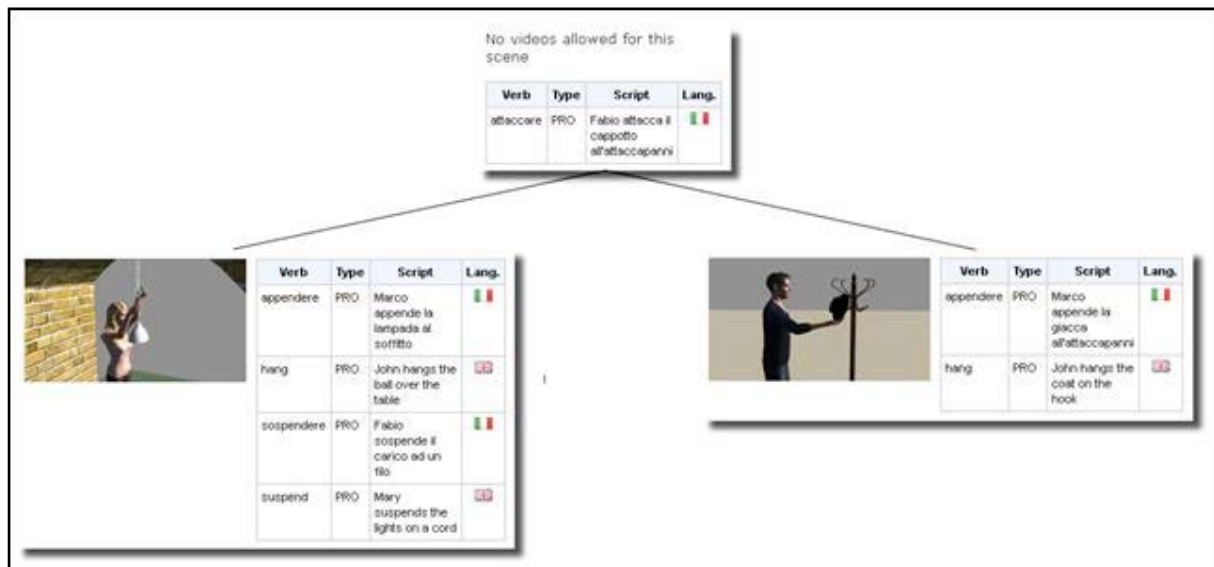


Fig. 6: Esempio di “Famiglia di Prototipi”: *attaccare*.

### 3.5 Troponimi e denominali

Una struttura dati così concepita consente anche di gestire agevolmente varie tipologie di iponimi, e la loro relazione con i tipi azionali dei verbi generali (fig. 7). Tra questi:

- troponimi, ovvero iponimi che esplicitano la modalità con cui l’azione viene compiuta dall’agente (es. *appiccicare* vs. *attaccare*);
- denominali, ovvero iponimi che esplicitano uno specifico materiale o oggetto di cui l’agente si serve per realizzare l’azione (es. *incollare, to glue, to tape* vs. *attaccare*).

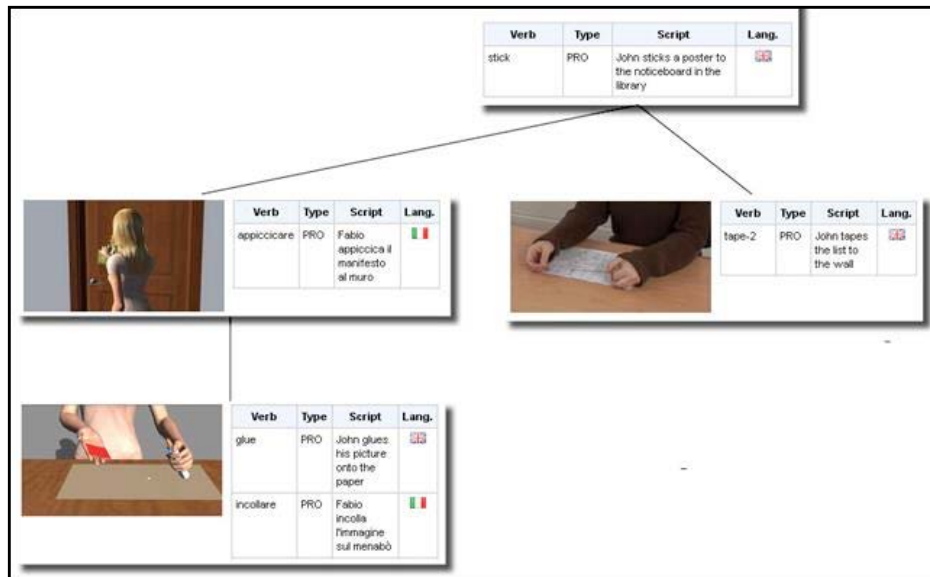


Fig. 7: Esempio di mapping di verbi denotativi e troponimi: to stick, appiccicare, to glue, incollare, to tape.

## 4 Conclusions

La versione 1.0 del database IMAGACT, rilasciata in data 1/09/2013, contiene 521 verbi ad alta e media frequenza per l'italiano e 550 per l'inglese, connessi ad una galleria di 1010 scene. La risorsa è interrogabile all'URL <http://www.imagact.it/>.

La metodologia illustrata ha permesso la generazione di tale ontologia al suo stadio attuale: le classi azionali individuate a partire dai lemmi delle due lingue sono state organizzate in una struttura dati coerente, conciliando la necessità di correttezza formale con la volontà di mantenere aderente la tipizzazione della variazione all'intuizione dei linguisti madrelingua che hanno prodotto l'annotazione. Per facilitare l'estensione della struttura dati ad altre lingue, il database è attualmente in fase di revisione e di semplificazione nell'ambito del progetto MODELACT ("From individuation to modelling in natural language ontology of action. Grounding the definition of action concepts on language infrastructures").

## 5 Bibliografia

- Fellbaum, Ch. (1998). *WordNet: an electronic lexical database*. Cambridge, MA: MIT Press.
- Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, Philadelphia, US.
- Gagliardi, G. (2014). *Validazione dell'Ontologia dell'Azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment (MCI)*. PhD thesis, Università degli Studi di Firenze, Italia.
- Frontini, F., De Felice, I., Khan, F., Russo, I., Monachini, M., Gagliardi, G. & Panunzi, A. (2012). Verb interpretation for basic action types: annotation, ontology induction and creation of prototypical scenes. In: M. Zock & R. Rapp, *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, CogALex III*, The COLING 2012 Organizing Committee, pp. 69-80.

- Moneglia, M. (1997). Prototypical vs. not-prototypical verbal predicates: ways of understanding and the semantic types of lexical meanings. In: *Vestnik Moskovskogo Universiteta (Moscow State University Bulletin)*, 2, pp.157-173.
- Moneglia, M. (*in press*). The Semantic variation of action verbs in multilingual Spontaneous speech Corpora. In: T. Raso, H. Mello (eds.), *Spoken Corpora and Linguistics Studies*, Amsterdam: Benjamin.
- Moneglia, M. & Panunzi, A. (2010). I verbi generali nei corpora di parlato. Un progetto di annotazione semantica cross-linguistica. In: I. Korzen & E. Cresti (eds.) *Language, Cognition and Identity. Extension of the Endocentric/Esocentric Typology*. Firenze: FUP, pp. 27-46.
- Moneglia, M., Monachini, M., Calbrese, O., Panunzi, A., Frontini, F., Gagliardi, G. & Russo, I. (2012). The IMAGACT cross-linguistic ontology of action. A new infrastructure for natural language disambiguation. In: N. Calzolari et al. (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation – LREC’12*, pp. 948-955.
- Rosch E. (1978). Principles of Categorization. In: Rosch E. and Lloyd B.B. (eds.), *Cognition and Categorization*. Hillsdale, NJ: Erlbaum. 27-48.

### **Acknowledgements**

Il progetto IMAGACT è stato finanziato dalla regione Toscana nell’ambito del programma PAR.FAS. (linea di azione 1.1.a.3) . Ulteriori ricerche sul database IMAGACT, incluso questo articolo, sono state realizzate grazie al contributo del progetto MODELACT (2013-2016), finanziato nell’ambito del programma nazionale Futuro in Ricerca.



# A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition

Ellert Thor Johannsson, Simonetta Battista  
A Dictionary of Old Norse Prose, University of Copenhagen  
nk950@hum.ku.dk, sb@hum.ku.dk

## Abstract

*Ordbog over det norrøne prosasprog/A Dictionary of Old Norse Prose (ONP)* is a historical dictionary project at the Department of Scandinavian Research at the University of Copenhagen and covers medieval Old Norse material found in prose texts from the oldest written documents to early modern times (Old Icelandic 1150-1540 and Old Norwegian 1150-1370). After the publication of the first four volumes of an intended thirteen volume printed edition, a decision was made to change the format of the dictionary to a digital edition available online. This new medium has had to accommodate the already published printed material as well as unprinted and not fully edited material. In this article, we discuss how the change in ONP's format has provided the user with some benefits in working with the dictionary material, but also some new challenges. We compare the features of the printed volumes to the features of the online version and in doing so address some key questions that relate to the perspective of the user and how the digital version can be further improved.

**Keywords:** historical dictionary; electronic publication; digitalization

## 1 Introduction

*Ordbog over det norrøne prosasprog/A Dictionary of Old Norse Prose (ONP)* was established in 1939 with the original intention of publishing a supplement to the already renowned Old Norse dictionaries of the 19<sup>th</sup> century, the works of Cleasby & Vigfusson (1874) and Johan Fritzner (1886, 1891, 1896). Furthermore ONP was limited to prose language, as a leading work on poetic language had then recently been published (Jónsson 1931) and therefore the poetic vocabulary was deemed sufficiently accounted for.

As the dictionary work progressed, the editorial staff soon realized that the project would fare better as an independent new dictionary based on its own principles and procedures. Plans were laid out for a new historical scholarly dictionary of Old Norse that would strive to represent the actual original medieval material by adhering to rigorous philological standard. This included retaining the spelling of scholarly text editions and using a system of references or sigla not only referring to a particular edition but also to each specific manuscript that edition is based upon.

For the first decades, the work consisted mostly of excerpting Old Norse texts and building an impressive collection of dictionary citations consisting of around 750.000 handwritten slips. When the col-

lection of citations was considered extensive enough to cover thoroughly the vocabulary from all known Old Norse Prose texts, plans were conceived to begin publication of series of printed volumes. In a publication celebrating the 25-year anniversary of the dictionary, the current chief editor stated the intentions of the dictionary staff for the publication to be complete within 25 years (Widding 1964: 21). In spite of this ambitious plan, the printed edition of the actual dictionary did not commence until 1989 with a volume of indices. This volume was the first in a series of estimated thirteen printed volumes scheduled for publication over the next decades. The printed publication continued with three more printed volumes (ONP1-3, covering the alphabet from *a-* to *em-*) at roughly five year intervals, the latest of which appeared in 2004. The remaining nine volumes would have thus likely taken around 45 years in production. Therefore, in 2005 a radical decision was made about the future of the dictionary. Instead of printed volumes, it should become a digital publication freely available on the web.

After the printed publication of the dictionary had been suspended, the work began on preparing the material for digitalization and electronic publication. This preparation work entailed various types of tasks including converting old reference sigla to the latest scholarly editions, scanning all citation slips under relevant headwords, as well as scanning scholarly editions and obtaining the necessary copyrights. This work finished in 2009 and in 2010 the first version of the digital ONP appeared online. Unlike the printed volumes, the digital edition is not a refined finished product. The decision to change the format of the dictionary also entailed making the dictionary material available online in a relatively raw form rather quickly. The first version of the digital edition consisted of two somewhat different components: on the one hand, the already published volumes in digital form and on the other hand the basic dictionary material, i.e. the collection of citations as handwritten slips, scanned under the appropriate lemma. The basic idea was that after making all the basic material accessible online the work of the editors would continue, gradually organizing the collection of citations, writing definitions and structuring the raw material into a more dictionary like format. In the long term, the online version aims to resemble the printed edition in terms of scope and attention to detail, although the nature of this new medium as well as the nature of the material has required a new approach and different editorial procedures from the ones applied to the printed publication.

This paper illustrates the difference between the printed edition and the digital edition of ONP with special focus on the user's perspective. Some of the questions that we address in our discussion are: What is required or expected of the user to be able to take advantage of the features of the dictionary? What are the pros and cons from the user's perspective of both paper and digital publication? How does the digital format offer the user different ways of accessing and working with the lexicographic material? What are the benefits and disadvantages of the current form of the dictionary? What improvements would be of further benefit to the user?

## 2 The basic characteristics of ONP

ONP has many features in common with similar dictionaries, but adds many layers to the information given. Figure 1 demonstrates what kind of information is displayed in the printed form of the dictionary and how it is structured.

dritr	261	262	drjúgmæltr
dritr <i>sb. m. (poet.)</i> <b>Gloss.:</b> <i>EJ -i-; CIV (drit); Bin -dritr; LP -i-; Fr4 -dritr; NO -i-; ÁBIM (drita); (Bl)</i>			
dritroði <i>sb. m. (poet.)</i> <b>Gloss.:</b> <i>CIVAdd -ó-; LP; AJ -ttr-; NO</i>			
<b>drit-skegg</b> <i>sb. n. [÷; -] □</i> <i>(derog.) “møgskæg” // (derog.) “muck-beard”:</i> þa mælti Merlin til spamanna konungs. Segid nv <u>dritskeggin</u> huat er unndir vattninv • <i>Bret AM 573 4<sup>o1</sup> 44v<sup>25</sup></i> (~ <i>lat. mendaces adultores HRB 383<sup>2</sup></i> )		3	<i>self-confident/self-important air, disdainfully, with a superior manner:</i> allra síðarst drattar einn blámaðr mjök <u>drjúgliga</u> svá sem stjórnari ferðarinnar <i>KlmB<sup>x</sup> 553<sup>27</sup></i> ; Þeir ríða <u>drjúgliga</u> fram um hinar þrjár fyrstu fylkingar, kveðja engan <i>KlmB<sup>x</sup> 226<sup>19</sup></i> <b>láta drjúgliga</b>
* <i>dritskeggingr</i> <i>sb. m. (Fr “Bret. 27 (12 v. l. 6)” ⇒ dritskeggingar Jón Sigurðsson 1849 [ANOH 1849] 12<sup>9</sup>, var. app.; cf. dritskeggin Bret AM 573 4<sup>o1</sup> 44v<sup>25</sup>)</i> <b>Gloss.:</b> <i>Fr; NO</i>		6	<i>optræde selvsikkert/overlegent // behave self-confidently/self-importantly:</i> <u>Drjúgliga</u> [var. digrliga <i>AM 62 fol “D<sup>1</sup>”</i> ] lætr landi en <i>ÞorvTÓT 94<sup>11</sup></i> ; Gunnarr reið við enn tólpta
<b>drit-skítr</b> <i>sb. m. (el. drit-skítr adj.) □</i> <i>(cogn.) en hann sendi apttur Andres <u>dritskit</u> [var. dritlióð <i>BöglEirsp 454<sup>26</sup></i>] • <i>Bögl81 276<sup>6</sup></i> (cf. Andres drith <i>278<sup>17</sup></i>)</i>		9	mann í Kirkjubæ ok kallaði út Otkel ... Skamkell ... mælti: Ek skal út ganga með þér ... Þykki mér þat ráð, at þú látir drjúgliga ( <i>ms. drygliga</i> ) • <i>NjR1908 111<sup>24</sup></i> → <i>ms. 26v<sup>16</sup></i> (cf. Þeim þótti undir því mest ... at þú hefðir látit sem drjúgligast, en ek gerða þik sem mestan mann af öllu <i>114<sup>20-22</sup></i> )
<i>dritskítr</i> <i>adj. → dritskítr sb. m.</i>		12	18
		15	<b>Gloss.:</b> <i>EJ (drjúgligr); CIV; Fr drjúgliga, drygliga; NO drjúgliga, drygliga</i>
			<b>drjúg·ligr, dryg·ligr</b> <i>adj. [compar. -ri, superl. -ast/-st-] □</i> 1) [til e-s] <i>tílstrækkelig, betydelig, effektiv // sufficient, substantial, effective:</i> Syn þec í orrosto diarvan oc ufælinn

Figure 1: From ONP3.

The main features are:

- two categories of lemma, normal vs. bold type; the actual entries are in bold, while words which are outside the corpus are shown in normal type. These can be poetical words (e.g. *dritr* and *dritroði* in Figure 1), foreign words which are not morphologically integrated into Old Norse, or the so-called “ghosts” (e.g. *dritskeggingr* in Figure 1). A “ghost” is a word found in other dictionaries, which is based on obsolete editions or misreading of a manuscript, and therefore has now “disappeared”;
- non-normalized orthography, i.e. the orthography of the relevant manuscript or scholarly edition is kept in the dictionary citations;
- two target languages: Danish and English;
- references to foreign parallel texts (esp. Latin);
- detailed system of sigla indicating not only reference to an edition but also the actual manuscript for each section of the text (in some cases different manuscripts are used within the same edition);
- morphological information (inflectional pattern and verb conjugation) based on texts (mainly the actual examples found in the dictionary database);
- syntactic information (especially verb complements and prepositional use);
- phrases and collocations;
- the citations are taken from actual Old Norse texts (editions or manuscripts) and are not modified in any way;

- the definitions are based on detailed analysis of the excerpted dictionary citations as well as secondary literature;
- references to glossaries and occasional references to secondary literature at the bottom of each dictionary entry.

### 3 The typical user and the prerequisites for using ONP

The user of ONP is typically a scholar of Old Norse language or literature, or medieval history and culture. This can be deduced from two “active” sources: the feedback forms on ONP’s website and specific requests addressed to the dictionary staff. Students of these fields also use ONP, especially those who are advanced enough in their studies to use original text material.

Using ONP has always required some degree of expertise. The dictionary is intended to be primarily used as an academic or scholarly research tool. The fundamental assumption is therefore that the user can read and understand the language and is familiar with non-normalized text editions, i.e. scholarly editions of Old Norse texts that seek to render the original spelling of the medieval source. Orthography in medieval texts can be highly irregular and often inconsistent, even within the same manuscript or document. This is a fact the user needs to be aware of.

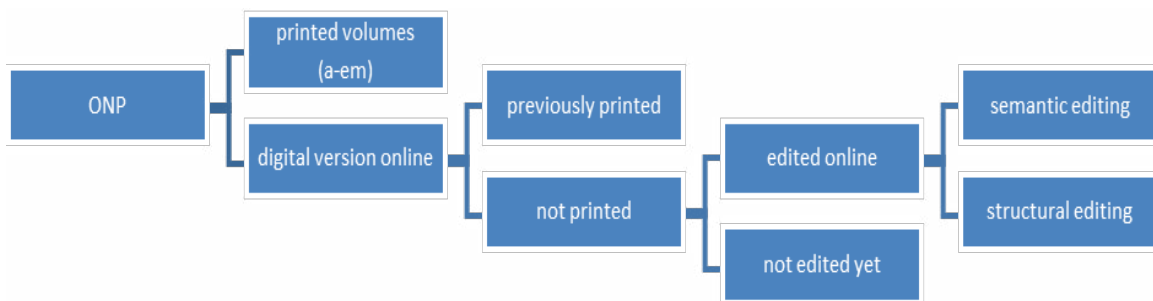
Related to this is the assumption that the user possesses knowledge to be able to “translate” or “convert” non-normalized forms into standard dictionary entry forms. This is a twofold task. First the user has to identify the normalized spelling of the word in question. Secondly the user has to know the basic form of the entry in order to look it up. For example when the user comes across forms like *diarvan* or *djorf* in a text edition he has to know that they are graphical and inflectional variations of the adjective *djarfr* ‘daring’, which is the normalized standard form of the headword (in the nominative singular masculine) to be able to look it up in the dictionary.

The user is also required to have some knowledge of grammar, or at least familiarity with the grammatical terminology and the presentation of grammatical information. The presentation of grammatical information in the dictionary entries is explained in some detail in the volume of indices (ONP:Registre 1989), which also includes a detailed list of all abbreviations ONP uses.



## 4 Comparing and contrasting the printed vs. the digital edition

The basic division of ONP into printed volumes and digital publication is fairly clear cut. As the publication of printed volumes has been suspended the digital version online consists of all the available material, including the previously printed material. This has already been mentioned above in section 1. However the online version is further divided into material that has been edited in the framework of the online version and material that has not been edited yet. Yet another level exists as the online edited material can be further divided into semantically edited material (primarily nouns) and structurally edited material (primarily verbs). Figure 2 gives an overview of this multi-level structure:



**Figure 2: The structure of ONP's published material.**

It is therefore not a straight forward task to compare the printed edition of the dictionary and the digital edition from the user's perspective. The multi-level structure of the material means that contrasting the printed edition vs. the digital edition gives a different picture depending on what exactly is being compared.

- If we compare a dictionary entry from one of the printed volumes with the same entry in the online version, we would find them almost identical, although the online version shows all available examples and not only the representative ones selected to appear in print. These “previously unpublished citations” appear outside of the structure of the dictionary article so the user needs to figure out where they belong in the article structure.
- If we compare a noun from the printed edition to a semantically similar noun found in the online version we would see some clear differences. There would be a difference between nouns that have been edited online and those, which have been not: the edited nouns have the structure of a dictionary entry with numbered definitions similar to the printed volumes. The citations are more accessible as they are typed up and correctly placed under the appropriate definition; the nouns that have not yet been edited are only available as a collection of all the citation slips under a normalized headword.
- When comparing verbs, the difference between the printed volumes and the online version is even more apparent since verbs have only been edited online for their structure, but not yet for semantic content. The user would find useful information about the verb constructions and verbal modifiers, but no definitions.

Looking at what these different types of comparisons have in common we can see that the main feature of the online version is the amount of actual examples made available. This is especially helpful when the user has good knowledge of the language and is able to himself find his way to a useful example by going through the list of citations. However there are still some structural features missing in many cases. Even though the large number of examples is beneficial to some it might be overwhelming to others, especially in those headwords where there are hundreds of examples. This can make searching for something specific very time consuming and in the absence of dictionary article structure, quite difficult.

## 5 Interacting with the dictionary material using the online version of ONP

There are several features of the online version that make accessing and working with the material easier than in the printed edition. Search in the printed edition is limited to alphabetical ordering of the lemmas. The online version offers a variety of tools to facilitate the search procedure. The main advantage is the possibility to tailor the search to the user's needs, e.g. search for all words that contain a particular derivational suffix, or compounds which have a common element, e.g. words where *bátr* 'boat' is the second part. There are other ways to make the search more precise, e.g. by limiting the results to certain parts of speech, or in case of nouns, grammatical gender.

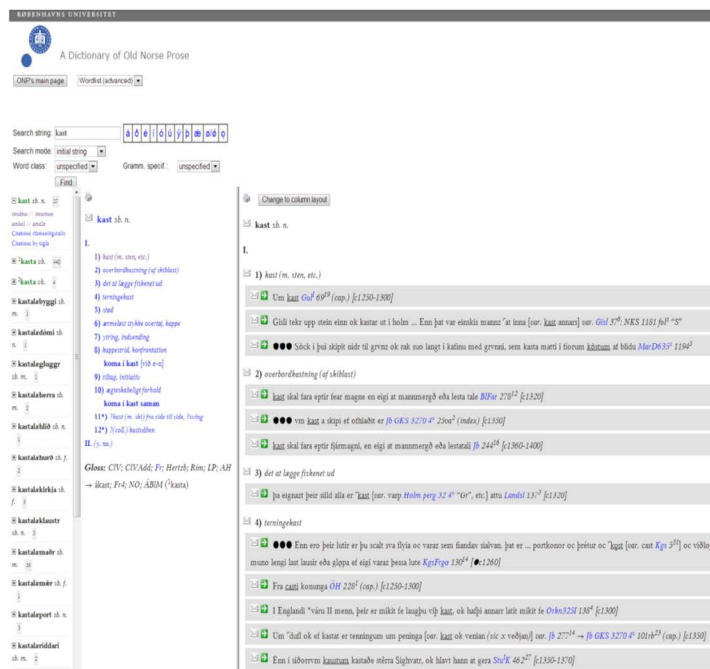


Figure 3: Example of an edited dictionary article in ONP Online.

A helpful search feature is the alphabetically ordered list of lemmas that appears on the left side of the screen when a search string is entered. This gives a quick overview of the words that have the same initial letters and might inspire the user to look up related words and compounds.

When it comes to material that has already been edited there are several additional features that facilitate working with the material. The structural overview (the middle column in Figure 2) is one such feature providing the user with a quick impression of different definitions from which to select the examples to view. This facilitates working through the list of citations, especially when headwords have a large number of citations associated with them.

The glossary section is located at the bottom of the structural overview. This list of abbreviations illustrates in which earlier dictionaries and glossaries the word is found. If the word is found under a different entry in these older works this is also indicated. Such glossary section is also a feature of the printed edition, but the online edition has the advantage of providing a hyperlink to the glossary or dictionary in question. Currently this feature is limited to the dictionary of Fritzner (1886, 1891, 1896): If the user clicks on the abbreviation *Fr.* a new tab opens displaying the relevant article.

On each level of the article an envelope icon is displayed. By clicking on this icon a new window will pop up giving the user the possibility to bring suggestions and comments to the attention of the editors. This can be done anonymously.

Once the user has chosen which citations to investigate first, she can click on a single definition and get a list of all relevant examples.

Clicking on the green arrow next to a specific citation will bring up a separate window:

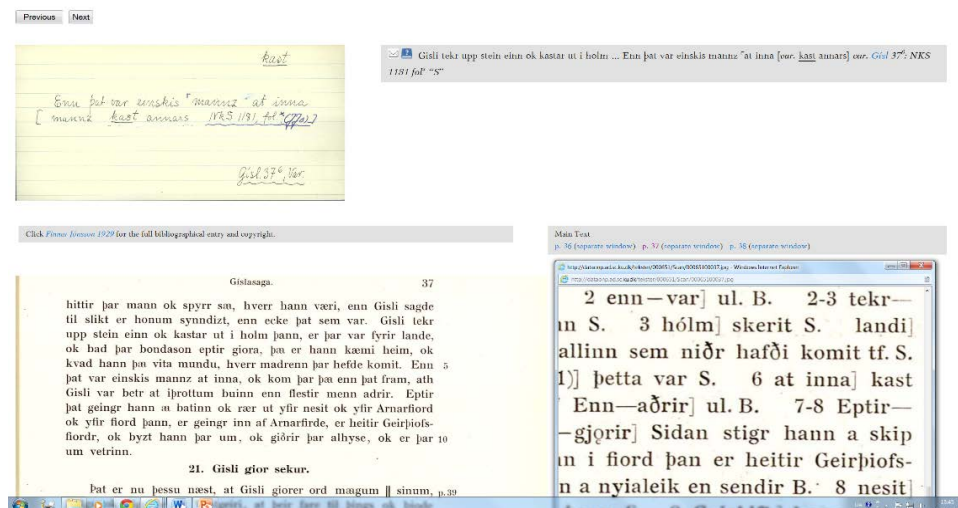


Figure 4: Closer look at a specific citation: citation slip, typed citation and scanned page from the edition (ONP Online).

Here the user will find access to a variety of additional information by further clicking on the relevant links. By clicking on the reference siglum the user has immediate access to the index entry of that particular text. This is displayed in yet a separate window. Here the user can quickly find out

more information about the text edition the dictionary refers to, i.e. what manuscript are used and where the text is found in those manuscripts (folio range). There might also be some other sources ONP has used in referring to this particular texts, such as other editions or even manuscripts that are not used by the editor in the scholarly edition the citation is taken from. This is the information that was also published in the printed volume of indices (ONP:Registre 1989). The user can seek further information about the scholarly edition by clicking on the title and on a separate page get a full entry from the ONP's bibliography. Once this information has been displayed the user has the option of browsing the bibliography, i.e. see what other scholarly editions or secondary literature this same editor might have helped create. The relative ease of access to all this secondary information can save scholarly users a lot of time as they often need the original citations to refer to in their own work.

Another feature that is of great benefit for the user is the access to the relevant scholarly edition itself, since ONP provides the possibility to view the actual printed page the citation is taken from. This is a huge improvement compared with the printed dictionary where the user had to get a hold of the actual book to get a closer look at the context of a specific citation. Almost all the works cited in the online edition are displayed in such a way. Due to copyright reasons the viewing of the editions is limited to three pages, i.e. the relevant page cited by the dictionary, the page immediately preceding it and the one immediately following it. This allows the user to gain a better grasp of the meaning of the word and further orient herself in the text the word is taken from.

## **5.1 Benefits of the digital edition**

We have already in our discussion touched upon some of the benefits and disadvantages of the digital edition of ONP from the perspective of the dictionary user. We will sum up the most important features that are of benefit to the user:

- **Multiple search capabilities:** Chance to tailor the search according to needs and interest of the user.
- **Access to all the material:** The user has access to all the raw dictionary material. All the citation slips are available as well as all edited dictionary entries. The user is able to make some use of ONP's collection of citations, even though he is not able to derive the full benefits of an edited dictionary entry.
- **Possibility of greater context:** The user has some access to the actual edition the dictionary citations are taken from and is able to browse through relevant parts of editions.
- **Interactive communication:** The online edition offers an easy way of bringing comments and suggestions to the attention of the dictionary staff.
- **In the edited part, citations are assigned to a definition:** The editor of the dictionary entry strives to assign each citation to a definition, but noting if the use of the word is ambiguous or possibly can be interpreted as belonging to a different definition from the one it is assigned to.

## 5.2 Additional benefits for the editors

In the previous discussion it has become evident that the digital edition of ONP has many benefits to the users in terms of quick and easy access to the relevant information. Some of those benefits also extend to the editors. There are mainly two that we wish to highlight. The first is that the editor receives feedback, suggestions or comments from users regarding the dictionary entries that have already been edited and made available on the web. The editor can then take the necessary measures to improve the article in question and make corrections or additions which will be made available next time the online version is updated. This is a new type of interaction between users and editors with potentially mutual benefits.

This is closely related to another benefit the digital edition gives to the editor, i.e. the possibility to correct and improve on his/her contribution to the dictionary, even after it has been published on the web. This is of course a huge change from printed publication, where a lot of the editor's work involved reading and re-reading his own work and works of others in order to make sure that no errors or misprints would find their way into the final published product. This process can be quite tedious and is not necessary to the same extent when the publishing platform is a digital version on the internet. In practical terms this means that now the editor can spend less time proof-reading and more time editing actual dictionary entries, thus increasing the productivity.

Both of these benefits, the possibility to interact with the user and the possibility to improve published dictionary entries, change the nature of the dictionary making process slightly. The process of editing and preparing the printed volumes of the dictionary was a collective effort where many editors were often involved in the production of an individual dictionary entry. The editing procedure for the digital publication is in its nature such that it relies less on the collective effort and more on the individual editor. As a result the dictionary entries are now signed with the initials of the editor in charge of each particular entry. Although the production of the dictionary is still very much a result of collaboration and teamwork, it is the individual responsibility of each editor to make sure his/her dictionary entries adhere to the rigorous standards ONP holds itself to and to respond to the user when the need to do so arises. This makes the editing work more transparent and gives the user a chance to bring suggestions directly to the person responsible for a particular headword if needed.

## 6 Room for improvement and future plans

Even though many aspects of the digital edition of ONP provide the user with increased possibilities to interact with the dictionary material and editors there is still very much room for improvement. In some respects the printed volumes of the dictionary provide the user with more consistent distribution of information, because of the discrepancies between levels of the online edition discussed in section 4 above. The main difference is between edited material and the unedited part where the infor-

mation is only accessible as a more loose structure collection of citations. However it is our hope that these discrepancies will gradually diminish as the editing work progresses and the dictionary material keeps getting updated. Currently around 140.000 citations have been semantically edited online and around 170.000 have been structurally edited. This means that around 300.000 citations are not yet edited.

## **6.1 Focus areas for improving ONP online**

One area where there is room for improvement involves the target language in the edited part of the digital edition. The current editorial procedure calls for the target language to be either English or Danish. In practice this means that in the latest dictionary entries published online, the definitions are monolingual and mostly in Danish. This limits the user base of the dictionary somewhat, although we hope to bring back the consistent bilingualism of the printed volumes in a not so distant future. Another area of focus involves reference to new editions. ONP has always tried to keep up with the latest scholarly advances in the field of Old Norse studies. If a new edition of a text appeared then all of the dictionary's citations involving that text would have to be updated to the new edition to reflect those advances. This means there is some internal inconsistencies in the printed volumes, i.e. some of the references in ONP3 did not exist when ONP1 came out. When the change to digital publication was put in place it was decided that point of reference should be the material as it was at that point in time. This means that there have been no updates to new editions after 2005. An historical scholarly dictionary like ONP needs to keep up with the latest research. Otherwise it slowly but surely will become obsolete. It is our hope that part of the continuous improvement of the digital edition will include updating the reference sigla to the latest available editions.

## **6.2 Potential future improvements of benefit to the user**

The current plan for ONP calls for the continuation of the editorial work and gradually assigning all unedited citation to structured dictionary entries available in the online version. In addition to completing the editing work, there are several other features, which will increase usability of the digital edition. Besides the ones already discussed, such as increasing the consistency between different levels of the online version as well as consistently providing the user with definition in two languages, we foresee other beneficial improvement that will become part of ONP online in the future.

The search options now are limited to the normalized lemma list. Part of the dictionary database includes a notation of by-forms and alternative forms, some of which are quite common although not searchable through the regular search options. A first step in increasing the search possibilities would be to link such non-normalized by-forms and alternative forms to their respective headwords and allow them to be part of the searchable material. The search options can be improved further. It is possible with relatively little effort to add a feature that searches not only in the list of lemmas, but also all citations that have been typed in. Of course, this feature would be rather primitive to begin with as

the citation texts are not lemmatized and not even normalized, but it definitely would be helpful to the user in certain cases.

Another feature that might in the long term become part of the online edition is a search feature, which would limit the search to certain periods, places (scriptoria), geographical areas (e.g. West Iceland) or even certain manuscripts (e.g. all texts found in a particular manuscript). The structure of the dictionary database allows for many kinds of searches, which are already available for the staff. These features are though still in a relatively early stage and have not been laid out in practical terms just yet but they would give the user new alternative ways to interact with the dictionary material.

## 7 References

- Cleasby, Richard & Gudbrand Vigfusson (1874). *An Icelandic-English Dictionary*, Oxford: Clarendon Press.
- Fritzner, Johan (1886, 1891, 1896). *Ordbog over Det gamle norske Sprog* 1-3, rev. ed., Kristiania: Den norske forlagsforening.
- Jónsson, Finnur (1931). *Lexicon poeticum antiquæ linguæ Septentrionalis / Ordbog over det norsk-islandske skjaldesprog*. Original edition by Sveinbjörn Egilsson. 2nd ed. Copenhagen: Kongelige nordiske oldskriftselskab.
- ONP Online. Accessed at: [www.onp.ku.dk](http://www.onp.ku.dk) [07/04/2014].
- ONP:Registre = Ordbog over det norrøne prosasprog : Registre / A Dictionary of Old Norse Prose: Indices. (1989). Copenhagen: The Arnamagnæan Commission.
- ONP1 = Helle Degnbol, Bent Chr. Jacobsen, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir. (1995). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose 1: a-bam*. Copenhagen: The Arnamagnæan Commission.
- ONP2 = James E. Knirk, Helle Degnbol, Bent Chr. Jacobsen, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir. (2000). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose 2: ban-da*. Copenhagen: The Arnamagnæan Commission.
- ONP3 = Helle Degnbol, Bent Chr. Jacobsen, James E. Knirk, Eva Rode, Christopher Sanders & Þorbjörg Helgadóttir. (2004). *Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose 3: de-em*. Copenhagen: The Arnamagnæan Commission.
- Widding, Ole (1964). *Den Arnamagnæanske Kommissions Ordbog, 1939-1964: Rapport og plan*, Copenhagen: G.E.C.GADS Forlag.





# Making a Learner's Dictionary of Academic English

Diana Lea  
Oxford University Press  
diana.lea@oup.com

## Abstract

This paper gives an account of the development of the *Oxford Learner's Dictionary of Academic English*, a dictionary for non-native-English-speaking students who are studying academic subjects at tertiary level through the medium of English. First, a corpus of academic English was created, using high-quality texts from a broad range of disciplines, maintaining a balance between textbooks, containing typical student reading material, and journal articles, modelling expert academic writing. Drawing on the corpus, and on previous research in the field, a core headword list of “general academic vocabulary” was drawn up. This list was continually supplemented with necessary defining words, complements, collocates, synonyms and antonyms of these words as the work of compiling the dictionary entries progressed. In compiling the entries, particular challenges were encountered in reconciling the academic and pedagogic requirements of the dictionary. This can be seen, for example, in decisions about sense division and the wording of definitions and also in the selection of example sentences from the corpus. Editors had to find ways to represent academic language faithfully whilst making it accessible to learners. The result is a genuinely academic learner's dictionary that should offer real help to learners with their academic writing.

**Keywords:** learner's dictionary; academic English; EAP; corpus

## 1 Introduction

Academic vocabulary has received considerable research attention, in particular with the effort to identify a core academic vocabulary, as distinct from general English vocabulary on the one hand and discipline-specific technical vocabulary on the other. Coxhead (2000) proposed the Academic Word List (AWL), a list of 570 word families, divided into ten sublists, found to account for around 10% of the words in a corpus of academic English, as opposed to 1.4% of the words in a fiction corpus. The AWL was generally well received by teachers and has been quite widely exploited in published materials (Coxhead 2011). More recently, however, Paquot (2010) and Gardner and Davies (2013) have proposed alternative lists, addressing some of the perceived shortcomings of the AWL, notably its exclusion of the 2,000 word families of the General Service List (West 1953) as already “known” to students at this level, and its construction around whole word families, regardless of discrepancies in frequency (Paquot 2010: 17) and even core meaning (Gardner and Davies 2013: 3) between different word family

members. Hyland and Tse (2007), however, have questioned the whole validity of a single, cross-disciplinary, core academic vocabulary, partly on the basis that the same words may be used in widely different ways in different disciplines.

In contrast, there has been much less attention paid to the idea of a dictionary of academic English, as opposed to a word list. Kosem (2008) surveyed a number of dictionaries marketed for university students, mostly aimed at native speakers, and concluded that, apart from supplementary material on academic writing, these dictionaries differ little in content from the general-purpose dictionaries on which they are based. Learners' dictionaries, such as the *Longman Dictionary of Contemporary English* and the *Oxford Advanced Learner's Dictionary* have started to acknowledge the interest in academic vocabulary and writing, by marking words in the AWL and including their own academic writing supplements, but they remain essentially dictionaries of general English. In this paper, I shall give an account of the *Oxford Learner's Dictionary of Academic English (OLDAE)* (2014), which I believe to be the first widely available, genuinely academic, learner's dictionary.<sup>1</sup>

I shall begin by outlining the principles and parameters that were established at the start of the project. I shall then describe some of the challenges that were encountered in the course of the project, notably building the academic corpus on which the dictionary is based; determining the headword list; and above all, reconciling the academic and pedagogic requirements, especially with regard to writing the definitions and selecting the example sentences. I shall then conclude by evaluating the achievements and limitations of the dictionary and suggesting some possible future developments.

## 2 Principles and Parameters

The principles and parameters of *OLDAE*, as a learner's dictionary of academic English, will be found to differ quite widely from those proposed by Kosem (2010). Although I agree with the general thesis that no one is a native speaker of academic English and that both native and non-native-English-speaking students should be viewed as "apprentice writers" when it comes to academic writing (Kosem 2010: 49), I believe the particular needs of these two groups are sufficiently different that a single dictionary cannot serve both equally well. *OLDAE* is therefore designed to serve the needs of non-native-English-speaking students of English for Academic Purposes (EAP), at a range of levels from the B1 student on a foundation course, to students at C2 level writing their Masters' dissertations. The dictionary is also, however, partly for practical reasons, much smaller in scale than that proposed by Kosem. It largely excludes general and technical English, focusing essentially on a core academic vocabulary across the disciplines, but taking a broad view of what this might encompass,

---

1 I am aware of the Louvain EAP Dictionary (Granger and Paquot 2010), an innovative online dictionary-cum-writing-aid, which can claim to be the very first learner's dictionary to be based on analysis of academic corpora. However, it is currently only available to staff and students at the Université catholique de Louvain; moreover, with only around 900 headwords, it may be arguably more a writing-aid-with-dictionary-entries than a complete dictionary in itself.

not attempting to identify a single definitive list for all students, but exploring the nuances of usage of so-called “general academic” words across different disciplines. It is specifically intended to help EAP students with their academic writing across a range of genres. The most fundamental principle underpinning the dictionary was that it should be based on a thorough analysis of genuine academic writing; this meant constructing a new corpus, the 85-million-word Oxford Corpus of Academic English (OCAE).

### 3 The Corpus

Output from a corpus can only ever be as representative and appropriate as the corpus itself; the content of OCAE therefore needed to match, as nearly as possible, the materials that target users of the dictionary would be reading and writing themselves. In terms of “reading” content, this was provided by higher education textbooks, mostly aimed at undergraduate level, which at 42 million words constituted just under half the corpus. “Writing” content was more challenging: ideally, what was needed was some 40-50 million words of very high-quality student essays and dissertations, but creating such a corpus was beyond our resources. Instead, we substituted expert academic writing from journals, monographs and handbooks, adding up to a further 43 million words; this assured the quality, and meant the findings on the usage of academic vocabulary would be sound. However, it did mean text of a much higher level than our students would be attempting to write, which made the selection of authentic but user-friendly example sentences rather more challenging – but I shall return to this point later. In terms of the balance of disciplines in the corpus, we tried to match this approximately to the profile of disciplines being studied by international students at English-medium universities. Figure 1 shows the breakdown of the corpus into different subject areas.

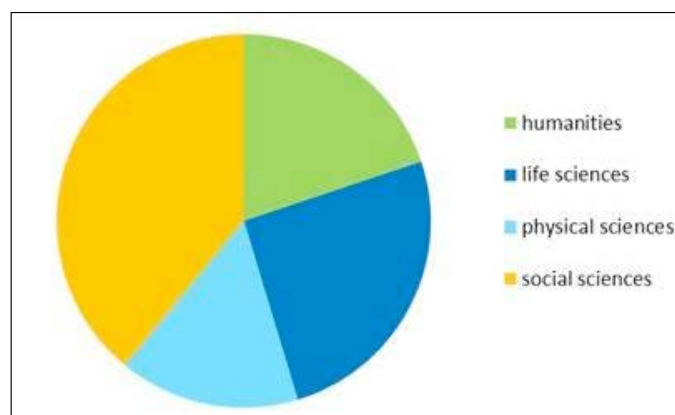


Figure 1: Breakdown of texts by subject area in the Oxford Corpus of Academic English.

Natural sciences (divided into life sciences and physical sciences) and social sciences each account for around 40% of the corpus, with the remaining 20% made up of humanities texts. The largest single disciplines were business and medicine, at around 8% each.

### The Headword List

The headword list was built up organically as the work of compiling the dictionary progressed. We began with a core list, comprised of words from the AWL, after checking them against the corpus. We added to this four word lists of our own, extracted from the four subcorpora of OCAE, as compared with a fiction corpus. As work began on compiling entries for these words, the headword list was rapidly augmented with necessary defining words, complements, collocations, synonyms and opposites of the words in the initial list. Collocations were an especially rich source of additional headwords. If dictionary users were to be enabled to use the core words productively, the linguistic contexts in which they could be used would be all-important. This meant generous treatment of collocations in the dictionary entries: for nearly 700 of the most important, collocationally prolific words, a separate section of the entry lists collocations in the style of a collocations dictionary (see Figure 2). These collocation entries then fed back into the main headword list of the dictionary as it was obviously important that no word should be listed as a collocation without being defined and exemplified in its own separate entry. In this way, the headword list expanded beyond a core of 3-4,000 academic words, to encompass higher-level words that are more context-specific, though still mostly at the subtechnical level, as well as presenting in appropriate academic contexts the functional words that are actually basic to all forms of discourse. The process of enhancing the headword list continued throughout the compiling and editing process and was completed by a trawl through the corpus for items of a certain level of frequency that had not been picked up and that might also warrant inclusion. It would have gone way beyond the scope of the project to include every word that a particular student might wish to look up; the aim was to give thorough coverage to the core words that all students would need, plus a generous helping of supplementary words that would be useful to many.

• **ADJECTIVE + FACTOR** important, major • key, critical, crucial • the main • significant • different, other • various • multiple • individual • external • additional • contributing, contributory • limiting • relevant • common • potential *This research demonstrated that for most working Americans the economy was still the most significant factor in their lives.* | contextual • structural • causal • situational • social • cultural • institutional • economic • socio-economic • demographic • environmental • psychological • genetic *There are deeper cultural and social factors influencing the adoption process.*

• **NOUN + FACTOR** risk • lifestyle *There are multiple genetic and lifestyle factors that contribute to the development of the disease.*

• **VERB + FACTOR** identify, determine • list • consider, assess, discuss • examine, investigate, explore, analyse • understand • include • highlight • address • ignore *The analysis was conducted to identify the factors associated with high job strain. ◊ Schmidlin et al. (1998) have examined risk factors for death during tornadoes.*

• **FACTOR + VERB** influence, affect, contribute to, play a role (in sth), impact (on sth) • drive • determine, shape • control • limit • account for, explain • be associated with • be involved • interact *Several other factors played a role in the decision-making. ◊ Collectively, these factors account for the negative effect of visual impairment on quality of life.*

**Figure 2: Collocations of *factor* from the *Oxford Learner's Dictionary of Academic English* (2014).**

## 5 Editorial Policy

There is insufficient space here to give a full account of the editorial policy of the dictionary. However, many of the challenges can be considered as different manifestations of the one central challenge: how to reconcile the academic and pedagogic requirements of the dictionary. We wished to make this both an academic dictionary and a learner's dictionary, but there were cases where a compromise was called for. I shall focus on two key aspects here: the definitions and the example sentences.

### 5.1 Definitions

The starting point for most of the definitions in the dictionary was the definitions in the 8<sup>th</sup> edition of the *Oxford Advanced Learner's Dictionary (OALD)* (2010). These definitions have the advantage of accessibility for learners, especially as they are written within a carefully controlled defining vocabulary of 3,000 words (in fact reduced to 2,300 for *OLDAE*). In many cases, *OALD* definitions were retained unchanged. However, there were important decisions to be made, often not only over the wording of definitions but over the content. One example is that of *variable* as both noun and adjective. The *OALD* entry (Figure 3) distinguishes two separate senses for the adjective and just one, coverall sense for the noun. The derivative *variably* is nested in the adjective, undefined.

*OLDAE*, however, recognizes that, for academic purposes, *variable* is a very important word, and EAP students need to know a lot more about it. The noun entry (Figure 4) separates out two further, much more specific

meanings that are important in an academic context: one that is relevant to all experimental sciences, and a basic meaning in mathematics that students from a wide range of disciplines will need to know. As well as conveying much more information than the *OALD* entry, this splitting of senses enables the definitions to be much more precisely worded: students have more to read in this entry, but the burden of interpreting and applying what they have read is actually lighter, because these definitions spell things out much more clearly. The adjective entry (Figure 5) distinguishes four separate meanings, each with different synonyms and antonyms, whilst *variably* (Figure 6) is recognized as an important academic word in its own right, not just a derivative of *variable*, and is given its own entry with two distinct meanings.

Figure 3: Entry for *variable* from the *Oxford Advanced Learner's Dictionary, 8<sup>th</sup> edition (2010)*.

**variable** **AW** /'veəriəbl; NAmE 'ver-; 'væ- / *adj., noun*  
**■ adj.** **1** often changing; likely to change **SYN** fluctuating: variable temperatures ◊ The acting is of variable quality (= some of it is good and some of it is bad). **➔** compare INVARIABLE **2** able to be changed: The drill has variable speed control. ◊ variable lighting **▶ vari-ably** **AW** /-iəbli/ *adv.*  
**■ noun** a situation, number or quantity that can vary or be varied: With so many variables, it is difficult to calculate the cost. ◊ The temperature remained constant while pressure was a variable in the experiment. **OPP** constant

Figure 4: Entry for *variable, noun* from the *Oxford Learner's Dictionary of Academic English (2014)*.

**variable<sup>1</sup>** **AWL** /'veəriəbl; NAmE 'veriəbl; 'vəriəbl/ *noun*  
**1** an element or a feature that is likely to vary or change: It is virtually impossible for any one model to take into account all of the many variables involved. **2** a property that is measured or observed in an experiment or a study; a property that is adjusted in an experiment: The key variables in this study are weight, cholesterol measurements and height. ◊ The following basic demographic variables were included in the model: gender, age and occupation. ◊ ~ of sth Age is an important explanatory variable of diverse consumption patterns and is expected to be a strong predictor of ICT ownership and use. **OPP** CONSTANT<sup>2</sup> (2)  
**➔** see also CATEGORICAL VARIABLE, CONTINUOUS VARIABLE, CONTROL VARIABLE, DEPENDENT VARIABLE, DUMMY VARIABLE, INDEPENDENT VARIABLE, LATENT VARIABLE, OUTCOME VARIABLE, PREDICTOR VARIABLE, RANDOM VARIABLE **3** (mathematics) a quantity in a calculation that can take any of a set of different NUMERICAL values, represented by a symbol such as *x*: The formulae show how the values of the variables *x* and *y* are calculated.

Figure 5: From the entry for *variable, adjective* from the *Oxford Learner's Dictionary of Academic English (2014)*.

**variable<sup>2</sup>** **AWL** /'veəriəbl; NAmE 'veriəbl; 'vəriəbl/ *adj.*  
**1** often changing; likely to change **SYN** FLUCTUATING: Variable costs vary according to the number of units of goods made or services sold. ◊ While rainfall is highly variable, it is generally distributed across two rainy seasons. ◊ ~ over sth In addition, soil moisture is influenced by precipitation, which is variable over time and space. **OPP** CONSTANT<sup>1</sup> (2) **2** not the same in all parts or cases; not having a fixed pattern **SYN** DIVERSE: Studies demonstrate variable rates of nitrogen fixation in microbial communities. ◊ ~ in sth Eggs are large and variable in size, depending on the species. ◊ ~ between/across/among sth Preferred habitats are variable between members of the family and range from temporary ponds to large rivers to swamps. **HELP** When **variable** is used to describe the quality of sth, the tone is slightly disapproving, meaning that some parts of it are good and some are bad **SYN** INCONSISTENT (3), MIXED (1): The quality of the pictures is variable, and some images might better have been omitted. **OPP** CONSISTENT (3), UNIFORM<sup>1</sup> **3** that can be changed to meet different needs or suit different conditions: A variable timer allows for close control of the final concrete temperature. ◊ Variable pay is associated with one economic outcome, change in productivity. **OPP** FIXED (1)  
**➔** compare INVARIABLE **4** (mathematics) (of a quantity) that can take any of a set of different NUMERICAL values, represented by a symbol such as *x*: In this paper, we study linear fractional differential equations with variable coefficients.



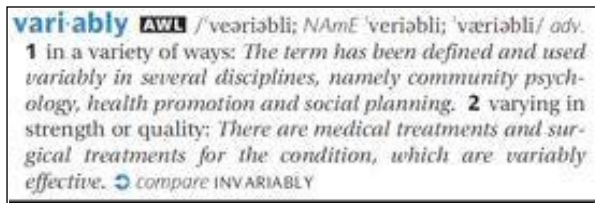


Figure 6: Entry for *variably* from the *Oxford Learner's Dictionary of Academic English* (2014).

A slightly different challenge is posed by a word like *recession* (Figure 7):

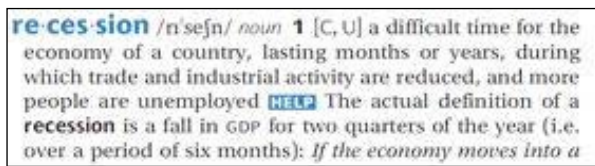


Figure 7: From the entry for *recession* from the *Oxford Learner's Dictionary of Academic English* (2014).

The main definition that comes first is closely based on the definition of this word offered in the *OALD*. However, as our economics adviser pointed out, it is not, strictly speaking, a definition at all, but a description. On closer inspection, this will be found to be true of many “definitions” offered in general learners’ dictionaries, and the dictionaries in general are all the better for it. They offer learners the degree of understanding they need in a form that is accessible to them. For the EAP student, however, the case is different. The student of economics (or history or geography or a number of related subjects) is not well served by a mere description of a recession, when it is in fact a very precisely defined economic term. Our solution was to offer the description first, followed by the “actual definition”, clearly signalled as such. (The economics adviser, it must be confessed, was not happy with this solution, and felt that only the exact definition should be offered; however, this was a view offered from an entirely academic perspective, with no concession to the particular needs of foreign learners, and so the editor’s view – that both general description and precise definition should be offered – prevailed.)

## 5.2 Example sentences

Selecting the example sentences was probably the most challenging aspect of compiling most entries and policy on this evolved over the course of the project, in some cases necessitating late revision of earlier compiled entries. Consultation with academics and EAP tutors at the planning stage impressed on the editors the need for extreme caution when lifting and editing examples from the corpus. Some were uncomfortable with the idea of editing corpus text at all. However, when faced with the reality of raw corpus text, set against the practical needs of the intended users of the dictionary, it became clear that many of the selected corpus examples would need some degree of editing to render them useful and appropriate for learners. Potential difficulties with unedited corpus text were numerous: very high-level vocabulary; difficult constructions; extremely long sentences; obscure and dist-

racting detail; general oddness. Editors also had to take into account the fact that the academic genres in the corpus – textbooks and journal articles – were not the genres that students themselves would be writing. Textbook examples were often tempting, as they were clear and accessible, but many textbooks employ a tone of “expert speaking to student” that would not be appropriate in a student essay.<sup>2</sup>

We initially approached the task of selecting example sentences from an academic corpus with the feeling that it was in some way a different task from selecting examples for a general learner’s dictionary from a general English corpus. Experience persuaded me, however, that this task was not in fact different in kind, though perhaps it was different in level of difficulty. The most useful examples are the most typical, which often means the most general:

Taylor makes the following argument:...

This approach yields dramatically lower estimates.

Several other factors played a role in the decision-making.

The most persuasive argument against this idea comes from Foster (2009).

Examples like this may not be taken directly from any one text; often they are a distillation of a number of different concordance lines, all of them very similar. Other examples – the majority – do contain context derived from a particular source text; and, where appropriate, may be taken from that text unedited. This helps them to feel more authentic; nonetheless, it is important that the context does not get in the way of understanding the linguistic point being presented in the example. The examples are intended to “feel authentic” but they cannot actually be authentic – even if completely unedited, they are inauthentic the moment they are taken from their context and set in italic type in a learner’s dictionary. Ultimately, though, the needs of the learner trump other considerations. Learners using this dictionary are not expected to immediately start writing fluent expert academic texts; what they need to acquire is a style that approaches more closely an appropriate academic style, whilst still being accessible from the level they are currently at.

## 6 Conclusion

The learner’s dictionary is not an academic genre but a pedagogic one. A learner’s dictionary of academic English needs to pay close attention to the rules and conventions of academic writing, and represent them as faithfully as it can, but the learner’s needs still take precedence. *OLDAE* is designed

---

2 The use of “we” is a case in point. Textbook writers use it frequently with the meaning of “you and I”, to include the (student) reader in a comment such as, “In this chapter, we shall see ...” This style is not employed in research articles, where experts are writing for other experts, and it is not to be recommended for students writing for a tutor or examiner. A usage note at the entry for *we* in the dictionary explains this point.



to meet the specific needs of tertiary level students writing assignments in English in a wide range of disciplines. It covers a generous “core” academic vocabulary, showing not only the meanings of words, but how to use them in context. We hope it will be a valuable new resource for students. Its limitations are largely those imposed by the relatively limited size and scope of the project. Coverage does not go much beyond the “subtechnical” level of vocabulary, but it is assumed that the technical vocabulary of the student’s own discipline will be explained as part of their subject course. The traditional format of print dictionary plus CD-ROM may also limit its appeal for some of today’s students. However, there is a lot of scope for presenting, combining and expanding the content in different ways to make it even more useful and accessible to a wider range of users. For example, a customizable online subscription model could allow users to combine *OLDAE* content with both more general content and more technical content from other dictionaries, according to the subject they are studying. To make this a reality would require rather more work, both editorial and technical, but it seems worth aspiring to.

## 7 References

- Coxhead, A. (2000). A New Academic Word List. In *TESOL Quarterly*. 34(2), pp. 213-238.
- Coxhead, A. (2011). The Academic Word List 10 years on: research and teaching implications. In *TESOL Quarterly*, 45(2), pp. 355-362.
- Gardner, D., Davies, M. (2013). A New Academic Vocabulary List. In *Applied Linguistics*. First published online August 2, 2013. Accessed at: [apli.oxfordjournals.org](http://apli.oxfordjournals.org) [25/03/14].
- Granger, S., Paquot, M. (2010). The Louvain EAP Dictionary (LEAD). In *Proceedings of the XIV EURALEX International Congress, 6-10 July 2010*. Leeuwarden, The Netherlands, pp. 321-326.
- Hyland, K., Tse, P. (2007). Is There an “Academic Vocabulary”? In *TESOL Quarterly*. 41(2), pp. 235-253.
- Kosem, I. (2008). Dictionaries for University Students: A Real Deal or Merely a Marketing Ploy? In *Proceedings of the XIII EURALEX International Congress, 15-19 July 2008*. Barcelona, Spain, pp. 1575-1584.
- Kosem, I. (2010). Designing a model for a corpus-driven dictionary of Academic English. PhD thesis. Aston University, Birmingham, UK.
- Longman Dictionary of Contemporary English*. 5<sup>th</sup> edition (2009). Harlow: Pearson Education Limited.
- Oxford Advanced Learner’s Dictionary*. 8<sup>th</sup> edition (2010). Oxford: Oxford University Press
- Oxford Learner’s Dictionary of Academic English* (2014). Oxford: Oxford University Press
- Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London, New York: Continuum International Publishing Group, pp. 11-17, 55-61.
- West, M. (1953) *A General Service List of English Words*. London: Longman, Green and Co.



# The Danish Thesaurus: Problems and Perspectives

Sanni Nimb, Lars Trap-Jensen, Henrik Lorentzen  
Society for Danish Language and Literature  
sn@dsl.dk, ltj@dsl.dk, hl@dsl.dk

## Abstract

In this paper, we present a new thesaurus for Danish and discuss some of the problems and decisions that the compilers have been faced with. The thesaurus is compared to other thesauri: Roget, Dornseiff and particularly to its smaller Danish predecessor Andersen. The different steps in the compilation process are outlined, with special attention being devoted to word ordering at the lowest level (alphabetical vs. semantic) and to the use of stylistic labels. In a comprehensive thesaurus, the conclusion is that semantic ordering is more useful for the user and that stylistic labels are necessary.

**Keywords:** thesaurus; onomasiological dictionary; semantic ordering; stylistic labels

During the last four years, a new Danish Thesaurus (*Den Danske Begrebsordbog*, DDB) has been compiled at the Society for Danish Language and Literature (DSL). Funded by the Carlsberg Foundation, the thesaurus is the first of its kind in 70 years and will at first be published as a printed book. The future plans are to publish it in an electronic version online at DSL's dictionary site *ordnet.dk* where it will be integrated with other dictionary resources at DSL. As it is completely based on, and systematically linked to, the approx. 100,000 lemmas and 135,000 word senses of the corpus-based The Danish Dictionary (DDO), we will have a thesaurus of modern Danish which contains all types of words, not least a wealth of compounds which is a common word type in a Germanic language like Danish. All types of fixed phrases and a number of frequent collocations from the DDO are given in the thesaurus. Well represented in the DDO, interjections and interjectional phrases are listed in the thesaurus together with semantically related verbs. Taxonomic vocabulary such as words for plants, animals, food, diseases, technical devices etc., is classified from a layman's point of view in common language, and all domains are thoroughly represented, including taboo areas. As an important side effect, the approach of common sense id numbers in the DDO and the DDB allows us to enrich a large number of the DDO lemmas and senses with supplementary information about synonyms and semantically related words in a future updated version of the DDO simply by making use of the thesaurus data. Experiments have likewise been carried out on the reuse of thesaurus data in the Danish WordNet, also based on the DDO and using the same sense id numbers (Nimb & Pedersen 2012; Nimb et al. 2013).

Unlike the comprehensive approach that we have adopted, some thesauri, for example the Norwegian Rosbach (2001) and our predecessor for Danish, *Dansk Begrebsordbog* (Andersen 1945, DB), focus instead on the core concepts of the language. We will discuss and compare these two approaches, focusing specifically on the difference between the DDB and the DB. Afterwards we will argue that a comprehensive approach has as consequence 1) that a semantic word ordering is necessary, and 2) that styli-

stic labels are necessary. But first a brief account of the different stages involved in the compilation of the DDB.

## 1 The Thesaurus-making Process

The basis of the compiling process of the DDB is an XML document representing all senses and their corresponding lemma forms in the DDO, supplied with all relevant semantic information from the dictionary: definition, domain, synonyms etc. On the basis of the sense unit document, the stock of words for the thesaurus is retrieved. As a case in point, consider the compilation of the section ‘Bicycling’. First, the lexicographer retrieves all sense units where the string *cykel* is part of the definition and all units of which the corresponding lemma string begins or ends in *cykel*, picks out the relevant sense units and inserts them into the thesaurus document. Synonyms and near synonyms from the DDO as well as words discovered by introspection are added and linked to the DDO if not already part of its stock of lemmata. The selection of words are grouped into semantic categories, such as *persons* riding a bike (cyclist, rider etc.), concrete *objects* (bikes, saddles etc.), *events* (to cycle, to ride a bike etc.) or *properties* of the persons, objects etc., headed by the best representative of the group (annotated in the structure). They may also be put into groups where only a thematic relation holds between the words; this option is mainly used for cases that are difficult to categorize: single words with few near synonyms but nevertheless belonging to the thematic section. In this way, we make sure that not only the prototypical words and senses are covered but also the grey or peripheral areas of the vocabulary or the radial members of the categories established (cf. Dirven & Verspoor 2004: 33).

All the groups are tagged with formalized information about their type of category, and within each semantic group, the words are presented in a logical order according to semantic criteria (see below). In the digital document, semantics is the sole basis for classification. In a later phase, data is converted in order to present the material in four main divisions according to word class for the printed thesaurus (see below). But initially, verbs and their verbal nouns belong to one group tagged with the type ‘act’ and sometimes also followed by interjections related to the act. Words designating properties are grouped together, no matter if the property is expressed in the form of an adjective (*happy*), a noun (*happiness*), a verbal expression (*to tread/walk on air*) or an adverbial (*on cloud nine*), here exemplified by English words and expressions.

If we again turn to the theme of bicycles, the section ‘Bicycling’ is one out of 29 sections in the chapter ‘Sports and leisure’, which in turn is one of 22 chapters that make up the full thesaurus. In total there are 888 sections in the thesaurus – some chapters have only 20 sections while others have more than 60. The chapters and sections were based on the division found in Dornseiff’s *Der deutsche Wortschatz nach Sachgruppen* (2004), but have been thoroughly adjusted in all the cases necessary. New sections were added for mental diseases, for contraception and abortion, for medicine, labour market and unemployment, to mention but a few. Other sections were removed, such as the German sections *Botē*

(‘messenger’, instead included in the section on communication), and *Autoindustri* (‘motor industry’), since they were either judged less important categories in modern Danish conceptualization or could easily be covered by other sections.

For various – not necessarily technical or logical – reasons, the grant was allocated for a printed dictionary. That is the reason why this paper is primarily concerned with the printed dictionary, but in all aspects of the compilation and editing process, data has been organized in a way that allows digital exploitation in a later phase, whether as an independent online dictionary or as an integrated semantic component of the DDO.

The manuscript for the printed thesaurus is produced on the basis of the thesaurus XML document. For each section, all words are extracted automatically and presented in four main groups according to word class (nouns, verbs, adjectives, remaining words) with clear marks (in the form of bullet points) of the shifts between semantic categories within each word class group, e.g. between the nouns for persons and the verbal nouns. The logical order within each semantic category, e.g. within the list of persons, respectively verbal nouns, is maintained in the text. The annotated initial words of each semantic group are automatically presented as keywords in the printed text. Some manual adjustment is still needed after the conversion, e.g. in the cases where a word end up being the only member of a group, or in case highlighted words conflict or have too large scope.

## 2 Comprehensive versus Restricted Thesauri

Where the aim of the DDB is to present all types of words in the DDO including many compounds, the editors of the predecessor DB chose to present only a restricted selection of Danish concepts. Andersen himself states that the vocabulary of the thesaurus was reduced to only a part of the vocabulary found in the contemporary monolingual Danish dictionary *Dictionary of the Danish Language* (ODS), and that he intentionally did not cover the vocabulary in detail, in contrast to the German thesauri which he also used as a model to improve the overall macro structure (Dornseiff’s 2nd edition from 1940 and 3rd edition from 1943), and also in contrast to the approach of the contemporary Swedish thesaurus (Bring 1930, cf. Andersen 1945: X-XI). This should be viewed in light of the fact that a large amount of already categorized words were at his disposal in a manuscript of c. 1,000 pages established during a period of 25 years by a schoolteacher in Copenhagen. The manuscript was considered suitable for publishing provided that a thorough revision and modernisation was carried out, for which Andersen was responsible. He left out a substantial amount of compounds and dialect words, as well as many words for natural objects and other concrete objects, the advantage of this being, he writes, a dictionary which is “klarere og mere hændig” (‘clearer and more handy’; Andersen 1945: X). Another recent Scandinavian thesaurus, Rosbach (2001) has a limited stock of words for Norwegian. In our opinion, the restricted collection of words might be a reason why neither Andersen (1945) nor Rosbach (2001) have become widely used dictionaries in their respective countries, compared to the

position of *Roget's Thesaurus* (2002) within the English-speaking community (Hüllen 2009: 40 and 44). Roget and also Dornseiff (2004) for German constitute valuable linguistic resources, offering a large variety of words and expressions as it is the main purpose of thesauri, namely to support language variation and provide linguistic inspiration to users in text producing situations (cf. Hüllen 2009: 29 and 46).

### 3 The Comprehensive Approach: Lexicographic Challenges

The comprehensive approach involves some lexicographic challenges since some of the methods of the restricted thesauri are not applicable when the contents increase, both in terms of the number of lexical units and the types of words and expressions included. The first challenge concerns the order in which the words are presented; the second concerns the treatment of words which are not part of the unmarked standard vocabulary.

#### 3.1 Semantic versus Alphabetical Word Ordering

The DB organizes words in word classes which are divided into different semantic groups (objects, events, persons etc.) separated by dashes. Persons are always preceded by double dashes as the final element of a noun group. Within each semantic group (between dashes) words are listed in alphabetical order, meaning that the first word is not necessarily a keyword for the whole group. For instance, words within the category 'milk' are presented in the order "Fløde, Kærnemælk, Mælk, skummet Mælk, Piskefløde, Sødmælk, Yoghurt" (cream, buttermilk, milk, skimmed milk, double cream, whole milk, yoghurt). In the case of words for 'water', we find "Brøndvand, Drikkevand, Gaasevin, Isvand, Kildevand, Kommunevand, Vand" (well water, drinking water, plain water, ice water, spring water, tap water, water). In both cases, the hypernyms (*milk* and *water*) are placed between different types of hyponyms and do not function as keywords for the established semantic category.

Dornseiff (1940 and 1943) used the same alphabetical presentation. In the 8th edition (2004), though, we find an important difference to this: a keyword, typically the hypernym, has been moved from its place in the alphabetical order to the initial position of the group. The only cue to the keyword function is the break in the alphabetical order, which may be challenging for the user. Consider the case of words for different types of dairy products in Dornseiff (2004): "Milch · Buttermilch · Joghurt · Kefir · Magermilch · Rahm · Sauermilch" (milk, buttermilk, yoghurt, kefir, skimmed milk, cream, curdled milk). One notes also that the choice of alphabetical order inevitably breaks down logical ordering, placing *skimmed milk* between different types of sour milk rather than next to *milk*, of which it is a direct hyponym. Another example from Dornseiff (2004) concerns words for coffee: "Kaffee · Blümchen · Cappuccino · Espresso · Lorke · Milchkaffee · Mokka · MuckeFuck" (coffee, weak coffee (informal), cappuccino, espresso, bad weak coffee (dialect), café au lait, mocha, coffee substitute/ersatz coffee (deroga-

tory)), where one could argue that the informal word *Blümchen* and the dialect word *Lorke* have too prominent positions and would according to logic be better placed after the different types of coffee. Where a restricted thesaurus, such as the DB, can get away with an alphabetically ordered list of a small amount of words within a semantic category, this is in our opinion not suitable for the comprehensive thesaurus. The higher the amount of words, the more disturbing the alphabetical order becomes. This is the case in Dornseiff (2004) where a given word is most likely to be semantically more closely related to the initial keyword of the whole semantic category than it is to its immediate neighbours. In that way, the semantics of one word cannot be used to understand the next word and thereby activate a forgotten word in the user's mind. Furthermore, the distance from a word to the keyword can easily get very long. By contrast, Roget (2002) presents words only in semantic order and has in fact done so from the very first edition (Hüllen 2009: 40), a principle also adopted by Bring (1930). Initial keywords at the highest level of the taxonomy are graphically highlighted in Roget (2002) by italic types.

*soft drink*, teetotal d., non-alcoholic beverage; water, drinking w., filtered w., eau potable, spring water, fountain; soda water, soda, ..., coffee, café au lait, café noir, black coffee, white coffee, decaffeinated coffee, decaf, Irish coffee, Turkish c., espresso, cappuccino, latte ...

**Example 1: Roget's Thesaurus (2002), soft drinks (excerpt).**

In the DDB, we implement the same type of semantic ordering as Roget. It relies entirely on the lexicographer's judgment (cf. Hüllen 2009: 29), following two principles which go hand in hand, as reflected in linguistic theories on prototypes. The first implies that the prototypical, or central, members of a category are presented before less prototypical, radial members (see for example Dirven & Verspoor 2004: 17 for a description). An example is the section in the DDB on furniture, where chairs and beds are placed before lamps and carpets. The second principle is based on the idea of basic level categories in a language, from which a division of the vocabulary into three conceptual levels may be derived: a generic, a basic and a specific level (Dirven & Verspoor 2004: 37). Following this principle, basic and general level terms will be placed before specific level terms in the thesaurus. In many cases, the general level term is used as the title of the section in question and the basic level terms are highlighted as keywords. Illustrated by English words, general level terms such as *animal*, *plant* and *furniture* constitute section titles and are in their corresponding sections presented before basic level terms such as *dog*, *tree* and *bed*. Both the general and the basic level terms (i.e. *animal* as well as *dog*) are marked as keywords and listed in the thesaurus before the specific terms, in this case kinds of dogs, trees or beds, for example *poodle*, *oak tree* and *double bed*. In the case of coffee, the DDB presents the words as seen in example 2.

**kaffe**, mokka (uformelt); espresso, café au lait, caffè latte, cappuccino, macchiato; filterkaffe, stempelkaffe, pulverkaffe, kolbekaffe, tyrkisk kaffe; sort kaffe; jordemoderkaffe (uformelt), mokka; en lille sort;

termokaffe; varmekaffe; mosevand (slang); en kop kaffe, kaffetår, refill; morgenkaffe, formiddagskaffe, eftermiddagskaffe, aftenkaffe

(**coffee**, mocha (informal); espresso, café au lait, caffè latte, cappuccino, macchiato; drip coffee, press pot coffee, instant coffee, coffee made in a coffee maker, Turkish coffee; black coffee; very strong coffee (informal), (a cup of) strong coffee; a cup of black coffee laced with spirits; thermo jug coffee; warmed-up coffee; bog water (slang); a cup of coffee, cup of coffee, refill; morning coffee, mid-morning coffee, afternoon coffee, evening coffee)

**Example 2: Coffee words in the DDB.**

An argument against the semantic ordering principle is that it becomes more difficult for users to find a specific, already known word as they cannot rely on the alphabet when browsing through a group. Instead, they must look via words that come closest in meaning in the text, and this is not always an easy task. To support overview and browsing as much as possible, we have chosen to highlight more keywords in the text than Roget does. In the case of soft drinks in example 1, also hypernyms at a lower level in the taxonomy will be highlighted, i.e. also *water*, *soda water* and *coffee*. Furthermore, keywords at the highest taxonomic level are presented in boldface in the printed DDB in order to obtain a clearer visual signal than the italics used in Roget. As a consequence, finally, the index is organized such that for each entry word the following information is given: entry, keyword, section indicated by its number, part-of-speech. This is similar to the solution in Roget and different from Dornseiff that refers only to section number and title.

### 3.2 Stylistic Labels

Comprehensive thesauri cover a much broader range of stylistic varieties, accentuating the need for stylistic labels. In this section we will discuss the types of information used in the DDB and compare it with DB, Roget and Dornseiff.

The DB does not have any information about register, for example we find several informal words for nose simply listed without any comment (*Mule* ‘muzzle’, *Næse* ‘nose’, *Næsebor* ‘nostril’, *Snabel* ‘trunk’, *Snude* ‘snout’, *Snydeskraft* ‘hooter, conk’, *Tryne* ‘snout’, *Tud* ‘snout, schnozzle’). Nor does Dornseiff use stylistic labels, which in our opinion sometimes leads to confusing lists of words, for example those concerning coffee mentioned above where dialect words (*Lorke* (bad weak coffee), informal words (*Blümchen* (weak coffee), and *MuckeFuck* (substitute/ersatz coffee) are presented alongside standard German words for coffee (*Kaffee*, *Cappuccino*, *Espresso*, *Milchkaffee*, *Mokka*). Roget, on the other hand, uses labels to some extent but gives no precise description as to the use of these. Since the DDB contains a substantial number of slang and informal words from the DDO, we have in line with Roget chosen to assign labels of the stylistic and temporal status to the words which are not part of the standard vocabulary. The labels we use are extracted from the information in the DDO but presented in simplified



form. The detailed set of values in the DDO are here converted into five types: four stylistic (derogatory, informal, slang, jocular) and one temporal (archaic). Regional language is rare in the DDO and therefore not included in the set of labels. The method of direct transfer, however, is problematic, due to the fact that the labels were originally applied in relation to other senses of the same word and maybe a few synonyms given in the DDO entry, not to a large group of synonymous expressions as it is in the DDB. Manual adjustment is therefore needed in many cases. Only in the event where a few labelled words occur between numerous unmarked words in a semantic group, can they be kept as they are, without further editing. In other cases, we need to adjust the text in order to achieve homogeneous information about linguistic style. This is particularly relevant where the labels of adjacent words clash when seen in context. In these cases, we harmonize the information by choosing one label to cover both, if possible. For example, *lampe* ('lamp') and *pære* ('light bulb') are two expressions for 'lamp' in Danish, the former labelled 'slang', the latter 'informal' in the DDO. In the DDB, both are labelled 'informal' as there is no clear-cut boundary between slang and informal language. In the case of large groups of synonymous expressions of the same stylistic value, the label of the initial keyword indicates that the following words have the same value. Information about temporal status is treated independently of other labels. Example 3 shows the case of derogatory words for women in Danish, before (1) and after (2) editing.

(1) **kælling** (neds.), kone, gås (neds.), tante (neds.), tøs (neds.), hundyr, furie, rappenskralde (neds.), strigle (neds.), mokke, hystade, skude (neds.), sæk, (neds.), tudse (neds.), so (neds.), smatso (neds.), klidmoster (neds.), kran (slang), madamme (neds.), sladretaske (neds.), rendemaske (slang, gammeldags), sladderkælling, havgasse (neds.), harpe (neds.), ribs (neds.)

(**bitch** (derogatory), woman, silly goose (derogatory), aunt (derogatory), hussy (derogatory), female animal, fury, shrew (derogatory), termagant (derogatory), fishwife (derogatory), battleaxe (derogatory), virago (derogatory), cow, hag (derogatory), gadabout, .. etc.

(2) **kælling** (neds.), gås, tante, tøs, hundyr, madamme, klidmoster, furie, mokke, harpe, strigle, ribs, hystade, rappenskralde, havgasse, sladderkælling, sladretaske, rendemaske (gammeldags), skude, kran, sæk, tudse, so, smatso

**bitch** (derogatory), silly goose, aunt, hussy, female animal, fury, shrew (old-fashioned), termagant, fishwife, battleaxe, virago, cow, hag, gadabout, .. etc.

**Example 3: Derogatory words for 'woman' in the DDB. The group is initiated by one label instead of keeping the automatically inserted labels from the DDO on each word.**

A manual adjustment is required in approx. 1/8 of the 888 semantic groups in the DDB, concentrated in certain semantic areas such as men, women, drinking alcohol, body parts, sexuality and bodily functions, but also physical punishment, conflict, scolding and others. For the larger part of the DDB, however, stylistic labels occur only sparsely and are typically kept the way they appear in the DDO.

## 4 Conclusion

It is a challenge to compile a comprehensive thesaurus which truly reflects the vocabulary of a semiological monolingual dictionary. Bringing such a project to a completion within a period of just a few years' time can only be successful by using computational methods and by means of a well-structured model which guides the lexicographer's categorization of the words and, maybe most importantly, offers ways of placing the many radial concepts of the language. The alphabetical ordering of the words adopted by previous works is impracticable and must be replaced by semantic guidelines to ensure a consistent logical order within the large vocabulary of each category. The close connection between the two dictionaries makes it possible to reuse data in various ways. In this paper, we have shown how stylistic labels from the dictionary can be transferred to the thesaurus, and in the future our plan is to extract information about the semantic relations between words in the opposite direction, from the thesaurus into the dictionary, in this way adding an onomasiological component to the way users may access dictionary data.

## 5 References

- Andersen, Harry (1945). *Dansk Begrebsordbog*. København: Munksgaard.
- Bring, S. C. (1930). *Svenskt ordförråd ordnat i begreppsklasser*. Stockholm: Hugo Gebers Förlag.
- DDO = *Den Danske Ordbog*. Accessed at: <http://ordnet.dk/ddo> [11/04/2014].
- Dirven, Rene & Marjolijn Verspoor (2004). *Cognitive Exploration of Language and Linguistics*. Philadelphia, PA, John Benjamins Publishing Company, USA.
- Dornseiff, Franz (2004). *Der deutsche Wortschatz nach Sachgruppen*, 8. Auflage, Berlin/New York: Walter de Gruyter.
- Dornseiff, Franz (1943). *Der deutsche Wortschatz nach Sachgruppen*, 3. Auflage.
- Dornseiff, Franz (1940). *Der deutsche Wortschatz nach Sachgruppen*, 2. Auflage.
- Hüllen, Werner (2009). Dictionaries of synonyms and thesauri. In A. P. Cowie: *The Oxford History of English Lexicography, vol. II, Specialized Dictionaries*. Oxford: Oxford University Press, pp. 25-46.
- Nimb, S. & B. S. Pedersen (2012). Towards a richer wordnet representation of properties - exploiting semantic and thematic information from thesauri. In *LREC 2012 Proceedings*. Istanbul, Turkey, pp. 3452-3456.
- Nimb, S., B. S. Pedersen, A. Braasch, N. H. Sørensen & T. Troelsgård (2013). Enriching a wordnet from a thesaurus. In *Workshop Proceedings on Lexical Semantic Resources for NLP from the 19th Nordic Conference on Computational Linguistics (NODALIDA)*. Linköping Electronic Conference Proceedings; Volume 85 (ISSN 1650-3740).
- ODS = *Ordbog over det danske Sprog*. Vol. 1-28 (1918-1956). Det Danske Sprog- og Litteraturselskab og Gyldendal. Online version at: <http://ordnet.dk/ods> [11/04/2014].

Roget, Peter Mark (2002). *Roget's Thesaurus*, 150th anniversary edition edited by George Davidson. London: Penguin.

Rosbach, Johan Hammond (2001). *Ord og begreper. Norsk tesaurus*. Oslo: Pax Forlag A/S.



# From a Dialect Dictionary to an Etymological One

Vilja Oja, Iris Metsmägi  
Institute of the Estonian Language  
vilja.oja@eki.ee, iris.metsmagi@eki.ee

## Abstract

Typically, loanwords from different sources are presented in different entries of a dialect dictionary, but the etymologies of the dialect words are often obscure. An etymologist, on the other hand, has to consider the phonetic shape and developments of a stem, its areal distribution and meanings in dialects to find out the etymology. What are the cooperative prospects of etymologists and dialectologists? The paper compares the presentation of dialect words in two dictionaries currently compiled on the web, namely, in the Estonian dialect dictionary and in the Estonian etymological dictionary. Despite the different specifics of the two dictionaries a number of similar problems have cropped up. Also, comparisons of the material have yielded essential information enabling solutions for both sides. Across dialects, a loanword often displays numerous phonetic variants, while a variant may easily sound untypical of the concrete dialect. The possible donor can be traced considering the occurrence of the dialect word in the traditional area of loanwords of a certain origin and the semantic relationship of the word with the presumed donor language. Cooperation between etymologists and dialectologists has contributed a lot to making a distinction between homonymy and polysemy, to identification of folk etymologies (words as well as semantic nuances), to distinguishing between separate loanwords and derivatives of the same stem, etc. A shared electronic environment enables bilateral specification of the linguistic material, if necessary.

**Keywords:** dialect vocabulary; etymology; homonymy; polysemy; semantic change; Estonian

## 1 Introduction

The study is focused on the presentation of dialect words in an etymological dictionary. In 2003 the Institute of the Estonian Language (IEL) launched the project of an Estonian Etymological Dictionary. The first edition of the Estonian etymological dictionary (EES – *Eesti etümoloogiasõnaraamat*) compiled at the Institute of the Estonian Language was published in spring 2012. This is an approximately small dictionary for a wide readership, with about 6600 entries. The entry list of the EES was based on the word stems contained in the Estonian normative dictionary (ÕS 2006). Thus it includes the stems of standard Estonian and a small selection of dialect vocabulary (see Metsmägi 2010). The work has revealed that quite a significant part of Estonian word stems still lack a satisfactory scholarly explanation of their origin. Now, an extended and much more thorough edition is being prepared on the basis of the EES. The entry list of the new Estonian etymological dictionary (henceforth: EED) will be aug-

mented by about 1500 additional stems: dialect words and other older stems (e.g. words denoting obsolete tools, archaic occupations etc.) and the latest loanwords. The large Estonian etymological dictionary previously available, the *Estnisches etymologisches Wörterbuch* by Julius Mägiste (1982–1983, 2nd edition 2000; 12 volumes, 4106 pp.; EEW) contains dialect words as well. Unfortunately, the author was unable to finish it or edit due to his death. For that reason, there are numerous misprints and defective entries. Another reason for the dictionary being incomplete is that in exile the author ceased to have access to the materials kept in Estonia.

Standard Estonian was developed on the basis of local dialects. Estonian dialects, however, are based on the vernaculars of several Finnic tribes, not just one. The evolution of the dialects was affected by both the local conditions and various socio-historical factors. Conscious and deliberate development of Estonian as a standard language for the whole Estonian territory was only started in the 19th century (see, e.g. Kask 1984). Estonian vocabulary cannot be unambiguously divided into dialect and standard words. Often a word has different meanings in different dialects, which – like unusual phonetics – may help in tracing the origin of the word. In addition, dialects are a treasury of archaic terms and obsolete expressions that do not belong to the modern standard. Some hints on the origin of Estonian dialect words can be found in the Estonian-German dictionary compiled by F. J. Wiedemann in the 19th century (Wiedemann 1973). The etymology of Estonian dialect vocabulary has also been discussed in monographic studies of loanwords (e.g. Ariste 1933, Koponen 1998, Must 2000, Vaba 1997).

The Archive of the Estonian dialects and Finno-Ugric languages (EMSUKA) at the Institute of the Estonian Language includes a collection of Estonian dialect vocabulary, which contains about three million paper slips carrying phonetic, morphological and semantic data of dialect words as well as usage examples. This lexical collection serves as basis for the Estonian dialect dictionary (EMS – *Eesti murrete sõnaraamat*) currently compiled. So far, 25 fascicles (5085 pages) of the EMS (*a – matkama*) have been published. As preliminary work for the EMS a concise dialect dictionary (VMS – *Väike murdesõnastik*) was compiled in the 1980s. The VMS represents a list of possible entry words for the EMS, together with their areal distribution and the approximate meaning of the non-standard words. The dictionary has been published both in print and electronically. At present both the Estonian dialect dictionary and the Estonian etymological dictionary are being compiled online using the EELex program created at the Institute of the Estonian Language. The program enables representation of database material in a book layout form. Both dictionaries will be published electronically. The electronic version enables linking the dictionaries with the rest of the online dictionaries of the Institute of the Estonian Language.

## 2 Word Selection

The dialect words to be added to the new Estonian etymological dictionary (EED) are drawn from the material stored in the Estonian dialect dictionaries EMS and VMS. The data will be elaborated using the collections of the Archive of the Estonian dialects (EMSUKA), if necessary. One major dilemma that relates to the preparation of any dictionary is what words to choose, what ones to leave out. The EMS is exhaustive, including the entire vocabulary available from all Estonian dialects. Some words have been recorded from a few sub-dialects only while some others are known practically throughout the Estonian territory. Limits have been set on the presentation of words not specific to dialect usage, e.g. foreign words, new standard words, special terms, slang and nursery language (see Must 1968, Oja 1996).

In the first place the entry list of the new Estonian etymological dictionary is to be supplemented by dialect words representing lexical peculiarities of larger regions, such as South Estonian dialects<sup>1</sup>, e.g. *kurst* ‘twisted handful of flax, hair, straw etc.’ < Latvian *gūrste* (Vaba 1997: 107); *kuuas* ‘axe handle’ < Latvian dial. *kuôts* ‘handle of a tool etc.’ (Vaba 1997: 108-109), or the Insular dialect of Estonian, e.g. *kurt* ‘(fir or pine) cone’ < Estonian-Swedish<sup>2</sup> *kott*, *kotte*, Old Swedish *grankuttar* ‘(fir or pine) cone’, cf. Finnish-Swedish *kort* id. (Ariste 1933: 69).

Old genuine stems, missing in the standard vocabulary but found in the dialects will be added to the new etymological dictionary even if sparsely recorded. For example, a rather widespread dialect word *kosk* ‘thick bark’ is possibly an ancient (Uralic) genuine stem to which Ugric and Samoyed equivalents have been suggested (SKES 222; SSA 1: 409–410; UEW 179–180). The word *heri* ~ *here* ‘linden bark’ used in some Estonian dialects (EMS II: 1026) belongs to the Finno-Ugric layer of the genuine stems (SKES 183; SSA 1: 345; UEW 148–149). The dialect word *kärg* ‘woodpecker’ has etymological equivalents in Finnic and Volgaic languages (SKES: 261; EEW: 1138; UEW: 652; SSA 1: 476). The genuine stems with obvious etymological equivalents only in Finnic as the closest cognate languages will be added to the entry list as well.

On the other hand, the entry list will be augmented by words that clearly testify to being borrowed. Among the dialect words there are some old loanwords that were borrowed in the period before Estonian had become a separate language, i.e. the Indo-European, Indo-Aryan, Baltic, Germanic, Scandinavian and Slavic (Old Russian) loanwords. In general, they have etymological equivalents in other Uralic resp. Finnic languages. For example South Estonian *mehiläne*, *mihiläne* or *mehine* ‘bee’ (see VMS 2: 22) and the other Finno-Ugric words with the same stem (e.g. Finnish *mehiläinen*, Hungarian *méh*, *mihe* etc.) have been borrowed from Proto-Indo-European or Proto-Indo-Iranian (SKES: 339; SSA 2: 156; UEW: 271). Another word for bee, *mesilane* and variants (VMS 2: 25), used in other dialects as well as in standard Estonian is a derivative from the noun *mesi* ‘honey’, which is another Indo-European or In-

1 About grouping the Estonian dialects see, e.g. Pajusalu 2003: 231.

2 In Middle Ages the coasts and islets of Estonia were populated by Swedes. Their dialect is called Estonian-Swedish (see e.g. Blumfeldt 1961).

do-Iranian loanword (EES 280–281; UEW: 273; SSA 2: 161). The noun *rend* used for a long dining table in the Insular dialect has been borrowed from Baltic via the now extinct Curonian language *rend* ‘table’ (Vaba 2009: 779–780). The Insular dialect word *vada* ‘seine net’ (in close cognate languages: Karelian and Veps *vada*, Livonian *vadā*, Finnish *vata*) has been borrowed from Proto-Germanic, cf. Swedish, Norwegian *vad*, Old Norse *vaðr*, Middle Low German (MLGm) *wade*, German *Wate* (LÄGLOS III: 381; SKES: 1671–1672; SSA 3: 417–418).

In South-East Estonian and Eastern dialects the word *mugel* (*mugla*, *mukl*) stands for ‘spent lye’ (Wiedemann 1973: 609; VMS 2: 35). This is a Russian loanword borrowed from Old Pskov dialect (EEW: 1558; Koponen 1998: 127; Ojansuu 1922: 139). In North and Central Veps the same word *mugl* stands for ‘lye’. The term for ‘soap’ used in eastern Finnic languages is a more recent loanword derived from the same Russian stem: Finnish, Karelian, Veps, Ingrian, Votic *muila*, *muil* etc. < Russian *mylo* (ALFE 1: 205–206; Kalima 1952: 123–124).

As a rule, the transparent loanwords belong to some group of younger loanwords, borrowed only after Estonian had become a separate language, i.e. Low German, (High) German, Swedish, Russian, Latvian and Finnish loanwords. For example *kink* ‘haunch; ham’ (EMS III: 165) < MLGm *schink(e)* ‘ham’ (SKES I: 195); *tohv* ‘kind of cloth’ < German *Stoff* ‘material; textile’ (EEW: 3202); *turslag* ‘strainer’ < German *Durchschlag* ‘strainer; sieve’ (EEW: 3373); *tutspard* ‘moustache’ < German *Stutzbart* ‘tile beard’ (EEW: 3383); *hisla* ‘very sour, over sour’ < Russian *kislyj* ‘sour’ (Must 2000: 97–98); *robotama* ‘work quickly and intensively, toil, drudge; work quickly but carelessly’ < Russian *rabótat*, dial. *rabotát* ‘to work’ (EEW: 2510; Must 2000: 333–334).

It is not uncommon that a stem of Indo-European origin has been borrowed into different Finnic languages and even to the dialects of the same language in different times and via different routes. Sometimes the words of the same origin have been borrowed repeatedly. For example North-Estonian *tulk* ‘interpreter, translator’ was borrowed from Old Russian, but the Finnish *tulkki* is a Scandinavian loan and the standard Estonian word *tõlk* id., is a newer loan from Russian (SKES: 1391; EEW: 3350; SSA 3: 324).

### 3 Entry Structure

In both dictionaries, an entry contains the following components: headword, grammatical information of the stem, dialectal variants with information on the regional distribution and word meaning or meanings. In addition the EMS gives usage examples of the word from different dialects. The etymological dictionary presents information on the origin of the word stem and the equivalents of the stem in cognate languages, with comments and bibliography. Both dictionaries make ample use of cross references and reference entries or subentries. In the electronic version, the references function as links to the referred entries.



In both dictionaries the headword is either a standard word or a dialect word in a standardized shape. The main form of an Estonian noun is nominative singular (or nominative plural in the case of *plurale tantum*) and the lemmatic form of a verb is the infinitive ending in *-ma*. In both dictionaries the headwords are ordered alphabetically, but in the EMS the place of *h*-initial words depends on the vowel following *h*–: *a-* (*ha-*), *e-* (*he-*), etc. as word-initial *h*– is lacking in most Estonian dialects.

The entry lists of both dictionaries have been built up on stem basis, in the way that the stems having different etymologies go to separate entries. In the dialect dictionary (EMS) the predictable phonetic variants of a word are all in the same entry, but irregular variants are presented in different ones. By way of exception, some irregular variants of loanwords are found together in the same entry (in more detail see Neetar 1992). In the etymological dictionary (EED) an entry will cover all words or stems originating in a common etymological source, i.e. derivatives, lexicalized inflected forms and stem variants (synchronic as well as diachronic ones). Only the separately borrowed derivatives of a stem, i.e. derivatives containing affixes of the donor language get separate EED entries. In the EMS, separate entries are provided for each suffixed derivative as well as for the compounds included. In the EED, only the compounds borrowed as a single item are given in separate entries, e.g. *leierkast* ‘barrel organ’ < German *Leierkasten*.

According to the specifics of either dictionary their main emphasis lies on different aspects. The major part of an EMS entry deals with the details of the dialectal variants of the word and with its use in dialects: meanings, sub-senses and examples. If necessary, a semantic group is provided with sub-heads for figurative senses, idioms and phrases. The etymological dictionary is focused on the origin of word stems. The etymology is described in the EED by cognate language equivalents and/or the loan source. The entry of EED will also be supplied with a bibliographical component, containing a survey of the etymological treatments of the entry stem (i.e. the references).

## 4 Some Crucial Problems

One of the trickiest issues facing the authors of either dictionary is classification of the linguistic material, in particular, choice of the entry word. Many similar stems make one wonder which of the words should be presented in the same entry and which should be given separate entries. In dialect dictionary, a polysemantic word gets several entries, the entry word *nukk*, for example, starts seven articles (VMS 2: 105). Collocations are usually presented under several entries (see Oja 1996). In an etymological dictionary, it is essential to discriminate between words originating from the same loan source and those originating from different ones. Thus a detailed analysis of dialectal variants and meanings is required. The different meanings of a word as well as different phonetic variants may originate from different sources. The sources in their turn may be mutually connected, e.g. there are parallel loans from different Germanic languages – Middle Low German, (High) German, and Swedish. Actually the situation is even more complicated, because the words borrowed from the local German

dialect, the so-called Baltic German, and from the local Swedish dialect called Estonian Swedish, should be treated as separate loanword groups as well. In addition, sometimes different meanings have been borrowed into Estonian from the same source but in different times. In these cases the word variants or meanings will be discussed in separate EED entries.

#### 4.1 Polysemy and Homonymy

In a general case, the senses of a polysemantic word are presented in one and the same entry in the Estonian etymological dictionary, but in the dialect dictionary they are sometimes found under different entry words. The latter case applies, e.g. to words *koot<sup>1</sup>* ‘flail, agricultural tool for manual threshing; mobile part of various objects’ (EMS III: 635–636) and *koot<sup>2</sup>* ‘part of animal foot; part of human foot (lower part of the leg; thigh); usually pl., facet. human feet’(EMS III: 636–637). Notably, the two senses differ quite radically in their geographic distribution: *koot<sup>1</sup>* mainly occurs in South Estonian dialects, and also in the Central and Eastern dialects and in the Pärnumaa region, but it is practically absent in the Insular dialect, Läänemaa region and in the North-Eastern Coastal dialects, whereas *koot<sup>2</sup>* occurs in the North Estonian areas, apart from single reports from South Estonian dialects. Etymologically speaking, the root is the same: < MLGm *kote*, *kute* ‘ankle; hoof; pastern’ (Saareste 1924: 204; EEW III: 946). Thus, its primary sense is ‘part of leg’, semantically transferred to ‘manual threshing tool’ in Estonian (Saareste 1924: 204). The geography of the word in the sense of a tool (*koot<sup>1</sup>*) is eloquent of the spread of an innovative two-piece threshing tool: it was first introduced in southern Estonia (Mark 1932: 369–370, ERL: 96), while the older one-piece tool called *vart* ‘threshing stick’ was favoured the longest in the West-Estonian islands and in some places in western Estonia (Manninen 1929: 45–46; ERL: 342). The South-East Estonian *prunt*’s (and variants) ‘skirt’ is a Latvian loanword < East-Latvian regional *bruņči* id. that probably originates from the colour word *brūns* ‘brown’ < MLGm *brūn* id. (Saareste 1924: 166; Vaba 1997: 168–169). As long as the Latvian dialect word was unknown the South Estonian term for skirt used to be associated with the Estonian dialect word *pruūt*, *proūt* ‘pleat’ (EEW: 2183).

#### 4.2 Folk Etymology and Semantic Changes

There are several ways of loanword adaptation. Being uncertain of the real semantic background of foreign terms people often associate them with a similarly sounding familiar word. This way folk etymology may change the phonetic shape as well as the meaning of the loanword beyond recognition. Such loanwords tend to have exceptionally many dialect variants that are poorly motivated phonetically. Semantic change in loanwords, being caused by cultural differences, local specifics, taboo, etc., will complicate semantic analysis as well as detection of the origin of the word. Note that in dialects a word may be used in its original meaning lacking in the standard language or in a sense that is closer to the original meaning than standard usage (Oja & Metsmägi 2013).

For example, a wallet or purse may be humorously called *tengelpung* in Estonian. The components of the compound word have been associated with the well-known loanwords *teng* ‘money’ (< Russian *denga* id.) and *pung* arch. ‘wallet, pouch’ (< MLGm *punge* ‘pocket’ or < Swedish *pung* ‘wallet, pouch’). Actually, the Estonian *tengelpung* (in dialects also *tenkelpuuh*, *tenkelpus* etc.) has been borrowed from the Baltic German dialect (< German *Denkelbuch* ‘old style paper notebook, pocketbook’ < *denken* ‘think, believe, plan, imagine’ + *Buch* ‘book’) (Ariste 1942: 20; Viires 1960: 158.)

The name of juniper (the plant in the genus *Juniperus*) is *kadahas* in Estonian. In the second half of the 19th and beginning of the 20th century a similar noun *kadahas* and the compound *kadaka/saks* (*saks* ‘squire, vulg. German’) or *kadaka/sakslane* (*sakslane* ‘German’) were used for Estonians who tried to look like Germans and spoke (usually incorrect) German (EMS II: 453). Although folk etymology would associate the Estonian words *hadakasaks*, *hadakasakslane*, *hadakas* ‘(half-) Germanized Estonian’ with ‘juniper’, the disdainful words have nothing to do with the tree. Instead, it is a loanword borrowed from the German compound word *Katensaße* ‘slum dweller, craftsman’ (< *Kate* ‘hut, shanty’ + *Saße* ‘place of residence’), which has been folk-etymologically modified to sound like certain familiar words (Saari 2004: 119–120). The word *katekismus* ‘catechism’ has dialect variants *katekeskmus(s)* and *katekeskmine* showing that the word has probably been folk-etymologically connected to the South-Estonian dialect word *kat’s* (vocalic stem *kate*) ‘two’ and *kesk(mine)* ‘middle; between’.

### 4.3 Derivatives or Separate Borrowings

Sometimes the question is if we have to do with a borrowed derivative, i.e., whether the derivative and the stem have been borrowed separately or not. The donor language may have been the source of the stem as well as of one or more of its derivatives. Some old Indo-European loanwords in the Finnic language group have undergone morphological and semantic adaptation to such an extent that they have come to be regarded as genuine native derivatives. For example, the structure of the Estonian *raudjas* ‘russet’ appears to be *raud* ‘iron’ + diminutive suffix *-jas*, but it is most likely a loan from a Baltic colour word, cf. Latvian *raūds*, *raūdis* ‘reddish brown’ from Proto-Indo-European *\*reud<sup>h</sup>-*, whereas the term *raud* ‘iron’ is a Germanic loan from Proto-Germanic *\*raudan-* (Oja 2004: 37–38).

Newer (esp. MLGm) loanwords display a lot of cases where a noun has been borrowed in parallel with a zero-affixed verb of the same stem, e.g. Estonian noun *hink* ‘gift, present’ < MLGm *schenke* ‘act of giving; (welcome) present, etc.’ (EEW: 834) and the verb *hinkima* ‘make a present, donate, give’ < MLGm *schenken* ‘make a present, to give’ (Ariste 1940a: 12; EEW: 834) or another example: *rööv* ‘robbery’ < MLGm *rōf* ‘robbery; booty’ (Ariste 1940b: 110) and *röövima* ‘to rob’ < MLGm *roven* ‘to rob’ (Ariste 1980: 34; SKES: 908; SSA 3: 122).

## 5 Conclusion

Word origin can be specified by following its dialectal variants as well as similar words in cognate languages and contact languages, considering both their areal distribution and meaning. The variation of the phonetic shape and the meaning of words in Estonian dialects may suggest different bor-

rowing times or travelling routes. The areal distribution of loanwords helps to pinpoint the centres of cultural innovations. An available dialect dictionary is of great help to word etymologization and etymological lexicography, offering concentrated and systematized information on the areal distribution, phonetic variation, and meanings of words. However, its entry list is most extensive, containing the whole dialect vocabulary, including compound words and derivatives, and thus an etymological lexicographer has to make a selection including the following: (1) dialect words representing lexical peculiarities of larger regions; (2) old genuine words; (3) loanwords with an obvious source; (4) derivatives borrowed separately from the stem.

A common issue in compiling both dictionaries is the arrangement of the highly variable linguistic data, in the face of polysemy and homonymy, folk etymology and semantic changes. Depending on the specifics of the dictionary, the final solutions may differ for the dialect dictionary and for the etymological dictionary. In a dialect dictionary the material cannot be presented systematically enough without considering word etymology. Hence an etymological dictionary makes an effective supplement to a dialect dictionary, helping to understand the background of the diversity of meanings and phonetic variation. Thus the best policy would probably be parallel compilation of the two dictionaries in a close cooperation of both teams, which is, however, extremely problematic to organize.

## 6 References

- ALFE = *Atlas Linguarum Fennicarum. Itämerensuomalainen kielikartasto. Läänemeresoome keeleatlas. Ostseeфинischer Sprachatlas. Лингвистический атлас прибалтийско-финских языков*. ALFE 1–3. Chief ed. T. Tuomi. Eds. S. Suhonen (1), T.-R. Viitso (2), V. Rjagoev (3). Suomalaisen Kirjallisuuden Seuran Toimituksia 800/1295. Kotimaisten kielten tutkimuskeskuksen julkaisuja 118/159. Helsinki: Suomalaisen Kirjallisuuden Seura, Kotimaisten kielten tutkimuskeskus 2004–2010.
- Ariste, P. (1933). *Eestirootsi laensõnad eesti murretes. Die Estlandschwedische Lehnwörter in der estnischen Sprache*. Acta et Commentationes Universitatis Dorpatensis B XXIX. Tartu.
- Ariste, P. (1940a). *Georg Mülleri saksa laensõnad* [Georg Müller's German loanwords]. Acta et Commentationes Universitatis Dorpatensis B XLVI.1. Tartu.
- Ariste, P. (1940b). Saksa laensõnadest 16. sajandi eesti kirjakeeles [About German loanwords in the 16th century literary Estonian]. In *Eesti Keel* 3–4. Tartu, pp. 108–112.
- Ariste, P. (1942). Etümoloogilisi märkmeid [Etymological notes] I. In *Acta et Commentationes Universitatis Dorpatensis* B XLIX. 1. Tartu, pp. 1–26.
- Ariste, P. (1980). Deutsche Lehnwörter im Wotischen. In *Spetsifitskie osobennosti leksiki i grammatiki ural'skih jazykov*. Acta et Commentationes Universitatis Tartuensis 517. Fenno-ugristica 6. Tartu, pp. 27–38.
- Blumfeldt, E. (1961). Estlandssvenskarnas historia [History of Estonian Swedes]. In: *En bok um Estlands svenskar* I. Stockholm: Kulturföreningen Svenska Odlingens Vänner, 63–178.
- EES = Metsmägi, I., Sedrik, M. & Soosaar, S.-E. (2012). *Eesti etümoloogiasõnaraamat* [Estonian Etymological Dictionary]. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.
- EEW = Mägiste, J. (1982–1983). *Estnisches etymologisches Wörterbuch*. Helsinki: Finnisch-Ugrische Gesellschaft.
- EMS = *Eesti murrete sõnaraamat* [Estonian Dialect Dictionary] (1994–2014). I–V (Fascicles 1–25), Eds. A. Haak, E. Juhkam, M.-L. Kalvik, M. Kendla, T. Laansalu, V. Lonn, H. Neetar, E. Niit, P. Norvik, V. Oja, V. Pall, E. Ross, A. Sepp, M.-E. Tirkkonen, J. Viikberg. Tallinn: Eesti Teaduste Akadeemia, Eesti Keele Instituut.

- ERL = Troska, G., Viires, A., Karu, E., Vahtre, L., Tõnurist, I. (2007). *Eesti rahvakultuuri leksikon*. Ed. A. Viires. 3., corrected and supplemented edition. [1. edition 1995, 2. edition 2000.] Tallinn: Eesti entsüklopeediakirjastus.
- Kask, A. (1984). *Eesti murded ja kirjakeel [Estonian dialects and literary language]*. Eesti NSV TA Emakeele Seltsi toimetised 16. Tallinn: Valgus.
- Koponen, E. (1998). *Etelävirron murteen sanaston alkuperä. Itämerensuomalaista etymologiaa [The Origin of the South-Estonian Dialect Vocabulary. Finnic etymologies]*. Suomalais-Ugrilaisen Seuran Toimituksia 230. Helsinki: Suomalais-Ugrilainen Seura.
- LÄGLOS = Kylstra, A. D., Hahmo, S.-L., Hofstra, T. & Nikkilä, O. (1991–2012). *Lexikon der älteren germanischen Lehnwörter in den ostseefinnischen Sprachen*. Amsterdam–Atlanta–New York: Rodopi.
- Manninen, I. (1929). Übersicht der ethnographischen Sammelarbeit in Eesti in den Jahren 1923–1926. In *Õpetatud Eesti Seltsi Aastaraamat. Sitzungsberichte der Gelehrten Estnischen Gesellschaft 1927*. Tartu: C. Matiesen, pp. 31–47.
- Mark, J. (1932). Über das Roggendreschen bei den Esten. Festvortrag, gehalten am 19. Januar 1931, dem 93. Jahrestage der Gesellschaft. In *Õpetatud Eesti Seltsi Aastaraamat. Sitzungsberichte der Gelehrten Estnischen Gesellschaft 1931*. Tartu: Õpetatud Eesti Selts. Gelehrte Estnische Gesellschaft.
- Metsmägi, I. (2010). Dialect materials in the Estonian Etymological Dictionary. In *Slavia Centralis III*, 1, pp. 196–204.
- Must, M. (1968). Über die Arbeiten am estnischen Dialektwörterbuch. In *Congressus Secundus Internationalis Fenno-Ugristarum Helsingiae habitus 23.–28. VIII 1965. Pars I, Acta Linguistica*. Helsinki: Societas Fenno-Ugrica, pp. 348–351.
- Must, M. (2000). *Vene laensõnad eesti murretes [Russian Loanwords in Estonian Dialects]*. Tallinn: Eesti Keele Sihtasutus.
- Neetar, H. (1992). Etymologisches im estnischen Dialektwörterbuch (EDW). In *Euralex '92. Proceedings I–II. Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, Finland*. Tampere, pp. 607–614.
- Oja, V. (1996). Word Combinations in the Estonian Dialect Dictionary. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, C. Rödger Pappmehl (eds.). *Euralex '96 Proceedings I–II. Papers submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg, Sweden*. Göteborg: Göteborg University, pp. 443–449.
- Oja, V. (2004). Some colour words with restricted reference. In *Latvijas Zinātņu Akadēmijas Vēstis. A daļa. Sociālās un humanitārās zinātnes 5*, pp. 37–42.
- Oja, V., Metsmägi, I. (2013). Laensõnade tähendusuhetest [Semantic relations of loanwords]. In H. Metslang, M. Langemets, M.-M. Sepper (eds.) *Eesti Rakenduslingvistika Ühingu aastaraamat. Estonian Papers in Applied Linguistics 9*. Tallinn: Eesti Rakenduslingvistika Ühing, pp. 181–194.
- Ojansuu, H. (1922). Eesti etimoloogiad [Estonian etymologies]. In *Eesti Keel* 4–5, pp. 137–139.
- Pajusalu, K. (2003). Estonian dialects. In *Estonian Language*. Ed. M. Ereht. Linguistica Uralica supplementary series 1. Tallinn: Estonian Academy Publishers, pp. 231–272.
- Saareste, A. (1924). *Leksikaalseist vahekordadest eesti murretes. Du sectionnement lexicologique dans les patois estoniens I*. Acta et Commentationes Universitatis Dorpatensis B VI.1. Tartu.
- Saari, H. (2004). *Keelehäälning. Eesti Raadio "Keeleminutid" 1975–1999*. Tallinn: Eesti Keele Instituut, Eesti Keele Sihtasutus.
- SKES = Toivonen, Y. H., Itkonen, E., Joki, A. J. & Peltola, R. (1955–1978). *Suomen kielen etymologinen sanakirja [Etymological Dictionary of Finnish]*. Lexica Societatis Fenno-Ugricae XII. Helsinki: Suomalais-Ugrilainen Seura.
- SSA = *Suomen sanojen alkuperä. Etymologinen sanakirja [The Origin of Finnish Words. An Etymological Dictionary]* (1992–2000). Eds. E. Itkonen, U.-M. Kulonen. Suomalaisen Kirjallisuuden Seuran toimituksia 556. Kotimaisten kielten tutkimuskeskuksen julkaisuja 62. Helsinki: Suomalaisen Kirjallisuuden Seura, Kotimaisten kielten tutkimuskeskus.

- Toivonen, Y. H. (1928). Zur Geschichte der finnisch-ugrischen inlautenden Affrikaten. In *Finnisch-Ugrische Forschungen* XIX, 1–270.
- UEW = Rédei, K. (1986–1988). *Uralisches etymologisches Wörterbuch*. Budapest: Akadémiai Kiadó.
- Vaba, L. (1997). *Uurimusi läti-eeesti keelesuhetest* [Studies on Latvian-Estonian Linguistic Relations]. Tallinn–Tampere: Eesti Keele Instituut, Tampereen yliopiston suomen kielen ja yleisen kielitieteen laitos.
- Vaba, L. (2009). *Rend ja laud*. Kisklauast söögilauaks [*Rend and laud*. From a Split Log to a Dining Table]. In *Keel ja Kirjandus* LII (10), pp. 779–784.
- Viires, A. (1960). *Hundilaut ja tengelpung*. In *Keel ja Kirjandus* III (3), pp. 157–158.
- VMS = V. Pall (ed.) *Väike murdesõnastik* [Concise dialect dictionary] (1982, 1989). I, II. Tallinn: Valgus.
- ÕS 2006 = Ereht, T., Leemets, T., Mäearu, S. & Raadik, M. *Eesti õigekeelsussõnaraamat 2006* [Estonian Normative Dictionary 2006]. Ed. T. Ereht. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus.
- Wiedemann, F. J. (1973 [1869]). *Eesti-saksa sõnaraamat*. *Estnisch-deutsches Wörterbuch*. Vierter unveränderter Druck nach der von Jakob Hurt redigierten [2.] Auflage [1893]. Tallinn: Valgus.

### **Acknowledgements**

The study has been funded by the Estonian Ministry of Education and Research (SF0050037s10) and Estonian Science Foundation (ETF9367).

# **Research on Dictionary Use**





# Wörterbuchbenutzung: Ergebnisse einer Umfrage bei italienischen DaF-Lernern

Carolina Flinz  
Universität Pisa  
c.flinz@ec.unipi.it

## Abstract

Die vorliegende empirische Untersuchung befasst sich mit einer Umfrage zur Wörterbuchbenutzung bei 41 Studentinnen und Studenten des *Dipartimento di Filologia, Letteratura e Linguistica* der Universität Pisa, dasselbe Department, an dem auch das deutsch-italienische sprachwissenschaftliche Online-Wörterbuch DIL erarbeitet worden ist (vgl. Flinz: 2011). Die schriftliche Umfrage wurde in Anlehnung an Hartmanns 5. Hypothese „*An analysis of users' needs should precede dictionary design*“ (1989) durchgeführt. Die wichtigsten Ergebnisse waren von großer Bedeutung für die Gestaltung der makro- und mikrostrukturellen Eigenschaften des Fachwörterbuches. Die Ergebnisse der Untersuchung und die daraus folgenden Reflektionen werden in thematischen Kernblöcken vorgestellt.

**Keywords:** Wörterbuchbenutzung; Umfrage; Fachwörterbücher; Online-Fachwörterbücher

## 1 Einleitung

Ziel des Beitrags ist es, die Ergebnisse einer schriftlichen Umfrage zu präsentieren, welche der Frage nachgeht, in welchen Situationen und zu welchen Zwecken italienische DaF-Lernende zweisprachige Sprachwörterbücher und insbesondere sprachwissenschaftliche Fachwörterbücher verwenden. Ferner soll untersucht werden, welche Einstellungen die Testpersonen über Online- und Printprodukte haben, welche Eigenschaften der Printprodukte sie positiv beurteilen und welche nicht. Nach einem Exkurs über den theoretischen Hintergrund (§2) der Untersuchung, die Wörterbuchbenutzung, werden die Analyse und die Kernblöcke der Umfrage vorgestellt (§3). Die wichtigsten Ergebnisse werden anschließend vorgestellt und diskutiert (§4). Abschließende Schlussfolgerungen beenden den Beitrag.

## 2 Stand der Forschung

Wörterbuchbenutzung als wissenschaftlich fundierte Disziplin hat sich im Rahmen der Lexikographie in den 80er Jahren entwickelt. Unterschiedliche Artikel und Beiträge zu diesem Thema (Hausmann et al. 1989; Ripfel/Wiegand 1988; Rossenbeck 2005; Welker 2010; Wiegand 1987/1998/2008/2010) bestätigen dieses wachsende Interesse, obgleich es laut Wiegand trotz seiner Bedeutung für die Er-

schaffung einer größeren Benutzereffizienz (1987:179) der am wenigsten erforschte Bereich der Lexikographie bleibt (2008:1). Neue Wörterbücher oder Neuauflagen sollten aus der wissenschaftlichen Erkenntnis dieser Praxis entstehen und einen höheren Nutzungswert haben (Wiegand 1987:179).

Die Forschung zur Online-Wörterbuchbenutzung ist hingegen noch in ihren Anfängen (vgl. Nesi 2000:845; Simonsen 2011:7) und die im Abstand von drei Jahren am IDS durchgeführte Studie „User-adaptive access and cross-references in elexiko (BZV elexiko)“ setzt es sich zum Ziel, diese Lücke zu füllen. Online-Wörterbücher sollten sich empirischer Analysen bedienen, die aufzeigen, welche spezifischen Gebrauchssituationen und Bedürfnisse vorhanden sind, wie die Wörterbücher tatsächlich benutzt werden und wie sie benutzerfreundlicher gestaltet werden könnten (vgl. Atkins/Varantola 1998; Hartmann 2000; Spitzer/Koplenig/Töpel 2012).

Vereinzelt sind Untersuchungen zu unterschiedlichen Fragestellungen der Wörterbuchbenutzungsforschung veröffentlicht worden. Sie widmen sich der:

- (1) Typologie der Wörterbücher nach Benutzungsmöglichkeiten (u.a. Engelberg/Lemnitzer 2009; Kühn 1989);
- (2) Problemlösung und Optimierungsvorschlägen (u.a. Wiegand 1995; Ripfel 1989a; Domínguez 2006; Kemmer 2010);
- (3) Erkundung typischer Benutzungssituationen (u.a. Kromann 1995; Tarp 2008) auch im Fachsprachbereich (Engelberg/Lemnitzer 2009).

Der Mangel an empirischen Untersuchungen wird aber weiterhin von einigen Arbeiten hervorgehoben (Hartmann: 2001; Kromann: 1995; Wiegand 2008). Um diesem Umstand entgegenzuwirken, sind 2012 / 2013 äußerst interessante Studien durchgeführt worden, im deutschen Raum die obenerwähnte Studie am IDS, im spanischen Raum die UDALPE-Befragung. Sie bedienen sich unterschiedlicher empirischer Methoden, wie der Beobachtung, der Befragung, des Protokolls, des Tests, des Experiments, des interpretativen Verfahrens, des „Simultaneous Feedback“ (de Schryver/Prinsloo 2000). Auch die Verbindung mehrerer Methoden wird als positiv eingestuft. Die Umfrage erweist sich jedoch durch ihre Verwendung in mehreren Studien als eine der meistbenutzten Methoden, um die Benutzungssituationen sowohl seitens des Fremdsprachenlerner (de Schryver/Prinsloo 2011) als auch seitens des Muttersprachlers (Retti 2004; Ekwa Ebanéga/Tomba Moussavou 2008) zu erkunden und um die Verwendung von Spezialwörterbüchern (Wang 2001; Muráth 2005; Taljard / Prinsloo 2011) und von Online-Wörterbüchern (vgl. elexiko) zu erforschen.

### 3 Die Untersuchung

Die vorliegende Untersuchung wurde 2012/2013 durchgeführt und befasst sich mit der Auswertung der Ergebnisse einer Umfrage bei 41 italophonen Studierenden des *Dipartimento di Filologia, Letteratura e Linguistica* der Universität Pisa, die Deutsch als Fremdsprachestudieren. Die Umfrage, dessen Hauptanliegen die Erkundung der Wörterbuchbenutzung einer DaF-Lerngruppe einer sprachwissenschaft-

lichen Fakultät ist, wurde in Anlehnung an Hartmanns 5. Hypothese „*An analysis of users' needs should precede dictionary design*“ (1989) durchgeführt, da diese grundlegend für das Projekt *DIL* ist<sup>1</sup>. Ein Teil der Ergebnisse wird in diesem Beitrag vorgestellt und problematisiert, mit entsprechenden Zusammenfassungen und Schlussfolgerungen.

Der Fragebogen in italienischer Sprache wurde den Probanden aufbereitet und hatte eine Ausfüllzeit von ca. 20 Minuten. Die 27 Fragen konzentrieren sich auf folgende Kernfrageblöcke: 1. Wörterbuchbenutzer (auch Soziodemografie), 2. Wörterbuchtyp, Wörterbuchbenutzungssituation, Bedürfnisse; 3. Wörterbuchformat; 4. Benutzung der Umtexte; 5. das sprachwissenschaftliche Fachwörterbuch und die mikrostrukturellen Eigenschaften. Die Fragen des ersten Blocks betreffen die persönlichen Daten der Befragten (Alter, Geschlecht, Muttersprache, Niveau der Deutschkenntnisse, Motivation zur Auswahl des Deutschen als Fremdsprache, das Erlernen anderer Fremdsprachen). Im zweiten Block werden die Gründe zur Benutzung des Wörterbuches, die typische Benutzungssituation, die Benutzungsfrequenz und die zu erfüllenden Bedürfnisse erforscht. Unterschiedliche Wörterbuchtypen (einsprachige, zweisprachige, fachliche, sprachwissenschaftliche und weitere) und ihre jeweilige Benutzungsfrequenz werden untersucht. Der dritte Block der Umfrage widmet sich dem Format, den Vor- und Nachteilen eines Online-Wörterbuchs, den verwendeten dynamischen Elementen und der Art der Verlinkung. Die Informationsstruktur, die vorkommenden Probleme und Optimierungsvorschläge werden ebenfalls analysiert. Es wird auch auf Copyright und Aktualisierung der Wörterbüchereingegangen. Der vierte Block betrifft die Umtexte und deren Benutzungsfrequenz: Ziele und Hauptmerkmale bestimmter Umtexte stehen im Mittelpunkt des Interesses. Im fünften Block wird spezifisch auf die Funktionen und Bedürfnisse des sprachwissenschaftlichen Fachwörterbuches eingegangen.

## 4 Die Untersuchung: Ergebnisse

Die Ergebnisse der Umfrage werden in den folgenden fünf vorgestellten Hauptfrageblöcken zusammengefasst: Wörterbuchbenutzer (4.1); Wörterbuchtyp und Bedürfnisse (4.2); Wörterbuchformat (4.3); Verwendung der Umtexte (4.4); Funktion und Bedürfnisse des sprachwissenschaftliches Fachwörterbuches (4.5).

---

1 DIL ist ein Deutsch-Italienisches Online-Fachwörterbuch der Linguistik, das an der Universität Pisa entwickelt worden ist und in ständiger Bearbeitung ist.

## 4.1 Persönliche Angaben

Die ersten sieben Fragen betreffen die persönlichen Daten der Befragten. Die Analyse der Daten hat Folgendes ergeben:

- (1) Die Gruppe besteht aus 23% Männer und 77% Frauen;
- (2) Die Altersgruppe kann wie folgt dargestellt werden:

	Männer	Frauen
19-21 Jahre	67%	74%
22-24 Jahre	22%	23%
Über 25 Jahre	11%	3%

**Tabelle 1: Altersaufteilung der Testpersonen.**

- (3) Die Muttersprache ist gemäß dem Umfrageziel hauptsächlich Italienisch, auch wenn 2 Studentinnen Albanisch und eine Studentin Russisch als Muttersprache haben;
- (4) Die befragten Studenten besuchen den ersten, den zweiten oder den dritten Jahrgang des Bachelorstudienganges;
- (5) Die Deutschkenntnisse der Testpersonen sind sehr unterschiedlich: die Mehrheit weist das Sprachniveau A1 (49%) auf; während nur 8% A2 hat; 38% der Studenten und Studentinnen besitzen das Mittelstufenniveau B1, während nur 5% B2erreicht haben . Kein einziger Student verfügt über ein Hochstufenniveau (C1 oder C2). Nur 18% der Versuchspersonen haben eine Zertifizierung der Fremdsprachenkenntnisse (Zertifikat B1).
- (6) Die Gründe zum Erlernen des Deutschen als Fremdsprache sind ziemlich homogen: 50% der befragten Personen haben Deutsch wegen möglicher Chancen auf dem Arbeitsmarkt gewählt, 23% wegen persönlicher Interessen (Liebe zur deutschen Sprache und Kultur; Bedeutung in der Europäischen Union etc.); 23% sind von der Fremdsprache fasziniert; 4% aus Neugierde, da sie schon weitere Fremdsprachen kennen;
- (7) Deutsch wird von fast allen Studenten als zweite (12%) oder dritte Fremdsprache (88%) gelernt. 98% sprechen Englisch als erste Fremdsprache und nur 2% Französisch. Die Aufteilung der zweiten erlernten Fremdsprache kann aus folgender Tabelle entnommen werden:

Französisch	Spanisch	Russisch	Polnisch	Keine weitere Fremdsprache nach Englisch
54%	28%	3%	3%	12%

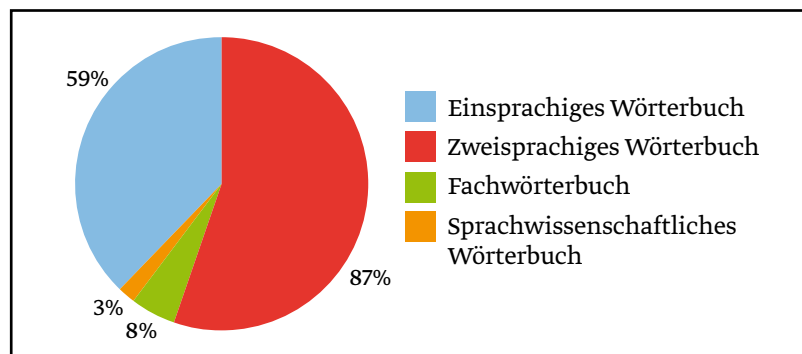
**Tabelle 2: Zweite erlernte Fremdsprache nach Englisch vor Deutsch.**

Zusammenfassend kann festgestellt werden, dass die Testgruppe in großem Maße aus Frauen besteht, die ein Durchschnittsalter zwischen 19 und 21 Jahren aufweisen. Sie besitzen entweder das Sprachniveau A1 oder B1, haben aber selten Prüfungen zur Bestätigung des erlangten Fremdsprachenniveaus

abgelegt. Die Motivation zur Wahl des Deutschen als Fremdsprache ist stark mit der Arbeitswelt verbunden; die zentrale Rolle Deutschlands in der EU ist einer der oft genannten Gründe. Deutsch wird meistens als dritte Fremdsprache nach Englisch und Französisch gewählt.

## 4.2 Wörterbuchtyp und Wörterbuchbenutzungssituation

Aus der Analyse der Items 7, 8, 9, 10, 11, 12, 13 kann ein Profil der verwendeten Wörter- und Fachwörterbücher seitens der Testpersonen geschaffen werden:



**Abbildung 1: Diagramm der von den Probanden benutzten Wörterbuchtypen.**

Die meisten Studentinnen und Studenten benutzen ein zweisprachiges Wörterbuch, Fachwörterbücher und Sprachwissenschaftliche Wörterbücher werden kaum verwendet; dies obwohl die Testpersonen auch sprachwissenschaftliche Kurse besuchen. Es wäre interessant nachzuforschen, welche Gründe es dafür gibt.

Die Frequenz der benutzten Wörterbuchtypen variiert sehr stark und kann der nachfolgenden Tabelle entnommen werden:

	oft	manchmal	selten	keine Angabe zur Frequenz
Einsprachiges Wörterbuch	26%	52%	22%	-
Zweisprachiges Wörterbuch	65%	20%	-	15%
Fachwörterbuch	-	33%	33%	34%
Sprachwissenschaftliches Wörterbuch	-	-	-	100%

**Tabelle 3: Darstellung der Verwendungsfrequenz der Wörterbuchtypen.**

Das zweisprachige Wörterbuch ist der meistbenutzte Wörterbuchtyp, wie man es bei Studenten, die Fremdsprachen lernen, erwartet hätte. Auch die Benutzungsfrequenz bestätigt dies, denn 65% der Probanden benutzen es oft. Negativ ist das Ergebnis zum sprachwissenschaftlichen Wörterbuch; aufgrund der Präsenz von Fächern wie Linguistik und Fremdsprachendidaktik im Studienplan der Studenten hätte man eine höhere Frequenz erwartet.

Als weitere verwendete Wörterbuchtypen werden von den Versuchspersonen Synonymwörterbücher (31%), Wörterbücher zur Rechtschreibung (3%) und zur Phraseologie (3%) genannt. Die Benutzung von Valenzwörterbüchern enthält keine Treffer.

Interessant ist die Beobachtung zu den Gründen oder Motivatoren, welche die Testpersonen dazu veranlasst haben sich der Wörterbücher zu bedienen: 51% nennen die Schule oder universitäre Institution (eine große Zahl von Probanden nennt die Deutschlehrerin oder Deutschprofessorin); 6% die Familie und 18% das persönliche Interesse mehr zu verstehen und sich selbstständiger zu entwickeln.

### 4.3 Wörterbuchformate

Die Ergebnisse zu den Fragen, die sich auf das Wörterbuchformat (Fragen 14, 15, 16) konzentrieren, sind in einem Zeitraum der Technologie und Multimedialität erstaunlich, denn noch 43% der Probanden benutzen Printwörterbücher, 8% CD-Rom Wörterbücher und nur 18% Online-Produkte. 20% der Versuchspersonen entscheiden sich für zwei Typen: Print- und Onlinewörterbücher (20%), CD-Rom und Onlinewörterbücher (8%) und Print- und CD-Romwörterbücher (3%).

Die Fragen (15 und 16) haben versucht, ein Bild über die möglichen Vorteile und Nachteile der obengenannten Formate zu schaffen. Als Vorteile werden folgende Aspekte genannt (von links nach rechts in einer absteigenden Skala):

	1.	2.	3.	4.	5.
Printformat	Zweckmäßigkeit	Vollständigkeit	Präzision	Zuverlässigkeit	Größere Anzahl von Einträgen und Beispielen
CD-Rom Format	Schnelligkeit	Vollständigkeit	-	-	-
Online-Format	Schnelligkeit	Jederzeit benutzbar	Einfachheit	Suche auch ohne Nennform / keine Bezahlung	Forum

**Tabelle 4: Vorteile der unterschiedlichen Formate.**

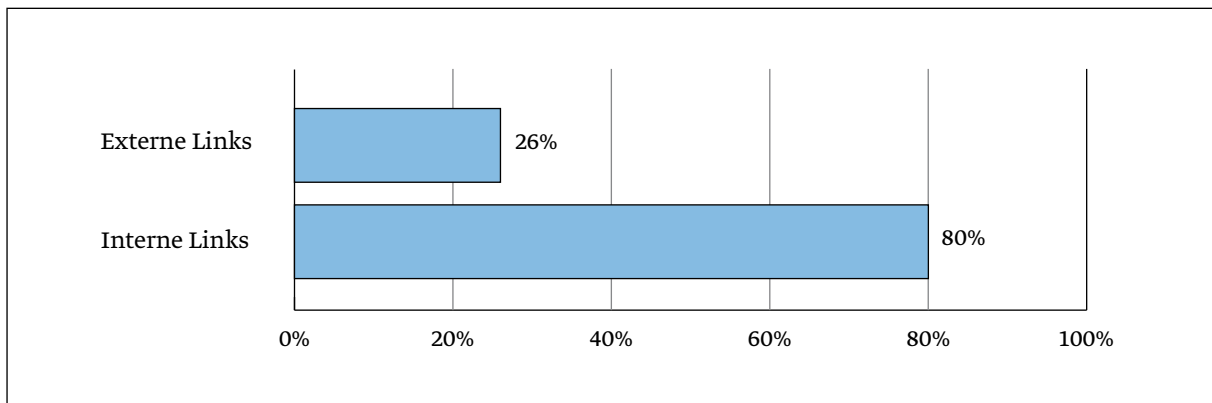
Als Nachteile:

	1.	2.	3.	4.
Printformat	Zeitverschwendung	keine Zweckmäßigkeit (Größe, Schwere)	Kosten	Niedrige Erfolgsquote
CD-Rom Format	Fehler	Genauigkeit / Präzision	Lücken	Keine automatische Aktualisierung
Online-Format	Genauigkeit / Präzision	Fehler	Internetverbindung	Vollständigkeit

**Tabelle 5: Nachteile der unterschiedlichen Formate.**

Wie man der Analyse entnehmen kann, widersprechen die mehrmals genannten Vorteile, wie Schnelligkeit und Einfachheit von der negativen Annahme, dass Online-Wörterbücher fehlerhafter, lückenhafter und oberflächlicher als die Printprodukte seien. Das traditionelle Printformat gilt trotz Zeitverlust und Problemen mit der Größe und Schwere nach wie vor als der zuverlässigste Typus.

Es wurden auch spezifische Items zu den Online-Wörterbüchern erarbeitet, wie Items 17, 18, 19, 20, 21, 22, 23, 24. Laut der Befragten sind die folgenden dynamischen Instrumente am wichtigsten: fortgeschrittene Suchmaschine (mit Schreibhilfen, Suche auch in den Definitionen, Suche durch Platzhalter etc.) mit 86%; einfache Suchmaschine mit 49%, alphabetische Leiste mit 13% und das Vorhandensein eines Forums (3%). Verlinkte Verweise werden von 74% der Probanden als positiv eingestuft; während nur 29% sie als negativ beurteilen und die Verlustgefahr (vgl. der Begriff „lost in hyperspace“), Langsamkeit, Unvollständigkeit als Gründe nennen. Interne Links (Verweise zu verbundenen Einträge etc.) werden, wie die nachfolgende Graphik veranschaulicht, als positiv bewertet:

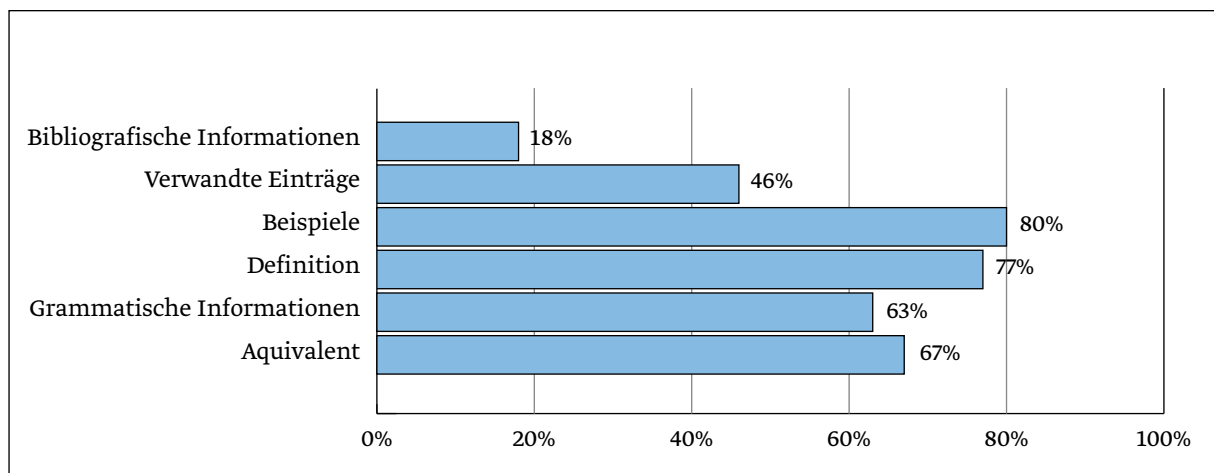


**Bild 2: Graphik zu den bevorzugten Links.**

Sehr interessant sind auch die Ergebnisse zu folgenden Instrumenten, die für Online-Produkte von großer Nützlichkeit sind, wie die Möglichkeit Kontakt mit den Autoren oder dem Wörterbuchteam aufzunehmen, um mögliche Fehler oder Lücken zu signalisieren, oder die Möglichkeit, den Eintrag zu drucken. Nur 10% der Befragten interessieren sich jeweils für beide Aspekte. Legitimierungsmerkmale haben nur für 28% der Versuchspersonen einen Wert, während die Aktualisierung der Wörterbücher (67%), die für viele Testpersonen einen ständigen Fortschritt gemäß der Entwicklung der Sprache bedeutet, von besonderer Wichtigkeit ist.

#### **4.4 Verwendung der Umtexte**

Die Frage 25 betrifft die Untersuchung zur Verwendung von Umtexten durch die Testpersonen, mit Berücksichtigung der Frequenzangaben (oft, manchmal, selten, nie). Die Ergebnisse bestätigen die Vorerwartungen, nämlich, dass wenige Umtexte verwendet werden:



**Bild 3: Graphik zu den verwendeten Umtexten.**

Die am meisten verwendeten Umtexte sind das Abkürzungsverzeichnis und das Register. Selten und kaum werden die Einleitung und die Benutzungshinweise gelesen, was auch zu der obengenannten Einschätzung der Online-Wörterbücher führen kann. Durch eine genauere Lektüre dieser Texte könnten eine irrtümliche Suche und eine falsche Beurteilung der Online-Produkte vermieden werden.

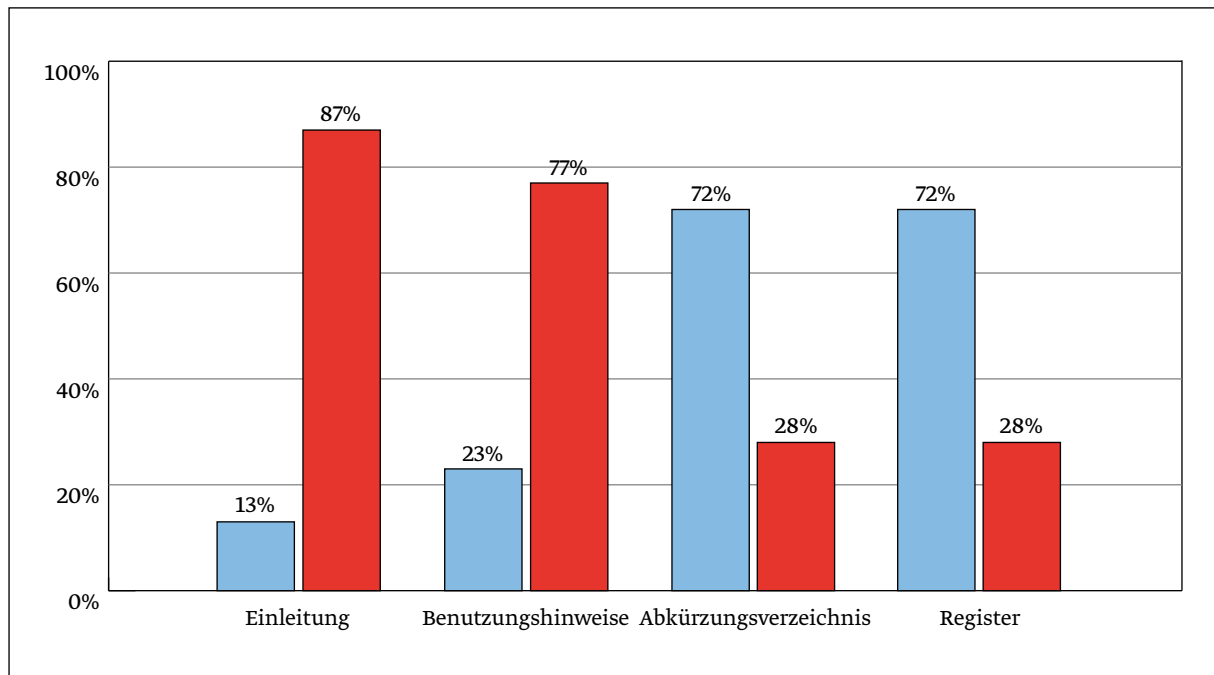
## **5 Funktion und Bedürfnisse des sprachwissenschaftlichen Fachwörterbuches**

Der fünfte Block konzentriert sich auf das sprachwissenschaftliche Wörterbuch. Trotz der negativen Ergebnisse bezüglich der Benutzung von sprachwissenschaftlichen Fachwörterbüchern würden die befragten Probanden, dieses Werkzeug zur Rate ziehen:

- (1) um einen Begriff /Text zu verstehen (66%);
- (2) um ein Wort oder einen Text zu übersetzen (33%);
- (3) um einen Text zu produzieren (21%);
- (4) um sich über ein Thema zu informieren (13%);
- (5) um bibliografische Informationen zu suchen (3%);
- (6) um ein Problem zu lösen (3%).



Folgende Angaben in der Mikrostruktur von sprachwissenschaftlichen Wörterbüchern werden von den Testpersonen signalisiert:



**Bild 4: Graphik zu den mikrostrukturellen Angaben eines sprachwissenschaftlichen Wörterbuches.**

Die Definition des Eintrages und das Vorhandensein von Beispielen werden als am wichtigsten eingeschätzt. Das Hinzufügen von grammatischen Informationen (wie Genus und Numerus), sowie des fremdsprachlichen Äquivalentes (eventuell auch mit grammatischen Informationen) sind auch in großen Maßen erwünscht. Die Angabe von verwandten Einträgen wird von 46% der Befragten hervorgehoben. Viel seltener wird das Hinzufügen von bibliographischen Informationen (18%) genannt. Als weitere Angaben werden folgende Bereiche genannt: Phonetik (54%); Kollokationen (54%); Rechtschreibung (46%); Synonyme und Antonyme (41%); Etymologie (31%); Markierung des Fachbereiches (18%). Eine weitere Angabe ist die phonetische Transkription.

## 6 Schlussbemerkungen

Die empirische Untersuchung liefert ein Bild über die Wörterbuchbenutzung der Probanden und über ihre Einstellung zu den unterschiedlichen Typen und Formaten von Wörterbüchern. Die Tatsache, dass die meisten Testpersonen zweisprachige Wörterbücher verwenden, ist keine Überraschung, erstaunlich ist jedoch die Tatsache, dass sie kaum Fachwörterbücher und insbesondere sprachwissenschaftliche lexikographische Produkte verwenden.

Die Ergebnisse zu den bevorzugten Formaten hätte man auch nicht erwartet: in einem Zeitalter, in dem Smartphones, Tablets und Internet im Alltag dominieren, ist die Zahl der benutzten Printprodukte jedoch sehr hoch. Die Vorteile und Nachteile der jeweiligen Formate bezeugen weiterhin, dass die Internetlexikographie noch einiges machen muss, um bessere und zuverlässigere Produkte zu entwickeln. Trotz wichtiger Projekte und Wörterbücher sind noch viele Produkte verbesserungsfähig und können sich nicht der soliden Basis wissenschaftlicher und theoretischer Erkenntnisse entziehen.

Die Analyse hat auch ergeben, dass Umtexte im Online-Medium (außer Abkürzungsverzeichnisse und Register) noch zu wenig benutzt werden, was vermutlich mit ihrer „Wissenschaftlichkeit“ in Zusammenhang gebracht werden kann. Das Augenmerk der Probanden auf das Copyright (oft mit Professionalität in Verbindung gesetzt) und insbesondere auf die zeitnahe Aktualisierung der Produkte deutet darauf hin, dass ein stärkerer Wunsch nach Produkten in ständiger Entwicklung besteht.

Interessant waren auch die Ergebnisse zu den mikrostrukturellen Eigenschaften eines sprachwissenschaftlichen Wörterbuchs: die Präsenz sowohl von grammatischen Informationen, Äquivalenzangaben (Sprachinformationen) als auch von Definitionen (Sachinformationen) deutet auf den Typ „Allbuch“ (Wiegand 1998: 762) hin. Der Wunsch, dass auch grammatische Informationen, phonetische Angaben zu den Äquivalenten sowie Gebrauchsbeispiele vorhanden sein sollten, zeigt, dass Fachwörterbücher diese mikrostrukturellen Eigenschaften stärker berücksichtigen sollten.

Außerdem sollte der Frage nachgegangen werden, ob und in wie fern die Studenten des Masterstudienengangs von diesen Ergebnissen abweichen.

## 7 Literaturangaben

- Atkins, B.T.S., Varantola, K. (1998). Language learners using dictionaries: The final report on the EURALEX/AILA research project on dictionary use. In *Using dictionaries: Studies of dictionary use by language learners and translators*, S. 83-122.
- De Schryver, G.M., Prinsloo, D.J. (2000). Dictionary-Making Process with ‘Simultaneous Feedback’ from the target Users to the Compilers. In U. Heid et al. (Hrsg.) *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000, Stuttgart, Germany, August 8th-12th, 2000*. Stuttgart: Niemeyer, S. 197-209.
- De Schryver, G.M., Prinsloo, D.J. (2011). Do Dictionaries Define on the Level of Their Target Users? A Case Study for Three Dutch Dictionaries. In *International Journal of Lexicography* 24/1, S. 5-28.
- Domínguez Vázquez, M.J. (2006). Von monolingualen Wörterbüchern zu kontrastiven Valenzwörterbüchern. Die Valenzwiedergabe unter der Lupe. In *Jahrbuch für Deutsch als Fremdsprache. Intercultural German Studies* 1/32, S. 231-241.
- Domínguez Vázquez, M.J., Balsa, M.M., Vidal Pérez, V. (2013). *Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen*. [http://www.academia.edu/4640188/Worterbuchbenutzung\\_Erwartungen\\_und\\_Bedurfnisse\\_Ergebnisse\\_einer\\_Umfrage\\_bei\\_Deutsch\\_lernenden\\_Hispanophonen](http://www.academia.edu/4640188/Worterbuchbenutzung_Erwartungen_und_Bedurfnisse_Ergebnisse_einer_Umfrage_bei_Deutsch_lernenden_Hispanophonen)
- Ekwa Ebanéga, G-M., Tomba Moussavou, F. (2008). Survey of Dictionary Use: Case Studies of Gabonese Students at the University of Stellenbosch and the Cape Peninsula University of Technology. In *Lexikos*, 18, S. 349-365.
- Engelberg, S., Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

- Flinz, C. (2011): DIL (Deutsch-Italienisches Wörterbuch der Linguistik). Vom Projekt zur Realität: Hinweise zum aktuellen Stand. In: *daf Werkstatt*, S. 185-200.
- Hartmann, R.R.K. (1989). Sociology of The Dictionary User: Hypotheses and Empirical Studies. In F.J. Hausmann et al. (Hrsg.) *Wörterbücher: ein Internationales Handbuch zur Lexikographie*. Bd. 1. Berlin, New York: de Gruyter, S. 649-657.
- Hartmann, R.R.K. (1999). Case Study: The Exeter University Survey of Dictionary Use [Thematic Report 2]. In R.R.K. Hartmann (Hrsg.) *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the TNP Sub- Project 9: Dictionaries*. Berlin: Thematic Network Project in the Area of Languages, S. 36-52. Accessed at: <http://www.fu-berlin.de/elc/TNPproducts/SP9dossier.doc> [07/03/2014].
- Hartmann, R.R.K. (2000). European Dictionary Culture. The Exeter Case Study of Dictionary Use among University Students, against the Wider Context of the Reports and Recommendations of the Thematic Network Project in the Area of Languages (1996- 1999). In U. Heid et al. (Hrsg.) *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000*. Stuttgart: Niemeyer, S. 385-391.
- Hartmann, R.R.K. (2001). *Teaching and Researching Lexicography*. London: Pearson.
- Hausmann, F.J., Reichmann, O., Wiegand, H.E., Zgusta, L. (Hg.) (1989). *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.1. Erster Teilband. Berlin, New York: de Gruyter.
- Kemmer, K. (2010). *Onlinewörterbücher in der Wörterbuchkritik* (OPAL 2-2010). Mannheim. (= Online publizierte Arbeiten zur Linguistik)
- Kromann, H.P. (1995). Deutsche Wörterbücher aus der Perspektive eines fremdsprachigen Benutzers. In H. Popp (Hrsg.) *Deutsch als Fremdsprache. An den Quellen eines Faches. Festschrift für Gerhard Helbig zum 65. Geburtstag*. München: Iudicium, S. 501-512.
- Kühn, P. (1989). Typologie der Wörterbücher nach Benutzungsmöglichkeiten. In F.J. Hausmann et al. (Hrsg.) *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.1. Erster Teilband. Berlin, New York: de Gruyter, S. 111-127.
- Müller-Spitzer, C., Koplenig, A., Töpel, A. (2012). Online dictionary use: Key findings from an empirical research project. In S. Granger, M. Paquot (Hg.) *Electronic Lexicography*. Oxford: Oxford University Press, S. 425-457.
- Muráth, J. (2005). Wörterbuchbenutzung von Fachübersetzerstudenten Ihre Erwartungen an ein Fachwörterbuch. In *Lexicographica*, 115, S. 401-415.
- Nesi, H. (2000). Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: the State of the Art. In U. Heid et al. (Hrsg.) *Proceedings of the Ninth EURALEX International Congress Stuttgart August 8.-12. 2000*. 1. Halbbd. Stuttgart: Niemeyer, S. 839-847.
- Retti, G. (2004). *Österreichisches Deutsch und Österreichisches Wörterbuch*. Accessed at: <http://gregor.retti.info/oewb/> and <http://gregor.retti.info/docs/retti1991/4.pdf> [07/03/2014].
- Ripfel, M. (1989a). Wörterbuchkritik eine empirische Analyse von Wörterbuchrezensionen. Tübingen: Niemeyer.
- Ripfel, M., Wiegand, H.E. (1988). Empirische Wörterbuchbenutzungsforschung. In *Studien zur neuhochdeutschen Lexikographie VI*. 2. Teilbd. Hildesheim: Olms, S. 91-520. (Germanistische Linguistik 87-90/1986)
- Rossenbeck, K. (2005). Die zweisprachige Fachlexikographie in der neueren und neuesten Wörterbuchforschung“. In *Lexicographica*, 21, S. 179-201.
- Sánchez Ramos, M. del Mar (2005). Research on Dictionary Use by Trainee Translators. In *Translation Journal*, 9.2. Accessed at: <http://www.proz.com/translation-articles/articles/227/1/Research-on-Dictionary-Use-by-Trainee-Translators> [19.03.2014].
- Simonsen, H.K. (2011). User Consultation Behaviour in Internet Dictionaries: An Eye-Tracking Study. In *Hermes. Journal of Language and Communication Studies*, 46, S. 75-101.

- Taljar, E., Prinsloo, D.J., Fricke, I. (2011). The use of LSP dictionaries in secondary schools? A South African case study. In *South African Journal of African Languages*, 31.1, S. 87-109.
- Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Max Niemeyer.
- Wang, W. (2001). *Zweisprachige Fachlexikographie. Benutzungsforschung, Typologie und mikrostrukturelle Konzeption*. Frankfurt a. M.: Lang.
- Welker, H.A. (2010). *Dictionary Use. A General Survey of Empirical Studies*. Brasilia: Author's Edition.
- Wiegand, H.E. (1987). Zur handlungstheoretischen Grundlegung der Wörterbuchbenutzungsforschung. In *Lexicographica*, 3, S. 178-227.
- Wiegand, H.E. (1995). Lexikographische Texte in einsprachigen Lernerwörterbüchern. Kritische Überlegungen anlässlich des Erscheinens von Langenscheidts „Grösswörterbuch Deutsch als Fremdsprache“. In H. Popp (Hrsg.) *Deutsch als Fremdsprache. An der Quellen eines Faches. Festschrift für Gerhard Helbig zum 65. Geburtstag*. München: Iudicium, S. 463-499.
- Wiegand, H.E. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. 1 Teilband. Berlin, New York: de Gruyter.
- Wiegand, H.E. (2008). Wörterbuchbenutzung bei der Übersetzung. Möglichkeiten ihrer Erforschung. In *Germanistische Linguistik*, 195-96, S. 1-43.
- Wiegand, H.E. (2010). Zur Methodologie der Systematischen Wörterbuchforschung: Ausgewählte Untersuchungs- und Darstellungsmethoden für die Wörterbuchform. In *Lexicographica*, 26, S. 249-330.

# Translation, Cultural Adaptation and Preliminary Psychometric Evaluation of the English Version of “Strategy Inventory for Dictionary Use” (S.I.D.U)

Gavriilidou Zoe  
Democritus University of Thrace  
zoegab@otenet.gr

## Abstract

The present paper reports results regarding the adaptation in English of the Strategy Inventory for Dictionary Use (S.I.D.U.) and the preliminary psychometric evaluation of the English version. S.I.D.U is a 36-item self-report questionnaire for assessing dictionary use strategies specifying four main areas of interest: a) dictionary use awareness skills, b) strategies for dictionary selection and acquaintance with dictionary conventions, c) lemmatization strategies, and finally d) look-up strategies. The original scale was translated from Greek into English, back-translated and reviewed. Cross-cultural adaptation included the experts' revision followed by its administration to 52 participants. Its internal consistency was .89. Similarly all four subscales showed good to excellent internal consistency (dictionary use awareness:  $\alpha=.83$ , dictionary selection and acquaintance:  $\alpha=.76$ , lemmatization:  $\alpha=.86$  and dictionary search:  $\alpha=.78$ ). Test-retest reliability ranged from fair to good for the total scale and its six-subcales

**Keywords:** Assessment; validation; pedagogical Lexicography; strategic dictionary use; dictionary selection strategies; dictionary acquaintance strategies; lemmatization strategies; dictionary search

## 1 Introduction

There is in recent literature a growing interest in pedagogical lexicography and more precisely in the study of *use situations* of dictionaries (Béjoint 1989; Gavriilidou 2010; Petrylaité, Vezyté, Vaskeliené 2008; Prichard 2008; Scholfield 2002), the *dictionary-using skills and strategies* (Ard 1982; Bogaards 1994; Diab 1990; Gavriilidou 2014; Hartmann 1994) or *the role of dictionary in pedagogical process* (Bensoussan 1983; Bogaards 1998; Hulstjin 1993; Nesi 1996). Relevant research (Gu & Johnson 1996; Nation 2001; Schmitt 1997) has also shown that dictionary use is an important vocabulary strategy that a) occurs successfully in conjunction with guessing (or inferencing) and note-taking, b) provides information about a specific item, and c) has a positive influence on the learner's acquisition process (Hulstijn 1993; Luppescu & Day 1993; Knight 1994; Laufer & Hadar 1997; Laufer & Hill 2000; Bruton 2007). No previous research, however, has focused on the relationship between *strategic dictionary use* and successful reading comprehension, production or vocabulary acquisition even though current trends in langua-

ge curricula design stress the importance of strategy use in language teaching. In the current study strategic dictionary use refers to a) the conscious awareness of when to use a dictionary and what type to use and b) the ability to employ efficient lemmatization and look-up strategies. Given that strategic dictionary use is beneficial in vocabulary acquisition, L2 Learning, reading comprehension (Scholfied 1982; Knight 1994; Hulstijn, Hollander and Greidanus 1996; Scholfield 1999; Prichard 2008) or text production (Nesi and Meara 1994; Fraser 1999; Elola, Rodríguez-García and Winfrey 2008), there is a major claim in pedagogical lexicography that strategic dictionary use should be taught and, as a result, dictionary users should develop a more effective strategic behavior while looking up words. However, strategic dictionary use instruction, while crucial, still remains a secondary concern in the relevant research. It is true, on the other hand, that the design of strategic dictionary use instruction programs has to be based on reliable data describing dictionary user's attitudes, preferences or strategies employed while they look up a word. In order to accurately explore the needs of dictionary users, measures that produce valid and reliable estimates of user's dictionary strategies must be identified. However, there was, until recently, no standardized instrument for profiling users' dictionary strategies in a valid and reliable way.

The purpose of this paper is to present the structure and characteristics of the English version of the newly developed self-report "Strategy Inventory for Dictionary Use" (S.I.D.U) (Gavriilidou 2011; 2013) which was first elaborated and standardized in Greek Language for profiling dictionary users in a valid and reliable manner. An assessment of strategic dictionary use which offers valid and reliable scores will further develop our understanding of the construct of strategic dictionary use and its relation to vocabulary acquisition, and text comprehension or production. The use of self-report instruments to investigate various aspects of individual learner differences is a common practice in the field of language learning research. However, although a given instrument may have been rigorously developed and subjected to various measures of reliability and validity, when it is translated into another language or used in a cultural setting different from the one originally intended, it must once again be rigorously examined. The cross-cultural adaptation of self-report questionnaires for use in a new country, culture or language necessitates use of a rigorous protocol in order to reach equivalence between the original source and target version. Furthermore, the items must not only be translated well linguistically, but also must be adapted culturally to maintain the content validity of the instrument at a conceptual level across different cultures (Beaton et al 2000). Thus, this paper also provides data about S.I.D.U's translation and cultural adaptation in English and focuses on following an appropriate adaptation protocol that would maximize the questionnaire's reliability and validity when used to compare the scores across cultures and languages. Finally the paper reports results regarding the instruments' reliability and validity.

## 2 The Strategy Inventory for Dictionary Use

The S.I.D.U (Gavriilidou 2011, 2013) is a standardized self-report instrument first elaborated in Greek for assessing the frequency with which the respondent uses different strategies or techniques during dictionary use. It can be administered since the age of 11 years old. It consists of 36 items with five Likert-scale responses of never or almost never true of me, generally not true of me, somewhat true of me, generally true of me, always or almost always true of me.

Given that strategic dictionary use is part of a larger construct of strategy use, we approached this instrument design following the development procedure of the Strategy Inventory for Language Learning (Oxford 1990). In order to develop the test's specification of the Greek version, all previous literature was consulted in detail and an exhaustive list including all reference skills cited in the literature was prepared. That list was used as a basis for item writing. More specifically two different kinds of research had been consulted: theoretical or empirical papers presenting detailed descriptions or taxonomies of the reference skills (or strategies) that dictionary users should demonstrate for a successful dictionary search (e.g. Béjoint 1981; Scholfield 1982 and 1999; Bogaards 1994; Roberts 1997; Hartmann 1999; Nesi 1999; Nation 2001; Hartmann and James 2002; Lew and Galas 2008) and empirical papers investigating the reference skills, misuse and errors of dictionary users during dictionary look-up (Béjoint and Moulin 1987; Maingay and Rundell 1987; Neubach and Cohen 1988; Nuccorini 1992 and 1994; Nesi and Meara 1994; Christianson 1997; Harvey and Yuill 1997; Wingate 2004; Elola, Rodríguez-García and Winfrey 2008; Petrylaitė, Vaškeliene, Vėžytė 2008). The method of multiple judges was adopted for the measurement of content validity of the pilot version of S.I.D.U. The measurement was carried out on a panel of 10 experts who were either University Professors specialized in Lexicology or Lexicography with a long experience in dictionary compilation or University Professors specialized in Language Teaching. The experts judged the relevance and usefulness of each one of the 52 candidate items of S.I.D.U and included 47 items in the pilot version.

To check the construct validity of the original S.I.D.U, principal component analysis with Varimax rotation on SPSS version 15 was adopted. A communality of .30 was set as a cut off for inclusion in the final analysis. Consequently, eleven items were excluded. The results showed that a total of 36 items loaded on four factors, accounting for the 51.7% of the variance. Based on these results, Gavriilidou (2013) organized the original S.I.D.U into four strategy subscales:

- Strategies which lead to a decision to use a dictionary in order to resolve a problem encountered inside or outside the class (dictionary use awareness) (items 1-14).
- Strategies which permit to select an appropriate dictionary type depending on the problem to be solved and guarantee the acquaintance with one's own dictionary (15-21).
- Lemmatization strategies, that is strategies which help finding the citation form of inflected forms found in the text. Users should be able to be based on morphological indices (stems, prefixes, suffixes, inflectional morphemes) of the unknown word that has been met in the text in order to make hypotheses about the look-up form of that word or should be acquainted with alphabetical sequencing otherwise lemmatization is not possible (22-29).

- Look up strategies, which control and facilitate the localization of the correct part of the entry where different meanings of the same word form are included (30-36)

Each of the four factors was considered according to previous literature and was named by the author. The test discriminated expert from non-expert users in all four categories of strategies ( $p < .001$ ). Internal consistency of the four subscales (dictionary use awareness, dictionary selection, acquaintance and lemmatization and dictionary search) and the overall scale of the S.I.D.U. were found to be excellent.

### **3 The Translation and Adaptation Protocol of S.I.D.U in English**

The process of adaptation of S.I.D.U into English was broken down into three steps: (a) the translation process, (b) cross-cultural verification and adaptation, and (c) verification of the psychometric properties of the instrument. The translation process consisted of the initial translations, synthesis of the translations and back translation. The second step included the expert committee review in the light of the focus group suggestions and other verification methods. Finally, in the third stage, the questionnaire was administered and its psychometric properties were verified.

The translation protocol was broken down into six stages: initial translation by two independent translators; synthesis of the translations during which any discrepancies between the two initial translations are resolved; back translation into the original language; expert committee review which should achieve semantic, idiomatic, experiential and conceptual equivalence; pretesting of the final version; and, finally, submission of final reports drawn for all the five stages to the coordinating committee (Beaton et al., 2000).

#### **3.1 The initial translation**

The first stage in the adaptation was the forward translation of S.I.D.U from Greek into English. Two bilingual translators living in Greece, whose mother tongue was English, one naïve and one informed about the purpose of the study produced two independent translations (T1 and T2). They also composed two independent written reports in which they explained the rationale of their translation choices as well as dubious phrases, uncertainties or encountered translation problems.

#### **3.2 The synthesis of the translations**

The two translations were compared and discrepancies reflecting ambiguities in the original instrument were noted. Then the two translators and the creator of the instrument worked on the first translator's (T1) and the second translator's (T2) versions and produced a synthesis of the two versions (T12) by discussing the translation of each of the 36 items. They also wrote a report describing the synthesis procedure, all cases discussed and all solutions adopted.



### **3.3 Back translation**

Two translators having Greek as mother tongue and ignoring the original version of S.I.D.U translated the T12 version back into Greek in order to verify that the translated T12 version in English reflects the same item content as the original Greek version. The two translators, who were not informed of the purpose of the study, handed in the two back translations (BT1 and BT2) as well as two independent reports documenting the back translation procedure.

### **3.4 Expert committee examination**

The expert committee consisted of two linguists, two lexicographers and the four translators. Its role was to examine all the relevant material (initial instrument, T1, T2, T12, BT1, BT2, and the five reports) and to review all the translations for resolving any discrepancy. Its final goal was to arrive to the final English version of the S.I.D.U. To do so, the experts counter-examined the source and target version of S.I.D.U checking the following: a) the semantic equivalence, that is if the words meant the same in Greek and English and whether there was any grammatical difficulties in English translation, b) the idiomatic equivalence, in other words the correct translation of idioms or collocations c) the experiential equivalence, in other words if all items expressed tasks which are experienced in the target culture d) the conceptual equivalence, that is if all the words hold the same conceptual meaning in the two cultures.

The committee produced the final English version of S.I.D.U and wrote a final report which they handed to the author of the instrument. This version was then used for collecting data in order to measure the psychometric properties of the instrument.

## **4 Reliability**

### **4.1 Sampling**

52 under graduate and post graduate students as well as professors of the department of Linguistics of the University of Chicago filled in the questionnaire.

### **4.2 Statistics**

To check the S.I.D.U's internal consistency a Cronbach's Alpha analysis was performed. To check the stability of S.I.D.U scores over time, test-retest data are reported and the intra-class correlation coefficient was computed.

## 5 Results

### 5.1 Internal Consistency

Based on the results of S.I.D.U, a total sum score of all 36 items was computed. Moreover, total scores in each subscale (dictionary use awareness, dictionary selection and acquaintance, lemmatization and dictionary search) were also computed. The total scale showed excellent reliability (Cronbach's  $\alpha = .89$ ). Similarly all four subscales showed good to excellent internal consistency (dictionary use awareness:  $\alpha = .83$ , dictionary selection and acquaintance:  $\alpha = .76$ , lemmatization:  $\alpha = .86$  and dictionary search:  $\alpha = .78$ ).

### 5.2 Test-retest reliability

Test-retest reliability for the total scale and the sub-scales ranged from fair to good (Total scale:  $r = .778$ ,  $p < .001$ , dictionary use awareness:  $r = .831$ ,  $p < .001$ , dictionary selection and acquaintance:  $r = .874$ ,  $p < .001$ , lemmatization:  $r = .761$ ,  $p < .001$ , dictionary search:  $r = .696$ ,) indicating that at least within the time frame considered here scores of S.I.D.U mirror stable individual differences.

## 6 Discussion

The present article reports findings concerning the validity and reliability of the translated and culturally adapted in English version of S.I.D.U. Like the original Greek version of the instrument whose internal consistency was found to be excellent (total scale:  $\alpha = .94$ , dictionary use awareness:  $\alpha = .90$ , dictionary selection and acquaintance:  $\alpha = .86$ , lemmatization:  $\alpha = .83$  and dictionary search:  $\alpha = .84$ ) (Gavriilidou 2013: 12), the translated version showed an excellent reliability for the total scale  $\alpha = .89$  and all four subscales (dictionary use awareness:  $\alpha = .83$ , dictionary selection and acquaintance:  $\alpha = .76$ , lemmatization:  $\alpha = .86$  and dictionary search:  $\alpha = .78$ ). Thus the paper, provides evidence for the English version of S.I.D.U as a useful and psychometrically sound measure of dictionary use strategies that may contribute to the scientific investigation of the strategies employed by dictionary users while choosing and using a dictionary, as well as for applied purposes such as the design of class interventions for raising strategic dictionary use. The purpose for developing the English version of S.I.D.U. was to provide a simple-to administer and reliable instrument for assessing strategic dictionary use cross-linguistically. The fact that the S.I.D.U. was found to be valid and reliable both in the Greek and English version is very promising in that regard.

The paper also records an appropriate adaptation protocol that would maximize the questionnaire's reliability and validity when used to compare the scores across cultures and languages. There were

attempts to reduce the potential biases that may occur during translation. Construct and item bias were recorded and were confronted appropriately in order to overcome the problem of measuring different constructs in different cultural groups or distorting the meaning of individual items. That is why “adaptation” and not “application” or “assembly” was selected as it allows for a solution to the afore-mentioned problems of bias. It can be concluded that the process of adapting the S.I.D.U from Greek into English recorded in this paper, however time consuming and costly, is the most effective way to produce an instrument for measuring the frequency of dictionary strategy use of dictionary users. It also allows for comparison of data and findings across nations as it provides the opportunity to examine dictionary strategies of those for whom there previously was no translated version of the S.I.D.U. The carefully planned and executed adaptation process ensures high instrument reliability and validity and offers other researchers interested in questionnaire adaptation a procedure that overcomes most of the problems entailed when instruments are used in different languages and cultures.

The major application of the English version of S.I.D.U. is to assess the dictionary use strategies employed by students or pupils in order to collect reliable data for the design of special curricula for dictionary use training. It can also be used to assess the improvement in dictionary use as a result of the application of these curricula in specific target groups. Furthermore, it can be used as an instrument of sample normalization in research focusing on the role of dictionary use in vocabulary acquisition and on the relationship between the dictionary use and successful reading comprehension or text production, ensuring that different samples of different researches include dictionary users with equivalent abilities in such a way that would yield comparable results. Finally, another possible use is for research purposes on pedagogical lexicography.

## **7 Conclusions-Limitations of the study**

The main contribution of the present empirical study is data about the psychometric properties of the English version of S.I.D.U and the proposed model of questionnaire adaptation, which involves methodological and theoretical considerations necessary for researchers who will adapt or develop relevant tests for various constructs. The proposed model covers empirical, methodological, and theoretical issues. Theoretical issues were addressed in the stage of construct definition. Methodologically, an approach for construct validation was suggested. In short, this model includes different steps and procedures for adapting or developing tests for questionnaires, while still being able to produce instruments that are valid and reliable.

However, it needs to be pointed out that the cultural adaptation procedure carried out in this study has focused on adults, students or professors. This was in line with our need to develop a screening instrument for this particular population, since most of the relevant studies focus on that population. Therefore, unlike the original Greek version of S.I.D.U which can be administered since the age of 11,

this particular translation may not be applicable to other age groups, and would need to be reviewed prior to generalized use.

Finally, the instruments' construct validity should be checked with a Factor Analysis using larger samples.

## 8 References

- Ard, J. (1982). The use of bilingual dictionaries by ESL students while writing. In *Review of Applied Linguistics* 58, pp. 1-27.
- Beaton, E. D., Bomardier, C., Guillemeni, F. & Bozi-Ferrz, M. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. In *Spine*, 25(24), pp. 3186-3191.
- Béjoint, H. (1981). The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. In *Journal of Applied Linguistics*, II (3), pp. 207-222.
- Béjoint, H. (1989). *The teaching of dictionary use: present state and future tasks*. Dictionaries. Berlin, New York: Walter de Gruyter. 1st vol. pp. 208-215.
- Béjoint, H. and Moulin, A. (1987). The place of the dictionary in an EFL programme. In A. Cowie (ed.), *The dictionary and the language learner: Papers from the EURALEX seminar at the University of Leeds*, 1-3 April 1985. Tübingen: Max Niemeyer Verlag, pp. 97-114.
- Bensoussan, M. (1983). Dictionaries and tests of EFL reading comprehension. In *English Language Teaching Journal* 37(4), pp. 341-345.
- Bogaards, P. (1994). Tuning the dictionary to the skills of intermediate learners. In *Fremdsprachen Lehren und Lernen* 23, pp. 192-205.
- Bogaards, P. (1998). Scanning long entries in Learner's Dictionaries. *Proceedings Euralex* 98, Liège, pp. 555-563.
- Bruton, A. (2007). Vocabulary learning from dictionary referencing and language feedback in EFL Translational Writing. In *Language Teaching Research*, 11, pp. 413-431.
- Christianson, K. (1997). Dictionary use by EFL writers: what really happens. In *Journal of second language writing*, 6, pp. 23-43.
- Diab, T. (1990). Pedagogical Lexicography: a case study of Arab nurses as dictionary users. *Lexicographica Series Maior* 31. Tübingen: Niemeyer.
- Elola, I., Rodríguez-García, V. and Winfrey, K. (2008). Dictionary use and vocabulary choices in L2 Writing. In *Estudios de lingüística Inglesa aplicada*, 8, pp. 3-89.
- Fraser, C. (1999). The Role of Consulting a Dictionary in Reading and Vocabulary Learning. In *Canadian Journal of Applied Linguistics*, 2(1-2), pp. 73-89.
- Gavriilidou, Z. (2010). Profiling Greek adult dictionary users. In *Studies of Greek Linguistics* 31, pp. 166-172.
- Gavriilidou, Z. (2011). Strategy inventory for dictionary use: Elaboration and Standardization. In Gavriilidou, Z., Efthymiou, A., Kambakis-Vougiouklis, P. and Thomadaki, E. (eds) *Proceedings of the 10th International Conference of Greek Linguistics* available at <http://www.icgl.gr>.
- Gavriilidou, Z. (2013). Development and validation of the Strategy Inventory for Dictionary Use (S.I.D.U). In *International Journal of Lexicography*, 26 (2), pp. 135-153.
- Gavriilidou, Z. (2014). User's abilities and performance in dictionary look up. In Lavidas, N., Alexiou, Th. and Sougari, A. (eds) *Major Trends in Theoretical and Applied Linguistics* vol. 2, De Gruyter Open, pp. 41-52 (available at <http://www.degruyter.com/viewbooktoc/product/422023>).
- Gu, P. Y & Johnson, R. K. (1996). Vocabulary learning strategies and language learning outcomes. In *Language Learning*, 46, pp. 643-679.

- Hartman, R. R. K. (1987). Four perspectives on dictionary use: A critical review of research methods. In A. Cowie (ed) *The dictionary and the language learner*, Tübingen, Max Niemeyer Verlag, pp. 11-28.
- Hartmann, R. R. K. (1994) Bilingualised versions of learners' dictionaries. *Fremdsprachen Lehren und Lernen* 23, pp. 206-220.
- Hartmann, R. R. K. and James, J. (2002). *Dictionary of Lexicography*, Routledge.
- Harvey, K. and Yuill, D. (1997). A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. In *Applied Linguistics*, 18(3), pp. 253- 278.
- Hulstijn, J. H. (1993). When do foreign language readers look up the meaning of unfamiliar words? The influence of task and learner variables. In *The Modern Language Journal* 7 (2), pp. 139-147.
- Hulstijn, J-H., Hollander, M. and Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use and re-occurrence of unknown words. In *The Modern Language Journal*, 80(3), pp. 327-339.
- Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. In *The Modern Language Journal*, 78, pp. 285-299.
- Laufer, B., & Hadar, L. (1997). Assessing the effectiveness of monolingual, bilingual and 'bilingualized' dictionaries in the comprehension and production of new words. In *The Modern Language Journal*, 81, pp. 189-196.
- Laufer, B., & Hill, M. (2000). What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? In *Language Learning and Technology*, 3, pp. 58-76.
- Lew, R. and Galas, K. (2008). Can dictionary use be taught? The effectiveness of lexicographic training for primary school level Polish learners of English. In E. Bernal and DeCesaris J. (eds.), *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 1273-1285.
- Lupescu, S. & Day, R. R. (1993). Reading, dictionaries and vocabulary learning. In *Language Learning*, 43, pp. 263-287.
- Maingay, S. and Rundell, M. (1987). Anticipating learners' errors: Implications for dictionary writers. In A. Cowie (ed.), *The dictionary and the language learner: Papers from the EURALEX seminar at the university of Leeds, 1-3 April 1985*. Tübingen: Max Niemeyer Verlag, pp. 128-35.
- Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press
- Nesi, H. (1996). The role of illustrative examples in productive dictionary use. In *Dictionaries* 17, pp. 198-206.
- Nesi, H., & Hail, R. (2002). A study of dictionary use by international students at a British university. In *International Journal of Lexicography*, 15(4), pp. 277-305.
- Nesi, H. and Meara, P. (1994). Patterns of misinterpretation in the productive use of EFL dictionary definitions. In *System* 22, pp. 1-15.
- Neubach, A. and Cohen, A. D. (1988). Processing strategies and problems encountered in the use of dictionaries. *Dictionaries: In The Journal of the Dictionary Society of North America*, 10, pp. 1-19.
- Nuccorini, S. (1992). Monitoring dictionary use. In H. Tommola (ed.), *EURALEX '92 Proceedings*, *Studia Translatologica*. Tampere, pp. 89-102.
- Nuccorini, S. (1994). On dictionary misuse. In W. Martin (ed.) *EURALEX 1994 Proceedings*. Vrije Universiteit Amsterdam, pp. 586-597.
- Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House / Harper and Row. Now Boston: Heinle and Heinle.
- Petrylaitė, R., T. Vėžytė, Vaškeliėnė, D. (2008). Changing skills of dictionary use. In *Studies about languages*, 12, pp. 77-82.
- Prichard, C. (2008). Evaluating L2 readers' vocabulary strategies and dictionary use. In *Reading in a foreign language*, 20(2), pp. 216-231.
- Roberts, R. (1997). Using Dictionaries Efficiently. In *38th Annual Conference of the American Translators Association*, San Francisco, California. 20 February 2012. <http://www.dico.uottawa.ca/articles-en.htm>.
- Schmitt, N. (1997). Vocabulary learning strategies». In Schmitt & McCarthy (eds) *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge: Cambridge University Press. pp. 199-227.

Scholfield, P. (1982). Using the English Dictionary for Comprehension. In *TESOL Quarterly* 16(2). pp. 185-194.  
 Scholfield, P. (1999). Dictionary Use in Reception, In *International Journal of Lexicography*, 12(1), pp. 13-34.  
 Scholfield, P. (2002). Why Shouldn't Monolingual Dictionaries be as Easy to Use as Bilingual Ones? [Retrieved June 10, 2010, from <http://www.longman.com/dictionaries/teachers/articles/p-scholfield-02.html>].  
 Wingate, U. 2004. Dictionary use - the need to teach strategies. In *Language Learning Journal*, 29, pp. 5-11.

### Acknowledgements

This study is part of the Thales project MIS 379335. It was held in the frame of the National Strategic Reference Frame (Ε.Σ.Π.Α) and was co-funded by resources of the European Union (European Social Fund) and national resources.

### Appendix: English version of S.I.D.U

Name (not surname):

Gender:

Date of birth:

Mother Tongue:

Career orientation:

This questionnaire will be used for research purposes and your contribution is very significant. Thank you for your help. Please read the following statements carefully and circle 1, 2, 3, 4 or 5 according to what is most true for you.

- (1) Never or almost never true of me.
- (2) Generally not true of me.
- (3) Somewhat true of me.
- (4) Generally true of me.
- (5) Always true of me.

I use a dictionary to find the meaning of a word	1	2	3	4	5
I use a dictionary to find the spelling of a word	1	2	3	4	5
I use a dictionary to find synonyms	1	2	3	4	5
I use a dictionary to find antonyms	1	2	3	4	5
I use a dictionary to check how a word is used	1	2	3	4	5
I use a dictionary to find the origin of a word	1	2	3	4	5
I use a dictionary to help myself in translation	1	2	3	4	5
I use a dictionary to find the syntax of a word	1	2	3	4	5
I use a dictionary to find the derivatives of a word	1	2	3	4	5
I use a dictionary to find word families	1	2	3	4	5
I use a dictionary to find the meaning of an expression	1	2	3	4	5

I use a dictionary at home	1	2	3	4	5
I use a dictionary when I read a text	1	2	3	4	5
I use a dictionary when I write a text	1	2	3	4	5
Before I buy a dictionary, I know the reason why I need it	1	2	3	4	5
Before I buy a dictionary at the bookshop, I glance through it to see what information it provides.	1	2	3	4	5
I choose a dictionary because it has a lot of entries and a lot of information in each entry.	1	2	3	4	5
I know what an etymological dictionary is and what it is used for	1	2	3	4	5
I know what a general dictionary is and what it is used for	1	2	3	4	5
I know what a bilingual dictionary is and what it is used for	1	2	3	4	5
I know what a dictionary of technical terms is and what it is used for	1	2	3	4	5
Before I use my new dictionary, I carefully read the introduction	1	2	3	4	5
Before I use my new dictionary, I carefully study the list of abbreviations	1	2	3	4	5
When I come across an unknown word in a text, I try to think in what form I should look it up in the dictionary.	1	2	3	4	5
When I can't locate a proverb or a set phrase in the entry where I thought I would find it, I begin a new search	1	2	3	4	5
When I hear a word I don't know, I consider various spelling possibilities and look it up accordingly	1	2	3	4	5
When I can't find a word where I thought I would find it, I begin a new search until I find it	1	2	3	4	5
To see how a word is used in spoken language, I use the usage labels provided in the entry	1	2	3	4	5
When I look up a word beginning with E, I search in the first quarter pages as E is one of the first letters of the alphabet	1	2	3	4	5
When I look up a word beginning with L, I open my dictionary in the middle	1	2	3	4	5
When I look up a word, I bear in mind its initial letter and then search where I believe this initial letter is in the dictionary.	1	2	3	4	5
When I look up a word, I simply open the dictionary and see if I am near the specific initial letter	1	2	3	4	5
When I look up a word, I constantly bear it in my mind during the search	1	2	3	4	5
When I realize that the word I am looking for has various different meanings, I go through them all one by one, assisted by the example sentences	1	2	3	4	5
When I find the word that I was searching for, I return to the text to confirm that the word matches the context	1	2	3	4	5
Before I use a word I found in the dictionary when writing a text, I read all the information on the grammar of that word (conjugation, syntax) to be sure of the correct usage.	1	2	3	4	5





# The Authentic Voices of Dictionary Users – Viewing Comments on an Online Learner’s Dictionary Before and After Revision

Ann-Kristin Hult  
Department of Swedish, University of Gothenburg  
ann-kristin.hult@svenska.gu.se

## Abstract

This paper deals with comment field data from two web questionnaire surveys, performed in 2007 and 2011, about the use and users of the *Swedish Lexin Dictionary* (SLD), an online monolingual learner’s dictionary. The SLD underwent comprehensive revision in between the two studies. In order to evaluate the update I compare the respondents’ comments on SLD before and after the revision. The two sets of comment field data are categorised by the types of comments expressed and by the information categories mentioned. As it turns out, respondents do seem happier with the new version of SLD. Generally, there are more positive comments after the revision than before and the information categories mentioned are more often set in positive contexts in the latter data. In addition, the majority of respondents in 2011 belong to the dictionary’s target group, which was not the case in 2007. Although respondents seem more satisfied with the new version of SLD, the two sets of comment field data contain largely the same kind of criticism, as for example in comments concerning (usually the lack of) lemmas and examples.

**Keywords:** dictionary use; learner’s dictionary; user comments

## 1 Introduction

Comment fields are often provided in questionnaires to give respondents the opportunity to expand upon what they have already reported. This article deals with comment field data from two web questionnaire surveys, performed in 2007 and 2011, about the use and users of the online monolingual *Swedish Lexin Dictionary*, SLD (<http://lexin2.nada.kth.se/lexin/>) (Hult 2008a; 2012). In the interval between the two studies the SLD underwent a comprehensive revision. One way of evaluating the success of the updating project is to compare respondents’ comments on SLD before and after the revision. Considering the improvements made, are the respondents in the 2011 study happier dictionary users than those in the 2007 study? In an attempt to answer this question I will first briefly describe the updating project. I will then categorise the two sets of comment field data according to the types of comments expressed. Thereafter, I will examine what information categories are mentioned in the

comments, and what is said about them. Finally, I will give voice to the users by quoting from the comment field data, focusing on the “Lemma”, “Meaning” and “Example” categories.

In the age of electronic dictionaries, users can and do play a much more active role in the process of dictionary making. Lexicographers may disagree on exactly how active users should be in this process (cf. Lew, in press). Be that as it may, since users are the consumers of our products, the least we can do is listen to what they have to say about them.

## 2 The Swedish Lexin Dictionary and the Updating Project

Lexin, short for “lexicon for immigrants”, was originally a series of paper dictionaries at the advanced beginner’s level for users with Swedish as a second language. Lexin includes one monolingual Swedish part, the SLD, which is the focus of this article, and about twenty bilingual dictionaries between Swedish and non-Scandinavian immigrant languages in Sweden such as Arabic, Russian and Somali. The Swedish material served as a basis for the bilingual dictionaries. The dictionaries are primarily aimed at recent immigrants to Sweden who are just beginning to learn Swedish. They are intended to be easy to use even by people with limited reading ability and little or no experience using dictionaries. Consequently, in comparison with many other dictionaries, a user-friendly layout is extremely important (Gellerstam 1999: 5). The SLD (and the entire Lexin series) is primarily intended for reception and secondarily for production. The lemma selection has been carefully adjusted to the needs of the dictionary’s target group. All lemmas have been augmented with comprehensive information typical of a general dictionary. In addition, SLD includes specific words covering special issues and institutions in Swedish society, along with factual information. About 1,800 elementary words are illustrated with pictures in a special section that covers 31 themes, such as *The Human Body (external and internal)* and *Cooking and Meals*.

The free online version of SLD and the Lexin series have been available since 1994. In addition to pictures, the online version features animations that demonstrate the meanings of 700 verbs. These are classified in fifteen main sections, such as *The Kitchen and Cooking*, *Emergency and Medical Service* and *Travel and Transport (City and Traffic)*. The animations make verbs otherwise difficult to make comprehensible through pictures – let alone definitions – much easier to understand. There are links from the dictionary entries to the pictures and animations. Use of the online version has steadily increased from the start. The site presently has about 20 million searches per month and the SLD has slightly more than one million searches per month (<http://lexin2.nada.kth.se/statistik/html>). The recent revision of the SLD encompassed both dictionary content and the website interface. Eventually, the bilingual dictionaries will also be revised.

The revision focused on strengthening the nature of the SLD as a learner’s dictionary, especially its function as a reception dictionary (Malmgren 2012: 456). It is the dictionary’s first comprehensive revision since it was released some thirty years ago. Naturally, one important task was to update the

lemma selection, which increased by about 3,000 lemmas. A great many synonyms and antonyms were also added. This was done mainly to improve users' chances of understanding the entry words: "one single synonym – even if not a perfect one – can make the difference between understanding and not understanding" (Hult et al. 2010: 807). The synonyms and antonyms also serve the purpose of increasing the users' vocabularies. Moreover, the update included a thorough adjustment subdividing the senses of each lemma and placing the examples in the immediate vicinity of the sense being illustrated. A great deal of effort was also dedicated to more explicitly describing particle verbs, which are often stumbling blocks for learners of Swedish. The structure of the entry after the definition was changed slightly. Compounds now come first, followed by self-explanatory syntactic examples and then idioms. Idioms are now given in bold, emphasising their status as sub-lemmas of a sort. Valency information – now somewhat more transparent – is found at the end of each numbered sense. Furthermore, although most examples in older editions of SLD are good, many infinitival phrases have been replaced by full sentences. In a learner's dictionary mainly intended for reception purposes, examples are extremely important. They must be self-explanatory, contain no difficult words and "[I]deally, they should evoke a little scene from everyday life, with prototypical actants" (Hult et al. 2010: 807). In the entry *väntar* ("wait(s)"), for example, instead of the infinitival phrase *vänta på bussen* ("wait for the bus"), the full sentence *de fick vänta på bussen i tio minuter* ("they had to wait for the bus for ten minutes") is given. In addition, all words in the dictionary are clickable.

The changes accomplished in the updating project are consistent with the conclusions of Pálfi & Tarp (2009) on the subject of learner's dictionaries. For instance, they recommend more synonyms and antonyms, more explicit valency information and lemmatisation of idioms. In his specific theories concerning learner's dictionaries, Tarp (2009: 199-200) also emphasises the importance of describing semantic relationships. This is realised in SLD through the copious links from the dictionary entries to the picture themes and, for verbs, the animations.

Let us now turn to the comment field data of the two web questionnaire surveys.

### 3 Types of Comments Expressed in the Two Sets of Comment Field Data

In the 2007 study, almost one third of the respondents, 110 out of 360, wrote something in the questionnaire's comment field. Interestingly, considering that the dictionary's target group is (relatively) recently arrived immigrants, almost 60% of respondents reported Swedish as their native language.

In terms of length, the shortest comment consists of one word and the longest of 130 words, with a median value of 14 words. In Table 1 the comments have been categorised into five types, each accompanied by a translated example. Comments belonging to more than one category were divided and placed into all appropriate types, which explains how the 110 comments in total turned into 152 parts.

Type of comment	Frequency		Examples (my translations)
	Absolute	Relative	
Criticism	53	35	<i>You only have easy words like “cat” and “dog”, but people know those words...I need explanations for more difficult words.</i>
Suggestions for improvement	48	31,5	<i>It would be fun if there were more technical terms or words related to certain subjects like biology, etc.</i>
Praise	30	20	<i>I think Lexin and [SLD] is a very good site, I use it often and it is the best look-up site I have ever found.</i>
Other	17	11	<i>I think all the items in question 8 are important to have full understanding in a context.</i>
Nonsense	4	2,5	<i>ROCKTHE WORLD</i>
<b>Total</b>	<b>152</b>	<b>100</b>	

**Table 1: The 2007 study: comment types, frequency and an example of each type, in descending order.**

Comments containing criticism are nearly as numerous as comments giving suggestions for improvement. Combined, they represent 66.5% of the comments. One fifth of the comments contain praise. The “Other” category covers comments of a more neutral nature or comments that did not directly relate to SLD. With a few exceptions, these comments contributed very little substantive information. The nonsense comments speak for themselves; they are few in number and contain no information of interest.

Let us now turn to the 2011 study. Nearly half the respondents, 371 out of 802, wrote something in the questionnaire’s comment field. Gratifyingly, a full 67% of the respondents belong to the dictionary’s target group: people whose native language is not Swedish and who arrived in Sweden fairly recently (during the 21<sup>st</sup> century).

In terms of length, the shortest comment consists of one word and the longest of 168 words, with a median of 17 words. In Table 2 the comments have been categorised into six types, each accompanied by a translated example. One column was added, called “Temporary change noticed”.<sup>1</sup> Again, comments belonging to more than one category were divided and placed in all appropriate types. In total, the 371 comments turned into 565 parts.

<sup>1</sup> The same week the questionnaire was displayed on the SLD website, the settings had been manipulated. Five categories had been removed from the dictionary articles with a view to indicating which categories users most preferred. Users were informed of the ongoing survey and encouraged to change the settings back to default if they wished. This change was mentioned in the comment field in about 15 questionnaires.

Type of comment	Frequency		Example (my translations)
	Absolute	Relative	
Praise	239	42	<i>Free service. Simple translation. Good examples on how to use words. (Originally written in English)</i>
Criticism	123	22	<i>You don't have enough words in Lexin.</i>
Suggestions for improvement	115	20	<i>Add more compounds, please!</i>
Other	69	12	<i>Didn't know about the pictures and videos.</i>
Temporary change noticed	15	3	<i>It's a pity you took away the examples and verb forms.</i>
Nonsense	4	1	<i>Huh.</i>
<b>Total</b>	<b>565</b>	<b>100</b>	

**Table 2: The 2011 study: comment types, frequency and an example of each type, in descending order.**

Compared to the 2007 study, the figures are distributed somewhat differently. Here, praise is the most frequent comment type, followed by criticism and suggestions for improvement. The positive comments (praise) are almost exactly equal in number to the suggestions for improvement and critical comments put together. Moreover, in relative figures there are twice as many positive comments compared with the 2007 study.

Notably, many comments ended up in the “Other” category. These are often part of longer comments; for instance, some respondents write something about themselves or their background (“I have dyslexia”), others say what other dictionaries they use (“I use Tyda instead”), and still others simply add a comment like “examples are important for immigrants” or “hope it remains free of charge”. Compared with the “Other” comments in the 2007 study, these comments often contribute more or less substantive information.

The data reported so far indicate that respondents are happier with the revised version of SLD than with its unrevised counterpart. Comparing the figures from before and after the updating project, we see that praise has increased by 13.2 percentage points, whereas criticism has decreased by 12.4 percentage points. Comments in the “Other” category have declined by 0.8 percentage points. Let us now move on to see what information categories users comment on and what they say about them.

## 4 Information Categories Commented On

In the 2007 version of SLD, before the revision, a dictionary article could have up to nine information categories. These are mentioned 137 times in the comment field data of the 2007 study, distributed as shown in Table 3.

Information category	Positive context	Negative context	Other	Total	
				Absolute	Relative
1. Lemma	9	91	4	104	76
2. Pronunciation	1	7	4	12	9
3. Part of speech	1	-	1	2	1,5
4. Inflection	-	1	1	2	1,5
5. Meaning	-	8	-	8	6
6. Compound	1	1	1	3	2
7. Example	-	3	-	3	2
8. Phrase	-	2	-	2	1,5
9. Picture theme	-	1	-	1	0,5
<b>Total</b>	<b>12</b>	<b>114</b>	<b>11</b>	<b>137</b>	<b>100</b>

**Table 3: Frequencies of SLD information categories in the 2007 study.**

“Lemma” is by far the information category most frequently commented on (76%). “Pronunciation” and “Meaning” are in second and third place with 12% and 8% of the comments, respectively. The remaining categories are all rarely commented on. As shown, occurrences in a negative context are most common.

In the 2011 revised version of SLD, a dictionary article may have up to thirteen information categories. These are mentioned 332 times in the comment field data of the 2011 study, distributed as shown in Table 4.

Information category	Positive context	Negative context	Other	Total	
				Absolute	Relative
1. Lemma	36	126	10	172	51,5
2. Pronunciation	4	12	2	18	5
3. Part of speech	-	1	-	1	0
4. Inflection	5	11	1	17	5
5. Abbreviation	1	1	-	2	0,5
6. Meaning	8	13	2	23	7,5
7. Antonym	1	5	-	6	2
8. Compound	-	13	-	13	4
9. Example	9	30	3	42	13
10. Valency	-	4	1	5	1,5
11. Phrase	4	15	1	20	6
12. Picture theme	4	3	2	9	3
13. Video sequence	1	1	2	4	1
<b>Total</b>	<b>73</b>	<b>235</b>	<b>24</b>	<b>332</b>	<b>100</b>

**Table 4: Frequencies of SLD information categories in the 2011 study.**

Again, “Lemma” is by far most the most frequently mentioned category in the comments. “Example” is in second place, followed by “Phrases”, “Pronunciation”, “Inflection” and “Meaning”. The least frequently mentioned are “Part of speech” and “Abbreviation”. The remaining categories received between four and thirteen comments. Again, occurrences in a negative context are the most common, but are significantly fewer than in the 2007 study, with a difference of 12.4 percentage points. Thus, analogue to that indicated in the previous section, respondents seem happier with SLD after the revision than before.

Now let us look at more authentic examples from the comment field data. I will focus on what the respondents said about the information categories “Lemma”, “Meaning” and “Example”.

## 4.1 Voices on Lemma

The typical praise related to lemma simply states that the respondent often finds the words they searched for in SLD, and are satisfied with what can be found. Practically all the critical comments concerning lemma in both studies complain about the dearth of them. Many respondents are critical of the absence of more difficult words. Here are a few examples concerning lemma taken from the two sets of data (my translations):

### Study 2007

- (1) Some words you search for aren’t there, even though they are real words.
- (2) Quite a few psychological or political terms/words aren’t there.
- (3) It is bad that there are no compounds or particle verbs, which are what we immigrants need the most help with.
- (4) I want the difficult words; the words included are mostly for beginners, or basic level.

### Study 2011

- (5) Many compounds are missing as well as special terms like *phonology* and *autism*.
- (6) Hard to find the meaning of long words.
- (7) Have a feeling only the most basic words are there, not other words.
- (8) Sometimes important words are missing.

Approximately 3,000 new lemmas were added in the revision, including many particle verbs, and many synonyms and antonyms were added to the dictionary articles. In this respect, we have fulfilled the respondents’ requests for more lemmas of different kinds. The figures speak for themselves: in 2007, 87.5% of the comments on lemma were negative, compared with 73% in 2011, while 8.7% of comments were positive in 2007 compared with 21% in 2011. Respondents are clearly less negative and more positive after the updating project. They remain, however, quite critical of the lemma selection and express more or less the same type of criticism in 2011 as in 2007. Indisputably, no dictionary will

ever succeed in fully satisfying users' need for more lemmas. A simple comment like "all words aren't there" is very telling in respect to users' expectations of dictionary's lemma selection. Along with the comments quoted above, this also reveals that many respondents do not clearly understand what type of dictionary the SLD is.

## 4.2 Voices on Meaning

In the 2007 study, there were no positive or neutral comments and only eight negative ones about the information category "Meaning". In the 2011 study, there are eight positive comments, thirteen negative comments and two neutral ones. Here are a few examples (my translations):

### Study 2007

- (9) More examples of the meaning of the word and how you can use the word.
- (10) A definition of *naturopathic practitioner* is missing and an understandable explanation of *recruit*.
- (11) Hard to find explanations of words.
- (12) You only have easy words like *cat* and *dog*, but people know those words. I need explanations for more difficult words.

### Study 2011

- (13) Would like to have more examples of possible useful prepositions with a word and what meaning you get if you use a particular preposition.
- (14) It should be possible to search for meaning/translation of expressions, not only single words.
- (15) Can you try to give the exact meaning of words or an exact synonym?
- (16) The definitions are good, short and concise and easy to understand.

The update included a thorough revision subdividing the senses of each lemma. Moreover, a great deal of effort was dedicated to more explicitly describing particle verbs which, as mentioned, often cause difficulties for learners of Swedish. Again, users' expectations are high, as they should be. In some respects, the respondents' views have been taken into account, for example in (9) and (13), and in other respects they have not, such as in (10). In a perfect world, all users would agree with the respondent quoted in (16).

## 4.3 Voices on Example

There were only three comments concerning "Example" in the 2007 study, all negative. Respondents want more examples, as expressed in the following comment: "have at least 2-3 examples for every word so you understand 100%". In the 2011 study, there are nine positive comments, thirty negative ones and three of a neutral nature. Here are a few examples (my translations):



- (17) Too few examples on how to use words, hard to find synonyms, antonyms and proverbs.
- (18) You should have more phrases and maybe a few more examples of each word or preposition.
- (19) Good examples where you often find exactly the inflection you are looking for.
- (20) Better if you could add more examples, particle verbs, past participle, proverbs, slang, which you have to learn and use every day.

As the senses of each lemma were subdivided, so were the examples, placing them in the immediate vicinity of the sense being illustrated. The update also added many morphological examples in terms of transparent compounds and several more, and more extensive, syntactical examples. In the updated version there is at least one example per lemma. This may not be enough in the opinion of the respondent quoted above, but examples are nonetheless much more numerous than was the case before the revision. What this respondent, and hopefully many others, acknowledges and appreciates is that the examples have been expanded to full sentences to more clearly illustrate the meaning of the lemma.

## 5 Conclusion

Taken as a whole, the users' comments on SLD can offer lexicographers valuable information about the needs of the dictionary's target group and might provide them with relatively concrete ideas on how to improve the dictionary. The question, however, was whether the respondents are happier with the SLD now than before the updating project. Comment field data from a web questionnaire survey before the update were compared with corresponding data from after the update. And, yes, respondents appear to be happier with the SLD now than before the revision. Firstly, there are significantly more positive comments after the revision than before. Secondly, the information categories commented on are more often set in positive contexts in the latter data. We also know that the majority of the respondents in 2011 belong to the dictionary's target group, in contrast to 2007. This suggests that the users who are happier with the SLD are also more representative of the intended users of the dictionary.

Even though respondents seem more satisfied with the updated version of SLD, the two sets of comment field data contain largely the same kind of criticism. Not unusually, many respondents have both good and bad things to say about the SLD. These circumstances indicate that one cannot have too much of a good thing. Or, as one respondent simply writes: "I think you can do it better, but it's not bad". Moreover, "Lemma" was the information category most frequently commented on, while the other categories were much less frequently commented on. Presumably, there is more information to be extracted from the comment field data.

It should be noted that the two sets of data were not fully comparable. The comments in 2007 add up to not quite one third of the number of comments in 2011. The first questionnaire contained ten questions and the second was extended to twenty questions. In addition, the texts preceding the comment field box were not identically worded.

## 6 Discussion

There are comments where respondents more or less explicitly demonstrate their awareness of what type of dictionary they are presently using, but there are also comments which clearly reveal the respondents' lack of understanding of what type of dictionary SLD is. This may not be a problem, since the dictionary is free and users are also free to go elsewhere if they are not satisfied with what they find. On the other hand, perhaps the purpose of SLD and all the Lexin dictionaries could be made more explicit on the website and on the internet as a whole. If not, at least the pictures and animations should be marketed since "for non-native speakers of the language, definitions, however skilfully written, are not usually the best way to convey meaning" (Lew, in press), and pictures and animations can concretise the meaning of a word in a very enlightening way. Surprisingly and somewhat disappointingly neither pictures nor animations are mentioned more than a handful of times in the comment field data. Admittedly, this does not necessarily mean that the respondents are unaware of these features, but unfortunately that might just be the case.

In the practice of dictionary making, professional lexicographers are undoubtedly the linguistic experts and users' potential contributions in this area may be limited. I would argue, however, that there are other equally important areas where users may very well have useful suggestions for improvement, particularly in relation to web-related issues like the user interface.

## 7 References

- Hult, A.-K. (2008a). Användarna bakom loggfilerna. Redovisning av en webbenkät i Lexin online Svenska ord. In: *LexicoNordica*, 15, pp. 73-91.
- Hult, A.-K. (2012). Old and New Study Methods Combined. Linking Web Questionnaire with Log Files from the Swedish Lexin Dictionary. In: R. Vatvedt Fjeld, J. M. Torjusén (eds.) *Proceedings of the 15<sup>th</sup> Euralex International Congress, Oslo, 7-11 August 2012*, pp. 922-928.
- Hult, A.-K., Malmgren, S.-G., Sköldberg, E. (2010). Lexin – a Report from a Recycling Lexicographic Project of the North. In: A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*, pp. 800-809.
- Gellerstam, M. (1999): LEXIN – lexikon för invandrare. In: *LexicoNordica* 6, pp. 3-17.
- Malmgren, S.-G. (2012). Från Svenska ord (Lexin) 3 till Svenska ord 4. In: B. Eaker, L. Larsson and A. Mattisson (eds.) *Nordiska studier i lexikografi 11. Rapport från Konferensen om lexikografi i Norden, Lund 24-27 May 2011*, pp. 454-465.

- Lew, Robert in press. User-Generated Content (UGC) in Online English Dictionaries. In: A. Klosa, A. Abel (eds.) *OPAL – Online publizierte Arbeiten zur Linguistik* 2013. (Preprint).  
*Lexin Online, Statistics*. Accessed at: <http://lexin2.nada.kth.se/statistik/html> [29/03/2014].
- Pálfi, L.-L., S. Tarp (2009). Lernerlexikographie in Skandinavien – Entwicklung, Kritik und Vorschläge. In: *Lexicographica. International Annual for Lexicography* 25/2009. Tübingen, pp. 135-154.
- Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography. Tübingen: Max Niemeyer.  
*Swedish Lexin Dictionary Online*. Accessed at: <http://lexin2.nada.kth.se/lexin> [29/03/2014].



# Mobile Lexicography: A Survey of the Mobile User Situation

Henrik Køhler Simonsen  
Copenhagen Business School  
hks.ibt@cbs.dk

## Abstract

Users are already mobile, but the question is to which extent knowledge-based dictionary apps are designed for the mobile user situation. The objective of this article is to analyse the characteristics of the mobile user situation and to look further into the stationary user situation and the mobile user situation. The analysis is based on an empirical survey involving ten medical doctors and a monolingual app designed to support cognitive lexicographic functions, cf. (Tarp 2006:61-64). In test A the doctors looked up five medical terms while sitting down at a desk and in test B the doctors looked up the same five medical terms while walking around a hospital bed. The data collected during the two tests include external and internal recordings, think-aloud data and interview data. The data were analysed by means of the information scientific star model, cf. (Simonsen 2011:565), and it was found that the information access success of the mobile user situation is lower than that of the stationary user situation, primarily because users navigate in the physical world and in the mobile device at the same time. The data also suggest that the mobile user situation is not fully compatible with for example knowledge acquisition.

**Keywords:** Mobile lexicography; mobile user situation; mobile user

## 1 Introduction and Problem

Today, most users are always on and always connected, cf. (Google 2013:2), which reports that 84% of us use smartphones while they do other things, and users today in fact use their mobile devices in a large number of situations.

Users are already mobile, but the question is to which extent knowledge-based dictionary apps are designed for the mobile user situation. The objective of this article is to analyse and discuss the characteristics of the mobile user situation with a view to putting the user back in focus in dictionary apps.

## 2 Methodology and Empirical Basis

Ten medical doctors were asked to look up five medical terms by means of the dictionary app Medicin.dk, which is a knowledge-based medical resource used by most health care persons (HCPs) in the Danish health care system. As many as 15,000 users regularly update the medical app Medicin.dk,

which indicates that the app is widely used by a variety of users. According to (Dolan 2012) everything in medicine is going mobile and both patients and physicians are changing behaviour in line with developments in health technology.

The test persons accessed the medical terms by means of the app Medicin.dk on an iPhone 4S, which was wirelessly connected to a PC by means of Reflector, cf. <http://www.airsquirrels.com/reflector/>. The medical doctors were asked to participate in two tests. In test A the test persons were asked to look up five medical terms while sitting down at a desk. In test B the ten test subjects were asked to look up the same five terms while slowly walking around a hospital bed.

The two tests were designed to imitate two typical user situations for many doctors: knowledge acquisition and knowledge checking prior to patient consultation and knowledge checking during a patient consultation.

The five tasks included looking up the five proper names Terbasmin (asthma), Tamoxifen (breast cancer), Antepsin (ulcer), Tredaptive (cholesterol) and Fludara (leukaemia) and can be summarized as follows:

- Task 1: Look up Terbasmin – to find information
- Task 2: Look up Tamoxifen – to find and extract information about side effects to be able to inform patient
- Task 3: Look up Antepsin – to find and extract information about dosage to be able to check prescription
- Task 4: Look up Tredaptive – to find and extract information about dosage to be able to inform patient
- Task 5: Look up Fludara – to find and check spelling of term to be able to write a text.

Both tests were recorded from the “inside” by means of Reflector, and at the same time the user activities were recorded from the “outside” by means of a digital camera. In addition to the recordings from the “inside” and the “outside”, the empirical basis also includes think-aloud-data, as the test persons were asked to think aloud and verbalize what they did and saw etc. To deduce additional qualitative comments, the empirical basis also includes interview data as the test persons were interviewed before and after the tests.

### 3 Theory

Related work with direct relevance for this survey includes a number of studies of how users interact with mobile devices, such as (Pedersen & Engrob 2008), (Church et al. 2009) and (Ehrler et al. 2013). In addition to theoretical considerations on user interaction and mobile devices, this work also includes

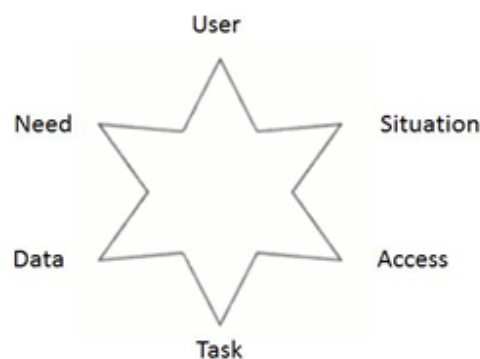
selected theoretical considerations on lexicography such as (Tarp 2006), (Verlinde et al. 2010) and (Simonsen 2011).

Pedersen & Engrob (2008) discusses a number of interesting usability tests with mobile devices. The objective of their work was to discuss which interaction technique was most suitable for mobile users. Pedersen & Engrob (2008) asked eight students to walk on a running machine while interacting with a PDA. The focus of their tests is not completely comparable with this survey, but it is highly relevant. Pedersen & Engrob (2008) found that the test persons used different interaction techniques in different user situations.

Another highly relevant contribution in this area is Church & Smyth (2009). Church & Smyth (2009) conducted a number of surveys of the mobile information needs of different users. Church & Smyth (2009) asked 20 test persons to participate in a four week long diary survey during which the test persons made notes on their mobile information needs and user situations. Church & Smyth found (2009) that user situations can be categorized in five overall categories such as Navigational, Informational, Transactional, Geographical and Personal Information Management. The informational need, cf. (Church & Smyth 2009:251) is the most important need and is focused on the goal of obtaining information about a topic.

Ehrler et al. (2013) reports on an evidence-based survey of user-interface design on handheld devices in health care. The usability test discussed by Ehrler et al. (2009) aimed at acquiring evidence about the quality of data recorded through interfaces on mobile devices and the test showed that the majority of test persons preferred the simpler models for data entry geared to the actual healthcare environment – a finding which was also clear on the basis of this survey.

Finally, a number of contributions on lexicography and information science are also relevant for this analysis. First of all the many contributions on the user and the lexicographic functions as discussed by for example (Tarp 2006) are necessary to understand the characteristics of the user. Furthermore, (Simonsen 2009), (Simonsen 2011a) and (Simonsen 2011b), who builds on (Verlinde et al. 2010), is relevant for the understanding of user research, the mobile user and mobile lexicography. Simonsen (2011a:565) makes the case for the information scientific star model as shown in figure 1 below.



**Figure 1: Information Scientific Star Model.**

The information scientific star model proposed by Simonsen (2011:565) is applied on the analysis of the mobile user situations below, and it is argued that modern dictionary app development should be based on these six factors. The above model builds on (Verlinde et al. 2010:5), who argues that “relevant data should be retrieved and processed according to the external situation that motivated consultation in the first place, and the information needed to change a state of affairs in the outside world should be operationalized”. Verlinde et al. (2010:5) make the case for a “lexicographic triangle” consisting of user, data and access, but it is argued that the analysis and design of information tools should be based on much more than that. Consequently, the information scientific star model was developed.

The star model includes the six dimensions: user, situation, access, task, data and need. Explicit considerations are required on the competency profile of the user, the user situation, the way the user accesses information, the type and complexity of the task, the type and complexity of data and the inherent need of the user. As the data suggest a number of these dimensions have been neglected during the design and development of the app Medicin.dk.

## 4 Results and Discussion

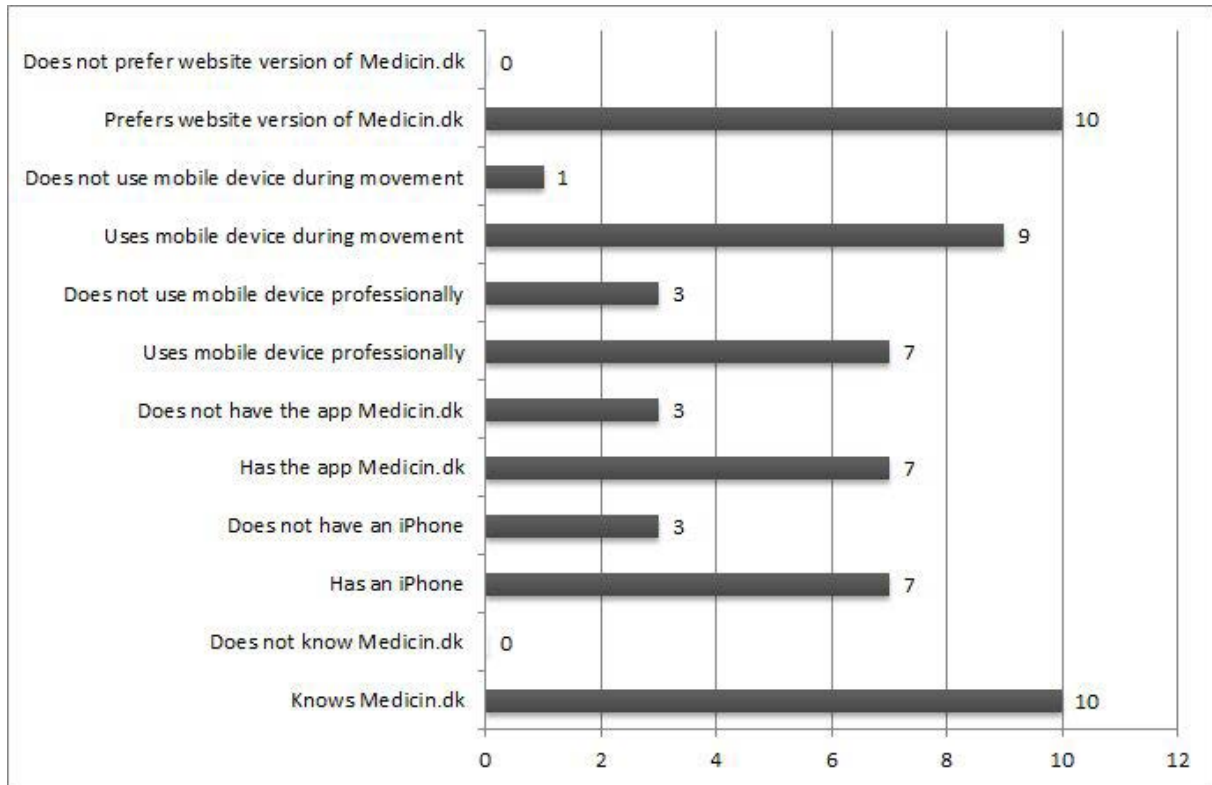
Ten medical doctors were asked to look up five medical terms by means of the dictionary app Medicin.dk and two tests were carried out. As already mentioned, the empirical data include ten recordings from the inside, ten from the outside, twenty think-aloud data recordings and interview data from all ten test persons.

An overview of some of the comparable answers offered by the test persons during the interviews is shown below in figure 2.

As will appear from figure 2 below, all ten medical doctors in fact prefer the website version of Medicin.dk when asked the question “Which platform and which user situation do you prefer”? This finding is in fact very much in line with (Ehrler et al. 2013), who argue that test persons seem to prefer the most simple data entry and data access model. The finding also seems to support the overall theoretical approach proposed by (Simonsen 2011), who makes the case for a balanced approach and focus in lexicography and information science. It may be argued that the finding is not that surprising, because medical doctors in public hospitals are not issued with a mobile device nor do their working conditions match knowledge acquisition by means of a mobile device. Furthermore, doctors often look for complex data and documents from many different sources and again the mobile device is not an obvious tool to use.

However, as will appear from figure 2 below seven out of ten doctors state that they in fact use their mobile devices professionally and nine out of ten doctors say that they use their mobile phone while moving around, so researching the mobile user situation is highly relevant.





**Figure 2.: Overview of Interview Data.**

The six dimensions described in the information scientific star model, cf. (Simonsen 2011:565) are encapsulated by the answer given by test person 4, who is a 52-year old female medical doctor. Test person 4 states “I prefer the website version of Medicin.dk, if my problem is complex. The app and the iPhone are handy, if I suddenly have a problem that I know can be solved by the app. However, if I need more knowledge I would rather use the website”.

What this statement seems to indicate is that the concrete task at hand more or less dictates the actual user situation and vice versa. Furthermore the task also dictates the amount and type of data sought by the user and in fact also the data access method needed. This correlation is in fact observable in most of the statements offered by the ten test persons, and the interview data show that the mobile user situation and cognitive lexicographic functions is not a perfect match. What is needed is an adaptable, dynamic and situational tool, which features seamless adaptation of data based on location-based services (LBS) and dynamic and situational presentation of data designed for the concrete task at hand and the competence profile of the user.

The quantitative test data from Test A and Test B also support these arguments, see figure 3 and 4 below. Figures 3 and 4 below show how the two tests were carried out and illustrate the stationary user situation and the mobile user situation.



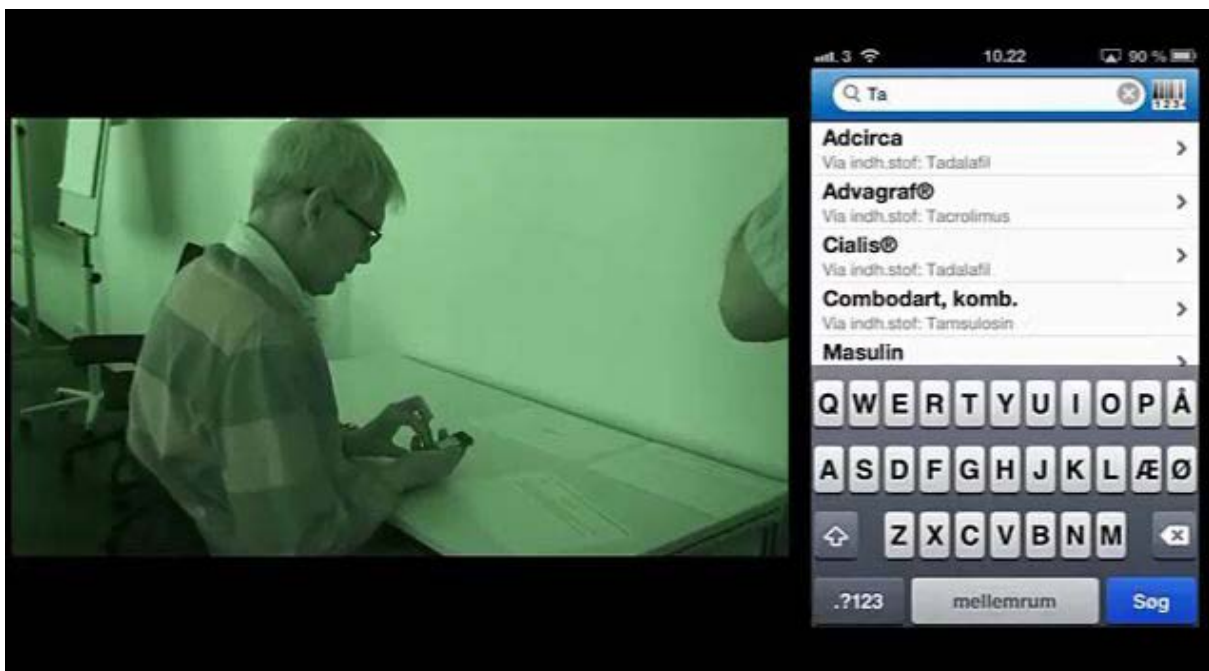
**Figure 3: Stationary Test.**



**Figure 4: Mobile Test.**

The numbers in figure 8 below are numerical representations of a systematic evaluation of each test person's information access success in each situation. As will appear from figure 5 the five columns list the five tasks that the ten doctors were asked to do during the stationary test and the mobile test. The term information access success covers an evaluation of the search speed, search quality, focus ability, device interaction ability of each test person on a scale from 1 to 10, where 1 is low information access success and 10 is high information access success. Each number thus represents an overall evaluation of each situation based on the many internal and external recordings. The interview data and think-aloud data also substantiate the numerical evaluations made.

Figure 5 below shows how test person 3 (TP3), who is a 62-year old medical doctor, solves the task "Look up Tamoxifen and extract information about side effects to be able to inform a patient about the most common side effects" in Test A - that is while he is sitting down at a desk.



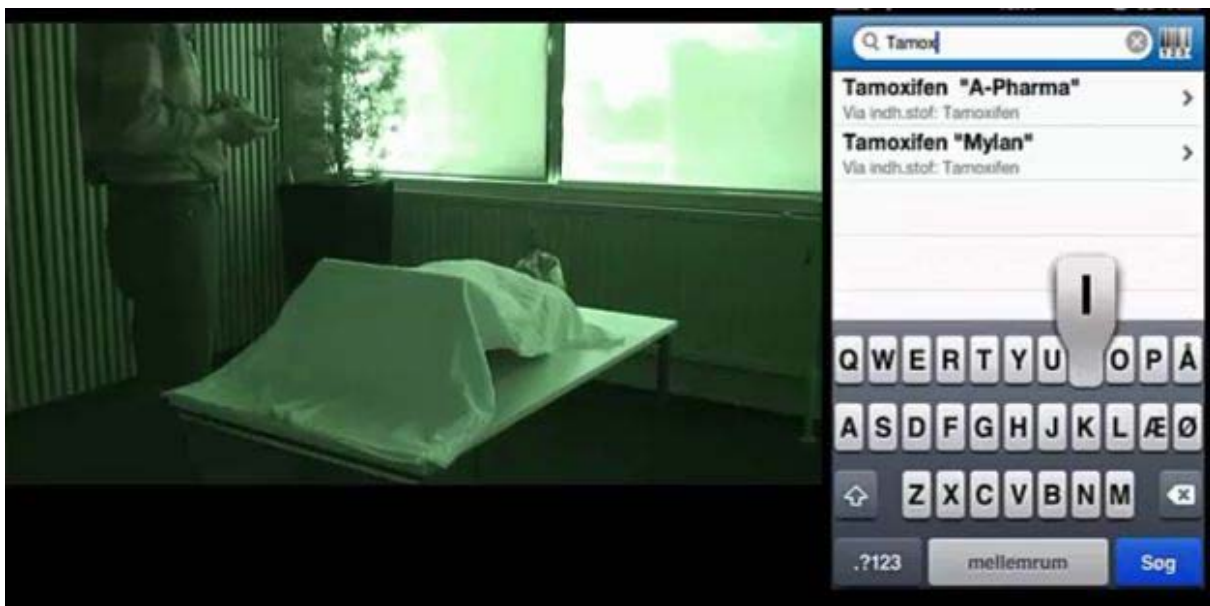
**Figure 5: TP3 solving Task 2 during Test A - Outside vs. Inside.**

Figure 5 is a snap shot of two video recordings, which originally were recorded at the same time, but they have been edited by means of a video editing tool so that the two recordings can be shown at the same time as a picture-in-picture video.

The entire edited recording shows how TP3 sits at the table in the left hand side of the picture interacting with the mobile device in the physical world, and in the right hand side of the picture TP3's search behaviour is shown from the inside. The left hand side video was recorded by means of a standard digital camera and the right hand side of the video was wirelessly recorded by means of Reflector, cf. <http://www.airsquirls.com/reflector/>.

The edited picture-in-picture video, which is based on aligned time codes to show a time-aligned video of the user situation seen from both the inside and the outside, gives a detailed picture of how TP3 solves a concrete task and it shows how a medical doctor uses a mobile phone while sitting down at a desk to look up complex medical information.

In comparison with the stationary user situation, Figure 6 below shows the mobile user situation, that is TP3 solving the same task (Task 2) during Test A. Again the actual user situation and user behaviour are recorded from the outside and the inside and Figure 11 is also a snap shot of an edited time-aligned, picture-in-picture video.



**Figure 6: TP3 solving Task 2 during Test B – Outside vs. Inside.**

Figure 6 above shows how TP3 walks around a “hospital bed” while solving task 2. The video gives a detailed picture of how TP3 uses a mobile phone to look up complex medical information while moving around at the same time.

A comparison of the two user situations shows that the access speed, that is from the moment the test person started the information access operation to the moment he ended the search operation, is higher during Test A than during Test B. That is in fact not surprising, because users can focus on the

search operation and the mobile device while sitting down, which is in contrast to the mobile user situation where users also have to allocate cognitive effort on navigating in the physical world.

The differences between the two user situations become clearer when the two recordings from the inside are edited and contrasted. Figure 7 below shows a snap shot of the time-aligned edited picture-in-picture video of how TP7 solved Task 2 while sitting on the left hand side (Test A) and walking around (Test B) on the right hand side.



Figure 7: TP7 solving Task 2 during Tests A and B – Inside.

The video shows that TP7 is much faster at locating the section on side effects while sitting down than while moving around. The information access speed is clearly higher when sitting down than when moving around. Another interesting fact is that TP7, just as three other test persons, chose to use the mobile device horizontally allowing the screen to show more text. This result also appeared for TP8, who also chose to use the mobile device in horizontal position. On the basis of these results it may be argued that users tend to use mobile devices like small computers while sitting down (the horizontal position), which in fact the video recordings from the outside also seem to document.

The many recordings from the inside and the outside are systematized and tabulated in Figure 8 below. The many numbers in figure 8 are numerical representations of a systematic evaluation of each test person's information access success in each situation. The term information access success covers an evaluation of three factors: search speed, search quality and device interaction ability.

The search speed was relatively easy to measure and is based on the time recorder in the many recordings. The measure used here was time.

It was far more difficult to precisely measure the search quality and device interaction ability. The evaluation of the search quality was partly based on an assessment of the quality of the search result

found by the test person. The evaluation was based on an analysis of the think-aloud data where the test person described what he did and found and an analysis of the video recordings. The most important measure in this analysis was the test person’s ability to quickly find the right information and verbalize it as think-aloud data. The measure used here was the ability to find the right information.

The device interaction ability was equally difficult to accurately measure. The evaluation of the test person’s ability to use the device effectively was partly based on an analysis of the video recordings and the think-aloud data, which together made it possible to describe each test person’s ability to use the device effectively. The measure used here was the ability to use the device effectively.

The data show that the information access success of the ten medical doctors was higher when they sat down at a desk than when they walked around a hospital bed. The data also seem to suggest that the task itself and the cognitive complexity of the information dictate the degree of information access success. In other words, simple and easy-to-find information correlates with high information access success while on the other hand complex and hard-to-find information yields lower information access success. This is clear when the two tasks “Find information” and “Find and extract information about dosage to be able to check prescription amount” are compared.

Task	Terbasmin		Tamoxifen		Antepsin		Tredaptive		Fludara	
	Moving	Sitting	Moving	Sitting	Moving	Sitting	Moving	Sitting	Moving	Sitting
T1	7	8	4	7	4	7	4	7	4	7
T2	3	4	2	5	2	5	2	5	2	5
T3	6	7	3	6	3	6	3	6	3	6
T4	7	8	4	7	4	7	4	7	4	7
T5	6	7	3	6	3	6	3	6	3	6
T6	2	3	2	3	2	3	2	3	2	3
T7	7	8	4	7	4	7	4	7	4	7
T8	7	7	4	7	4	7	4	7	2	7
T9	3	8	2	7	3	5	3	6	4	7
T10	4	6	3	7	2	7	4	7	2	6
Total	52	66	31	62	31	60	33	61	30	61

Figure 8: Overview of Test Data.

When asked the question “Do you use your mobile device while moving?” test person 3 states “Yes – when I suddenly think of a medical question that I would like to look up, but I also use my mobile phone in other situations”. This answer is somewhat in contrast to the answer provided by test per-



son 5, who states “No – not really. I mostly use my mobile phone when I am sitting down because I think the screen is too small and my fingers are too big for the key pad screen”.

Interestingly, test person 5 is a 61-year old male medical doctor, and is the oldest test person, which seems to indicate that age plays a role in mobile information access behaviour as does age in the discussion of digital natives, who are tech-savvy young people to whom digital technology is an integrated part of their lives, cf. (Prensky 2001) for a detailed discussion of digital natives.

It was also found that the information access speed and quality of the mobile, punctual user situation is somewhat lower than the stationary, punctual user situation. The many recordings from both the inside and the outside clearly show that the test persons need to navigate both in the physical world and in the mobile device user interface. They stop walking during the interaction with the mobile device, because they also need to look up and navigate in the room.

When asked the question “What do you think of the mobile user situation?” test person 1 states “I do not think that there is a big difference between moving around and sitting. Okay – maybe you spend more time on the search operations when you walk around, because you have to look up and see where you are” and test person 1 states “As long as I stop up and stand still I actually think it works fine”.

To substantiate the argument about information access speed, test person 7 is much faster at locating the section on side effects while sitting down than while moving around. The information access speed is clearly higher when sitting down than when moving around. Another interesting fact is that test person 7, just as three other test persons, chose to use the mobile device horizontally allowing the screen to show more text. This result also appeared for test person 8, who also chose to use the mobile device in horizontal position. On the basis of these results it may be argued that users tend to use mobile devices like small computers while sitting down (the horizontal position), which in fact the video recordings from the outside also seem to document.

The 5-inch screen size of a standard smartphone such as the iPhone is simply not enough. Size does matter when it comes to successful information access and the layout and design of dictionaries has always been relevant for lexicography, cf. for example (Almind 2005) and (Almind & Bergholtz 2007). This is very much still the case as the data presented above suggest. The problem is that the human-mobile interaction is not optimal. The input device (the finger) and the small letters shown on the 5-inch screen are not a perfect match as one of the test persons surveyed actually also verbalize. Obviously, it would be logical just to call for bigger screens, but that would be naïve because smartphones are in fact supposed to be small. However, HUD technology may at some point allow us to display dictionary data in HUD format (Head-up Display) where information is visually size enhanced and relayed to the user surroundings, but it will be some years before that technology becomes commoditized. A number of theoretical contributions discuss mobile design and mobile usability, for example (Budiou & Nielsen 2013), who make a very strong case for more usability research in mobile design, (Cerejo 2012), who discuss the many elements of the mobile user experience, including the more social and personal elements of mobile user personas

and of course also (Nielsen 2011), who offers a myriad of practical and easy-to-use instructions on mobile design.

However, what we can do at this point is to design the actual dictionary app in such a way that intelligent search engines and easy-to-use interfaces facilitate easy information access. As the data of the survey suggest, simple search engines with a simple search field and a simple TOC-like display of the dictionary data are preferred by most users. Scrolling through large text blocks reduces the information access success of the ten medical doctors surveyed.

It was also found that the information access success of the ten medical doctors was drastically reduced in cognitive user situations, that is when they were asked to solve cognitively-based problems like task 2, task 3 and task 4, which were all about locating complex information with a view to making decisions as to side effects, dosage and how to take the medicine etc.

This finding is also expressed by test person 7 who states “If I have to look a little bit deeper into a question then I clearly prefer the computer. I would definitely use the computer if I were to prescribe medicine that I have never used before”. In other words, it was found that the mobile user situation and cognitive lexicographic functions is not a perfect match.

All this in fact seems to suggest that mobile lexicography needs to reinvent itself and take into account the six dimensions proposed above by (Simonsen 2009). This contention seems to be supported by (Church & Smyth 2009:255-256), who state that: “...mobile users are on-the-move and as such are interested in locating different types of content. We found context to be a very influential factor in many mobile information scenarios and as such argued for the need for new types of context-sensitive mobile interfaces that take full advantage of temporal, location, and preference-based contexts”.

A similar argument was made by (Leroyer & Kruse 2011: 411-415), who describe a pragmatic data presentation and user interface in a French/Danish Real Estate e-Dictionary. Leroyer & Kruse (2011) make the case for a situational user interface, which definitely is the way forward in mobile lexicography.

However, mobile lexicography should not only be based on temporal and situational dimensions. Mobile lexicography is different from Internet lexicography and very much different from paper lexicography. Mobile lexicography is unique, because the user very often is mobile and on the move when using his device.

That very fact calls for new theoretical considerations and on the basis of the empirical data and the discussion above the following mobile lexicography principles can be identified.

### **Mobile user principle**

The mobile user is on the move and needs and accesses information while on the go. This makes the mobile user punctual, impatient, imprecise and preoccupied with other things.

### **Mobile situation principle**

The mobile user situation is characterized by being volatile, punctual and by often taking place while the user does other things. The mobile user typically checks knowledge and performs simple search

ches. The mobile user situation primarily supports simple, punctual, communicative lexicographic functions, and is not suited to support complex, cognitive lexicographic functions.

### **Mobile data access principle**

The mobile user navigates in both the physical world and in the user interface of the mobile device at the same time. This calls for a very simple and easy-to-use data access method for example a very intelligent semasiological search engine or even better a voice-activated search engine like Siri in an iPhone.

### **Mobile data principle**

The mobile user situation also dictates the type and complexity of the mobile data. The size of the user interface and the punctuality of the user situation mean that complex data and long text segments are not optimum mobile data.

## **5 Conclusion**

This article discussed the mobile user situation, and it was demonstrated that medical doctors prefer the website version instead of the app version. It was also found that the information access success of the mobile user situation is lower than that of the stationary user situation, primarily because users are required to navigate in both the physical room and in the mobile device. It was also found that the mobile user situation is not at all suitable for solving cognitive lexicographic problems such as for example knowledge acquisition etc.

On the basis of the survey it is argued that classic lexicographic virtues such as attention to the characteristics of the user situation, the task, the type of user and the presentation of data seem to be in demand in app development. The data provided in a dictionary app must be adapted to the mobile user situation and the data access structure of the app should take into account the limitations of the mobile user situation and should be task-dependent. The empirical data and the discussion led to the formulation of four principles on mobile lexicography.

Users are already mobile, but lexicography does not seem to be up to speed with the users. A dictionary app should satisfy concrete and potential lexicographic needs. Consequently, further research in mobile lexicography is needed – to put the user back in focus.

## **6 References**

Almind, Richard (2005): Designing Internet Dictionaries. In: *Hermes, Journal of Linguistics no 34-2005*, pp.37-54.



- Almind, Richard & Bergenholtz, Henning (2007): Klæder skaber folk: Om layout i ordbøger. In: *Hermes – Journal of Language and Communication Studies no 39-2007*, pp. 31-47.
- Budiu, R., & Nielsen, J. (2013). In Rimerman S., Walker A. M. (Eds.), *Mobile usability*. Berkeley: The Nielsen Norman Group.
- Church, Karen & Smyth, Barry (2009): Understanding the intent behind mobile information needs. In: *IUI 2009 International Conference on Intelligent User Interfaces*, pp. 247-256.
- Cerejo, L. (2012). *The elements of the mobile user experience. Mobile design patterns* (1st ed., pp. 5-20). Freiburg, Germany: Smashing Media GmbH.
- Dolan, Pamela Lewis (2012): Everything in medicine is going mobile. In: *amednews.com*. Accessed at <http://www.ama-assn.org/amednews/m/2012/03/26/bsa0326.htm> [04/07/2012].
- Ehrler, Frederich, Walesa, Magali, Sarrey, Evelyne, Lovis, Christian (2013): Evidence-based User-Interface Design. In: *Studies in health technology and informatics*, pp. 57-61.
- Google (2013): *Our Mobile Planet*: Accessed at: <http://services.google.com/fh/files/misc/omp-2013-dk-en.pdf> [01/09/2013].
- Leroyer, Patrick & Kruse, Liselotte (2011): Ejendomsordbogen fransk/dansk: ny integreret e-ordbog. In: *Nordiska studier i lexicografi 11 - 2011*, pp. 405-417.
- Nielsen, Jakob (2011): Mobile Usability Update. In: *useit.com*. Accessed at: <http://www.useit.com/alertbox/mobile-usability.html> [04/07/2012].
- Pedersen, Anders Fritz & Engrob, Jan Hyldgaard (2008): Interaktion under bevægelse: Et komparativt studie af interaktionsteknikker til arbejde med komplekse data på håndholdte enheder, AAU, 60 sider.
- Premsky, M. (2001): Digital natives, digital immigrants part 1. *On the Horizon*, 9(5), 1-6: Accessed at <http://www.emeraldinsight.com/journals.htm?issn=1074-8121> 1 [01/04/2014].
- Reflectorapp.com (2012): Accessed at: <http://www.airsquirls.com/reflector/> [02/01/2012].
- Simonsen, Henrik Køhler (2009): Se - og du skal finde: en eyetrack-undersøgelse med særlig fokus på de leksikografiske funktioner. In: *Nordiske studier i lexicografi 11. Rapport fra Konference om lexicografi i Norden. Finland 3.-5. juni 2009*. Tampere: Nordisk forening for lexicografi 2009, pp. 274-288.
- Simonsen, Henrik Køhler (2011a): User Consultation Behaviour in Internet Dictionaries: An Eye-Tracking Study. In: *Hermes - Journal of Language and Communication in Business*, 46-2011, pp. 75-102.
- Simonsen, Henrik Køhler (2011b): Et informationsvidenskabeligt serviceeftersyn af Medicin.dk. In: *Nordiska studier i lexicografi 11 - 2011*, pp. 563-574.
- Tarp, Sven (2006): *Leksikografien i grænselandet mellem viden og ikke-viden: Generel lexicografisk teori med særlig henblik på lørnerlexicografi*. Doktorafhandling. ASB.
- Verlinde, Serge, Leroyer, Patrick, Binon, Jean, (2010): Search and You Will Find. From Stand-Alone Lexicographic Tools to User Driven Task and Problem-Oriented Multifunctional Leximats. In: *International Journal of Lexicography*, Vol. 23, Issue (1) 2010. S. 1-17.



# Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht

Martina Nied Curcio  
Università degli Studi Roma Tre  
martina.nied@uniroma3.it

## Abstract

Die Benutzung von Smartphones, iPhones, iPads und Tablets im Fremdsprachenunterricht scheint den Studierenden den schnellen und unmittelbaren Gebrauch von Online-Wörterbüchern zu garantieren, sowie unbegrenzte Recherchemöglichkeiten zu bieten, so dass sprachliche Schwierigkeiten direkt überwunden werden können. Doch sieht es danach aus, als würden die Fremdsprachenstudierenden das Potential nicht ausnützen und sich auf zweisprachige Wörterbücher und Übersetzungsprogramme konzentrieren. Auch ihr Benutzerverhalten in Bezug auf Online-Wörterbücher scheint dem der Verwendung von Print-Wörterbüchern ähnlich zu sein. Eine empirische Untersuchung mit italienischen Studierenden der Germanistik zur Smartphone-Benutzung zeigt, wie und wofür sie hinsichtlich lexikalischer Fragen das Smartphone verwenden, welches Benutzerverhalten sie generell, aber auch in spezifischen Übersetzungsaufgaben, an den Tag legen und wie sie selbst über ihre Benutzung mit dem Smartphone bei bestimmten Schwierigkeiten reflektieren. Die Ergebnisse vermitteln erste Eindrücke und zeigen Trends auf, die Ausgangspunkt für weitere umfangreichere Forschungen sein können, die ihrerseits der Fremdsprachendidaktik und der lexikographischen Praxis wichtige Impulse geben können. Interessant in diesem Zusammenhang sind die „ricerche incrociate“ („cross research“) und die Konsultation von multilingualen Online-Wörterbüchern mit der Konsultation des Englischen als „Sandwichsprache“.

**Keywords:** Benutzung von Smartphones; Online-Wörterbücher; Recherchekompetenz; Lexikographie und Fremdsprachendidaktik

## 1 Einleitung

Noch bis vor wenigen Jahren haben die Studierenden einer Fremdsprache im Fremdsprachenunterricht spezifische lexikalische Lücken beim Umkodieren in die Fremdsprache mit einem bilingualen Wörterbuch (oft sogar in Taschenbuchformat (sic!), vgl. Nied Curcio 2011) versucht zu schließen. Heute holen sie während des Fremdsprachenunterrichts spontan ihr Smartphone<sup>1</sup> hervor, um sprachli-

---

1 Der Begriff *Smartphone* wird hier als Hyperonym für *iPhone*, *smartphone*, *iPad* und *Tablet-Computer* verwendet, d.h. sämtliche Produkte, die in Form eines Mobiltelefons oder Mini-Computers die Funktionalität und Konnektivität eines Computers übernehmen.

che Schwierigkeiten anhand von Online-Informationen oder mit Hilfe von Apps zu überwinden.<sup>2</sup> Es ist offensichtlich, dass die Verwendung von Smartphones den Fremdsprachenlernenden die verschiedensten, fast unbegrenzten Möglichkeiten bietet, um bestehende sprachliche Schwierigkeiten in kürzester Zeit überwinden zu können. Auch die Möglichkeit, die Struktur des Online-Wörterbuchs durch Notizen, Lesezeichen, Verweise, Schlagwörter u.a. den individuellen Bedürfnissen anzupassen, die Wörterbucheinheiten zu aktualisieren, zu erweitern, zu diskutieren, sind verlockend. Im Bereich der Wörterbuchbenutzungsforschung liegen bisher – in Bezug auf die Übersetzung und die Fremdsprachendidaktik – noch relativ wenige empirische Forschungen vor (vgl. Mackintosh 1998; Nied Curcio 2011), auch wenn sich die Publikationen, die die Nutzerperspektive in den Vordergrund stellen, in den letzten Jahren vermehrten (Tarp 2011; Taljard, Prinsloo & Fricke 2011; Boonmoh 2012; de Schryver, Prinsloo 2011; Domínguez Vázquez, Mirazo & Vidal 2013, Müller-Spitzer 2013). Es hat sich mittlerweile bestätigt, dass Online-Wörterbücher häufiger als Printwörterbücher verwendet werden; die ersten Beiträge zu empirischen, nutzerorientierten, Untersuchungen sind erschienen.<sup>3</sup> Publikationen zur spezifischen Verwendung von Smartphones hinsichtlich lexikographischer Fragen in der Fremdsprachendidaktik sind mir derzeit nicht bekannt. Interessant ist m.E. eine spezifische Analyse zur Benutzung von Smartphones gerade deshalb, weil dessen Verwendung über eine „reine“ Wörterbuchbenutzung hinausgeht, weitere Informationsquellen einschließt und einen Einblick in die generelle Recherchekompetenz der Fremdsprachenlernenden gibt, denn die Lernenden sind in ihrer Wahl der Recherchertools und der Informationsquelle (z.B. Online-Wörterbücher bzw. Wörterbuch-Portale, Enzyklopädien, Suchmaschinen, Foren, usw.) völlig frei. Meiner Beobachtung nach benutzen die Fremdsprachenlernenden zurzeit wie selbstverständlich das Smartphone und es scheint, als würden sie es auch während des Fremdsprachenunterrichts verwenden,<sup>4</sup> und sich dabei auf den Gebrauch von zweisprachigen Internet-Wörterbüchern – und dort auf die immediate Suche nach dem passenden Äquivalent in der anderen Sprache – zu konzentrieren; das technische Potential und seine Recherchemöglichkeiten werden anscheinend nicht ausgenutzt. Es sieht demnach aus, als wäre die Benutzung des Smartphones ähnlich der Verwendung von zweisprachigen Print-Wörterbüchern, d.h. die Lernenden verwendeten insbesondere bilinguale Wörterbücher (vgl. Engelberg & Lemnitzer 2004; Albrecht 2005; Corda & Marelllo 2004: 82), sie würden meist nur das 1. Übersetzungsäquivalent in Be-

---

2 Das Nachschlagen in Printwörterbüchern wurde immer als Last empfunden (vgl. Hulstijn, Hollander & Greidanus 1996 in Engelberg & Lemnitzer 2004: 81) und ihre Verwendung scheint nun durch Online-Wörterbücher (vgl. Domínguez, Mirazo & Vidal 2013) abgelöst worden zu sein. Das Smartphone und die Möglichkeit, schnell und (fast) überall auf online-Informationen zurückgriffen zu können, scheint diesen Trend voranzutreiben.

3 Für einen Überblick über den Gebrauch von Online-Wörterbüchern s. Möhrs & Müller-Spitzer (2013), spezifische empirische Untersuchungen wurden von Domínguez, Mirazo & Vidal (2013) und Domínguez Vázquez, Mollica & Nied Curcio (2014) durchgeführt.

4 Diese Tatsache ist nicht selbstverständlich, war doch der Gebrauch von Wörterbüchern während des Fremdsprachenunterrichts lange Zeit untersagt, da das Wörterbuch zu sehr in Verbindung mit der Grammatik-Übersetzungsmethode gebracht wurde und viele Lehrpersonen zudem der Meinung waren, die Kenntnisse der Fremdsprache ohne Wörterbuch abprüfen zu müssen. Außerdem gaben sie den einsprachigen Wörterbüchern den Vorrang und die Verwendung von Smartphones an pädagogischen Institutionen ist zudem oft verboten.

tracht (vgl. Atkins & Rundell 2008) ziehen und schenken eventuellen metasprachlichen Notationen keine Beachtung (vgl. Nied Curcio 2011: 204; Dóminguez Vázquez, Mollica & Nied Curcio 2014). Um genauere Informationen zum Gebrauch des Smartphones während des Fremdsprachenunterrichts zu erhalten, wurde von mir im März 2014 eine empirische Untersuchung im akademischen DaF-Unterricht in Italien durchgeführt. Diese erste Untersuchung verfolgt ein wichtiges Ziel, nämlich auszuloten, welche Forschungsfragen in Bezug auf den Gebrauch von Online-Wörterbüchern, Smartphones und der Recherchekompetenz von Fremdsprachenlernenden noch offen sind und welche spezifischen empirischen Untersuchungen im Bereich der Wörterbuchbenutzungsforschung in Zukunft angegangen werden sollten, um daraufhin wichtige Impulse sowohl der lexikographischen Praxis als auch der Fremdsprachendidaktik zu geben. Im nächsten Kapitel (2.) wird auf das Experiment näher eingegangen; danach (Kapitel 3) werden die Ergebnisse exemplarisch präsentiert, um schließlich (Kap. 4) die Forschungsmöglichkeiten aufzuzeigen.

## 2 Ziel und Design der empirischen Untersuchung zum Gebrauch des Smartphones

Die empirische Untersuchung, die aus einer Kombination von Umfrage und spezifischen Aufgaben mit retrospektiven Fragen<sup>5</sup> bestand, mit Studierenden des Deutschen als Fremdsprache an der Universität Roma Tre durchgeführt wurde, sollte (neben den oben genannten Zielen) auch Auskunft darüber geben, in welchen kommunikativen und Lernsituationen sie das Smartphone für lexikographische Fragestellungen nutzen, bei welchen sprachlichen Schwierigkeiten sie auf Online-Informationen zurückgreifen, welche Internetseiten oder/und Online-Wörterbücher sie verwenden und ob sie den dort gefundenen Informationen vertrauen. In Bezug auf zweisprachige deutsch-italienische bzw. mehrsprachige Internetwörterbücher sollte herausgefunden werden, welche sie verwenden<sup>6</sup> und warum? Nutzen Sie auch Übersetzungstools/-programme oder Suchmaschinen?<sup>7</sup> Werden auch einsprachige Internetwörterbücher des Deutschen wie Das digitale Wörterbuch

---

5 *Retrospektive Fragen* (vgl. Faerch & Kasper 1987; Flick 2007) eignen sich gut, um herauszufinden, welche Überlegungen die Probanden bei einer bestimmten auszuführenden Handlung anstellen. Die Auswertung kann sowohl quantitativ als auch qualitativ erfolgen, da sie die klassische methodologische Dichotomie zwischen qualitativen und quantitativen Methoden überwindet, was vor allem im Bereich des Fremdspracherwerbs und der Fremdsprachendidaktik wünschenswert ist (vgl. Aguado 2009).

Für präzisere Ergebnisse zu den einzelnen Aufgaben wäre es besser, bei jeder einzelnen Aufgabe retrospektive Fragen zu stellen, was jedoch den Fragebogen beträchtlich verlängert und die Konzentration der Probanden strapaziert. Das *Think-Aloud-Protocol*, vgl. Flick 2007) wäre m.E. für weitere Forschungen die beste Möglichkeit, um genauere Daten zu erhalten.

6 Bekannte deutsch-italienische Internetwörterbücher sind: *Leo* (dict.leo.org/itde/index\_de), *Pons* (de.pons.eu/italienisch-deutsch/), *bab.la* (de.bab.la/woerterbuch/deutsch-italienisch/), *dict.cc* (deit.dict.cc/), *dicios* (www.it/dicios.com) oder ELDIT, ein lexikographisches Großprojekt der EURAC in Bozen (www.eurac.edu/ELDIT) [4.4.2014].

7 bspw. *Google*-Übersetzer translate.google.de oder die Kombination eines mehrsprachigen Wörterbuchs und Suchmaschinen wie z.B. *linguee* (www.linguee.de) [4.4.2014]

der deutschen Sprache des 20. Jahrhunderts,<sup>8</sup> die deutsche Seite von free dictionary<sup>9</sup> oder das einsprachige Duden-Online-<sup>10</sup> oder das italienische einsprachige italienische *Treccani*-Wörterbuch<sup>11</sup> spezifische Synonym-Wörterbücher<sup>12</sup> oder spezielle DaF-Wörterbücher<sup>13</sup> konsultiert? Nutzen sie Wortschatzportale<sup>14</sup> und Wortschatz-Glossare, wie bspw. das Glossar vom Goethe-Institut<sup>15</sup> oder Internet-Seiten zur deutschen Sprache wie canoo.net, oder Online-Enzyklopädien, Abbildungen, Videos, bestimmte Online-Texte oder Korpora? All diese Möglichkeiten stehen den Lernenden frei zu Verfügung.

Hierzu wurden folgende konkrete Fragen gestellt, die von den Probanden nacheinander (ohne vor- oder zurückzublättern) beantwortet wurden.<sup>16</sup> Die Fragen beinhalteten Entscheidungsfragen (1., 6., 12., 16., 18., 22., 24., 26.), Multiple-Choice- (1., 11,) und offene Fragen (2., 3., 4., : 5., 6., 7., 8., 9., 10., 13., 14., 15., 17., 19., 20., 21., 22., 23., 25.):

- (1) Benutzt du das Smartphone während des Fremdsprachenunterrichts, um bestimmte Informationen zu suchen? JA/ Nein. Wenn ja: häufig?/manchmal/ selten?
- (2) Warum nutzt du es/ nutzt du es nicht?
- (3) In welcher Situation suchst Informationen (z.B. während einer Grammatik-Übung, eines Lesetextes, Schreibaufgabe, um Vokabeln nachzuschlagen, um Vokabeln zu lernen,...)?
- (4) Suchst du im Internet oder benutzt du eine bestimmte App? Wenn du eine App verwendest, welche? Wenn du im Internet suchst, auf welchen Seiten suchst du normalerweise?
- (5) Benutzt du auch Online-Wörterbücher? Wenn ja, welche?
- (6) Benutzt du Google-Translator oder ein anderes Übersetzungsprogramm? JA/ Nein.
- (7) Welche Internetseite oder App ist deiner Meinung nach die beste, um nach der *Bedeutung eines deutschen Wortes* zu suchen? Warum?
- (8) Welche Internetseite oder App ist deiner Meinung nach die beste, um *ein italienisches Wort* ins Deutsche zu *übersetzen*?<sup>17</sup> Warum?
- (9) Traust du den Informationen, die du findest? Warum ja/nein?
- (10) Gibt es auch Probleme/ Schwierigkeiten? Wenn ja, welche?
- (11) Benutzt du das Smartphone auch in authentischen Kommunikationssituationen? Zum Beispiel
  - wenn du mit deutschsprachigen Personen zusammen bist?
  - wenn du ein Video in deutscher Sprache anschaust?

---

8 [www.dwds.de](http://www.dwds.de) [4.4.2014]

9 [de.thefreedictionary.com/](http://de.thefreedictionary.com/) [10.11.2013]

10 [www.duden.de](http://www.duden.de) (10.11.2013)

11 [www.treccani.it/vocabolario/dizionario/](http://www.treccani.it/vocabolario/dizionario/) [10.04.2014]

12 [www.duden.de/rechtschreibung/Synonymwoerterbuch](http://www.duden.de/rechtschreibung/Synonymwoerterbuch) oder [synonyme.woxikon.de/](http://synonyme.woxikon.de/) [10.11.2013]

13 [de.pons.eu/deutsch-als-fremdsprache/](http://de.pons.eu/deutsch-als-fremdsprache/) [10.11.2013]

14 [wortschatz.uni-leipzig.de/](http://wortschatz.uni-leipzig.de/)[10.11.2013]

15 [www.goethe.de/z/jetzt/dejwort/dejwort.htm](http://www.goethe.de/z/jetzt/dejwort/dejwort.htm) [10.11.2013]

16 Die Fragen wurden in italienischer Sprache formuliert. Sie werden hier direkt in deutscher Sprache wiedergegeben. Es wurde bewusst die *Du*-Form gewählt, da im italienischen Kontext die Form des *Lei* (,Sie') eine zu große psychologische Distanz schafft. Die Form des *voi* (,ihr') schien mir zu unpersönlich und generell.

17 Während es bei Frage 7 um die Konsultation in Bezug auf die sprachliche Rezeption geht, fokussiert die Frage 8 eine Handlung in Bezug auf die sprachliche Produktion.

- wenn du deutsches Radio hörst?
- um Vokabeln zu lernen?
- wenn du einen Text/ Buch/ online-Informationen auf Deutsch liest?
- Ich benutze ein Online-Wörterbuch auch, um nur darin zu „lesen“/„herumzublättern“ ohne ein bestimmtes Ziel
- Anderes: .....

Während die ersten 11 Fragen allgemein gehalten wurden und auf der Erinnerung der Lernenden basierten, wurden für die nächsten Fragen konkrete Aufgaben formuliert, um herauszufinden, welche Schwierigkeiten die Probanden bei einer bestimmten Handlungssituation im Fremdsprachenunterricht haben und wie sie dabei mit dem Smartphone umgehen. Zuerst sollten die Probanden 5 Sätze mit polysemen italienischen Verben ins Deutsche übersetzen:<sup>18</sup>

- A. *Devo chiedere al mio capo* (,Ich muss meinen Chef *fragen.*‘),
- B. *Mi ha chiesto di te* (,Er/Sie hat mich *nach dir gefragt.*‘),
- C. *Mi ha chiesto un favore* (,Er/Sie hat mich *um einen Gefallen gebeten.*‘),
- D. *Per Natale mia figlia mi ha chiesto un viaggio* (,Für Weihnachten hat sich meine Tochter von mir eine Reise *gewünscht/hat mich um eine Reise gebeten.*‘) und
- E. *Mi chiede sempre cose impossibili* (,Sie *verlangt* immer Unmögliches von mir‘/‘Sie fragte immer unmögliche Sachen‘/,Sie fragt (mich) immer nach unmöglichen Sachen‘).

Direkt anschließend wurden retrospektive Fragen gestellt:

- (12) Hast du das Smartphone benutzt, um nach Informationen zu suchen? JA/ Nein.
- (13) Wenn ja, für welche/n Satz/ Sätze?
- (14) Warum/ Was hast du gesucht?
- (15) Wo hast du gesucht (Internetseiten, Online-Wörterbücher,...)?
- (16) Bist du auf Schwierigkeiten gestoßen bei den Informationen, die du gefunden hast bzw. bei dem Wörterbuch, das du benutzt hast?
- (17) Wenn ja, welche?
- (18) Danach sollten die Studierenden jeweils vier italienische bzw. deutsche Wörter und jeweils einen Satz in die andere Sprache übersetzen. Die Wörter waren: *F. merenda*, *G. bamboccioni*, *H. raccomandazioni*, *I. ISTAT* sowie *J. Wendehals*, *K. Quark*, *L. hartzten*, *M. hdl*. Zu den Wörtern gehörten Kulturspezifika, Neologismen und Abkürzungen, z.T. aus der gesprochenen Sprache, die bei einer Übersetzung in eine andere Sprache Schwierigkeiten bereiten, und zudem oft nicht in (Online-)

---

18 Da die Studierenden mit der Übertragung polysemer italienischer Verben nachweislich Schwierigkeiten haben (vgl. Nied Curcio 2005), eignet sich diese Aufgabe besonders gut, um herauszufinden, ob das Wörterbuch, das die Studierenden verwenden, verschiedene Übersetzungsäquivalente angibt und ob metasprachliche Notationen zur Valenz und zum Kasus bei der Disambiguierung der Lesart des polysemen Verbs helfen können, bzw. ob die Studierenden diese Informationen – vorausgesetzt sie sind vorhanden – überhaupt berücksichtigen.

Wörterbüchern. Bspw. ist *hartzen* nicht in den bekanntesten deutsch-italienischen Wörterbüchern *Leo*, *Pons*, *canoo.net*, *dicios*,<sup>19</sup> vertreten, während es jedoch bspw. in der online-Enzyklädie *Wikipedia* erklärt wird und somit durch eine einfache Suche gefunden werden kann. In *duden.de* werden sogar Wortbedeutung („von Hartz IV leben“), Gebrauch (Jargon), Wortart, Worttrennung, Aussprache, Deklination und ein Beispiel mit der übertragenen Bedeutung im Jugendjargon (gestern Abend war ich nur am Hartzen (*konnte mich zu keiner Arbeit, Tätigkeit überwinden*)) angegeben. Bei den Sätzen ging es um umgangssprachliche Ausdrücke, die sehr häufig sind, aber in der Lexikographie oft vernachlässigt werden: N. „*Kommst du noch auf einen Absacker mit?*“ und O. „*È stato stroncato da un infarto mentre lavorava..*“ Die Schwierigkeit im Satz N. ist das Verständnis der Bedeutung von *Absacker*, im Satz O. geht es um den produktiven Transfer des Wortes *stroncare* und des gesamten Satzes mit einem Verb im Passiv und einem im Gerundio.<sup>20</sup> Die Online-Wörterbücher *pons*, *leo*, *bab.la*<sup>21</sup> und *Eldit* enthalten keinen Eintrag für *Absacker* und geben nicht das richtige Äquivalent für *stroncato* an (*Eldit* enthält keinen Eintrag für *stroncare*), *dicios* enthält in einer Liste die Äquivalente *abbrechen*, *dahinraffen*, *erliegen*, *zerstören*, *zunichtemachen* (das 2. und 3. wäre in diesem Kontext richtig), ohne jedoch weitere Zusatzinformationen zu geben, damit der Benutzer das richtige Äquivalent auswählen und den Satz korrekt produzieren könnte. *Google Übersetzer* gibt für den ersten Satz folgende Übersetzung an: \**Sarà ancora con un drink*, was in die richtige Richtung geht, jedoch grammatisch nicht korrekt ist, während wenn man nur das Wort *Absacker* eingibt, das Wort *berretto da notte* („Schlafmütze“) erscheint, was kontextuell selbstverständlich ebenfalls keinen Sinn ergibt.<sup>22</sup> Für den zweiten Satz gibt die Übersetzungsfunktion von *Google* im Deutschen \**wurde von einem Herzinfarkt schlug während der Arbeit* an, sucht man nur nach dem Wort *stroncato*, so erscheint *verrissen*.<sup>23</sup>

Auch zu diesen Übersetzungsaufgaben wurde retrospektiv gefragt (wie 12-17) :

- (19) Hast du das Smartphone benutzt, um nach Informationen zu suchen? JA/ Nein.
- (20) Wenn ja, für welche/n Satz/ Sätze?
- (21) Warum/ Was hast du gesucht?
- (22) Wo hast du gesucht (Internetseiten, Online-Wörterbücher,...)?
- (23) Bist du auf Schwierigkeiten gestoßen bei den Informationen, die du gefunden hast bzw. bei dem Wörterbuch, das du benutzt hast?
- (24) Wenn ja, welche?

19 [it.dicios.com/](http://it.dicios.com/) [10.04.2014]

20 Das italienische *Gerundio* hat keine exakte Entsprechung im Deutschen und muss i.d.R. entweder durch eine Nominal- oder Verbalphrase im Deutschen wiedergegeben werden (s. Sattler 2008).

21 [www.bab.la](http://www.bab.la) [10.04.2014]

22 Zu vermuten ist, dass es sich hier um eine inkorrekte, wörtliche – und nicht figurative – Übertragung des engl. „nightcap“ („Schlummertrunk“) geht.

23 Aus offensichtlichen Platzgründen muss hier auf eine umfassende und detaillierte Präsentation der Ergebnisse verzichtet werden; die gewählten Beispiele sollen exemplarisch die Problematik aufzeigen.



Interessant in diesem Zusammenhang ist natürlich nicht nur die subjektive Reflexion der Probanden bezüglich ihrer eigenen Wörterbuchbenutzung, sondern auch das Ergebnis selbst. Wurden die Aufgaben korrekt ausgeführt bzw. welche Fehler wurden gemacht? War die Konsultation des Smartphones objektiv erfolgreich und wenn nicht, welche Benutzungsfehler gehen aus einem Vergleich der Ergebnisse und der Benutzerhandlung<sup>24</sup> hervor?

Zum Abschluss wurden von mir noch zwei allgemeine Fragen zur Lexikographie eingefügt:

(25) Möchtest du mehr über Wörterbücher erfahren? Ja/Nein.

(26) Wenn ja, was genau?

(27) Möchtest du mehr über Recherchemöglichkeiten im Internet (einsprachige deutsche Online-Wörterbücher, zweisprachige Online-Wörterbücher, Glossare, Korpora, Internetportale, spezifische Internetseiten,...) erfahren?

### 3 Die Ergebnisse der Untersuchung<sup>25</sup>

#### 3.1 Die Benutzung des Smartphones im Fremdsprachenunterricht

An der Untersuchung nahmen 36 Germanistikstudierende<sup>26</sup> teil. 32 Probanden hatten ein Smartphone/I-phone mitgebracht, zwei ein Tablet und zwei Probanden waren nicht im Besitz eines Smartphones und haben das ihres Nachbarn benutzt. Die Auswertung erfolgte sowohl quantitativ als auch qualitativ; die Ergebnisse sollen im Folgenden mit ihrem prozentualen Anteil zusammenfassend beschrieben werden.<sup>27</sup>

Von den Probanden benutzen 80% ihr Smartphone generell während des Fremdsprachenunterrichts, um bestimmte Informationen hinsichtlich des Unterrichts zu suchen. 38% der Studierenden suchen nach Wörtern für die schriftliche Textrezeption (und 32% für die Textproduktion; 18% verwenden es, um Vokabeln zu lernen. Fast immer geht es dabei um Wörter: „per cercare il significato delle parole“

24 Zur Klassifikation der Wörterbuchbenutzungshandlungen sowie der Benutzungsfehler vgl. Wiegand 1998.

25 Die Ergebnisse können hier aus offensichtlichen Platzgründen nur stark gekürzt und ausschnitthaft präsentiert werden. Auch auf Graphiken, die für eine bessere Übersichtlichkeit dienen könnten, muss aus dem gleichen Grund verzichtet werden.

26 Von den 36 Studierenden waren 34 Muttersprachler Italienisch, eine spanischsprechende Studierende und eine Studentin mit Rumänisch als Muttersprache. 28 der Studierenden (78%) lernten seit Beginn ihres Universitätsstudiums Deutsch, d.h. seit 2 Jahren und haben somit noch relativ geringe Deutschkenntnisse (Niveaustufe A2-B1 nach dem Gemeinsamen Europäischen Referenzrahmen); Die anderen 8 Studierenden (22%) hatten Deutsch als Fach in der Schule und lernen die Sprache seit 5-7 Jahren. Die Zeit war von mir nicht limitiert, das Design jedoch für eine Bearbeitung innerhalb 90 min. ausgelegt. Die Studierenden gaben nach 75-90 min. ab.

27 Es ist selbstverständlich, dass eine Untersuchung mit 36 Studierenden nur eine erste Idee geben kann und in ihrer Quantität keinesfalls eine valide Aussagekraft erlangt. Trotzdem zeigt sie interessante Ergebnisse auf (vgl. 3) und weist m.E. auch auf Potential für weitere umfangreichere Forschungsmöglichkeiten hin (vgl. 4.).

(,um die Bedeutung von Wörtern nachzuschlage‘), „durante una produzione scritta, per cercare dei vocaboli“ (,während der Textproduktion, um die Wörter nachzuschlagen‘) oder „per studiare dei vocaboli“ (,beim/zum Vokabellernen‘). Nur zwei Probanden nannten auch „espressioni“ (,Ausdrücke‘) und „modi di dire“ (,Redewendungen‘), die sie nachschlagen.

Fast alle Studierenden gebrauchen sowohl das Internet, als auch verschiedene Apps. Sie geben den Apps den Vorrang, wenn sie die Bedeutung eines konkreten Wortes suchen und es schnell gehen soll. An 1. Stelle bei den Apps steht das *Pons*-Wörterbuch und an 2. Stelle *Google Übersetzer*, gemeinsam mit dem Portal *WordReference*.<sup>28</sup> Bei der Internet-Recherche geben Sie *Google Übersetzer* den Vorrang und an 2. Position stehen die *Google*-Suchfunktion und *Wikipedia*. Seltener genannt werden *Pons*, *Youtube* und *WordReference*. Es wird deutlich, dass das wichtigste Kriterium die schnelle und direkte Suche ist.

Bei der Frage 5, in der konkret nach der Verwendung von Online-Wörterbüchern gefragt wird, geben 34 Studierende (94%) an, dass sie welche benutzen: Die drei am häufigsten genannten sind: *WordReference* (12), *pons* (11) und *Eldit* (11). Außerdem wurde 4 Mal *bab.la* sowie je 3 Mal *Leo* und *Larousse*<sup>29</sup> angeführt, die auch eine zweisprachige Wörterbuchversion Italienisch↔Deutsch beinhalten. Von drei Probanden wurde auch *Google Übersetzer* als Wörterbuch aufgezählt (sic!). Nur vereinzelt wurden einsprachige Wörterbücher (*Duden*, *Treccani* für das Italienische) oder Portale wie *Collins*<sup>30</sup>, *Reverso*<sup>31</sup>, *Urban dictionary*<sup>32</sup>, *dictionary reference*<sup>33</sup>, und das italienisch-deutsche Wörterbuch des *Corriere della Sera* erwähnt. Die spezifische Frage, ob sie *Google Übersetzer* verwenden, beantworteten 25 (69%) der Germanistikstudierenden mit Ja.

Die Antworten auf die Frage 7. *Welche Internetseite oder App ist deiner Meinung nach die beste, um nach der Bedeutung eines deutschen Wortes zu suchen? Warum?* zeigen folgende Ergebnisse.

Zehn der Studierende (28%) geben *Eldit* den Vorrang, gefolgt von *Pons* (19%). An dritter Stelle der Präferenzen stehen *Leo* und *Larousse*. Zweimal wurde auch *Google Übersetzer* angeführt. Interessant ist, dass die Studierenden zwar verschiedene Internetseiten/Apps aufgezählt haben, aber trotzdem die gleichen Gründe für ihre Wahl nennen, und zwar:

- (1) Zufriedenstellende Beispielsätze (30%)
- (2) Das Wort wird im Kontext angeführt (19%)
- (3) Erwähnung von typischen Ausdrücken, Idiomen (11%)
- (4) Grammatische Informationen, z.B. Konjugation, Valenz (8%)
- (5) Leichter und schneller Zugriff (8%)
- (6) Präzise und komplett (5%)
- (7) Liste von Übersetzungsäquivalenten (5%)

28 [www.wordreference.com/](http://www.wordreference.com/) [10.04.2014]

29 [www.larousse.fr/dictionnaires/allemand-italien](http://www.larousse.fr/dictionnaires/allemand-italien) [10.04.2014]

30 [www.collinsdictionary.com/dictionary/italian-english/dizionario](http://www.collinsdictionary.com/dictionary/italian-english/dizionario) [10.04.2014]

31 [dizionario.reverso.net/](http://dizionario.reverso.net/) [10.04.2014]

32 [www.urbandictionary.com/](http://www.urbandictionary.com/) [10.04.2014]

33 [dictionary.reference.com/](http://dictionary.reference.com/) [10.04.2014]

Acht Probanden (22%) beantworteten die Frage nicht oder gaben an, dass sie keine Antwort wussten. Dieser Anteil liegt bei Frage 8., in der nach der besten Internetseite/ App zum *Übersetzen* gefragt wurde, sogar bei 12 (33%). Auch die Ergebnisse im Vergleich zu 7. sind unterschiedlich. Als die beste Seite/App wird *Pons* zitiert (22%), da der Zugriff leicht sei, die Beispiele zufriedenstellend seien, das Wort im Gebrauchskontext stehe, es viele Kollokationen und (idiomatische) Ausdrücke gebe und grammatische Zusatzinformationen (auch die syntaktische Struktur) vorhanden seien. Interessant ist, dass an 2. Stelle das Print-Wörterbuch liegt (14%) – obwohl sich die Frage eigentlich nur auf Online-Wörterbücher bezog. An 3. Stelle liegt *Leo* (11%). Zwei Studierende waren der Meinung, dass kein Wörterbuch besser als das andere sei, sondern dass die „ricerche incrociate“ bzw. cross research oder „überkreuzte“ Recherchen, bei denen man hin- und herspringt, vergleicht und abwägt, was das Beste wäre. Eine weitere interessante Aussage war die einer Studentin, die meinte, dass für sie *WordReference* die beste Internetseite sei, da sie bei der Übersetzung vom Italienischen ins Deutsche (was ja ihre Fremdsprache ist) via Englisch übersetze, und nicht direkt Italienisch>Deutsch.

Bei Frage 9., ob die Studierenden den Informationen, die sie finden, trauen, gab es folgende Ergebnisse:

Ja (absolut): 19%

Ja, aber ich suche noch an anderer Stelle: 22%

Ja, meistens: 15%

Ja, ziemlich: 8%

Nicht immer: 14%

Nein, ich suche noch an anderer Stelle: 22%

Der Prozentsatz der Studierenden, die sich auf die Informationen im Internet mehr oder weniger verlassen – wenn auch mit Einschränkung – beträgt doch 64%. Gründe dafür waren, dass ihnen die Seite von der Lehrperson empfohlen wurde, dass sie die Seite kennen oder dass sie gute Erfahrungen damit gemacht hatten. 44% der Probanden (mehr oder weniger von den Online-Informationen überzeugt) suchen noch an anderer Stelle und nannten als Grund ihre Zweifel aufgrund von negativen Erfahrungen. Drei Probanden, die sich auf die Informationen im Internet nicht verlassen, trauen nur dem Print-Wörterbuch.

67% der Probanden gaben – in Bezug auf die Schwierigkeiten, die sie mit Online-Informationen haben – an, dass sie sich unsicher fühlen und Zweifel haben, da sie nie wissen, *welches* Äquivalent in einem bestimmten Kontext das richtige ist<sup>34</sup>. Diese Unsicherheit zeige sich auch, wenn sie verschiedene Internetseiten konsultierten und die gefundenen Informationen nicht übereinstimmten. Das Problem ist demnach die Auswahl des richtigen Übersetzungsäquivalents, v.a. gerade dann, wenn zu wenige Übersetzungsäquivalente/ Bedeutungsvarianten angegeben würden und die Informationen zu vage seien. Ihrer Meinung nach mangle es an Kriterien, die ihnen bei der Orientierung helfen, sowohl auf der Mikro- als auch auf der Makrostruktur. Mangel an Informationen wurde von ihnen auch

---

34 Dies gilt sicherlich insbesondere für die sprachliche Produktion.

in Bezug auf die metasprachliche Notation angegeben. Zudem fehlten grammatische Informationen wie z.B. Präpositionen, die von Nomen und Verben regiert werden („mancano delle informazioni grammaticali (es. le preposizioni che reggono certi nomi/ verbi)“). Weitere Mängel wurden auch in Bezug auf Phraseologismen, Wortkompositionen<sup>35</sup> und Fachbegriffe genannt.

Auch in authentischen Situationen, in denen sie mit Deutsch in Kontakt sind, wird das Smartphone von ihnen verwendet: 75% der Probanden nutzen es v.a. beim Lesen eines Textes/Buches oder von Online-Informationen auf Deutsch (d.h. Rezeption), 69% der Studierenden verwenden es während sie ein Video/einen Film in deutscher Sprache anschauen (d.h. Rezeption) und 64% lernen damit Vokabeln. Nur selten wird darin ohne Ziel „gelesen“/„herumgeblättert“ (22%). In Anwesenheit von Muttersprachlern wenden sich die Studierenden i.d.R. an diese (11%), da sie als Experten betrachtet werden und als Autorität über dem Wörterbuch stehen. Dass sie von dieser Möglichkeit Gebrauch machen, gab eine Probandin auch als Antwort bei Frage 9 an, in der es um vertrauensvolle Informationen ging: „Non sempre mi fido quindi provo a consultare più fonti o chiedo a persone più competenti (amici madrelingua).“ (‘nicht immer traue ich [den Informationen], deshalb versuche ich mehrere Quellen heranzuziehen oder frage Leute, die kompetenter sind (Muttersprachler)’). Am wenigsten benutzen sie es beim Radiohören (5%).<sup>36</sup>

### 3.2 Wie Wörterbuchnutzung bei spezifischen Übersetzungsaufgaben

In diesem Kapitel werden die Ergebnisse vorgestellt, die sich aus der Analyse der Übersetzung der fünf Sätze A.-E. mit dem polysemen italienischen Verb *chiedere* (s.2.), der Wörter (F.-M.) und der beiden Sätze N. und O. Sie werden zudem in Bezug auf ihre Korrektheit und die Konsultation des Wörterbuchs von Seiten der Germanistikstudierenden, bzw. die Reflexion der Studierenden gebracht.

#### 3.2.1 Die Übersetzungen des italienischen Verbs *chiedere* und die Benutzung des Smartphones

Auch wenn man berücksichtigt, dass die Studierenden noch kein hohes Sprachniveau haben, machen die Ergebnisse sehr nachdenklich, denn der höchste Prozentsatz an Korrektheit bei einem Satz lag nur bei 31% für Satz A, was auf große Schwierigkeiten von Seiten der Studierenden hinweist.

31% A. *Devo chiedere al mio capo* (‘Ich muss meinen Chef fragen.’)

35 Dieser Mangel kann nicht unbedingt den Wörterbüchern zugeschrieben werden; die Komposition, auch ad-hoc, ist charakteristisch für die deutsche Sprache, und nicht jede Komposition kann in einem Wörterbuch aufgeführt werden, auch wenn Online-Wörterbücher weniger als Print-Wörterbücher das Platzproblem haben und „auf dem Laufenden“ sein könnten. Hier würde eine zusätzliche Suche mit Suchmaschinen weiter helfen, um die Bedeutung des Wortes zumindest im Kontext auflösen zu können. Aber auch eine sprachliche Dekomposition mit einer erneuten Suche im Wörterbuch kann zu einem positiven Ergebnis führen.

36 Diese Frage war zu sehr auf das Radio limitiert, und müsste bei einer zukünftigen Untersuchung generell auf Hörtexte erweitert werden, so dass auch das Musikhören einbezogen wird. Eine Studentin gab von sich aus an, dass sie beim Musikhören das Smartphone zur Konsultation von unbekanntem Wörtern verwende. (Bei einer spezifischeren Fragestellung wäre diese Möglichkeit sicher öfters in Betracht gezogen worden.)

- 22% B. *Mi ha chiesto di te* („Er/Sie hat mich *nach* dir *gefragt*.“)  
28% C. *Mi ha chiesto un favore* („Er/Sie hat mich *um* einen Gefallen *gebeten*.“)  
11% D. *Per Natale mia figlia mi ha chiesto un viaggio* („Für Weihnachten hat sich meine Tochter von mir eine Reise *gewünscht*/hat mich *um* eine Reise *gebeten*.“)  
20% E. *Mi chiede sempre cose impossibili* („Sie *verlangt* immer Unmögliches von mir/“Sie *fragte* immer unmögliche Sachen/„Sie *fragt* (mich) immer *nach* unmöglichen Sachen“).

Für die Entscheidung, ob der Satz als richtig interpretiert wurde, war die Tatsache, ob

- (1) Die Probanden das richtige Übersetzungsäquivalent gefunden haben,
- (2) ob sie auf die grammatischen Zusatzinformationen geachtet hatten (falls diese vorhanden waren). Es ging hier insbesondere um die Valenz und den Kasus.<sup>37</sup>

Die Schwierigkeit der Übersetzung lag bei Satz A. insbesondere in der kontrastiven Valenz: Im Italienischen braucht das Verb *chiedere* ein Subjekt und ein indirektes Objekt, während das deutsche Verb *fragen* ein Subjekt und eine Akkusativergänzung realisieren muss. Die Fehler lagen zu 80% bei der Valenz, z.B. *\*Ich muss meinem Chef fragen*, aber auch in der Wahl des falschen Verbs (20%), wie *\*Ich muss meinen Chef anfragen* oder *\*Ich muss meinen Chef nachfragen*. Satz A war der Satz, zu dem das Smartphone am wenigsten verwendet wurde. Vermutlich waren sich einige Probanden der kontrastiven Valenz nicht bewusst, haben sich nur auf die Bedeutungsentsprechung konzentriert, und nicht nachgeprüft. Sie haben parallel zur italienischen Valenz den deutschen Satz konstruiert. Die Konsultation des Smartphones lag von allen recherchierten Sätzen bei A. am niedrigsten: nur 44% der Studierenden haben das Smartphone für die Übersetzung dieses Satzes verwendet.

Bei Satz B. war es wiederum die Valenz. Nicht nur die italienische Präposition *di*, sondern auch das direkte Objekt im Italienischen wurde oft nicht adäquat im Deutschen wiedergegeben. Sätze wie *\*Sie fragt an mich über dich./ \*Sie hat mich über dich gefragt./ \*Er hat von dir mir gefragt./ \*Er fragte mich nach dich*. waren häufig. Der Satz *\* \*Er fragte mich über dich*. wurde von denjenigen geschrieben, die *google Übersetzer* verwendet hatten und – da sie auf Fehler gestoßen waren – in Anlehnung an den dort aufzufinden Satz *\*Er fragt mich über Sie*. umformuliert haben: „Ho cercato l’intera frase da Google Translator ma nella traduzione ho rilevato alcuni errori.“ („Ich habe den ganzen Satz in Google Translator gesucht aber in der Übersetzung habe ich einige Fehler bemerkt.“). 77% der Probanden haben sich des Smartphones hier bedient und es ist zu bemerken, dass auch diejenigen, die in einem Online-Wörterbuch wie z.B. *Leo*, nachgeschaut haben, Fehler gemacht haben, was zeigt, dass sie sich bei den verschiedenen Übersetzungsäquivalenten nicht orientieren konnten, denn *Leo* gibt für *chiedere di qu./qc.* vier Äquivalente an (*sich<sup>Akk</sup> nach jmdm/etw. erkundigen; nach jmdm./ etw. fragen; nach jmd. Fragen* und *nach jmd. verlangen*). Doch auch eine Recherche in *Pons*, wo *chiedere notizie di qu – sich nach jmdm erkundigen* steht, sind die

37 Dabei muss erwähnt werden, dass morphologische Fehler wie bspw. eine falsche Konjugation (*\*Er bitt mich um unmögliche Dinge*) sowie Fehler in der Satzstellung (*\*Er bat um einen Gefallen mich*) nicht berücksichtigt wurden, da der Schwerpunkt der Untersuchung darauf lag, ob die Probanden das richtige Übersetzungsäquivalent finden und auf die grammatischen Zusatzinformationen achten.

Sätze nicht korrekt, da die Probanden nicht auf *jmdm* geachtet haben und dadurch keine Dativergänzung realisiert haben: \**Er hat sich nach dich erkundigt*.

Obwohl der Satz C. ein Funktionsverb bzw. Phraseologismus in beiden Sprachen enthielt (*chiedere un favore – um einen Gefallen bitten*), war die Fehlerquote etwas geringer als bei B, lag aber immer noch hoch. 83% der Probanden haben diesen Satz im Smartphone recherchiert. Wenn sie nach dem gesamten Ausdruck in einem Online-Wörterbuch (oder den gesamten Satz in *Google* Übersetzer gesucht hatten), dann kamen sie zum richtigen Satz. Einige suchten jedoch nur nach dem Wort *favore* und schrieben deshalb \**Sie/Er hat mir einen Gefallen gefragt*.

Satz D war der Satz, bei dem die Fehlerquote am höchsten war (72%). 64% der Probanden haben dafür das Smartphone verwendet, d.h. weniger als bei Satz C. Die Einträge sowohl in den gängigen Online-Wörterbüchern als auch in Übersetzungsprogrammen weisen die Studierenden nicht darauf hin, dass es sich hier um den Kontext eines *Geschenks* geht. Der Satz wurde nur von denjenigen richtig übersetzt, die entweder schon mehrere Jahre Deutsch lernen oder die einen Ausdruck für *chiedere un regalo* (‘um ein Geschenk bitten’) gesucht hatten.

Der fünfte Satz (E.) wurde ebenfalls von 80% der Studierenden nicht korrekt übersetzt. 66% konsultierten das Smartphone hierzu. Auch wenn man gestehen muss, dass die Übersetzung des Satzes ins Deutsche für Lernende, die in der Mehrheit erst seit 2 Jahren Deutsch lernen, sehr schwierig war, muss doch unterstrichen werden, dass es – auch wenn sie *fragen* als Verb verwendeten, meist die Valenzangaben (wie bei A.) nicht richtig realisiert wurden.

Generell kann gesagt werden, dass 35 von 36 Probanden (!) das Smartphone für fast alle Sätze verwendet haben! – und trotzdem lag die Fehlerquote zwischen 69-89%! Bei der Suche nach Informationen haben sie – wie auch unter 3.1. beschrieben – nur die bekanntesten Online-Wörterbücher zu Rate gezogen oder mit Übersetzungsprogrammen gearbeitet; es wurden keine weiteren Hilfsmittel und Recherchemöglichkeiten – wie unter 2. beschrieben – in Betracht gezogen. Hier liegt sicher schon der erste Benutzerfehler<sup>38</sup>, nämlich der Wörterbuchwahlfehler. Wie man jedoch an der Relation zwischen Häufigkeit der Benutzung des Smartphones und Fehlerquote ablesen, begehen die Studierenden auch eine Menge Handlungsfehler, v.a. sind es die Nichtbeachtung von grammatischen Zusatzinformationen (Verbergänzung (z.B. Akkusativ- oder Dativergänzung), regiertem Kasus einer Präpositivergänzung (z.B. einen Gefallen bitten + um+*Akk*), aber auch die Suche n u r nach einzelnen *Wörtern*, und nicht nach dem gesamten Phraseologismus. Diese Fehler führten i.d.R. immer zu einem nicht korrekten Satz. Die Umkehrung bedeutet jedoch nicht immer, dass bei Beachtung dieser Informationen der Satz korrekt übersetzt wird, denn das Online-Wörterbuch selbst bietet, wie wir gesehen haben, nicht immer eine wirkliche Unterstützung, denn manchmal gibt es nur bedingt grammatische Zusatzinformationen oder/und es gibt mehrere Übersetzungsäquivalente, so dass die Benutzer keine Orientierungshilfe für die Auswahl des richtigen Äquivalents finden.

---

38 Zu Wörterbuchbenutzerfehler s. Wiegand 1998: 519.



### 3.2.2 Die Übersetzung von Neologismen und die Benutzung des Smartphones

Alle Probanden (100%) gaben an, dass sie ihr Smartphone bei den Übersetzungsaufgaben F.-O. benutzt hatten, um nach bestimmten Informationen in Bezug zu suchen. 33 davon (92%) haben zugegeben, dass sie bei der Aufgabenbewältigung gleichzeitig Schwierigkeiten mit den Internetseiten oder Online-Wörterbüchern hatten, die sie für diese Aufgabenstellung konsultiert hatten. Diese Schwierigkeiten, Zweifel und Unsicherheiten werden von ihnen oft auch geäußert (s. Frage 23): Die häufigste Antwort war: „non ho trovato alcune parole“ (‚einige Wörter habe ich nicht gefunden‘). Weitere Kommentare waren bspw. „risultava strana la traduzione“ (‚die Übersetzung war seltsam‘), „più traduzioni discordanti“ (‚mehrere gegensätzliche Übersetzungen‘) und „sempre quella di riuscire a capire/trovare la giusta traduzione“ (‚immer das gleiche [Problem], nämlich zu verstehen, welche die richtige Übersetzung ist‘).

Bei der Beschreibung der Ergebnisse in Bezug auf die Übersetzungsäquivalente der italienischen und deutschen Wörter in F.-M. möchte ich mich auf die deutschen Neologismen *hartzen*, *Absacker* und die Abkürzung *hdl*, sowie den umgangssprachlichen italienischen Ausdruck *stroncare* (‚sterben‘) in Zusammenhang mit *Herzinfarkt* im Satz O. konzentrieren. Der Grund für die Auswahl der deutschen Wörter liegt darin, dass für *hartzen* und *hdl* in den Online-Wörterbüchern *Pons* und *Leo* kein Eintrag zu finden ist (für *Absacker* gibt es zwar die Entsprechung *bicchiere della staffa*, doch meinten die Studierenden, dass sie nicht wüssten, worum es ginge, was darauf hinweist, dass es ein weniger gebräuchlicher Ausdruck als im Deutschen ist). Hätten die Probanden jedoch in der *Pons*-Textübersetzung geschaut, so hätten sie den korrekt übersetzten Satz *unitevi a me per un bicchierino?* gefunden, was der deutschen Bedeutung ziemlich nahe kommt. Um das richtige Äquivalent zu finden, waren die Studierenden darauf angewiesen, nicht nur das Online-Wörterbuch zu verwenden, sondern weitere (auch sehr simple) Recherchen zu tätigen. Zu einer erfolgreichen Recherche führten hier die Konsultation der Suchmaschine von *Google* im Allgemeinen, da man bei *hartzen* auf die *Wikipedia*-Seite geführt wird und dort eine Bedeutungserklärung findet. Bei der *Google*-Suche nach *Absacker* erscheint an oberster Stelle die Webseite *duden.de* mit der Erklärung „am Ende eines Zusammenseins oder vor dem Schlafengehen getrunkenes letztes Glas eines alkoholischen Getränks“, was für einen Deutschlernenden auch mit keinem hohen Deutschniveau m.E. verständlich ist. Für eine erfolgreiche Recherche zu *hdl* muss man zuerst verstehen, dass es sich um eine Abkürzung handelt und dann in *Google* z.B. „Abkürzung hdl“ eingeben, um *hab dich lieb* zu finden und dann das entsprechende *tvb* zu realisieren.<sup>39</sup> Auch wenn die weiteren Recherchen von den Probanden keine komplexen Recherchekompetenzen abverlangt haben, so war es für eine erfolgreiche Übersetzung doch wichtig, verschiedene Informationsquellen zu konsultieren, evtl. zwischen Wörterbüchern und Internetseiten zu vergleichen. Sie konnten sich nicht nur auf zweisprachige Wörterbücher verlassen, wie sie das gerne tun, sondern mussten auch

39 Bei einer einfachen Recherche „hdl“ in der Suchmaschine gelangt man zu „high density lipoprotein“. Die Abkürzung sollte bei einer weiteren Untersuchung in einen Kontext eingebaut werden, auch wenn von den Studierenden bemerkt wurde, dass es ihnen klar war, dass es sich nicht um das Protein handeln könne.

mit Informationen im einsprachigen umgehen können und gezielt – im Fall von *hdl* – nach einer Abkürzung suchen. Der Prozentsatz an richtigen Lösungen war extrem niedrig: nur 4 Studierende (11%) haben die Bedeutung von *hartzen* gefunden. Drei der Probanden schrieben „essere disoccupato“ („arbeitslos‘ sein) und eine Studierende, die jedoch schon seit 7 Jahren Deutsch lernt, gab an: „essere disoccupato (arbeitslos sein) parola gergale, introdotto nel 2009, neologismo“ („arbeitslos sein (...) umgangssprachlicher Ausdruck/Jargon, 2009 eingeführt, Neologismus). Die Abkürzung *hdl* wurde von sieben Probanden (19%) als *tvb/ti voglio bene/ti amo* wiedergegeben. Die „erfolgreichen“ Studierenden haben meist gegoogelt und eine „ricerca incrociata“ („cross research‘, überkreuzte Recherche‘) durchgeführt, d.h. auf verschiedenen Webseiten und Online-Wörterbüchern nachgeschaut, verglichen und sich für eine der Übersetzungsmöglichkeiten entschieden. Der deutsche Satz *Kommst du mit, einen Absacker trinken?* wurde von 19% der Probanden korrekt übertragen. Der italienische Satz, bei dem die Hauptschwierigkeit im Verb *stroncare* (hier ugs.: ‚sterben‘) lag, hatte eine Erfolgsquote von 16%. Als Exempel für eine erfolgreiche Übersetzung sollen folgende zwei Kommentare als Zitate dienen: „mi sono aiutata da google che mi ha dato indicazioni su neologismi“ (‘ich habe Google zu Hilfe genommen, wo ich Hinweise auf Neologismen gefunden habe‘) und „Il primo dizionario utilizzato non era in grado di tradurre alcune parole che, attraverso una ricerca sul web, sono riuscita a trovare tramite i collegamenti alle varie pagine“ („Das erste Wörterbuch, das ich benutzt habe, konnte einige Wörter nicht übersetzen; über eine Webrecherche und den verschiedenen Verbindungen (Vergleichen) ist es mir dann gelunge.“ Eine weitere Strategie, die i.d.R. zu Erfolg führte, war die Suche mit Hilfe des Englischen als „Sandwichsprache“,<sup>40</sup> d.h. die Lernenden verwenden das Englische als Zwischensprache, um zur anderen Sprache zu gelangen, sowohl Deutsch>Englisch>Italienisch als auch viceversa: „Ho avuto difficoltà con „hartzen“ la cui traduzione era disponibile solo in inglese sotto forma di slang. Per „Absacker“ ho dovuto cercare da tedesco a inglese e poi da inglese a italiano.“ Interessant wäre eine noch präzisere Angabe zu *Google*, d.h. wo und was sie gegoogelt haben, ob sie auch z.B. Bilder verwendet haben – eine Strategie, die gerade für das Verstehen von Kulturspezifika sehr sinnvoll ist. Nur ein Student, der den Satz *\*Vieni ancora a bere un Absacker* geschrieben hat, hatte in *Google*-Bilder recherchiert und eine Flasche gefunden, auf deren Etikett der Name Absacker stand, so dass er dachte, das Getränk hieße so: „Sono andato su google e dalle immagini ho visto cosa era.“ (‘Ich bin auf Google gegangen und auf den Bildern habe ich gesehen was es war.‘) Eine eindeutige Relation besteht bei den Sätzen N. und O. zwischen dem Gebrauch von *Googler*-Übersetzer und nicht korrekten Übersetzungen, evident bei Sätzen wie *\*Wurde von einem Herzinfarkt schlug während der Arbeit*. Hatte der Proband jedoch Zweifel an der Übersetzung und wendete linguistische Strategie, wie die des Vereinfachens und Umformulierens an, so war die Chance auf Erfolg groß, wie bei den folgenden Fällen: „Absacker, però la traduzione non mi soddisfaceva perciò l’ho interpretata a modo mio, cercando altre

40 Ich definiere diesen Begriff – in Anlehnung an die *Sandwich-Technik* als didaktische Methode (vgl. Butzkamm 2004) – als die Sprache, die bei der lexikographischen Recherche zwischen einem Sprachenpaar, bzw. zwischen einer Ausgangs- und einer Zielsprache eingeschoben wird, mittels deren Bedeutung das entsprechende Übersetzungsäquivalent gefunden wird.



combinazioni di parole per una traduzione più esatta.“ (‘A., aber die Übersetzung hat mich nicht überzeugt, also habe ich es auf meine Art interpretiert und andere Kombinationen von Wörtern für eine exaktere Übersetzung gesucht’) und „stroncato: è stato difficile trovare un corrispettivo tedesco e ho avuto bisogno di semplificare la frase in ambito di traduzione.“ (‘,stroncato: es war schwierig ein deutsches Äquivalent zu finden und ich musste den Satz bei der Übersetzung vereinfachen’). Es zeigt sich, dass positive Resultate auch bei relativ niedrigem Sprachniveau auch dann erreicht werden können, wenn die Recherchekompetenz gut ausgebildet ist, ein gesundes Misstrauen der dargebotenen Übersetzungsmöglichkeit gegenüber besteht, das Sprachbewusstsein hoch ist und so zum Einsatz von linguistischen Strategien führen kann.

## 4 Zusammenfassung und Forschungsausblick

Wenn man bedenkt, dass die Studierenden mit ihrem Smartphone unbegrenzte Recherchemöglichkeiten hatten und dass sie bequem und schnell an Informationen gelangen konnten, ist es doch verwunderlich, dass

- (1) sie nur einige wenige Internetseiten (meist Übersetzungsprogramme) und Online-Wörterbücher für die konkrete Übersetzung verwendet haben. Es scheint, als würden sie nur wenige kennen. Interessant ist, dass sie zwar mehr Online-Seiten bzw. -wörterbücher aufgezählt haben (Fragen 1.-11. der Untersuchung), als sie an späterer Stelle konkret für die Aufgaben benutzt haben (was aber auch daran liegen kann, dass sie innerhalb des Seminars von 90 min. fertig sein wollten, auch wenn kein Zeit-Limit angegeben war).
- (2) ihr Benutzungsverhalten dem des Print-Wörterbuchs ähnelt, bspw. dass nur wenige Studierende einsprachige Online-Wörterbücher wie *Duden* oder *Treccani* verwenden, dass sie zu schnellen Ergebnissen kommen wollen und das erstbeste Tool nutzen, das sie finden können (bei der *Google*-Recherche nehmen sie das an erster Stelle stehende Wörterbuch, im Wörterbuchartikel nehmen sie das 1. Äquivalent) und machen sich nicht die Mühe, weiter zu suchen – nach dem Motto: das Erste ist gut genug. Außerdem sind sie auf die Suche nach Wörtern fixiert; sie recherchieren nicht nach der Bedeutung eines Wortkomplexes, d.h. sie suchen keine Phraseologismen. Metasprachliche Notationen, wie bspw. Kasus und Valenz, werden „übersehen“ bzw. nicht berücksichtigt.
- (3) viele der Studierenden (noch) kein Vertrauen in die Online-Recherche haben und sich noch recht orientierungslos in Bezug auf lexikalische Lernschwierigkeiten im Internet bewegen, andererseits aber auch Studierende gibt, die Übersetzungsprogramme verwenden, ohne sich darüber Gedanken zu machen.

All diese Ergebnisse zeigen wieder einmal – wie schon bei Forschungen zu Print-Wörterbüchern –, dass generell die Lexikographie und im Besonderen die Wörterbuchbenutzung in die Fremdsprachendidaktik integriert werden müssen. Wie auch die Antworten der Studierenden auf die Frage 26.

zeigen, wünschen sich 35 Studierende (97%), mehr über Recherchemöglichkeiten im Internet zu erfahren, welche Online-Wörterbücher es gibt, wie verschiedene Online-Wörterbücher konzipiert wurden, wie sie strukturiert sind, wie man sich orientiert und wie man zuverlässige Informationen erkennt, und v.a. wie man sie erfolgreich benutzt: „come si usa senza sbagliare“ (‘wie man sie gebraucht ohne Fehler zu machen‘).

In Bezug auf zukünftige Forschungsmöglichkeiten geben m.E. insbesondere die „ricerche incrociate“ und der Verwendung des Englischen als „Sandwichsprache“ Anlass, neue Ideen für weitere umfassendere Untersuchungen zu entwickeln.

- (1) Hinsichtlich der überkreuzten Recherche wäre es bspw. interessant, in einer umfassenden Studie herauszufinden, welche Online-Informationen die Smartphone-Benutzer zu Rate ziehen, wie sie konkret von einer Internseite/einem Wörterbuch zum anderen wechseln, wie sie vergleichen, welche Schwierigkeiten und Zweifel sie dabei haben und welche Informationen die Wahl eines bestimmten Übersetzungsäquivalentes beeinflussen. Als Methode wäre m.E. das *Think-Aloud-Protocol* (Lautes Denken-Protokoll) die beste Möglichkeit, mit der die Gedankengänge des Smartphone-Benutzers während dieser Recherche am besten zum Vorschein kämen und für detaillierte Analysen auch aufgezeichnet werden könnten.
- (2) Meines Wissens noch nicht erforscht ist die Tatsache, dass sich Studierende für ein bestimmtes Sprachenpaar des Englischen als „Sandwichsprache“ bedienen. Man könnte z.B. Studierende, die gewohnheitsmäßig mit *WordReference* arbeiten, bei ihrer Smartphone-Benutzung beobachten und ihr Lautes Denken gleichzeitig aufzeichnen. Man würde von ihren sprachlichen Schwierigkeiten Genaueres erfahren und wie sie mit ihrem Smartphone diesbezüglich umgehen, um Lösungen für ihre Schwierigkeiten zu suchen, d.h. wo suchen sie, was suchen sie, wie suchen sie, wie wägen sie ab, wie gelingt ihnen der Transfer zwischen den Sprachen, usw. Ich kann mir gut vorstellen, dass sich hier neue Erkenntnisse ergeben, die für die Konzeption und Ausarbeitung zukünftiger Online-Wörterbücher und Apps von großer Bedeutung sein werden. Außerdem bin ich überzeugt – nicht zuletzt wegen der zunehmenden Mehrsprachigkeit –, dass Englisch im Bereich der Wörterbuchbenutzung zwischen zwei Sprachenpaaren immer häufiger als „Sandwichsprache“ fungieren und bewusst von den Studierenden als Strategie eingesetzt wird. Diese Ergebnisse aus der Wörterbuchbenutzungsforschung in der Fremdsprachendidaktik sollten in die zukünftige Schreibung von Online-Wörterbüchern unbedingt einfließen und gerade im Bereich der zweisprachigen Lexikographie überdacht werden.

## 5 Literatur

- Aguado, K. (2009). Möglichkeiten und Grenzen mehrmethodischer empirischer Fremdsprachenlehr- und -lernforschung. In B. Baumann, S. Hoffmann, M. Nied Curcio (Hgg.). *Qualitative Forschung in Deutsch als Fremdsprache*. Frankfurt: Lang, S. 13-22.
- Albrecht, Jörn (2005). *Übersetzung und Linguistik*. Tübingen: Narr.
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bimmel, P., Van de Ven, M. (2000). Man nehme ein Wörterbuch... *Fremdsprache Deutsch. Übersetzen im Deutschunterricht*, 23, S. 38-39.
- Boonmoh, A. (2012). E-dictionary Use under the Spotlight: Students' Use of Pocket Electronic Dictionaries for Writing. *Lexikos. Journal of the African Association for Lexicography*, S. 43-68.
- Butzkamm, W. (2004). Lust zum Lehren, Lust zum Lernen. Eine neue Methodik für den Fremdsprachenunterricht. Tübingen: Francke.
- Corda, A., Marelllo, C. (2004). *Lessico. Insegnarlo e impararlo*. Perugia: Guerra.
- de Schryver, G.-M., Prinsloo, D.J. (2011). Do Dictionaries Define on the Level of their Target Users? A Case Study for Three Dutch Dictionaries, *International Journal of Lexicography*, 21(1), S. 5-28.
- Domínguez Vázquez, M. J., Mirazo Balsa, M. & Vidal Pérez, V. (2013). Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen. In M. J. Domínguez Vázquez (Hg.). *Trends in der deutsch-spanischen Lexikographie*, 135-172. Frankfurt: Lang.
- Domínguez Vázquez, M.-J., Mollica F. & Nied Curcio, M. (2014, im Druck). Simplex-Verben im Italienischen und Spanischen vs. Präfix- und Partikelverben im Deutschen. Eine Untersuchung zum Gebrauch von Online-Wörterbüchern bei der Übersetzung. In M.-J. Domínguez Vázquez, F. Mollica & M. Nied Curcio. *Zweisprachige Lexikographie zwischen Translation und Didaktik*. Berlin, New York: de Gruyter (= Lexicographica: Series Maior).
- Engelberg, S., Lemnitzer, L. (2004). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.
- Faerch, C., Kasper, G. (Hgg.) (1978). *Introspection in Second Language Research*. Clevedon Philadelphia: Multilingual Matters LTD.
- Flick, U. (2007). *Qualitative Sozialforschung. Eine Einführung*. Reinbek bei Hamburg: Rowohlt.
- Mackintosh, K. (1998). An empirical study of dictionary use in L2-L1 translation. In S. Atkins (Hg.): *Using Dictionaries*. Tübingen: Niemeyer, S. 123-149.
- Möhrs, Ch., Müller-Spitzer, C. (2013). *Elektronische Lexikografie*. Tübingen: Groos.
- Müller-Spitzer, C. (2013). Contexts of dictionary use. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, & M. Tuulik (Hgg.). *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Institute for Applied Slovene Studies/Eesti Keele Instituut, S. 1-15. [http://eki.ee/elex2013/proceedings/eLex2013\\_01\\_Mueller-Spitzer.pdf](http://eki.ee/elex2013/proceedings/eLex2013_01_Mueller-Spitzer.pdf) [5.4.2014]
- Müller-Spitzer, Carolin (Hg.) (2014): *Using Online Dictionaries*. Berlin, New York: de Gruyter. (= Lexicographica: Series Maior 145).
- Müller-Spitzer, Carolin & Koplenig, Alexander (2014, im Druck). Requisitos y expectativas de un buen diccionario online. Resultados de estudios empíricos en la investigación sobre el uso de diccionarios con especial atención a los traductores. In M. J. Domínguez Vázquez, X. G. Guinovart & C. Valcárcel Riveiro (Hgg.). *Lexicografía románica, Vol. 2. Aproximaciones a la lexicografía moderna y contrastiva*. Berlin, New York: de Gruyter.
- Müller-Spitzer, Koplenig & Töpel (2012). Online dictionary use. Key findings from an empirical research project. In: S. Granger, M. Paqot (Hgg.). *Electronic lexicography*. Oxford: Oxford University Press, S. 426-457.

- Nied Curcio, M. (2005). Verbale Polysemie und ihre Schwierigkeiten im DaF-Erwerb. In D. Di Meola, A. Hornung & L. Rega. *Perspektiven Eins. Tagungsakten der Tagung ‚Deutsche Sprachwissenschaft in Italien‘ vom 6./7. Februar 2004*. Roma: Istituto Italiano di Studi Germanici, S. 195 – 211.
- Nied Curcio, Martina (2011). Der Gebrauch von Wörterbüchern im DaF-Unterricht. Am Beispiel von Übersetzungsübungen. In P. Katelhön, J. Settinieri (Hgg.): *Wortschatz, Wörterbücher und L2-Erwerb*, Wien: Praesens, S. 181– 204.
- Sattler, W. (2008). Rund ums Gerund. Das Gerundio und seine Wiedergabe im Deutschen. In M. Nied Curcio. *Ausgewählte Phänomene zur Kontrastiven Linguistik Italienisch – Deutsch. Ein Lehr- und Übungsbuch für italienische DaF-Studierende*. Milano: Franco Angeli, S. 98-117.
- Taljad, E., Prinsloo, D. & Fricke, I. (2011). The use of LSP dictionaries in secondary schools? a South African case study. In *South African Journal of African Languages*, 31(1), S. 87-109.
- Tarp, S. (2011). Lexicographical and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. In H. Bergenholtz, P.A. Fuertes-Olivera (Hgg.): *e-Lexicography. The Internet, Digital Initiatives and Lexicography*. London, New York: Continuum, S. 54-70.
- Wiegand, H. E. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. 2 Bände. Berlin, New York: de Gruyter.

# Dictionary Users do Look up Frequent and Socially Relevant Words. Two Log File Analyses.

Sascha Wolfer, Alexander Koplenig, Peter Meyer, Carolin Müller-Spitzer  
Institut für Deutsche Sprache, Mannheim, Germany  
wolfer@ids-mannheim.de

## Abstract

We start by trying to answer a question that has already been asked by de Schryver et al. (2006): Do dictionary users (frequently) look up words that are frequent in a corpus. Contrary to their results, our results that are based on the analysis of log files from two different online dictionaries indicate that users indeed look up frequent words frequently. When combining frequency information from the Mannheim German Reference Corpus and information about the number of visits in the Digital Dictionary of the German Language as well as the German language edition of Wiktionary, a clear connection between corpus and look-up frequencies can be observed. In a follow-up study, we show that another important factor for the look-up frequency of a word is its temporal social relevance. To make this effect visible, we propose a de-trending method where we control both frequency effects and overall look-up trends.

**Keywords:** research into dictionary use; frequency; corpus; social relevance; log file analysis

## 1 Introduction<sup>1</sup>

In this paper, we use the 2012 log-files of two German online dictionaries (Digital Dictionary of the German Language and the German language edition of Wiktionary) and the 100,000 most frequent words in the Mannheim German Reference Corpus (Deutsches Referenzkorpus, DEREKO) from 2009 (Kupietz et al., 2010) to answer the question of whether dictionary users really do look up frequent words, first asked by de Schryver et al. (2006). The research question standing behind is whether it actually makes sense to select words based on frequency, or, in other words, if it is a reasonable strategy to prefer words that are more frequent over words that are not so frequent. Answering this question is especially important when it comes to building up a completely new general dictionary from scratch and the lexicographer has to compile a headword list. By using an approach to the comparison of log-files and corpus data which is completely different from that of the aforementioned authors, we provide empirical evidence that indicates – contrary to the results of de Schryver et al. and Verlinde & Binon (2010) – that the corpus frequency of a word can indeed be an important factor in determining

---

<sup>1</sup> In Koplenig, Meyer, & Müller-Spitzer (2014) we present and discuss the results of this study in more detail.

what online dictionary users look up. In addition, we incorporate word class information readily available in Wiktionary into our analysis to improve our results considerably. In a follow-up study, we show that (temporal) social relevance of particular words can influence look-up behaviour considerably. For the latter study, we used the 2013 log files of the German language edition of Wiktionary.

## 2 Previous research

To understand whether including words based on frequency of usage considerations makes sense, it is a reasonable strategy to check whether dictionary users actually look up frequent words. Of course, in this specific case, it is not possible to design a survey (or an experiment) and ask potential users whether they prefer to look up frequent words or something like that. That is why de Schryver and his colleagues (2006) compared a corpus frequency list with a frequency list obtained from log-files. The aim of De Schryver et al.'s study was to find out whether dictionary users look up frequent words. Due to the nature of the statistical method they used, de Schryver et al. (2006) actually tried to answer two different questions: do dictionary users look up frequent words frequently? And, do dictionary users look up less frequent words less frequently? (cf. Koplein, Meyer, & Müller-Spitzer, 2014: 232). The result of their study is part of the title of their paper: "On the Overestimation of the Value of Corpus-based Lexicography". Verlinde & Binon (2010: 1148) replicated the study of de Schryver et al. (2006) using the same methodological approach and essentially came to the same conclusion.<sup>2</sup>

In this paper, we will try to show why de Schryver et al.'s straightforward approach is rather problematic due to the distribution of the linguistic data that is used. In this context we suggest a completely different approach and show that dictionary users do indeed look up frequent words (sometimes even frequently). In a follow-up study, we present a case study that suggests that, as soon as frequency information is partialled out, analyses of log-files can also reveal information about the (sometimes very short-lived) social importance of particular words.

## 3 The Data

### *Corpus data*

When we look at the DEREWO list, a word list compiled using the DEREKO corpus, and plot the relative frequency against the rank, we receive a typical Zipfian pattern. This means that we have a handful of word forms that have a very high frequency and an overwhelming majority of word forms that

---

2 In contrast, **Henrik** Lorentzen, Nicolai H. Sørensen and Lars Trap-Jensen in their talk at the e-lexicography conference 2013 also came to the conclusion that frequent words in a corpus are also frequently looked up in a dictionary (Talk: "An odd couple - corpus frequency and look-up frequency: what relationship?" Video available at <http://eki.ee/elex2013/videos/> [last access on 02/04/2014]).

have a very low frequency. Or, in other words, our DEREWO list consists of 3,227,479,836 word form tokens. The 200 most frequent word form types in the list make exactly half of those tokens.

### Log-files

The Wiktionary log file types are roughly 8 w times as many as the DWDS log file types. To make the results both comparable and more intuitive, we rescaled the data by multiplying the raw frequency of a query by 1,000,000, dividing it by the sum of all query tokens and rounding the resulting value. We then removed all queries with a value smaller than one. Thus, the resulting variable is measured in a unit that we would like to call *poms* (per one million searches). For example, a value of 8 means that the corresponding phrase is searched for 8 times per one million search requests. Table 1 summarizes the resulting distribution.

Category ( <i>poms</i> )	Wiktionary log-files (%)	DWDS log-files (%)
1	57.94	57.30
2 - 10	33.71	31.15
11 - 49	6.69	9.09
50 - 500	1.63	2.44
500 +	0.03	0.02
<b>Total</b>	<b>100.00 (abs. 185,071)</b>	<b>100.00 (abs. 156,478)</b>

**Table 1: Categorized relative frequency of the log file data.**

Category	X searches <i>poms</i>	Wiktionary log-files (%)	DWDS log-files (%)
regular	at least 1	100.00	100.00
frequent	at least 2	42.06	42.70
very frequent	at least 11	8.35	11.55

**Table 2: Definition of the categories used in the subsequent analysis and relative log file distribution.**

Table 1 shows two things: firstly, the Wiktionary and the DWDS log-files are quite comparable on the *poms*-scale; secondly, just like the corpus data, the log-files are heavily right skewed. More than half of all query types consist of phrases only searched for once *poms*. When we cumulate the first two categories, we can state for both the Wiktionary and the DWDS data that 90% of the queries are requested 1 up to 10 times *poms*. So there is only a small fraction of all phrases in the log-files that are searched for more frequently.

## 4 Data analysis

In the previous section, we described the data and presented a new unit of measurement called *poms*. If we think about our research question again – whether dictionary users look up frequent words (frequently) – it is necessary to find an appropriate method for analyzing the data using this unit. For example, we could regress the log file frequency (in *poms*) on the corpus frequency, but an ordinary least squares (OLS) regression implies a linear relationship between the explanatory and the response variable, which is clearly not given. (Log-)Transforming both variables does not solve our problem, either, and this is in any case seldom a good strategy (O’Hara & Kotze, 2010). We could use the appropriate models for count data such as Poisson regression or negative binomial regression, but, as Baayen (2001, 2008: 222-236) demonstrates at length, we still have to face the problem of a very large number of rare events (LNRE), which is typical for word frequency distributions. And even if we could fit such a model, it would remain far from clear what this would imply for our initial lexicographical question. Using the standard Pearson formula to correlate the corpus and the log-file data suffers from the same nonlinearity problem as the OLS approach. Therefore de Schryver et al. (2006) implicitly used the nonparametric Spearman rank correlation coefficient which is essentially just the Pearson correlation between ranked variables. We believe that this is still not the best solution, mainly because, on a conceptual level, ranking the corpus and log-file data implies that subsequent ranks are equidistant in frequency, which is clearly not the case. Again, the inherent Zipfian character of the distribution explains why the ranks are far from equidistant. For example, the difference in frequency between the first and the second rank is 251,480, whereas the difference between the 3000th and 3001th is only 5. Nevertheless the Spearman rank correlation coefficient treats the differences as equal<sup>3</sup>.

In the last section, we grouped the log-files (cf. Table 1) into *poms* categories. As a possible solution to the problems we just outlined, we now use this grouping again and stipulate the following categories: if a word form is searched for at least once *poms*, it is searched for regularly, if it is searched for at least twice, we call it frequent, and if it is searched for more than 10 times, it is very frequently searched for. Table 2 sums up the resulting values. Please keep in mind that according to this definition, a very frequent search term also belongs to the regular and the frequent categories. Our definition is, of course, rather arbitrary and mainly has an illustrative function, but due to the Zipfian distribution of the data, only a minority of the searches (roughly 4 out of 10) occur more than once *poms* and even fewer words (roughly 1 out of 10, roughly 8 percent for Wiktionary, roughly 12 percent for the DWDS) are searched for more than ten times *poms* (cf. Table 1). Therefore this definition at least approximates the distribution of the log file data. Nevertheless, instead of using the categories presented in the first column of Table 2, we could also use the second column to label the categories. So it must be borne in mind that the labels merely have an illustrative function.

---

<sup>3</sup> In principle, we could use another similarity metric, for example the cosine measure (i.e. the normalized dot product, cf. Jurafsky & Martin, 2009: 699), but as in the case of using a count regression model, we are not sure what the value of the coefficient would actually imply both theoretically and practically.



We then wrote a Stata program that starts with the first ten DEREKO ranks and then increases the included ranks one rank at a time. At every step, the program calculates how many of the included word forms appear in the DWDS and Wiktionary log-files regularly, frequently, and very frequently (scaled to percentage). Table 3 summarizes the results for 6 data points.

Included DEREKO ranks	DWDS (%)			Wiktionary (%)		
	regular	frequent	very frequent	regular	frequent	very frequent
10	100.0	100.0	100.0	100.0	100.0	100.0
200	100.0	99.0	87.5	99.5	99.5	86.5
2,000	96.9	91.0	67.6	98.4	96.0	64.9
10,000	85.5	72.9	47.5	86.3	75.3	40.2
15,000	80.3	66.5	41.8	77.4	66.1	33.7
30,000	69.4	54.6	31.3	62.7	50.9	23.4

**Table 3: Relationship between corpus rank and log file data.**

In this table, the relationship between the corpus rank and the log file data becomes obvious: the more DEREKO ranks we include, the smaller the percentage of those word forms appearing regularly/frequently/very frequently in both the DWDS and the Wiktionary log-files. Let us assume for example that we prepare a dictionary of the 2,000 most frequent DEREKO word forms; our analysis of the DWDS and the Wiktionary data tells us that 96.9 % of those word forms are searched for regularly in DWDS, 91.0 % are searched for frequently and 67.6 % are searched for very frequently. For Wiktionary, these figures are a bit higher.

Figure 2 plots this result for the DWDS and the Wiktionary log-files separately. It comes as no surprise that the curve is different for the three categories, being steepest for the very frequent category, since this type of log file data only makes up a small fraction of the data. To further improve our analysis, we looked at the word forms that are absent in both the DWDS and the Wiktionary log-files but that are present in the unlemmatised DEREKO corpus data. There is a roughly 60% overlap, which means that 6 out of ten word forms missing in the DWDS data are also missing in the Wiktionary data. To understand this remarkable figure, we tried to find out more about the words that are missing in the log-files but are present in the corpus data. In our talk, we can present how we used this data to improve our analyses.

We would like to provide an additional impression of our results by asking what proportion of all search requests (tokens) could be covered with such a corpus-based strategy. If we again use the example of the first 15,000 DEREKO most frequent word forms, then around half of all DWDS search requests that occur regularly or frequently (poms) are covered, while around two-thirds of all very frequent requests are successful. If we included the 30,000 most frequent DEREKO words, roughly two-thirds of the regular and frequent and 80.0% of the very frequent DWDS search requests would be covered in

the dictionary. In other words, this means if we included the 30,000 most frequent DEREKO word forms, the vast majority of requests would be successful.

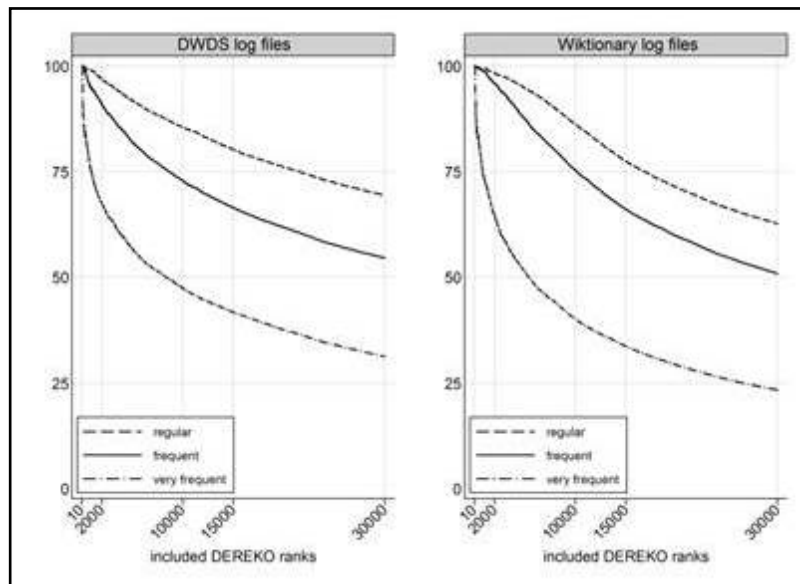


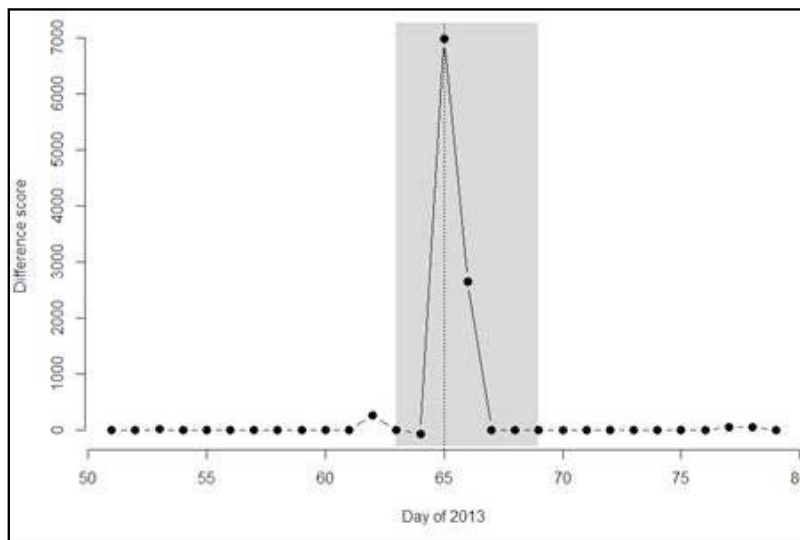
Figure 1: Percentage of search requests which appear in the DWDS/Wiktionary log-files as a function of the DEREKO rank.

## 5 Social relevance and dictionary usage

The previous sections showed why we think that it is a good idea to include frequently occurring words in dictionaries. Although it is clear that there is a strong and reliable relationship between frequency of occurrence and look-up frequency, it is also quite clear that the former is not the *only* predictor of the latter. To investigate further contributing factors, we consulted log files from the German Wiktionary in the year 2013 and aggregated the hourly Wiktionary log files to weekly datasets to keep the computations manageable. We excluded all pages with titles longer than 80 characters. We also excluded all pages that were visited fewer than one time in one million visits. The overall aim of this study is to identify points in time where certain words are looked up extraordinarily often. To achieve this, we need to control for the overall trend of look-up frequency of each word. It is no surprise that look-up frequency varies over time. Words are looked up more or less often during the course of a year. This variation can be captured by the overall trend *within* a word's visits. By controlling for these long-term trends, we also capture general look-up differences *between* words that stem i.a. from the frequency effects outlined above. What we are currently interested in are rather short-term 'peaks' in the number of visits a specific word receives. The number of visits a specific word receives is the sum of the overall trend for this word and 'noise' which is not captured by this trend (cf. Beckett, 2013:92-95,103,109). This noise, or - informally speaking - what is left over after the overall trend has been considered is exactly the kind of data we are interested in. To extract this variable, we fitted a Tukey

smoother using running medians of length 3<sup>4</sup>. This smoother captures the trend. The variable we are going to use is the difference between this smoother and the actual visits, we call this the difference score or residual visits. Using this technique, we can look beyond the effect of frequency and overall tendencies of a specific word. In other words, this technique enables us to identify extraordinary look-up behavior for individual words in individual points in time<sup>5</sup>. To extract interesting words, we rank words by their smooth-difference score. All highly ranking words have especially high proportions of unexplained variance in visits per one million visits in the respective week.

We will now describe two headwords with noticeable difference scores to provide a first impression of our results. The word “Furor” (English “furor”, “rage”) takes rank 14 of the ordered list of difference scores. The differences between smoothed and observed visits per one million visits are constantly around zero. However, in week 10 of 2013 (04/03/2013 to 10/03/2013), its difference scores go up to 2,210 with a total of 4,687 visits (for all other weeks except week 10, the mean of raw visits for “Furor” is 60.7). In this week, German president Joachim Gauck used the word “Tugendfuror” (roughly: “furor/rage of virtue”) in a debate on sexism in Germany. His whole comment, and especially the word “Tugendfuror” was subject of public discussion in Germany throughout the media. Figure 3 shows the difference scores of “Furor” on a daily basis for one month (20/02/2013 to 20/03/2013). One can clearly see how residual searches *poms* rise for one critical day (06/03/2013) and then take one to two days to ‘normalize’ again to a residual value around zero.



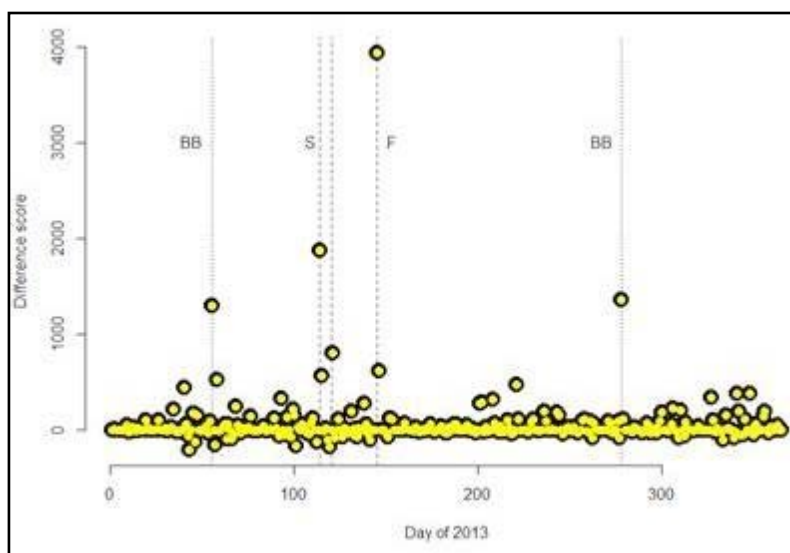
**Figure 2: Difference between smoothed and observed visits per 1 million visits for “Furor” between 20/02/2013 and 20/03/2013. Week 10 is highlighted.**

4 To do this, we used the default behavior of the function *smooth()* provided by the ‘stats’ package of the statistical programming language R (R Core Team, 2013).

5 Of course, these differences can also take negative values. Indeed, many of them do. This means that a word received less searches *poms* in a particular week than would be expected given the word’s overall trend. If we normalize these differences by dividing the difference score by the smoothed visits per one million visits, we can also compare words between one another. In this paper, we did not apply this operation.

Interestingly, Joachim Gauck used the word in an interview<sup>6</sup> already published on 03/03/2014. Though one can see a minimal rise on that day, it took three days until other major newspapers picked up the debate because other voices alluded to potential problems of Gauck’s choice of words<sup>7</sup>. Obviously, quite a lot of people then wondered what the head of the compound “Tugendfuror” actually means and referred to Wiktionary during those days. “Furor” is a good example of a temporarily socially relevant word. Actually, subject of discussion was the lexical meaning of “Furor” and its connections to other, potentially pejorative words. So, the discussion was lexico-semantic in nature and it comes as no surprise that people tended to look up the word the German president used.

However, there are other noticeable words in certain periods of time, which are not directly related to discussions in society or politics that are lexical in nature. Figure 4 shows the residual searches poms of the word “Borussia” in time. “Borussia” is Latin for “Prussia” and part of the name of several German sports clubs. The most prominent ones are football clubs.



**Figure 3: Difference between smoothed and observed visits per 1 million visits for “Borussia” for the whole year 2013. Vertical lines indicate football matches (BB: Borussia Mönchengladbach vs. Borussia Dortmund, S: UEFA Champions League semi-finals, F: Final).**

Peaks are identifiable in the difference scores for “Borussia” over time; symbolizing temporarily increased look-ups for “Borussia” in Wiktionary that cannot be explained by frequency of occurrence or overall search preferences alone. Each dashed vertical line in Figure 4 represents one match in the knockout phase of the UEFA Champions League (CL) competition with the participation of Borussia Dortmund. Look-ups of “Borussia” sharply increase around match days. For the semi-finals (“S”) and especially the all-important final match (“F”), residual searches poms increase sharply around match

6 See <http://www.spiegel.de/politik/deutschland/sexismus-debatte-gauck-beklagt-tugendfuror-im-fall-bruederle-a-886578.html> [last access on 01/04/2014].

7 See <http://www.sueddeutsche.de/politik/sexismus-debatte-als-tugendfuror-aufschrei-wegen-gauck-1.1616310> [last access on 02/04/2014].

days. There are two other vertical lines (“BB”) which do not mark a match day in the CL. BB marks 24/02/2013 and 05/10/2013, the days Borussia Dortmund competed against Borussia Mönchengladbach in the German first division. This match is associated with increased difference scores, too. In contrast, no other match in the German first division did lead to increased residual searches for “Borussia”. Obviously, the popularity and importance of the CL competition led to repeated increases in the social relevance of the term “Borussia”. Also, when both Borussias competed against each other in the national championship, public interest in the somewhat cryptic name part was also increased. In comparison to the “Furor” case presented above, the look-up behavior concerning “Borussia” is more surprising. There is no lexico-semantic debate involved that could persuade people to look up “Borussia”. Increased media coverage and general public awareness concerning a football club alone seems sufficient to trigger noticeable increases in look-up behavior. This is a remarkable and important observation for research into dictionary use. Another example is the word “larmoyant” (English “lachrymose” meaning “tearfully sentimental”) which was used in a sports commentary in an exhibition match between the French and German football national team. Here, the commentator described one specific German national as being too “larmoyant” which led to sharply increased lookups within the same hour (which is the minimum temporal resolution available for the Wiktionary log files). There are several more interesting cases extractable from the Wiktionary log files that we cannot report here. Social relevance in other cases was induced by a variety of social contexts like TV game shows and even astronomical events like a solstice. Explaining why residual look-ups increased in a specific timeframe is interesting and it definitely points to the fact that the social context directly influences look-up frequencies in internet dictionaries – all in a very short time frame. In future research, however, we want to operationalize social relevance of words in a large-scale, automatized way. Such a measure would enable us to correlate these two measures not only over singular cases but many words. Furthermore, one could identify social contexts that are especially capable (or others that are not capable at all) to trigger look-up peaks in online dictionaries. This line of research could also contribute to the overall question of this paper: Which words should be included in dictionaries? Certainly, words that are socially highly relevant over long periods of time are good candidates.

## 6 Conclusion

In general, the use of a corpus for linguistic purposes is based on one assumption:

“It is common practice of corpus linguistics to assume that the frequency distributions of tokens and types of linguistic phenomena in corpora have – to put it as generally as possible – some kind of significance. Essentially more frequently occurring structures are believed to hold a more prominent place, not only in actual discourse but also in the linguistic system, than those occurring less often” (Schmid, 2010: 101).

We hope that we have provided evidence which shows that, based on this assumption, corpus information can also be used fruitfully when it comes to deciding which words to include in a dictionary. This corpus-based strategy is no “magic answer”. We simply think it is the best one there is, given that there are no other systematic alternatives.

Beyond that, social relevance of words or their (extraordinary) presence in social discourse seems to be a highly relevant factor in this context.

## 7 References

- Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.
- Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge University Press.
- Beckett, S. (2013). *Introduction to Time Series Using Stata*. College Station: Stata Press.
- De Schryver, G.-M., Joffe, D., Joffe, P., & Hillewaert, S. (2006). Do dictionary users really look up frequent words?—on the overestimation of the value of corpus-based lexicography. *Lexikos*, 16, 67–83.
- Jurafsky, D. & Marti, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational Linguistics, and speech recognition*. Upper Saddle River: Pearson Education (US).
- Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, D. Tapias, M. Rosner, S. Piperidis, J. Odjik, J. Mariani, ... K. Choukri (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10)* (pp. 1848–1854). Valetta, Malta: European Language Resources Association (ELRA).
- Koplenig, A., Mayer, P., & Müller-Spitzer, C. (2014). Dictionary users do look up frequent words. A log file analysis. In: C. Müller-Spitzer (Ed.). *Using online dictionaries* (pp. 229–250). Berlin, New York: de Gruyter. (Lexicografica: Series Maior 145).
- O’Hara, R.B. & Kotze, D.J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2), 118–122.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org> (last accessed 02/04/2014).
- Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (pp. 101–133). Berlin, New York: de Gruyter.
- Verlinde, S., & Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1144–1151). Ljouwert: Afûk.

### Acknowledgements

We are very grateful to the DWDS team for providing us with their log-files.

# La performance dell'utente apprendente di italiano LS/L2 e la microstruttura dei dizionari: sussidi per lo sviluppo della Lessicografia Pedagogica

Angela Maria Tenório Zucchi  
Università di San Paolo, Brasile  
angelazucchi@usp.br; angelatz@gmail.com

## Abstract

Questo paper presenta dei risultati quantitativi e qualitativi di una ricerca empirica sulla comprensione di unità lessicali contestualizzate e l'uso dei dizionari, svolta con studenti brasiliani di lingua italiana come lingua straniera presso la Facoltà di Lettere dell'Università di San Paolo. Con una metodologia innovativa nell'usare immagini come risposta nelle alternative a scelta multipla e nell'organizzare i dati raccolti in particolari schede a partire dal comportamento dei tre gruppi di studenti (con dizionario monolingue, con dizionario bilingue e senza dizionario), questo esperimento ha reso possibile verificare statisticamente l'efficacia dell'uso dei dizionari, sia bilingui o monolingui, per la comprensione di parole di uso comune, ma a volte non presenti nei manuali didattici, presentate in testi scritti in italiano. Attraverso la scheda denominata FLICULAD, si è potuto controllare in modo sistematico quali sono state le informazioni contenenti la microstruttura di ogni unità lessicale analizzata che hanno contribuito o meno, secondo l'opinione dello studente, ad arrivare alla comprensione del significato di quella unità in quel contesto.

**Keywords:** comprensione; dizionario monolingue; dizionario bilingue; ricerca empirica

## 1 Introduzione

Il tema dell'edizione 2014 del Congresso Internazionale EURALEX, "Il focus sull'utente", è particolarmente in sintonia con la ricerca<sup>1</sup> da me realizzata e parzialmente presentata in lingua portoghese nell'edizione del XIV EURALEX Conference, in Olanda. In quell'occasione sono stati esposti la metodologia, con esempi, e i risultati statistici (Zucchi, 2010). Adesso invece saranno presentati i risultati legati all'aspetto qualitativo della ricerca: dati che riguardano il modo in cui le informazioni nella microstruttura dei dizionari possono o meno facilitare la comprensione, dati emersi dal contributo degli informanti, e le relative implicazioni per la Lessicografia Pedagogica, teorica e pratica (Welker, 2008; Xatara et al. 2011).

---

1 Ricerca per la Tesi di Dottorato presentata al Programma di Post-Laurea del Dipartimento di Linguistica/FFLCH/USP.

La ragione che ha motivato la ricerca consiste nel fatto che, nonostante il ruolo dei dizionari nell'ambiente scolastico sia riconosciuto, si vedono tuttora pubblicate e commentate opinioni diverse quanto alla efficacia del suo uso nell'apprendimento di una lingua straniera (LS) o lingua seconda (L2); inoltre viene spesso espressa e discussa l'idea che il dizionario monolingue sia più adeguato degli altri all'acquisizione delle lingue (Nunes e Finatto, 2007:40; Corda e Marengo, 2004:98). Pertanto, si è voluto controllare, attraverso test basati su lettura e verifica, la reale comprensione di determinate unità lessicali (UL) contestualizzate, con l'uso di un dizionario monolingue (DM - De Mauro, 2000), di un dizionario bilingue (DB - Michaelis, 2003) - ambedue in formato elettronico e allora disponibili *on-line* gratuitamente - o senza l'uso di dizionario (SD).

Da questa ricerca, oltre al risultato quantitativo, si sono potuti raccogliere dati precisi riguardo alle informazioni presenti nella microstruttura dei dizionari considerate utili dagli studenti, per la comprensione delle unità lessicali oggetto dello studio ed evidenziate nei testi letti. Inoltre, si è potuta verificare la maggiore o minore efficacia di alcune delle risorse dei dizionari elettronici. Per chiarire meglio, si riassumeranno qui brevemente la metodologia e i risultati statistici; in seguito, verrà fornita parte dei risultati qualitativi oggetto di questo intervento, derivati dalle osservazioni dei partecipanti sulle informazioni lessicografiche.

## 2 Metodologia<sup>2</sup>, organizzazione dei dati e risultati statistici

I soggetti della ricerca empirica, realizzata per quasi quattro mesi nel primo semestre del 2008, erano ventiquattro studenti brasiliani del corso di laurea in Lettere/Lingue della Facoltà di Filosofia, Lettere e Scienze Umane (FFLCH), della Università di San Paolo (USP), indirizzo di lingua, letteratura e cultura italiana, livelli corrispondenti a A2 e B1 (QCER, 2001), della stessa Facoltà a cui appartiene come docente chi scrive. Alla FFLCH, gli studenti cominciano le lezioni di italiano nel secondo anno della laurea in Lettere/Lingue e, in generale, senza nessuna conoscenza previa di questa lingua straniera. Al primo anno frequentano sei ore di lezioni di lingua italiana obbligatorie e possono frequentare discipline di cultura italiana come complementari. Quelli che hanno collaborato con la ricerca volontariamente erano studenti del primo e del quinto semestre di lingua italiana e nessuno di loro era stato prima in Italia o aveva genitori o dei parenti con cui parlare in italiano.

Prima della realizzazione del test si è spiegato ai partecipanti lo scopo della ricerca e quali sarebbero state le condizioni di lettura per ogni testo presentato. Per conoscere gli informanti, a ciascuno è stato chiesto di compilare un questionario con domande che riguardavano il suo profilo di studente, il suo uso di dizionari e la sua competenza nella consultazione di un dizionario.

Il test consisteva nella lettura di quattro testi autentici (si presenta il testo 4 in appendice) o adattati in lingua italiana, di differente argomento, in cui venivano evidenziate complessivamente quaranta

---

2 Un riassunto in inglese della metodologia e dei risultati di questa ricerca è presentato in Welker (2010)



unità lessicali (qui elencate in appendice), di cui cinque nei due primi testi e quindici negli ultimi due. È importante dire che tutte le UL scelte fanno parte del vocabolario di base dell'italiano, presentato nelle liste del *DAIC - Dizionario Avanzato dell'Italiano Corrente*, di Tullio De Mauro (1997). La loro scelta è stata fatta a seconda della loro disponibilità nei testi e della possibilità della loro rappresentazione in immagine, perché questo è stato il criterio per la verifica della comprensione come si vedrà in seguito. Ci sono voluti due mesi di preparazione per l'elaborazione del test finale e prima di applicarlo definitivamente, è stata realizzata una prova con un piccolo gruppo di studenti di altri corsi di italiano e poi sono stati fatti i dovuti cambiamenti nel test.

Nella ricerca, queste UL sono state denominate "parole-stimolo" e non sono state presentate in un elenco nella loro forma lemmatizzata, ma solo direttamente nel testo, flesse o no, in modo evidenziato in neretto e numerate. Dopo aver letto il testo, il collaboratore riceveva il foglio del test, su cui doveva segnalare l'alternativa corretta (fra A, B, C o D), corrispondente al significato relativo al testo di ogni UL evidenziata, seguendo i passi più avanti descritti. Queste alternative erano costituite da immagini (disegni o fotografie) che alludevano al significato della UL in quel dato testo, secondo il modello metodologico di "stimolo verbale/ reazione non-verbale" (Morales, 1994), si è dunque cercato di selezionare prima dei sostantivi concreti, ma si è verificato possibile creare anche l'immagine di alcuni verbi e sostantivi astratti e in inoltre due unità fraseologiche (*servizio di posate* e *capelli legati*). Fra le alternative, una indicava la giusta corrispondenza e le altre tre erano distrattori, alternative non corrette riguardo al dato enunciato. Si è considerato tuttavia che i distrattori dovevano essere plausibili.

Al collaboratore veniva richiesto di seguire i seguenti passi: 1. leggere prima il testo e osservare le UL li evidenziate; 2. dichiarare se la UL gli era già nota (nel plico per le risposte del test c'era un foglio per ogni UL); 3. scrivere una possibile traduzione della UL in portoghese; 4. consultare il dizionario (DM o DB a seconda del gruppo; passo non incluso nel test per il gruppo senza dizionari - SD); 5. osservare le figure e segnalare l'alternativa corretta; 6. scrivere nell'apposito spazio le informazioni lette sul dizionario che gli erano state di aiuto e quelle che, al contrario, gli avevano reso più difficile la comprensione della UL. La ricercatrice seguiva i passi degli studenti e chiedeva loro di rispondere nel modo richiesto spiegando che la ricerca aveva obiettivi statistici, che i parametri dovevano essere rispettati e che si voleva verificare se le informazioni fornite dai dizionari erano utili. È stato spiegato che il test non aveva lo scopo di testare la loro competenza linguistica, bensì di verificare i risultati dell'uso dei due dizionari (o la sua mancanza) per la comprensione di quelle determinate unità lessicali contestualizzate.

Il test era uguale per tutti e tre gruppi, con eccezione della consultazione del dizionario per il gruppo senza dizionario, al quale non venivano richiesti i passi 5 e 6. Gli informanti erano ventiquattro, di sesso femminile e maschile, distribuiti ugualmente nei tre gruppi (DM, DB e SD), con età media di ventiquattro anni, tutti di madrelingua portoghese, tutti studenti di Lettere/Lingue, indirizzo Lingua e Letteratura Italiana, livelli A2 e B1 (QCER).

La raccolta dei dati è avvenuta nell'università, durante incontri individuali o in piccoli gruppi. I dizionari utilizzati erano allora disponibili in rete gratuitamente: il monolingue De Mauro, edizioni Para-

via (2000), ora fuori consultazione, e il bilingue Michaelis (2003) *on-line*, ancora disponibile. La scelta di dizionari gratuiti in rete è dovuta al frequente uso di questo tipo di dizionari da parte degli studenti e allora specialmente di questi due.

I dati raccolti sono stati organizzati in fogli Excel con le risposte degli studenti alle alternative, la loro traduzione per ogni UL, il modo in cui digitavano le UL nel campo di ricerca, la voce completa data dal dizionario per ogni UL e le informazioni contenenti in essa separate in colonne, poi, nelle righe riferenti a queste informazioni, le osservazioni degli studenti su quello che serviva loro di ausilio o meno per la comprensione. Si è creata a questo fine una scheda denominata FLICULAD (*Ficha lexicográfica informacional da compreensão de unidade lexical com auxílio de dicionário* – Scheda lessicografica informazionale della comprensione di unità lessicale con l'ausilio di dizionario) che ha facilitato l'analisi statistica e poi l'analisi qualitativa dei dati.

I fogli sono stati inviati al *Centro de Estatística Aplicada* (CEA), Centro di Statistica Applicata, dell'Istituto di Matematica della stessa università (IME-USP), e sottoposti ad analisi statistica (tecniche utilizzate: analisi descrittiva unidimensionale, multidimensionale e test di ipotesi non parametriche). Da tale analisi si sono ricavate due tipologie di risultati: 1. relazione tra il profilo degli informanti (dettagliato attraverso il questionario) e la loro performance in termini di percentuale di risposte adeguate; 2. relazione tra tipo di dizionario utilizzato, o suo mancato uso, e percentuale di risposte adeguate (per ogni testo separatamente).

L'informazione più importante emersa da questi risultati è che tutti e due i gruppi che hanno fatto uso del dizionario per la comprensione delle unità lessicali segnalate nel testo hanno avuto più successo del gruppo senza dizionario, per tutti e quattro i testi, come si può verificare nelle tabelle in appendice. L'analisi inferenziale non ha mostrato evidenze che indicassero che l'uso del dizionario monolingue fosse superiore a quello bilingue o viceversa. Il gruppo che ha usato il dizionario monolingue ha avuto una performance migliore in due testi (n.1, n.4), mentre il gruppo con il dizionario bilingue ha avuto una performance migliore nei confronti di altri due (n.2, n.3). In futuro sarà realizzata un'analisi sui vari aspetti dei testi e dei co-testi (tipologia testuale, campo semantico, argomento, ordine di apparizione delle parole ecc) e sulle UL che hanno portato a questo risultato. Da questi risultati statistici si può concludere che in ambedue i dizionari esistono caratteristiche che facilitano la comprensione e altre che la ostacolano. Alcune di queste caratteristiche sono state menzionate dagli studenti e a partire di tali dichiarazioni messe in relazione con i risultati corretti si è fatta l'analisi qualitativa descritta qui di seguito.

### **3 Dati ricavati dalle dichiarazioni degli studenti e dalle loro azioni**

In questa e nella sezione seguente si presentano i risultati dell'analisi fatta a partire dai dati raccolti negli esercizi e posteriormente organizzati nella scheda FLICULAD citata prima, che non viene ripro-

dotta in questo articolo per mancanza di spazio. Ogni UL aveva la sua scheda in foglio Excel, con le informazioni riguardanti gruppi e dizionari DM e DB, dove sono state riprodotte le voci lessicografiche di ogni UL e le corrispondenti informazioni ricavate dalle azioni - digitazione nel campo di ricerca e scelta dell'alternativa - e dichiarazioni (passo n.6 nella metodologia) dei sedici studenti dei due gruppi riguardo alla comprensione di tale UL con l'uso dei dizionari monolingui e bilingui.

Il foglio è stato diviso in due parti orizzontalmente, per il DM e per il DB. Per ogni dizionario, si è riprodotta la voce completa dell'UL in un'unica cella del foglio e in seguito, nella stessa riga, si sono scomposte le stesse informazioni in celle separate: divisione in sillabe; marca d'uso; categoria grammaticale; esempi; polirematiche (denominazione del DM). Anche la definizione è stata scomposta in celle per forma equivalente e per semi identificanti ogni caratteristica, questi ultimi divisi ancora fra *semi descrittivi* e *semi applicativi* (denominazione di Pottier, 1970).

La prima colonna del foglio conteneva l'identificazione degli studenti come **DMLI1** (gruppo dizionario monolingue, studente 1 di Lingua Italiana I) successivamente per gli otto studenti del gruppo e sotto la parte riguardante il dizionario bilingue **DBLI1** (gruppo dizionario bilingue, studente 1 di Lingua Italiana I). La seconda colonna indicava l'alternativa scelta dallo studente: in neretto quelle indicanti l'alternativa giusta. Nella terza colonna venivano riprodotte le digitazioni<sup>3</sup> degli studenti inserite nel campo di ricerca lemma del dizionario elettronico. Sulla riga corrispondente ad ogni studente seguendo le informazioni di ogni colonna (i dati lessicografici), nelle celle è stata segnalata la lettera **F** quando quell'informazione del dizionario (i semi, l'esempio ecc.) ha reso facile la comprensione, secondo quello che ha scritto lo studente sull'apposito spazio nel compilare l'esercizio, oppure **D** quando il dato ha reso la comprensione difficile e sono rimaste vuote le celle corrispondenti alle informazioni che non sono state citate. Le 'marche d'uso', per esempio, non sono state notate da nessuno studente. Nel compilare le schede, oltre a creare spazi uniformi e comparabili, si è lasciato uno spazio perché si annotassero le informazioni che i discenti ritenevano rilevanti nella loro ricerca della UL.

Le alternative scelte e le dichiarazioni dei discenti riguardanti ognuna delle quaranta UL e i due gruppi di studenti (DM e DB) organizzati in un unico foglio hanno permesso una visualizzazione che ha favorito l'analisi dei dati per confrontarli secondo i risultati positivi, le azioni dello studente e le informazioni lessicografiche dei dizionari usati.

Per quanto riguarda il gruppo degli studenti che hanno usato il dizionario bilingue, l'uso del DB è risultato efficiente e efficace tutte le volte in cui esso offriva come primo equivalente quello che corrispondeva a una delle figure presentate e che lo studente riusciva subito ad identificare come pertinente al contesto letto. Invece, non è stato efficace per la comprensione di alcune unità lessicali che presentavano caratteristiche particolari, come quelle riguardanti gli utensili di cucina *mestolo* e *tegame* (testo n.1). Nel primo caso, invece di *mestolo* il DB conteneva la parola *mestola* e il suo equivalente in portoghese *escumadeira*, che fa parte dello stesso campo semantico, ma è un utensile con funzioni diverse. Dal contesto si deduceva facilmente che la mestola non era l'utensile adeguato perché non può

---

3 È stato richiesto agli studenti di scrivere in un apposito spazio del foglio di risposta il modo in cui digitavano la UL. Per esempio, la ricerca del lemma corrispondente al verbo coniugato *marcisco* ha rivelato molte difficoltà.

contenere l'acqua; nonostante ciò, alcuni studenti nella risposta hanno selezionato la sua immagine, fatto che può dimostrare maggiore fiducia nel dizionario che nel contesto, o semplicemente un caso di disattenzione. Altri invece hanno scritto che il dizionario ha reso la comprensione più difficile. Nel caso di *tegame*, il DB forniva equivalenti generici come *panela*, *caçarola* [IT: *pentola*, *casseruola*], ma questi equivalenti non aiutavano a scegliere fra le quattro figure di pentole diverse presentate, fra le quali appunto il *tegame*, caratterizzato dai bordi bassi. In questo caso, gli studenti hanno annotato che sarebbe stata utile una descrizione dei tipi di pentole, perché non ne conoscevano le differenze.

Si è verificato che la presenza di dettagli descrittivi nella definizione del DM è stata di gran aiuto agli utenti. Per esempio, gli studenti hanno dichiarato che il tratto distintivo *di bordi bassi* nella voce *tegame* li ha condotti alla figura di una pentola con i bordi bassi e che nella definizione di *mestolo* le informazioni *di metallo di forma piuttosto incavata* hanno facilitato la comprensione di quale utensile si trattasse.

Da un'altra parte, l'eccesso di informazioni tecnico-scientifiche è stato indicato dai consultatori del DM come un ostacolo alla comprensione, mentre la presenza di dati culturali li ha favoriti. Per esempio, la voce *alloro* (testo n.1) presenta il primo significato relativo all'albero *alto fino a dieci metri, con foglie coriacee aromatiche [...]*, e nella seconda accezione presenta l'uso culturale relativo al *simbolo di sapienza, di gloria [...]*. L'italiano e il portoghese, ambedue lingue di origine latina, condividono l'immagine metaforica delle foglie di alloro e ciò facilita la comprensione meglio delle caratteristiche dell'albero. Tuttavia, alcuni studenti hanno indicato che sarebbe stata utile una descrizione più precisa della forma delle foglie, giacché davanti alle fotografie del test (A. *ruta*; B. *prezzemolo*; C. *rosmarino*; D. *alloro*) chi non era un po' esperto di cucina non era in grado di distinguerle. In modo analogo, i semi descrittivi *con [...] iridescenze sulla testa e sul collo* nella voce di *piccione* sono stati indicati come causa di difficoltà, mentre l'esempio dato come unità polirematica (nomenclatura di De Mauro per le unità fraseologiche) *piccione viaggiatore* è stata indicata come informazione che facilitava la comprensione.

Dalle azioni degli studenti, si è osservato che, riguardo a tutti e due i dizionari elettronici utilizzati, il fatto che accanto allo spazio di digitazione non fosse presentato il lemmario è stato un fattore sfavorevole per l'apprendente di italiano LS, perché a volte questo utente scrive la parola in modo sbagliato o non ne conosce la forma lemmatizzata presente nei dizionari (singolare, maschile, verbi all'infinito). Per esempio, la UL *piccioni* (testo n.1) è stata digitata spesso in iversi modi, come *picciono*, *picciona*, *piccio*, prima di arrivare alla forma base *piccione*. Se ci fosse stato il lemmario sullo schermo, sarebbe stata più facilmente identificata. Nello stesso modo, si verificata molta difficoltà con il verbo coniugato *marciscono*, prima di arrivare a *marcire*. Con lo sviluppo della tecnologia, si spera che i dizionari elettronici possano condurre l'utente alla pagina del lemma anche se viene digitato un verbo coniugato.

D'altro canto, alcuni comportamenti dei discenti hanno ostacolato l'efficace consultazione, come per esempio, tanto nel gruppo DM quanto nel DB, il non proseguire nella lettura completa della voce quando il significato adeguato al testo era uno degli ultimi (es. *busta* e *carrello*, testo n.3). Questo fatto è stato osservato anche in uno studio condotto da Cote Gonzales e Tejedor Martinez (2007) nell'ambito dell'insegnamento dell'inglese in Spagna.

Un'altra difficoltà si è avuta quando la forma flessa di un verbo corrispondeva a un lemma indipendente, come nel caso di *inginocchiata* (testo n.4), sostantivo femminile (un tipo di finestra con inferriata curva e sporgente nella parte inferiore) omografo del participio passato femminile del verbo *inginocchiare* (significato corretto nel testo). In questo caso, alcuni studenti hanno letto soltanto la voce non adeguata e hanno segnalato la definizione come difficoltosa. Anche in questo caso, la presenza del lemmario sarebbe stata utile. Proficua sarebbe stata anche una nota con l'informazione sulla forma omografa del verbo.

## 4 Contenuto e ordine di informazioni nella microstruttura

Non si intende discorrere su ogni informazione contenuta nella microstruttura dei dizionari consultati, ma soltanto su quelle che hanno contribuito o meno alla comprensione delle unità lessicali evidenziate nel testo.

Nei riguardi del DB, anche se la microstruttura è più semplice di quella del DM, è doveroso notare che si è avuto il cento per cento di risposte corrette quando il DB ha presentato un equivalente pertinente al significato del testo e alla rappresentazione della UL e lo studente è riuscito a trovarlo digitando la forma giusta e scegliendo l'accezione adatta.

Una caratteristica che riguarda la micro e la macrostruttura del dizionario bilingue utilizzato ha influenzato la performance degli studenti: l'organizzazione delle parole omonime e polisemiche. Anzi che presentare le parole omonime in differenti lemmi con i loro equivalenti, il DB presenta un unico lemma con tutti i suoi equivalenti, riguardanti le diverse accezioni in portoghese, distinti solo da numeri in neretto (**1. 2. 3.**). Questa caratteristica è stata segnalata dagli studenti come un fattore di difficoltà. La scelta del lessicografo è giustificabile in un dizionario cartaceo, dove esiste il problema dello spazio fisico e raggruppare i significati è una soluzione efficace. Tuttavia in un dizionario elettronico, dove il problema dello spazio non esiste, questa scelta non è la più adeguata, come si è comprovato in questa ricerca.

Dai risultati del gruppo con il dizionario monolingue (DM), si è notato che in tutte le definizioni delle UL in cui c'erano *semi applicativi* indicanti cioè 'qual è l'uso', 'a cosa serve', la presenza di questi semi ha contribuito alla comprensione, secondo l'opinione degli informanti. Gli utenti che hanno notato questi semi nella definizione hanno avuto successo nella comprensione, come nei casi delle UL *piccioni* - sema applicativo: *allevato per le carni gustose* - e *alloro* - sema applicativo: [*le foglie di tale pianta*] *usate in cucina per il tipico gradevole aroma* (testo numero 1).

È stato accennato prima l'insuccesso discente dovuto al non proseguire fino alla fine nella lettura della voce, tuttavia è da segnalare che è auspicabile l'ordine dal più alto al più basso uso nella sequenza delle accezioni, in modo che, *alloro* sia descritto prima come foglie e dopo come albero.

La relazione (osservata nella scheda FLICULAD prima citata) fra il numero di risposte corrette e le dichiarazioni degli studenti riguardo alla microstruttura delle UL mostra quanto l'organizzazione di un'opera lessicografica, specie se in formato elettronico, possa influire sulla comprensione del lemma. Un altro dato importante da considerare è anche l'atteggiamento dell'utente, cioè le sue azioni, che possono o no contribuire al conseguimento del successo nella ricerca di un vocabolo. Su questo punto spetta al professore di italiano LS/L2 il compito di portare il dizionario, "il tesoro della lingua", in aula, far vedere agli studenti le possibilità di consultazioni e sviluppare insieme la loro competenza di utenti.

## 5 Considerazioni finali

Queste sono solo alcune delle osservazioni ricavate dai risultati ottenuti, che non hanno la pretesa di essere una critica lessicografica ai dizionari utilizzati, ma che possono dare un ausilio a lessicografi, docenti e studenti di italiano LS/L2 nell'elaborazione di nuove opere lessicografiche, nell'affermazione del ruolo dei dizionari in aula, bilingui o monolingui, giacché si è verificato che i risultati della comprensione quando si parte dal contesto (gruppo senza dizionario) sono stati meno efficaci di quelli raggiunti con l'uso dei dizionari, e grazie all'azione del docente nell'insegnare a sfruttare bene questo importante strumento.

Le ricerche empiriche sull'uso dei dizionari offrono nuovi orizzonti per l'elaborazione di opere lessicografiche, principalmente quelle con obiettivi pedagogici, perché, una volta individuate e sistematizzate le attitudini e le opinioni degli utenti, è possibile perfezionare la stesura delle voci e riflettere sull'inclusione o meno di certe informazioni nella micro- e macrostruttura del dizionario.

## 6 Riferimenti bibliografici

- De Mauro, T. (1997) DAIC - Dizionario Avanzato dell'Italiano Corrente. Torino: Paravia, 1997.
- De Mauro, T. (2000), *Il Dizionario Italiano On-Line*, prima disponibile <http://old.demauroparavia.it/> [10/2009]
- Michaelis Dicionário Escolar Italiano-Português* (2003) <http://michaelis.uol.com.br/escolar/italiano/index.php> [10/2009].
- Corda, A., Marellò, C. (2004) *Lessico insegnarlo e impararlo*. 1° Ed. 1999, Torino, Paravia e 2° Ed. Perugia, Guerra.
- Cote Gonzalez, M., Tejedor Martinez, C. (2007) Dictionary use and translation activities in the classroom. In: Welker, H.A. (Org.) *Horizontes De Linguística Aplicada*, Ano 6, N.2. Brasília, UnB.
- Morales, H. L. (1994) *Métodos De Investigación Linguística*. Salamanca, Ediciones Colegio De España.
- Nunes, P.A., Finatto, M.J.B. (2007) Dicionários monolíngues para aprendizes de inglês como língua estrangeira: alguns elementos para o Professor. In: Welker, H.A. (Org.) *Horizontes De Linguística Aplicada*, Ano 6, No.2. Brasília, UnB.
- Welker, H.A. (2008) *Panorama Geral da Lexicografia Pedagógica*. Brasília, Thesaurus Editora.

- Welker, H. A. (2010) *Dictionary Use. A general survey of empirical studies*. Brasília, Author's Edition. [http://ppla.unb.br/hawelker/images/stories/professores/documentos/2010\\_Dictionary\\_use.pdf](http://ppla.unb.br/hawelker/images/stories/professores/documentos/2010_Dictionary_use.pdf) [04/2014]
- Xatara, C. et al [Org.] (2011) *Dicionários na teoria e na prática como e para quem são feitos*. São Paulo, Parábola.
- Pottier, B. (1978) *Linguística Geral - Teoria e Descrição*. Rio de Janeiro, Presença.
- Zucchi, A. M. T. (2010a) *O dicionário nos estudos de língua estrangeiras: os efeitos de seu uso na compreensão escrita em italiano*. Tese de Doutorado. PPG Semiótica e Linguística Geral, FFLCH, USP. Orientadora: Profa. Dra. Maria Aparecida Barbosa. São Paulo.  
[http://dedalus.usp.r/F/5862IAKNNJG93L2EM1K6XSADLDTS9SR5UD7XBEG214VAPINLP7-55951?func=full-set-set&set\\_number=061963&set\\_entry=000001&format=999](http://dedalus.usp.r/F/5862IAKNNJG93L2EM1K6XSADLDTS9SR5UD7XBEG214VAPINLP7-55951?func=full-set-set&set_number=061963&set_entry=000001&format=999)
- Zucchi, A.M.T. (2010b) *O uso de dicionários na compreensão escrita em italiano LE*. In Dykstra, A. e Schoonheim, T. *Proceedings of the 14th EURALEX International Congress*. Leeuwarden, Friske Akademy.  
[http://www.euralex.org/proceedings-toc/euralex\\_2010/](http://www.euralex.org/proceedings-toc/euralex_2010/) [04/2014]

### Appendix 1 - Le unità lessicali

		Le unità lessicali (UL) evidenziate, qui presentate solo nel co-testo
TESTO 1	1	mia madre rosola i <b>piccioni</b> in olio
	2	i piccioni in olio in un <b>tegame</b> , li sgocciola
	3	soffrigge un <b>trito</b> abbondante di cipolla
	4	una foglia di <b>alloro</b> e un po' di vino
	5	aggiunge due <b>mestoli</b> di acqua calda
TESTO 2	1	si siede a <b>capotavola</b> e si accorge
	2	si accorge di aver dimenticato la <b>dentiera</b> . Girandosi
	3	infilato una mano in <b>tasca</b> ed averne tolto una dentiera
	4	fa uso anche dello <b>stuzzicadenti</b> . Alla fine
	5	Non sono un dentista. Sono un <b>becchino</b> ."
TESTO 3	1	rifiuti solidi che non <b>marciscono</b> per i solidi
	2	di attirare <b>topi</b> e scarafaggi.
	3	di attirare topi e <b>scarafaggi</b> . Se questo solido
	4	io sono una che non <b>spreca</b> . Non sono una mangiona
	5	il mio vitto è pasta o <b>riso</b> , pane, formaggio
	6	un po' di <b>affettati</b> , e poi gli avanzi
	7	e poi gli avanzi li do alla <b>cagna</b> della vicina.
	8	le <b>scatole</b> dei regali, invece, le tengo io per organizzare le mie cose
	9	uso le <b>bottiglie</b> come vasi.
	10	avevo una <b>busta</b> per l'umido
	11	fino a quando <b>scendevo</b> ,
	12	mi portavo con il <b>carrello</b> le mie buste
	13	attaccate dei <b>manifesti</b> informativi
	14	con un <b>timbro</b> ufficiale del Comune
	15	invece di sprecare <b>soldi</b> in corsi professionali di mestieri



TESTO 4	1	quando vede sua madre, Anna, <b>inginocchiata</b> davanti a um mucchio di foto
	2	inginocchiata davanti a um <b>mucchio</b> di foto
	3	mi ero messa a fare ordine nei <b>cassetti</b> e ho trovato delle foto
	4	era così buffo, con quel <b>pizzo</b> da diavoletto
	5	e il suo inseparabile <b>berretto</b> rosso...
	6	quando noleggiammo il <b>motoscafo</b> e ce ne andammo
	7	con papà che non sapeva manovrare il <b>timone</b> e ci sballottava
	8	mah, sembro una suora con le <b>ballerine</b> , gli occhiali, i capelli legati
	9	mah, sembro una suora con le ballerine, gli <b>occhiali</b> , i capelli legati
	10	con le ballerine, gli occhiali, i <b>capelli legati</b> , senza un filo di trucco
	11	dai, vieni qua, <b>bacia</b> la tua vecchia mamma...
	12	devo scappare, sennò il direttore <b>si arrabbia</b> e chi lo sente dopo!
	13	gli ho detto che la prendevi oggi la <b>valigia</b> , eh!
	14	voglio chiedergli in prestito il <b>servizio di posate</b> francese per la cena
	15	chiedigli di renderti l' <b>impermeabile</b> che hai lasciato nella sua macchina.

#### Appendix 2: Alcuni dei risultati statistici (CEA/USP)

Testo	proporzione di risposte corrette	
	media	ds
1	0,64	0,20
2	0,83	0,25
3	0,83	0,13
4	0,76	0,15

Tabella 1: Proporzioni di risposte corrette d'accordo con il testo analizzato.

Testo 1	proporzione di risposte corrette	
	media	ds
DM	0,75	0,14
DB	0,70	0,15
SD	0,48	0,18

Tabella 2: Proporzioni di risposte corrette nel testo 1 d'accordo con l'uso del dizionario.



Testo 2		
Gruppi e dizionario	proporzione di risposte corrette	
	media	ds
DM	0,90	0,19
DB	0,93	0,10
SD	0,68	0,34

Tabella 3: Proporzioni di risposte corrette nel testo 2 d'accordo con l'uso del dizionario.

Testo 3		
Gruppi e dizionario	proporzione di risposte corrette	
	media	ds
DM	0,84	0,14
DB	0,88	0,09
SD	0,76	0,15

Tabella 4: Proporzioni di risposte corrette nel testo 3 d'accordo con l'uso del dizionario.

Testo 4		
Gruppi e dizionario	proporzione di risposte corrette	
	media	ds
DM	0,83	0,09
DB	0,79	0,15
SD	0,67	0,16

Tabella 5: Proporzioni di risposte corrette nel testo 4 d'accordo con l'uso del dizionario.

### Appendix 3 - Esempio di uno dei quattro testi

#### Testo 4

#### Vecchie foto e ricordi

Sono le otto di mattina, Lidia si affretta alla porta perché è leggermente in ritardo per il lavoro, quando vede sua madre, Anna, **inginocchiata** (1) davanti a un **mucchio** (2) di foto, così assorta che non si accorge della figlia. Lidia, curiosa, le si avvicina:

Lidia: Mamma, che stai facendo? Buongiorno...

Anna: Buongiorno tesoro... niente, mi ero messa a fare ordine nei **cassetti** (3) e ho trovato delle foto che non vedevo da anni...

Lidia: Ah... fai vedere anche a me.

*Curiosa, Lidia si avvicina e comincia a guardare le foto che Anna ha in mano.*

*Lidia: Guarda guarda! Ma questo non è Franco?*

*Anna: Proprio lui! Te lo ricordi? Eri così piccola...*

*Lidia: Piccola! Avrò avuto 14 anni... ma già, per te sono piccola anche adesso... Certo che mi ricordo di Franco! E come potrei dimenticarlo? Era così buffo, con quel **pizzo (4)** da diavoletto e il suo inseparabile **berretto (5)** rosso...*

*Anna (mostrando una foto) - Questa è quella volta che facemmo la gita in Corsica con papà, ti ricordi?*

*Lidia: Certo! Quando noleggiammo il **motoscafo (6)** e ce ne andammo tutti a Saint Florent, con papà che non sapeva manovrare il **timone (7)** e ci sbalottava tutti da una parte all'altra! Ti ricordi quando ci rovesciammo addosso il caffè appena fatto perché lui fece una manovra brusca e nessuno riuscì a restare in piedi? Come si arrabiò Franco quando vide che si era macchiato tutto il suo candido completo da marinaio! Me lo ricordo ancora!*

*Anna: E guardati in questa foto di scuola! Certo che eri proprio carina!*

*Lidia: Carina! Mah, sembro una suora, con le **ballerine (8)**, gli **occhiali (9)**, i **capelli legati (10)**, senza un filo di trucco...*

*Anna: La bellezza dell'asino... dai, vieni qua, **bacia (11)** la tua vecchia mamma...*

*Lidia: Va bene, va bene (la bacia; poi guarda l'orologio); uh, come è tardi! Devo scappare, sennò il direttore **si arrabbia (12)** e chi lo sente dopo! Ciao mamma.*

*Anna: Ah! E non dimenticarti di passare da Gianni! Gli ho detto che la prendevi oggi la **valigia (13)** eh!*

*Lidia: Sì, sì, stai tranquilla. Voglio anche chiedergli in prestito il **servizio di posate (14)** francese per la cena di sabato...*

*Anna: Allora, già che ci sei, chiedigli di renderti l'**impermeabile (15)** che hai lasciato nella sua macchina.*

*Lidia: È vero... l'impermeabile!! E chi ci pensava più!*

*Anna: L'impermeabile ti ci vuole proprio ora, che comincia a piovere parecchio.*

*Lidia: Sì, mamma, non ti preoccupare, lo prendo lo prendo. Ciao, e buoni ricordi!*

*Anna: Buona giornata cara, a più tardi*

# **Lexicography and Language Technologies**



# ALiquot – Atlante della Lingua Italiana QUOTidiana

Michele Castellarin, Fabio Tosques  
Humboldt-Universität zu Berlin  
michele\_castellarin@yahoo.it, ftosques@gmail.com

## Abstract

Der folgende Beitrag erklärt und stellt die Methoden und Resultate dar, die im Projekt ALIQUOT (Atlante della Lingua Italiana QUOTidiana – Atlas der italienischen Alltagssprache) verwendet bzw. gewonnen wurden, um die italienische Alltagssprache zu untersuchen. Das Ziel des Projekts besteht darin, geolinguistische Karten zu erzeugen, die die enorme Vielfalt der regionalen und lokalen Varianten der italienischen Sprache im täglichen Gebrauch sichtbar machen. Die Daten für ALIQUOT werden auf der Basis der indirekten Methode mit einem Online-Fragebogen anonym erhoben. Nur wenige persönliche Daten (Alter, Geschlecht, Wohnort, Beruf, Bildungsgrad usw.) werden abgefragt. Mit Hilfe dieser Daten sollen spätere soziolinguistische Analysen ermöglicht werden. Sämtliche geolinguistische Karten werden online publiziert und sind für jeden frei zugänglich. Wir möchten der Forschung und Lehre mit ALIQUOT ein nützliches Tool in die Hand geben, welches sprachliche Phänomene und Variationen des Italienischen präzise darstellt. ALIQUOT ist damit der erste Sprachatlas, der exakte und umfangreiche Karten der italienischen Alltagssprache anschaulich präsentiert.

**Keywords:** Geolinguistik; Geosynonym; Digitaler Sprachatlas; Alltagssprache; Regionalismen; Dialekte; sprachliche Variation

## 1 Das Projekt

*Abbiocco, papagna, papazze* oder *cicagna*? Und wie sagt man bei Euch, wenn einen nach dem Essen die typische Müdigkeit überkommt? Das fragte am 10. September 2013 die Moderatorin Isabella Eleodori vom Radiosender *R101* ihre Hörer in einer Sendung nach 15:00 Uhr. Die Antworten der Hörer kamen aus ganz Italien, von Bozen bis Palermo. Nicht ganz klar wurde in der kurzen Umfrage, ob es sich bei den Antworten um Dialektismen oder um Regionalismen bzw. um Alltagssprache handelt. In der Tat ist es häufig nicht ganz einfach zu entscheiden, ob eine bestimmte Bezeichnung nun dialektal, regional oder alltagssprachlich gefärbt ist.

Der enorme Reichtum im Bereich Wortschatz, der häufig von Region zu Region variiert sowie die Variationen im Bereich der morphosyntaktischen und phonologischen Strukturen werden unter der Bezeichnung *italiano regionale* zusammengefasst. Mit diesem Konzept werden alle diatopischen Variationen subsumiert, die sich vom Standarditalienischen unterscheiden und für die Variationen in Italien grundlegend sind.

Die Verknüpfung von sprachlicher Variation und die Darstellung im geographischen Raum wird schon seit vielen Jahrzehnten – besonders in diversen Sprachatlanten – durchgeführt. Dabei ist zu beachten, dass es zwar eine große Zahl von Untersuchungen zu den Dialekten und Minderheitensprachen in Italien gibt, hingegen die regionalen Varianten eher stiefkindlich behandelt werden.

Erstmals wurde das Konzept des *italiano regionale* 1939 von Devoto eingeführt (vgl. Devoto 1939). Dabei handelt es sich um das Ergebnis, welches durch die (sprachliche) Einigung Italiens entstanden ist, als die Nationalsprache auf die vielen dialektalen Varianten traf: „quante sono le regioni italiane, altrettanti sono i tipi di italiano regionale che si vanno costituendo“ (Devoto 1939: 60 cit. in Cerruti 2009: 18-19). Soziale Klassen, die bis dahin ausschließlich Dialekt gesprochen haben wurden zu „creatrici di lingua“ (Cerruti 2009: 18).

Zu den ersten wissenschaftlich-systematischen Untersuchungen zu diatopsichen Variationen des italienischen Wortschatzes zählt die 1956 von Robert Rüegg verfasste Dissertationsschrift „Zur Wortgeographie der italienischen Umgangssprache“ (vgl. Rüegg 1956).

Mit dem Konzept der Regionalismen in Italien beschäftigten sich beispielsweise De Mauro (1963), De Felice (1977), Sobrero (1996) sowie Untersuchungen wie jene von Antonini und Moretti (2000), Sobrero und Miglietta (2006), Cerruti (2007) oder Poggi Salani (2010). Wie schon Devoto in seiner Untersuchung, kommen auch diese Autoren zu dem Ergebnis, dass es sich beim *italiano regionale* um eine Art „Zwischensprache“ (Interlingua) handelt, die zwischen Dialekt und Standardsprache anzusiedeln ist. Sobrero und Miglietta definieren diesen Typ als eine Sprache, die von phonetischen, morphologischen, syntaktischen und lexikalischen Erscheinungen der lokalen Dialekte beeinflusst wird.

Für Telmon hingegen, der von einer „nuova dialettizzazione“ (Telmon 1990: 14) spricht, handelt es sich bei den Regionalismen um Wörter „che provengono dal fondo lessicale del dialetto (o dei dialetti) e, trovandosi in un contesto italiano, sono adattate al sistema morfo(no)lessicale [sic!] dell'italiano, quale risulta da analoghe trasferenze ai diversi livelli“ (Telmon 1990: 14-15).

Bevor das *italilano regionale* die sprachwissenschaftliche Forschung erreichte, verzeichneten einzelne Wörterbücher diatopische Varianten des täglichen Gebrauchs. Dazu zählen beispielsweise das in den 1950er Jahren erschienen *Dizionario Moderno delle parole che non si trovano nei dizionari comuni* von Alfredo Panzini (1950), das *Grande Dizionario della Lingua Italiana* von Salvatore Battaglia (1972), das *Grande Dizionario Italiano dell'Uso* (GRADIT) (cfr. De Mauro 1999) und das *Vocabolario della Lingua Italiana* (Treccani 2009).

Was sowohl in den Forschungsarbeiten wie auch in den Wörterbüchern fehlt, sind Karten, auf denen die verschiedenen Varietäten im Raum visualisiert werden. Ein erster Versuch, diese Lücke zu schließen und Geosynonyme auf einer Karte zu präsentieren, wurde in *Il Vocabolario della Lingua Italiana 2009* (Treccani 2009) realisiert. Die 100 Karten, von *accendere* bis *vigile urbano*, die im Anhang des Treccani-Wörterbuchs zu finden sind, zeigen einen ersten Ansatz, wie Geosynonyme verzeichnet werden können. Problematisch ist hier jedoch, dass die verschiedenen Bezeichnungen nach Regionen eingetragen wurden und Sprachgrenzen in der Realität nur selten mit den Regionengrenzen übereinstimmen.

Einen anderen Weg der Visualisierung geht das von Elspaß und Möller (vgl. Möller, Elspaß 2008) vor gut zehn Jahren initiierte Projekt *Atlas zur deutschen Alltagssprache* (AdA). Dort werden keine kompletten Regionen nach einer bestimmten Antwort eingefärbt, sondern die Antworten der einzelnen Orte werden als verschiedenfarbige Punkte auf der Karte sichtbar. Möglich wird dies auch durch die von ihnen verwendete Technik, da hier mit Hilfe von Software die geographische Information so ausgewertet wird, dass die Antworten der Informanten problemlos auf den Karten verortet werden können. Um eine ähnliche Darstellung, wie wir sie im AdA vorfinden, auch für Italien zu ermöglichen, wurde, ausgehend von den im Treccani verzeichneten Karten, die Idee des *Atlante della Lingua Italiana QUOTIdiana* (ALQUOT) geboren. Sinn und Zweck des Projekts ALQUOT besteht darin, die enorme lexikalische Vielfalt des italienischen Sprachraums anschaulich darzustellen und der Öffentlichkeit einen digitalen interaktiven Sprachatlas zur Verfügung zu stellen. Dafür werden seit Anfang 2013 alltags-sprachliche Bezeichnungen zu den unterschiedlichsten Begriffen abgefragt.

In der ersten Fragerunde, die vom 1. Januar bis zum 31. Juli 2013 durchgeführt wurde, wurde nach regionalen Varianten zu den folgenden zehn Lexemen gefragt: *marinare la scuola, lavorare, schiaffo, anguria, topo, fidanzato/a, calzetto, appendiabito, immondizia* und *pungere*. Die Datenerhebung erfolgte bzw. erfolgt ausschließlich online, d.h. die Nutzer sind aufgerufen, einen webbasierten Fragebogen auszufüllen. Dabei haben sie die Möglichkeit, aus vorgegebenen Antworten auszuwählen oder – falls keine davon zutreffen sollte – eine Antwort in das dafür vorgesehene Textfeld („altro“) einzutragen.

Die einzelnen Fragen der Fragerunden sind thematisch kategorisiert, z.B. Obst und Gemüse (*anguria, melone, fagiolini*) oder Gegenstände des täglichen Gebrauchs (*grempiule, balcone, immondizia, appendiabito*) etc. Die Kategorisierungen sollen die spätere Veröffentlichung in thematisch gegliederte Wörterbücher oder Atlanten ermöglichen.

Bei der Auswahl der indirekten Erhebungsmethode stützen wir uns auf die Erfahrungen Eichhoffs, da besonders bei Wortschatzuntersuchungen die „Vorteile der schriftlichen Befragung [= indirekten Befragung] voll zur Geltung [kommen], während die Nachteile das Datenmaterial in seinen wesentlichen Aspekten nicht berühren (Eichhoff 1982: 550). Besonders die Möglichkeit, viele Informanten in kurzer Zeit gewinnen zu können, war ausschlaggebend für diese Form der Erhebung. Nachteile, wie Spektrum der Informanten, Datenqualität, Datenschutz usw., die im Allgemeinen bei online-Erhebungen angeführt werden, scheinen nicht zuzutreffen. Das zeigen auch die jahrelangen Erfahrungen die Elspaß und Möller im Projekt AdA gesammelt haben (vgl. Elspaß, Möller 2006).

Nach Abschluss der ersten Fragerunde konnten wir feststellen, dass die Daten konsistent sind und besonders alle Altersgruppen der Gesellschaft erfasst wurden (vgl. Abbildung 1). Auch der Vergleich unserer Karten mit jenen von *Il Vocabolario della lingua italiana* (Treccani 2009) macht deutlich, dass die Antworten der Informanten weitgehend mit den dort verzeichneten übereinstimmen.

Sowohl die Erhebung der Daten als auch die Präsentation derselben erfolgt mit Hilfe von eigenständig entwickelter Software. Dabei wurde stets darauf geachtet, dass möglichst einfache Techniken eingesetzt werden, die es nicht erfordern, dass der Nutzer zusätzliche Plug-ins o.ä. installieren muss. Als Lösung hat sich hier der sogenannte LAMP-Stack (Linux, Apache, MySQL, PHP) angeboten, da hier

freie Tools zur Anwendung kommen und diese sich seit langer Zeit etabliert haben. Zusätzlich setzen wir auf modernste Techniken wie HTML5, CSS3 und verschiedene JavaScript Bibliotheken wie jQuery usw. Die Daten werden direkt nach dem Absenden des Fragebogens in einer relationalen Datenbank gespeichert.

Die Informantensuche erfolgt in erster Linie durch das Anschreiben von Freunden, Bekannten, Kollegen, Studenten und Universitäten, mit der Bitte, den Fragebogen auszufüllen und diesen an möglichst viele Freunde und Bekannte weiterzuleiten. Die Suche erfolgt somit nach dem bekannten Schneeballprinzip. Darüber hinaus nutzen wir für den Kontakt mit den Informanten das am häufigsten genutzte soziale Netzwerk *Facebook*. Das gibt uns die Möglichkeit, unsere Teilnehmer über aktuelle und neue Entwicklungen, die das Projekt betreffen, zu informieren und durch deren Wertungen und Kommentare neue Informanten zu gewinnen. Die vorhandene Technik erlaubt es uns, die eingegebenen Daten permanent zu überprüfen und zu visualisieren. Wo wir noch während der Fragerunde große Lücken in der Datenmatrix entdecken, schreiben wir gezielt Kommunen, Provinzen, kulturelle Einrichtungen und Schulen an. Immer mit der Bitte, den Fragebogen auszufüllen und unsere Umfrage weiterzuleiten.

Obschon wir bei der Erhebung der Daten nur auf das Internet setzen, hat sich gezeigt, dass alle Altersgruppen gut vertreten sind und die Verteilung zwischen männlichen und weiblichen Informanten ausgewogen ist (vgl. Abbildung 1). Trotz der im Grunde zufälligen Auswahl der Informanten konnten wir feststellen, dass die Daten konsistent und vertrauenswürdig sind.

Neben den eigentlichen Fragen zum Lexikon, müssen die Informanten auch einige sozio-demographische Fragen beantworten. Dazu zählen beispielsweise: Wohnort, Postleitzahl, Geburtsort, Alter, Ausbildung, Beruf, Geschlecht sowie der Geburtsort und Wohnort der Eltern. Dank dieser Daten ist es möglich, sozio-demographische Aussagen zu den Informanten zu machen und diese mit den Antworten zu verknüpfen, wodurch u.a. auch soziolinguistische Karten erzeugt werden können.

Die Präsentation der Daten (vgl. Castellarin, Tosques 2012) erfolgt ebenfalls ohne Installation von zusätzlicher Software seitens des Nutzers. Hier wird auf die, für unsere Zwecke völlig ausreichenden Möglichkeiten zurückgegriffen, die von Google Maps mit Hilfe der speziellen API (Application Programming Interface) zur Verfügung gestellt werden. Google Maps bietet schließlich nicht nur die Präsentation der Daten in einer großen Karte an, sondern ermöglicht auch, spezielle Regionen, Provinzen oder Städte heranzuzoomen. Da wir nach der Postleitzahl fragen, erhalten wir mit Hilfe der Zoomfunktion die Möglichkeit, in größeren Städten sehr detaillierte Karten zu erzeugen.<sup>1</sup>

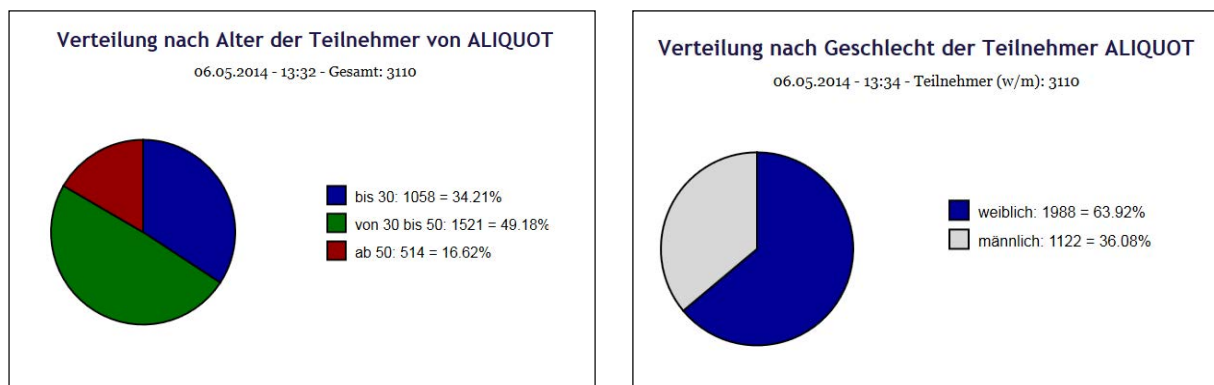
---

1 Dabei zeigte sich, dass in den Städten tatsächlich Varianten auftreten, die sich bestimmten Stadtvierteln zuordnen lassen (vgl. Abbildung 7).



## 2 Ergebnisse

In den folgenden Abschnitten werden exemplarisch fünf Karten aus den drei Fragerunden vorgestellt und zum Teil mit jenen des Treccani-Wörterbuchs verglichen. Auf den Webseiten des Projekts (<http://www.atlante-aliquot.de>) stehen mit dem Ende der dritten Fragerunde am 30. Juni 2014 Interessierten 41 Karten zur Verfügung. Bisher (Stand: 05/2014) haben insgesamt 3.110 Informanten teilgenommen (1. Runde: 867; 2. Runde: 1406; 3. und aktuelle Runde 837)<sup>2</sup>. Dabei zeigt sich, wie schon erwähnt, dass einerseits alle Altersgruppen hinreichend repräsentiert sind (vgl. Abbildung 1 links) und auch die Verteilung zwischen weiblichen und männlichen Teilnehmern ausgewogen ist (vgl. Abbildung 1 rechts).



**Abbildung 1: Verteilung nach Alter (links) und nach Geschlecht (rechts) der Teilnehmer von ALIQUOT.**

Beim Vergleich mit den Treccani-Karten geht es weniger um einen direkten Vergleich der Daten als um einen ersten Eindruck. Es ist uns durchaus bewusst, dass sich unsere und die Treccani-Karten nur mit Einschränkungen vergleichen lassen. Haben die Treccani-Karten (vgl. Abbildung 2 und 3 rechts) modellhaften Charakter und bilden die Realität eher synoptisch ab, kommen unsere Karten (vgl. Abbildung 2 und 3 links, 4, 5 und 6), denen eine völlig andere Datenbasis zu Grunde liegt, der sprachlichen Realität weitaus näher auch wenn diese natürlich immer noch komplexer ist als die, die sich – sei es in gedruckter, sei es in elektronischer Form – darstellen lässt. So zeigen die Treccani-Karten die Verwendung der Geosynonyme pauschal in den Grenzen der Regionen Italiens. Von einer regionen-basierten Darstellung haben wir von Anfang an Abstand genommen, da die administrativen Grenzen nur selten deckungsgleich sind mit den Sprachgrenzen. Nichtsdestotrotz sind die Treccani-Karten für uns eine große Hilfe: erstens bei der Auswahl der Geosynonyme für vergangene, laufende und zukünftige Fragerunden und zweitens für einen ersten Eindruck von der Qualität der im Projekt ALIQUOT erhobenen Daten.

2 Vgl. [http://www.atlante-aliquot.de/aliquot/showall\\_locations.php](http://www.atlante-aliquot.de/aliquot/showall_locations.php).

Die von uns generierten Karten stehen somit nicht in direkter Konkurrenz zu den Treccani-Karten. Letztere dienen dem Projekt als Ausgangsbasis für die Präsentation der in Italien verwendeten Alltagssprache. Da es sich bei der Alltagssprache um die am häufigsten verwendete und – besonders im Vergleich zu den Dialekten – weit weniger untersuchten Sprachform handelt, legt das Projekt ALIQUOT ebenfalls den Focus auf deren Verwendung, Formenreichtum und detaillierten Darstellung<sup>3</sup>.

## 2.1 Fragerunde 1: Karte zu *lavorare* („arbeiten“)

Die für ALIQUOT erhobenen Geosynonyme für *lavorare* bestätigen die allseits bekannte Dreiteilung der Halbinsel, wie sie in der Literatur, z.B. in De Felice (1984), Coveri, Benucci, Diadoro (1998) oder Grassi, Sobrero, Telmon 2003, in Sprachatlanten wie beispielsweise dem *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS, vgl. Jaberg, Jud 1928-1940) oder dem *VIVAio Acustico delle Lingue e dei Dialetti d'Italia* (Kattenbusch 1999 ff.) sowie im *Vocabolario della lingua italiana* Treccani (2009) beschrieben wurde. Durchgehend wird dort der Sprachraum in einen nördlichen (*lavorare*) und einen südlichen (*faticare*) unterteilt. In der dritten und kleinsten Zone, die Sizilien, Sardinien, den äußersten Westen Piemonts und Ligurien betreffen, benutzen die Sprecher Formen von *travagliare*.

Die im Treccani Wörterbuch publizierte Karte zu *lavorare* zeigt klar und deutlich, wo die drei Zonen der Verteilung der Geosynonyme von *lavorare* liegen (vgl. Abbildung 2 rechts).

Die im Projekt ALIQUOT erhobenen Daten zu *lavorare* (vgl. Abbildung 2 links) unterscheiden sich im Ergebnis nur leicht von jenen, die im Treccani Wörterbuch dargestellt werden.

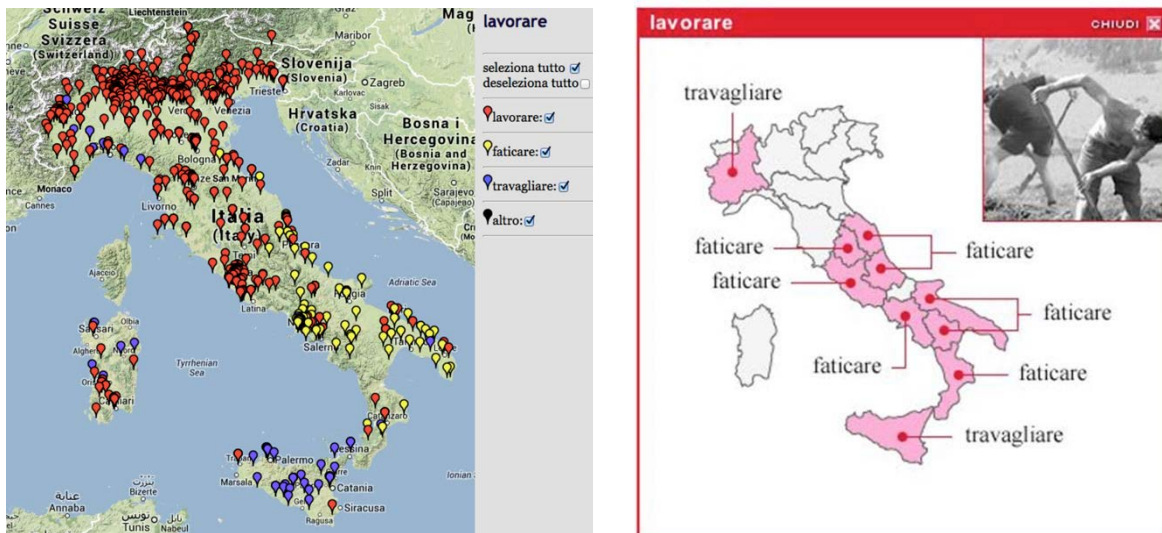


Abbildung 2: ALIQUOT-Karte (links) und Treccani-Karte (rechts) der Geosynonyme für das Lexem *lavorare*.

3 Wir möchten darauf hinweisen, dass die hier vorgestellten Karten teilweise etwas vereinfacht wurden. Dies hängt damit zusammen, dass das Punktenetz von ALIQUOT sehr engmaschig ist und die Karten aufgrund der zahlreichen Varianten und der hohen Teilnehmerzahl ohne graphische Aufbereitung ansonsten in gedruckter Form in diesem Größenformat nur schwer lesbar wären.

Die Karte scheint in jedem Fall die Annahme zu bestätigen, dass der Dialekt auf die Formen der Alltagssprache einwirkt. Dazu bedarf es jedoch weiterer Annahmen. Eine erste betrifft die offensichtliche Ausdehnung der Standardsprache, d.h. des Lexems *lavorare* bis in den Süden der Halbinsel. Während beispielsweise die wenigen Punkte von *faticare* und *travagliare*, die außerhalb der typischen Sprachgrenzen liegen, vermutlich auf den familiären Ursprung des Informanten zurückzuführen sind, trifft dies für *lavorare* nicht zu. Schon ein flüchtiger Blick zeigt, dass *lavorare* die dialektale Form *travagliare* in Piemont und Sardinien so gut wie ersetzt hat und auf dem besten Wege zu sein scheint, die regionale Form in Zentren wie Neapel, Bari, Lecce usw. zu verdrängen.

Der Gebrauch eines Geosynonyms anstelle eines anderen kann – was der hier beschriebene Fall zeigt – sehr wahrscheinlich auf eine systematische Italianisierung zurückgeführt werden. Dieser Prozess hat mit der politischen Einigung Italiens begonnen. So scheinen die südlichen Varianten nach und nach vom Standarditalienischen verdrängt zu werden.

Voraussagen, die das „große Rätsel der Sprachwissenschaft“ (Nützel 2009: 85), den Sprachwandel, betreffen sind häufig schwer zu treffen und können postwendend unseriös werden. Dennoch möchten wir anhand der dargestellten Situation kurz über die Zukunft der beiden Lexeme *faticare* und *travagliare* nachdenken. Die gegenwärtige Konstellation zeigt, dass die hochsprachliche Variante *lavorare* sich auch im Süden offensichtlich weiter verbreiten wird und damit eine immer größere Bedeutung im Wortschatz des täglichen Gebrauchs der Sprecher einnehmen wird.

## 2.2 Fragerunde 1: Karte zu *fidanzato/a* („Verlobter/Verlobte“)

Bezüglich der Geosynonyme von *fidanzato/a* werden im Treccani-Wörterbuch drei zusätzliche Varianten aufgezeigt (vgl. Abbildung 3 rechts). Im Norden der Halbinsel Formen von *moroso/a*, in Umbrien *frego/a* und in Kalabrien und Sizilien *zito/a*. Alle weiteren Regionen bleiben weiß, was vom Leser so verstanden werden könnte, dass hier die standardsprachliche Form *fidanzato/a* verwendet wird.<sup>4</sup> Die ALIQUOT-Karte zeigt eine sprachliche Situation, die komplexer ist als jene, die im Treccani-Wörterbuch verzeichnet ist. Ein erster Blick auf unsere Karte verdeutlicht, dass *zito/a* eine größere Verbreitung erfährt als auf der Treccani-Karte, d.h. neben Kalabrien und Sizilien auch die Basilikata und Teile Apuliens.

Auch bei der Verbreitung des Lexems *moroso/a* werden Unterschiede deutlich. Während im Treccani-Wörterbuch ganz Norditalien – ohne Trentino-Südtirol – genannt wird, zeigt die ALIQUOT-Karte hier und da Unterschiede: wird im Veneto ausschließlich *moroso/a* für die „geliebte Person“ verwendet, ist in Piemont und in der Lombardei *fidanzato/a* die gebräuchlichste Form. Auch bei der Verbreitung in Mittelitalien werden Unterschiede deutlich: so wird in der Emilia-Romagna großteils *moro-*

---

4 Dabei ist zu beachten, dass die Karten im Treccani-Wörterbuch keinen Anspruch auf Vollständigkeit haben. Das zeigt ein Blick auf alle 100 Karten. Häufig bleiben Regionen weiß, obschon dort eigentlich Varianten des Standards zu erwarten wären. Sehr deutlich wird dies z.B. bei der Region Molise, wo laut Treccani durchgehend die hochsprachliche Variante verwendet wird.

*so/a* verwendet. Das Lexem *frego/a*, welches im Wörterbuch von Treccani für Umbrien angegeben wird, ist auf der ALIQUOT-Karte nur sporadisch in dieser Region vorhanden.<sup>5</sup>

Ein anderes Geosynonym, welches bei Treccani nicht erwähnt wird, laut der Informanten von ALIQUOT aber vor allem in den großen Zentren wie Rom, Neapel, Cagliari, Florenz, Mailand und Turin zu finden ist, ist *ragazzo/a*. Ebenfalls nur bei ALIQUOT verzeichnet ist das Geosynonym *sposo/a*, das auch in anderen Teilen Südtaliens zu erwarten wäre.

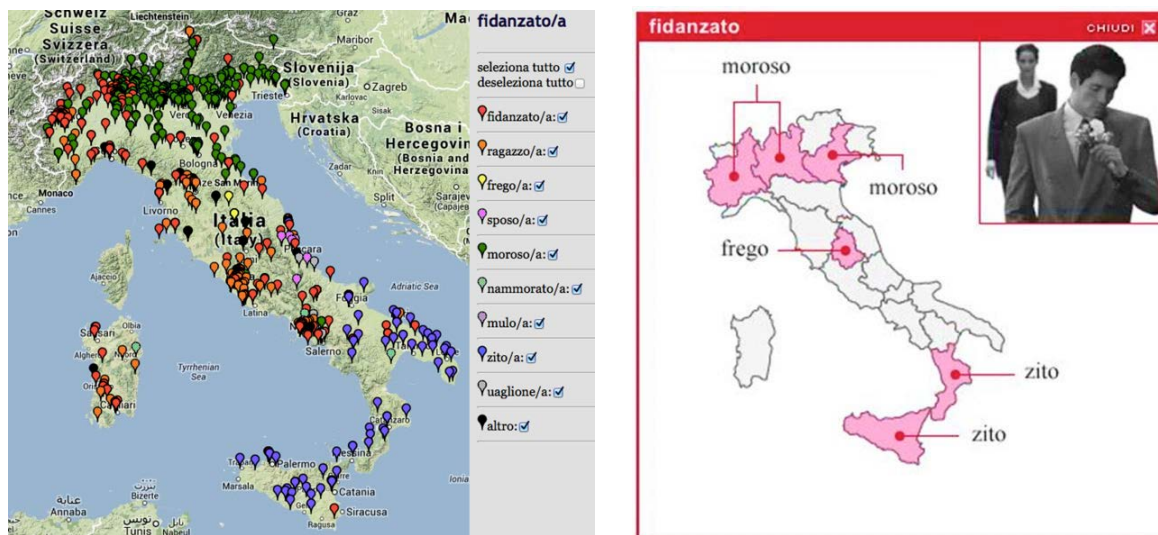


Abbildung 3: ALIQUOT-Karte (links) und Treccani-Karte (rechts) der Geosynonyme für das Lexem *fidanzato/a*.

Die im Projekt ALIQUOT erhobenen Daten zeigen somit für *fidanzato/a* und *ragazzo/a* ein anderes Bild als das Treccani-Wörterbuch. *Fidanzato/a* wird – wie zu erwarten – in der Toskana gebraucht, zeigt jedoch auch eine beachtliche Verbreitung auf der gesamten Halbinsel. Besonders in Rom, Neapel, Pescara, Sardinien, im westlichen Piemont, in Ligurien, Venetien, Friaul, im Trentino und etwas weniger in Apulien und Sizilien.

Ähnlich sieht es auch für den Terminus *ragazzo/a* aus. Ursprünglich vor allem in Rom verwendet, hat der Begriff eine beachtliche Ausdehnung erfahren, wenn auch nicht in dem Maße wie *fidanzato/a*. Formen von *ragazzo/a* werden, wie auf der Karte zu erkennen ist, auch in den großen Zentren Italiens verwendet. In Turin haben beispielsweise *ragazzo/a* und *fidanzato/a* die zu erwartende Form *moroso/a* vollständig verdrängt.

5 Dabei darf nicht unerwähnt bleiben, dass die Region Umbrien in der ersten Fragerunde nicht flächendeckend beantwortet wurde.

## 2.3 Fragerunde 2: Karte zu *padre/papà* („Vater/Papa“)

Eine Karte zu den Geosynonymen für *padre* ist im Treccani-Wörterbuch nicht vorhanden. Daher können unserer Ergebnisse nicht verglichen werden. Grossomodo existieren für *padre* fünf Geosynonyme in Italien (vgl. Abbildung 4). In der Toskana und in den Marken hören wir hauptsächlich Formen von *babbo*. In der Basilikata und in Apulien (ohne Salento) ist die häufigste Form *attane*. Im Salento selbst wird *sire* verwendet und seltener *tata*. Die Informanten ALIQUOTs antworteten im Norden Italiens hauptsächlich mit dem Terminus *papà*. Bei genauerer Betrachtung der Karte fällt auf, dass das Lexem *padre* in der Alltagssprache so gut wie unbekannt ist. Nur sehr wenige Informanten haben von den fast 1500 Teilnehmern die hochsprachliche Form *padre* als gebräuchlichste angegeben.

Sprachlich lassen sich auf der Karte für *padre* deutliche Zonen erkennen. Sofort ins Auge fällt das am häufigsten gebrauchte Lexem *papà*. In der Toskana und im Norden Sardinien wird durchgehend *babbo* verwendet.

In Süditalien konkurrieren neben der häufigsten Form *papà* verschiedene Geosynonyme: *attane*, *tata* und *sire*.



Abbildung 4: ALIQUOT-Karte für die Geosynonyme des Lexems *padre*.

## 2.4 Fragerunde 2: Karte zu *gomma da masticare* („Kaugummi“)

Als letzte lexikalische Karte möchten wir im folgenden die Ergebnisse für *gomma da masticare* vorstellen. Dabei handelt es sich um einen relativ jungen Begriff, der erst nach dem Zweiten Weltkrieg in den italienischen Wortschatz Eingang gefunden hat.

Die Karte (vgl. Abbildung 5) stellt eine Besonderheit dar, da hier eine Fremdsprache (englisch) auf die einheimische Sprache bzw. den einheimischen Dialekt trifft. Deutlich wird, dass die englische Be-



zeichnung und die englische Aussprache (*chewing-gum*) nur wenig vertreten sind. Die meisten Sprecher ziehen eine wörtliche Übersetzung des englischen Terminus dem Original vor und benutzen Formen von *gomma da masticare* oder einfach die Kurzform *gomma*. Diese beiden Formen verteilen sich über den ganzen Sprachraum. Eine beachtliche Teilnehmerzahl, die über ganz Italien verteilt ist, antwortet mit phonetischen und/oder morphologischen Varianten der englischen Form. So wird in manchen Gegenden aus *chewing-gum ciunga* oder *cevingum* [ˈtʃe:viŋɡum]. Daneben finden sich auch Lehnübersetzungen wie *cingomma*, *gingomma* oder *cincingomma*. Sehr weit verbreitet sind die beiden Formen *cicca* (von Norditalien entlang der Adriaküste bis nach Apulien) und *cicles* (besonders in Turin und im Großraum Bologna).



Abbildung 5: ALIQUOT-Karte für die Geosynonyme des Lexems *gomma da masticare*.

## 2.5 Fragerunde 3: Karte zum transitiven Gebrauch intransitiver Verben

Wir möchten den vorliegenden Beitrag nutzen, um vorausschauend die Ergebnisse zu betrachten, die den transitiven Gebrauch von intransitiven Verben untersucht.<sup>6</sup> Ein flüchtiger Blick auf die Karte (vgl. Abbildung 6) verdeutlicht eine Zweiteilung des Untersuchungsgebiets. Im Norden und in Mittelitalien (Toskana, Marken, Umbrien) werden Aussagen wie *scendimi le chiavi* („bring mir den Schlüssel nach unten“) als grammatisch falsch interpretiert. Im täglichen Gebrauch werden sie nicht verwendet. Südlich der Linie Neapel-Pescara werden die Aussagen als korrekt empfunden und in der Alltagssprache gebraucht.

6 Vorausschauend deshalb, da es eine Frage betrifft, die Teil der aktuellen dritten Fragerunde ist, die vom 01. Januar 2014 bis zum 30. Juni 2014 läuft. Zum Zeitpunkt der Erstellung der Karte haben ca. 850 Personen geantwortet. Wie auf der Karte zu sehen ist, sind die Punkte so gut verteilt, dass eine erste Aussage getroffen werden kann.

Offen bleiben noch die Antworten *accettabili ma non usate* („akzeptabel aber nicht verwendet“) und *inaccettabili ma usate* („inakzeptabel aber verwendet“). Der erste Fall trifft nur auf Nord- und Mittelitalien zu, wo – wie gesehen – der Großteil der Sprecher Aussagen dieses Typs als ungrammatisch komplett ablehnt. Sprecher, die die Meinung vertreten, solche Ausdrücke seien *inaccettabili ma usate* sollten sich konsequenterweise nur südlich der oben genannten Linie finden.

Überraschend ist vielleicht die Wahl der Antwort *inaccettabili ma usate* in den norditalienischen Zentren wie Mailand, Turin, Venedig und Bologna. Eine Einordnung der dort gegebenen Antworten erfordert jedoch weitere Untersuchungen, die besonders die soziolinguistischen Aspekte sowie die sozio-kulturellen Umstände der Informanten genauer unter die Lupe nehmen. Da sich diese Antwort auf die großen Städte beschränkt, sollte hier die interne Migration vom Süden nach Norden genauer untersucht werden.

Eine allgemein gültige Erklärung für dieses Phänomen wurde bisher noch nicht gefunden. Die Antworten lassen sich vielleicht auf das Sprachbewusstsein der Italiener zurückführen. Während für Norditaliener Aussagen wie *esci il cane* („führ den Hund Gassi“) zwar unüblich sind und nicht gebraucht werden, sind für Süditalien solche Aussagen typisch.

Aussagen wie diese werden von einer großen Sprechergruppe im Süden als akzeptabel empfunden und sie benutzen sie auch im täglichen Gebrauch. Interessanterweise findet sich in unseren Antworten im Süden auch eine Gruppe, der durchaus bewusst ist, dass es sich um eine inkorrekte Form handelt, diese aber dennoch Verwendung findet.

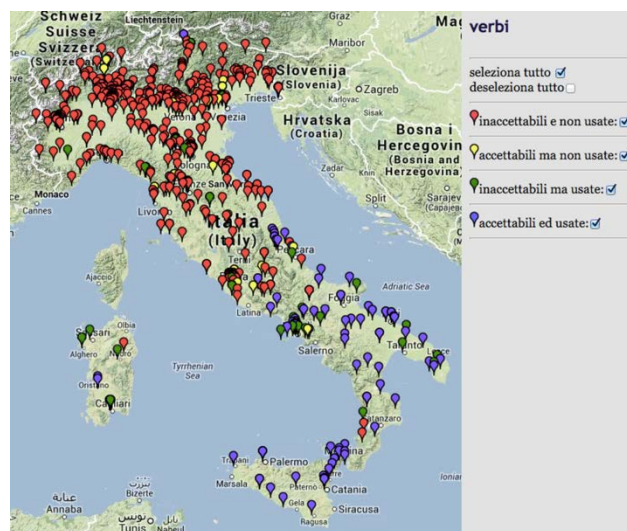


Abbildung 6: ALIQUOT-Karte für den transitiven Gebrauch intransitiver Verben.

### 3 Ausblick und Zukunftsperspektiven

Die Entscheidung, die Karten von ALIQUOT ausschließlich online zu publizieren, wurde aus zwei Gründen gefällt: Erstens können wir so schnell und unkompliziert die Karten schon kurz nach dem Ende eines Fragezyklus publizieren. Zweitens haben wir damit die Möglichkeit, Probleme der Visualisierung, die mit der Komplexität der Karten einhergehen, relativ einfach zu lösen.

Nicht unwesentlich ist, dass es sich bei den ALIQUOT-Karten um interaktive Karten handelt und nicht um statische. So erlaubt uns die Technik, einzelne Geosynonyme an- und abzuwählen, wodurch die Lesbarkeit und der Erkenntnisgewinn bezüglich der Verteilung im Raum enorm verbessert wird.

Der Technik, namentlich von Google Maps, verdanken wir auch die schon erwähnte Zoomfunktion, da dabei Variationen innerhalb einer Stadt sichtbar gemacht werden können. Hilfreich ist, dass wir nach den Postleitzahlen fragen und diese als Referenzpunkt für die Darstellung dient (vgl. Abbildung 7). Auf der Karte *anguria* zeigt sich, dass im Großraum Rom die Informanten am häufigsten mit dem für den mittellitalienischen Raum typischen Lexem *cocomero* (blau) antworten. In den nördlichen Stadtvierteln (Montemario, Vittorio, Nomentano) hingegen wurde erstaunlicherweise mit *anguria* (gelb) geantwortet, ein Lexem, das v.a. in Norditalien zu erwarten wäre. Weshalb auch in Rom mit der nördlichen Variante (*anguria*) geantwortet wird, obschon wir in Mittelitalien ausschließlich *cocomero* vorfinden und damit keine Transitionszone vorhanden ist, erfordert eine genauere Untersuchung. In jedem Fall zeigte sich, dass in den genannten Stadtvierteln im Norden Roms auch bei anderen unserer Fragen eher mit einem standarditalienischen Lexem geantwortet wurde als mit dem für Rom und Umgebung typischen.

Ziel von ALIQUOT ist, in der Zukunft detaillierte Studien zur diatopischen und diastratischen Variation der italienischen Alltagssprache auf der Halbinsel und in der italienischen Schweiz vorzunehmen. Dafür erheben wir bei den Nutzern einige sozio-demographische Daten wie das Alter, Geschlecht, die Schulbildung usw. Eine Verbindung der sozio-demographischen Daten mit den linguistischen führt somit zu neuen Ergebnissen, die im Bereich der Alltagssprache noch nicht genauer untersucht wurden.



Abbildung 7: Varianten zur Frage *anguria* innerhalb Roms (*cocomero* blau, *anguria* gelb).



Schließlich erlaubt uns die Technik auch den Ursachen für den Gebrauch bestimmter Geosynonyme auf den Grund zu gehen. Häufig liegen den alltagssprachlichen Bezeichnungen dialektale Formen zu Grunde, die glücklicherweise in den Sprachatlanten vorbildlich dokumentiert sind. So können die Daten von Sprachatlanten wie dem AIS oder dem ALI mit den Daten von ALIQUOT korreliert werden, wodurch die Interferenzen zwischen Dialekt und Alltagssprache verdeutlicht werden.

Zuletzt soll noch die Möglichkeit erwähnt werden, die Karten von ALIQUOT in einem gedruckten Atlas zusammenzufassen oder diese als Basis für zukünftige *joint-ventures* mit Lexikographen oder Lexikologen zur Verfügung zu stellen und damit die verschiedenen Kompetenzen optimal zu vereinen. So könnten erstmalig detaillierte Karten zur italienischen Alltagssprache entstehen, die dem Formenreichtum derselben würdig wären.

## 4 Literatur

- AdA: Elspaß, S., Möller, R. (2001). *Atlas zur deutschen Alltagssprache*. [<http://www.atlas-alltagssprache.de>].
- AIS: Jaberg, K., Jud, J. (1928–40). *Sprach- und Sachatlas Italiens und der Südschweiz*. 8 vol. Zofingen: Ringier.
- ALI: Bartoli, M. G. (1995). *Atlante linguistico italiano*. Roma: Istituto Poligrafo e Zecca dello Stato.
- ALIQUOT: Castellarin, M., Tosques, F. (2013): *Atlante della Lingua Italiana QUOTidiana*. [<http://www.atlante-aliquot.de>].
- Antonini, F., Moretti, B. (2000). *Le immagini dell'italiano regionale. La variazione linguistica nelle valutazioni dei giovani ticinesi*. Locarno: Dadò.
- Battaglia, S. (1972). *Grande Dizionario della Lingua Italiana*. Vol. 21. Torino: UTET.
- Castellarin, M., Tosques, F. (2012). ALIQUOT – L'Atlante della Lingua Italiana QUOTidiana. In *Rivista Italiana di Dialettologia. Lingue dialetti società*. XXXVI, pp. 245-262.
- Cerruti, M. (2009). *Strutture dell'italiano regionale. Morfosintassi di una varietà diatopica in prospettiva sociolinguistica*. Frankfurt a. M. u.a.: Peter Lang.
- Cerruti, M. (2007). Sulla caratterizzazione aspettuale e la variabilità sociale d'uso di alcune perifrasi verbali diatopicamente marcate. In *Archivio Glottologico Italiano* a. 92, n. 2, S. 203-247.
- Coveri, L., Benucci, A. & Diadori, P. (1988). *Le varietà dell'italiano: manuale di sociolinguistica italiana; con documenti e verifiche*. Roma: Bonacci.
- De Felice, E. (1977). Definizione del rango, nazionale o regionale, dei geosinonimi italiani, in Italiano d'oggi. Lingua nazionale e varietà regionali. In *Atti del Convegno internazionale di studio* (Trieste, 27-29 maggio 1975). Trieste: Lint, S. 109-117.
- De Felice, E. (1984). *Le parole d'oggi: il lessico quotidiano, religioso, intellettuale, politico, economico, scientifico, dell'arte e dei media*. Milano: Mondadori.
- De Mauro, T. (1963). *Storia linguistica dell'Italia unita*. Bari: Laterza.
- Devoto, G. (1939). La norma linguistica nei libri scolastici. In *Lingua nostra*, 1. Frankfurt a. M.: Vittorio Klostermann, S. 57-61.
- Eichhoff, J. (1982). Erhebung von Sprachdaten durch schriftliche Befragung. In W. Besch u.a. (Hrsg.) *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. 1. Halbband. Berlin u.a.: De Gruyter, S. 549-553.
- Elspaß, S., Möller, R. (2006). Internet-Exploration: Zu den Chancen, die eine Online-Erhebung regional gefärbter Alltagssprache bietet. In *Osnabrücker Beiträge zur Sprachtheorie* 71: S. 141-156.
- GRADIT: De Mauro, T. (a cura di) (1999). *Grande Dizionario Italiano dell'uso*. Torino: UTET.
- Grassi, C., Sobrero, A. A. & Telmon, T. (2003), *Introduzione alla dialettologia italiana*. Bari: Laterza.

- Möller, R., Elspaß, S. (2008). Erhebung dialektographischer Daten per Internet: Ein Atlasprojekt zur deutschen Alltagssprache. In S. Elspaß, W. König (a cura di). *Sprachgeographie digital. Die neue Generation der Sprachatlanten (mit 80 Karten)*, Hildesheim u.a.: Olms. S. 115-132.
- Nützel, N. (2009). *Sprache oder Was den Mensch zum Menschen macht*. München: cbt Verlag.
- Panzini, A. (1950). *Dizionario moderno delle parole che non si trovano nei dizionari comuni*. Milano: Hoepli.
- Poggi Salani, T. (2010). Italiano regionale. In *Enciclopedia dell'italiano 2010*. Treccani, [[http://www.treccani.it/enciclopedia/italianoregionale\\_\(Enciclopedia\\_dell'Italiano\)/#](http://www.treccani.it/enciclopedia/italianoregionale_(Enciclopedia_dell'Italiano)/#)] [04/03/2014].
- Rüegg, Robert (1956). *Zur Wortgeographie der italienischen Umgangssprache*. Köln: Kölner Romanistische Arbeiten.
- Sobrero, A. (1996). *Introduzione all'italiano contemporaneo : la variazione e gli usi*. Roma: Laterza.
- Sobrero, A., Miglietta, A. (2006). *Introduzione alla linguistica italiana*. Bari: Laterza.
- Telmon, T. (1990). *Guida allo studio degli italiani regionali*. Alessandria: Edizioni dell'Orso.
- TRECCANI = Treccani (2009). *Il vocabolario della lingua italiana*. Roma: Treccani.
- VIVALDI: Kattenbusch, D. (1999): *VIVAio Acustico delle Lingue e dei Dialetti d'Italia*. [<http://www2.hu-berlin.de/vivaldi>].

# Applying a Word-sense Induction System to the Automatic Extraction of Diverse Dictionary Examples

Paul Cook<sup>1</sup>, Michael Rundell<sup>2</sup>, Jey Han Lau<sup>3</sup>, and Timothy Baldwin<sup>1</sup>

<sup>1</sup>Department of Computing and Information Systems, The University of Melbourne

<sup>2</sup>Lexicography Masterclass and Macmillan Dictionaries

<sup>3</sup>Department of Philosophy, King's College London

paulcook@unimelb.edu.au, michael.rundell@lexmasterclass.com, jeyhan.lau@gmail.com, tb@ldwin.net

## Abstract

There have been many recent efforts to automate or semi-automate parts of the process of compiling a dictionary, including building headword lists and identifying collocations. The result of these efforts has been both to make lexicographers' work more efficient, and to improve dictionaries by introducing more systematicity into the process of their construction. One task that has already been semi-automated is that of finding good dictionary examples, and a system for this, GDEX, is readily available in the Sketch Engine. An ideal system, however, would be able to automatically retrieve candidate examples of a particular sense of a word, which is beyond the current scope of GDEX. In this paper, as a step towards this ambitious goal, we propose and evaluate a method for applying a 'word-sense induction' system to automatically extract examples that exhibit a greater diversity of usages of a target word than can currently be obtained through GDEX. We then discuss the future prospects for systems that are able to automatically select candidate dictionary examples for a particular word sense.

**Keywords:** dictionary examples; word-sense induction; computational lexicography

## 1 Automating Lexicography

A major challenge facing contemporary lexicography – in the commercial sector, at least – is the goal of maximizing the potential of digital media and abundant language data, against a background of limited financial resources. As a response to this challenging climate, there has been significant progress in the automation of some of the tasks involved in compiling dictionaries and lexical databases – an approach which has the potential to deliver reduced development costs together with improved coverage of lexicographically-relevant facts (e.g., Rundell & Kilgarriff 2011, Cook et al. 2013, Kosem, Gantar & Krek 2013).

One component of this project is the automatic retrieval from corpus data of 'good' dictionary examples, whether for presenting editors with a shortlist of appropriate candidates, populating a dictio-

nary database with examples, or providing the end-user with a range of instances of words in context (Kilgarriff et al. 2008; Kosem, Husák & McCarthy 2011). Notwithstanding these advances, there is scope for improvement in two areas. First, example-finding software does not yet routinely achieve the contextual diversity that characterizes example-sets selected by skilled lexicographers. Secondly, it does not attempt the difficult but critical task of mapping corpus instances onto dictionary senses. Some current dictionaries provide a range of (automatically-retrieved) examples to complement the manually-selected ones in the dictionary. This approach can be found, for example, in the 5th edition of the *Longman Dictionary of Contemporary English (LDOCE)* (<http://ldoce.longmandictionariesonline.com/dict/SearchEntry.html>), where users can opt to see up to ten ‘Examples from the corpus’, and in *Wordnik* ([www.wordnik.com](http://www.wordnik.com)), where numerous authentic examples are shown on the right-hand side of the screen. Google Translate (<http://translate.google.com>) has a somewhat similar feature. But in none of these cases are the examples attached to word senses: in *LDOCE*, the entry for *pool* (noun) includes a random assortment of examples, including references to swimming pools, pools of investment funds, and even football pools. Ideally, we need a system which automatically selects optimally-diverse examples for a polysemous word (so that users are offered examples exhibiting the full contextual range of the word’s behaviour) and matches the examples to the individual dictionary senses whose meaning they instantiate.

In this paper we report an experiment in which a ‘word sense induction’ methodology is applied to extracting corpus examples in a way that fulfills the first of these goals – identifying examples showing the diversity of contexts in which a word is used. We conclude by discussing the prospects for using the output of the word-sense induction system to map the ‘induced senses’ the system discovers in the corpus to dictionary senses.

## 2 Diversifying Example Sentences with Word-sense Induction

In this section we describe the GDEX method for identifying good dictionary examples and a recently-presented word-sense induction (WSI) system, and then propose a method to combine these two technologies to automatically select more-diverse dictionary examples.

### 2.1 GDEX

GDEX (Kilgarriff et al. 2008) is a system for automatically selecting good dictionary examples from a corpus. Sentences containing a given target word are scored based on a number of heuristics about what makes a sentence a good dictionary example, such as sentence length, the position of the target word in the sentence, and the other words occurring in the sentence. Given a query for a particular word, GDEX then returns the top-scoring sentences in a corpus, which can be manually examined for selection as examples. GDEX has become a standard lexicographical tool, and is available for use with

many corpora in the Sketch Engine (SkE, <http://www.sketchengine.co.uk/>, Kilgarriff & Tugwell 2002). Adaptations of GDEX (Kosem, Gantar & Krek 2011) have incorporated simple notions of diversity to avoid selecting duplicate or very similar sentences. We propose a more sophisticated notion of diversity targeted at selecting a set of candidate example sentences exhibiting a wider range of senses of the target word.

## 2.2 Word-sense Induction

WSI is ‘the task of automatically grouping the usages of a given word in a corpus according to sense, such that all usages exhibiting a particular sense are in the same group’ (Cook et al. 2013). Crucially this grouping is done without reference to a pre-existing sense inventory. WSI is the automatic counterpart to the manual lexicographic process of word sense disambiguation (WSD, Atkins & Rundell 2008: 269).<sup>1</sup> Although the notion of word sense is of course controversial, it is nevertheless standard for dictionaries to carve up the meanings of a word into senses, even though dictionaries will vary in terms of the sense distinctions made for a given polysemous word.

Topic modeling (Blei, Ng & Jordan 2003) is a computational technique for automatically discovering latent structure in a corpus that has recently been successfully applied to a wide range of NLP tasks. A typical topic model automatically ‘learns’ the topics in a corpus, and the mixture of topics in each document in the corpus. Each topic is represented as a probability distribution over words; each document is represented as a probability distribution over topics.

Lau et al. (2012) present a WSI system based on topic modeling. Rather than building a topic model for an entire corpus, they build a separate model for each target word. In this model the “documents” are short contexts – typically 3 sentences – containing a usage of the target word. There is not necessarily a correspondence between the topics in a topic model and topics in the sense of the subject of a text (although a topic model for a corpus will often learn topics that do indeed correspond to this more common usage of *topic*). In Lau et al.’s WSI methodology, the topics in the topic model are interpreted as word senses.

In traditional topic models (e.g., latent Dirichlet allocation, Blei, Ng & Jordan 2003) the number of topics to be learned must be specified manually in advance. For WSI this would mean that the number of senses for each word would need to be set by hand. Words of course differ with respect to their polysemy, and an appropriate number of senses could only be determined based on corpus analysis. Lau et al. therefore use hierarchical Dirichlet process (Teh et al. 2006), a type of topic model that also automatically learns the appropriate number of topics for a given document collection.

---

1 It is worth clarifying a terminological difference between the lexicographic and natural language processing (NLP) communities. In NLP, WSD refers specifically to the task of selecting the most appropriate sense, from a given sense inventory, for a given instance of a word in context. In lexicography WSD typically refers to identifying the various senses of a word, i.e., constructing a sense inventory, based on corpus evidence.

In Lau et al.’s model, each ‘document’ – typically consisting of a sentence including an instance of the target word, and one sentence of context before and after – is represented as a bag-of-words, i.e., word order is ignored, but the frequency of words in the context is maintained. Because the immediate context surrounding a word can be highly informative of that word’s sense, positional word features that encode the specific three words occurring to the left and right of the target word are also included. In this document representation stopwords are ignored, and all other words are lemmatized. An example of the representation used is given in Table 1. For reasons of brevity the ‘document’ in this example is a single sentence (whereas it would typically be three sentences). An example of the senses induced by the model for the lemma *box* (n) is given in Table 2. Recall that each sense is a probability distribution over words; here the top-10 most likely terms for each sense are shown. Examining these terms allows us to roughly interpret the senses the system has induced. For example, senses 1 and 2 seem to correspond to usages of *box* in the context of sports and elections, respectively. Random samples of 5 corpus instances corresponding to induced senses 1 and 4 for *box* (n) are given in Table 3. For sense 1, all of the usages seem to correspond to an area of a sports ground, although the first four relate to soccer, but the last relates to tennis. For sense 4, the model has identified usages related to the entertainment industry, but includes a mixture of usages of the expressions *box office* and *box set*, as well as other usages such as the final example.

Target Lemma	<i>box</i> (n)
‘Document’	The Flames had a two-man advantage near the end of the second period when Lacroix and McSorley were in the penalty <b>box</b> for kneeling and unsportsmanlike conduct, respectively.
Bag-of-words Features	flame, two-man, advantage, near, end, second, period, lacroix, mcsorley, penalty, knee, unsportsmanlike, conduct, respectively
Positional Word Features	lacroix_#-3, mcsorley_#-2, penalty_#-1, knee_#+1, unsportsmanlike_#+2, conduct_#+3

**Table 1: An example of the topic model features.**

Sense Number	Top-10 Terms
1	box minute @card@ game ball goal score shot penalty play
2	box @card@ ballot ballot_#-1 police official election vote find party
3	box @card@ company computer digital cable converter black_#-1 black converter_#-1
4	office box office_#+1 million @card@ film movie weekend dollar ticket
5	box @card@ n’t look find small think put room store
6	box cereal recipe cut outlet post fat gram vip wire
7	shanker teller lionel penn crush ferry jesus julie maine marshal

**Table 2: The top-10 terms for each of the senses induced for the lemma *box* (n).**

Sense Number	Usage
1	<p>Bolivia added an extra insurance goal in the 80<sup>th</sup> when Ronaldo Garcia sent a long blast from outside the <b>box</b> into the upper corner.</p> <p>Arsenal were left nursing a justifiable grievance over the referee’s failure to award a second-half penalty when Kuyt unbalanced Alexander Hleb with a tug from behind as the midfielder wriggled into space deep in the <b>box</b>.</p> <p>Thiago made it 3-0 in the 79<sup>th</sup> minute with a powerful left foot drive from the edge of the <b>box</b>.</p> <p>Masami Ihari headed away a corner from Juninho but only as far as Zinho, who let fly with a spectacular volley from the edge of the <b>box</b> that gave no chance at all.</p> <p>He’s placing it in the <b>box</b> beautifully, hitting closer to the line with lots of aces.</p>
4	<p>And while Joel is paying himself a backhanded compliment, especially in the context of B-sides, live tracks and rarities <b>box</b> set, it got me thinking.</p> <p>It’s the harsh command of the <b>box</b> office, demanding a big seller and the heck with all else.</p> <p>“For the most part, it’ll help us,” says Bobbie Welch, <b>box</b> office manager for New Mexico State University, which has hosted ZZ Top, Guns N’ Roses and Paul McCartney.</p> <p>For those seeking more kid-friendly fare, the “Spotlight Collection” discs – including a sixth released Tuesday (\$27) – offer more than 30 family-appropriate installments from the Golden <b>box</b> sets.</p> <p>Only 17 percent of reviews were positive, according to RottenTomatoes.com, but 82 percent of audience survey respondents checked off the “excellent” or “very good” <b>boxes</b>, according to Sony.</p>

**Table 3: Usages corresponding to induced senses 1 and 4 of the lemma *box* (n).**

Lau, Cook & Baldwin (2013a,b) recently showed this WSI methodology to be the overall best performing system on two recent SemEval WSI shared tasks (Jurgens & Klapaftis 2013; Navigli & Vannella 2013). Cook et al. (2013) demonstrated that this system can be applied as a lexicographical tool for finding new word-senses. We therefore adopt this WSI system here for diversifying automatically-selected dictionary examples.

### 2.3 Diversification

For a given target lemma, we obtain the top-100 GDEX examples for a corpus from SkE. We further obtain a random sample of up to 50k usages of the target from the same corpus. In each case we extract

the sentence containing the usage of the target, and one sentence of context on either side. Following Lau et al. (2012) we remove stopwords and lemmatize the tokens in the context. We then run the WSI system on these usages of the target lemma. The WSI system outputs a label indicating the induced sense number of each target instance. These induced senses correspond to groups of usages that exhibit the same sense, according to the WSI system, not dictionary senses.

We then use these induced sense labels to diversify the top-100 GDEX examples. To do so, we repeatedly iterate through the top-100 GDEX sentences (i.e., we consider each sentence in turn, one by one). For each pass over the sentences, we select the best GDEX sentence (according to GDEX's ranking) for each induced sense, which has not been selected in a previous pass. We repeat this until all sentences have been selected. In the subsequent analysis we compare the top-5 GDEX examples to the top-5 examples produced by this diversification procedure.

### 3 Analysis

For this preliminary analysis we selected 98 target lemmas to analyse: 54 from a recent SemEval WSI task (Jurgens & Klapaftis 2013) and 44 additional medium-polysemy lemmas. We extracted GDEX sentences, and the additional randomly-selected usages, from the ukWaC (Ferraresi et al. 2008).

We ran the GDEX diversification procedure described in the previous section for each target lemma. The top-5 GDEX sentences for the target lemma exhibited varying numbers of induced senses (as determined by the WSI software). However, if the top-5 GDEX sentences already exhibit many (e.g., 4 or 5) induced senses, then our diversification procedure has little or no impact. We therefore focused our analysis on lemmas where the top-5 GDEX usages exhibited less diversity. Crucially such cases can easily be automatically identified. Here we discuss the findings for twelve lemmas whose top-5 GDEX sentences were the least diverse, exhibiting just two induced senses of the target lemma in each case. (In no case did the top-5 GDEX usages exhibit just one induced sense.)

For each lemma, two sets of five example sentences were prepared: (1) the top-5 GDEX sentences, and (2) the top-5 sentences from our diversification procedure. These sets of sentences were presented to a professional lexicographer (the second author of this paper) who was asked to judge which set of examples was better. Crucially the lexicographer did not know which method the sets of examples corresponded to. For eight lemmas the examples produced through our new diversification procedure were selected as better; in the remaining four cases the default GDEX examples were chosen. To give an idea of the potential of our method, we discuss the output of the two systems for two of the target lemmas which were analysed: *exploitation* and *bitter*.



Top-5 sentences from GDEX:

- (1) It must be a world where humanity has been liberated from social distress, brutal **exploitation** and war.
- (2) A new minister replaces the old one, the daily grind of **exploitation** resumes.
- (3) And Engage, wittingly or not, is aiding it in this **exploitation**.
- (4) It's not trade we're against, it's **exploitation** and unchecked power.
- (5) Others have seen Napster as little more than payback for decades of record company **exploitati-**  
**on** of artists.

Top-5 sentences from our new diversification approach:

- (6) It must be a world where humanity has been liberated from social distress, brutal **exploitation** and war.
- (7) Others have seen Napster as little more than payback for decades of record company **exploitati-**  
**on** of artists.
- (8) Delivery will focus upon the development and **exploitation** of repertoire through the workplace.
- (9) Requirement 24 - Identify all auditable events that may be used in exploitation of known covert storage channels.
- (10) Workers can refer young people they think are vulnerable or at risk of homelessness and sexual **exploitation**.

The examples that are shared by the two sets are both good. However, the second and third sentences from default GDEX are weak because they offer little context for interpretation and include anaphora. Although the final sentence for default GDEX (sentence 4) is acceptable, the diversified sentences provide a better snapshot of the word overall, in that they cover the more neutral sense of exploitation (sentence 4) and include an example of *sexual exploitation* (sentence 5).

Top-5 sentences from GDEX:

- (1) If cows eat too many carrots, their milk tastes **bitter**.
- (2) It's so easy for us to be **bitter**.
- (3) Men obeyed their base immediate motives until the world grew unendurably **bitter**.
- (4) In his destroyed Urim, its lament is **bitter**.
- (5) No; and henceforth I can never trust his word. **Bitter**, bitter confession!

Top-5 sentences from our new diversification approach:

- (1) If cows eat too many carrots, their milk tastes **bitter**.
- (2) It's so easy for us to be **bitter**.
- (3) In his destroyed Urim, its lament is **bitter**.
- (4) The rivalry between Soka Gakkei and Aum Shinrikyo was **bitter**.
- (5) Joe Frazier never felt more **bitter** about defeat, and continues even today to hate his great rival.

In this case, three of the examples appear in both sets, and their quality varies: the first in each set is good; the second acceptable, if a little short on context; the third (*In his destroyed Urim...*) is weak, and would certainly not make an appropriate example sentence for a pedagogical dictionary (or any other dictionary, for that matter). But if we compare the two examples unique to each set, those in the second set (sentences 4 and 5) are clearly more suitable as dictionary examples than those in the first (3 and 5). *Rivalry* has a high saliency score as a collocate of *bitter* (though *disappointment* would have been even better, assuming the goal is maximum typicality). And the addition of a collocating preposition in the last example (with *about*, the most frequent preposition appearing with *bitter*) provides additional diversity.

## 4 Discussion

Software for selecting dictionary examples could be improved if it were optimised to select diverse examples for a polysemous word, such that the examples show the full range of usage for that word. In this paper we have proposed a novel method for automatically selecting a more diverse set of dictionary examples from a corpus than can currently be obtained using GDEX. We carried out a small-scale preliminary evaluation of this method, and found that – in terms of diversity – our approach outperformed GDEX for eight out of twelve lemmas analyzed. The results are encouraging rather than conclusive. But with further improvements based on what we have learned through this experiment, this new method could be applied to real lexicographic tasks – either for providing editors with candidate lists from which to select examples for a dictionary, or for automatically providing dictionary users with additional examples. In either case, the outcome should be a set of examples exhibiting a more diverse range of usages than current software tools usually supply.

Systems for selecting examples would be further improved if they were able to match automatically-identified examples to corresponding dictionary senses. Lau et al. (2014) recently proposed a method to link the senses induced by the same WSI system used here to senses in a dictionary. They evaluated this method in the context of identifying the relative frequencies of the senses of a given word in a corpus, and showed it to perform comparably to previously-proposed approaches for this task (McCarthy et al. 2007).<sup>2</sup> In future work, we intend to combine this method for linking induced senses to dictionary senses with the approach described in this paper, in order to identify good examples at the level of *word senses*, as opposed to *lemmas*. WSI remains a very difficult task for current natural language processing technologies. Crucially, in the context of identifying sense-specific dictionary examples, it might not be necessary to correctly identify the dictionary sense of every corpus instance of a target word. Instead, it might suffice to identify the dictionary sense corresponding to

---

2 Very interestingly, but of less relevance to the present paper, they also showed that this method has the potential to identify dictionary senses that are unattested in a corpus, and senses that are induced by the WSI system but not listed in a dictionary.

those instances where the system is highly confident of its prediction, and to then apply GDEX to select good dictionary examples amongst those instances. We are therefore optimistic about the future possibility of automatically adding sense specific examples to dictionaries, although much work remains to be done.

The WSI system of Lau et al. (2012) lies at the core of the method presented in this paper. To encourage further research on WSI and its applications, Lau, Cook & Baldwin (2013a,b) made this system publicly available under a license which permits its use for commercial purposes (<https://github.com/jhlau/hdp-wsi>). We hope that others will make use of this software to consider further applications of topic modeling and WSI in computational lexicography.

## 5 References

- Atkins, B. T. S. and Rundell, R. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Cook, P., Lau, J. H., Rundell, M., McCarthy, D. & Baldwin, T. (2013). ‘A lexicographic appraisal of an automatic approach for detecting new word-senses’, in Kosem et al. 2013: 49-65.
- Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, pages 47-54, Marrakech, Morocco.
- Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290-299, Atlanta, USA.
- Kilgarriff, A. and Tugwell, D. (2002). ‘Sketching words’. In Corréard, M.-H., editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, pages 125-137. Euralex, Grenoble, France.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008) ‘GDEX: Automatically Finding Good Dictionary Examples in a Corpus’, in Bernal, E. and DeCesaris, J. (Eds) *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra: 425-433.
- Kosem, I., Gantar, P., & Krek, S. (2013). ‘Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing’, in Kosem et al. 2013: 32-48.
- Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) (2013). *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.
- Kosem, I., Husák, M., & McCarthy, D. (2011). ‘GDEX for Slovene’. In I. Kosem, K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New applications for new users, Proceedings of eLex 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies: 151-159.
- Lau, J. H., Cook, P., and Baldwin, T. (2013a). unimelb: Topic modelling-based word sense induction. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 307-311, Atlanta, USA.
- Lau, J. H., Cook, P., and Baldwin, T. (2013b). unimelb: Topic modelling-based word sense induction for web snippet clustering. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2:*

- Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 217–221, Atlanta, USA.
- Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 591–601, Avignon, France.
- Lau, J. H., Cook, P., McCarthy, D., Gella, S., and Baldwin, T. (2014). Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA.
- McCarthy, D., Koeling, R., Weeds, J., and Carroll J., (2007). Unsupervised Acquisition of Predominant Word Senses. *Computational Linguistics*, 33(4):553–590.
- Navigli, R. and Vannella, D. (2013). Semeval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, USA.
- Rundell, M. and Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In Meunier F., De Cock S., Gilquin G. and Paquot M. (Eds), *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Benjamins: 257-281.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

### **Acknowledgements**

This work was supported in part by the Australian Research Council.

# Cross-linking Austrian dialectal Dictionaries through formalized Meanings

Thierry Declerck, Eveline Wandl-Vogt  
DFKI GmbH, Language Technology Lab, Germany;  
Austrian Academy of Sciences, ICLTT, Austria  
declerck@dfki.de, eveline.wandl-vogt@oeaw.ac.at

## Abstract

This paper deals with the formalization of aspects of definitions used in dialectal dictionaries. We focus on the way “meanings” are encoded in such dictionaries, an essential position for many users and lexicographers, and describe how such an encoding can be re-used for cross-linking entries of on-line (dialectal) dictionaries. In this contribution we describe in some details experiments we made in this respect with two Austrian dialectal dictionaries: The dictionary of Bavarian dialects of Austria (“Wörterbuch der bairischen Mundarten in Österreich”, WBÖ) and the dictionary of the Viennese dialect (“Wörterbuch der Wiener Mundart”, WWM), which we ported into the SKOS and *lemon* models in order to publish them in the Linguistic Linked Open Data cloud. We show how this approach is not only appropriate for supporting the automation of the cross-linking of dialectal dictionaries, but also for linking entries of the dialectal dictionaries to other types of lexical and encyclopaedic resources in the web

**Keywords:** Dialectal lexicography; Semantic Web; Linguistic Linked Open Data; Austrian dialects

## 1 Introduction

In the context of recent work dedicated to porting the dictionary of Bavarian dialects of Austria (“Wörterbuch der bairischen Mundarten in Österreich”, WBÖ)<sup>1</sup> and the dictionary of the Viennese dialect (“Wörterbuch der Wiener Mundart”, WWM)<sup>2</sup> onto representation formats supporting their publication in the Linked Open Data (LOD) framework<sup>3</sup>, and more specifically in the Linguistic Linked Open Data cloud<sup>4</sup>, we got our attention directed to the investigation on how this approach could support an automation of the cross-linking of such dialectal language resources. For this, we focused on the way “meanings” are encoded in the selected dictionaries, an essential position for many users and

---

1 See <http://www.oeaw.ac.at/dinamlex/WBOE.html> and (Wandl-Vogt 2005; Wandl-Vogt 2008). See also (Declerck & Wandl-Vogt 2013) for a description of the approach adopted for porting WBÖ to the SKOS representation language.

2 See (Hornung & Grüner 2002).

3 See <http://linkeddata.org/>

4 See <http://linguistics.okfn.org/resources/lod/>

lexicographers. We take advantage here of a property of dialectal dictionaries concerning the expression of meanings of entries: Although conceived as monolingual reference work, dialectal dictionaries share with bilingual dictionaries the fact that they express the meanings of their entries in a different language. The meta-language for expressing the meanings of entries in both WBÖ and WWM is Standard German and Austrian German, as can be seen for example in the WBÖ entry “Puss”: one of its meanings is expressed by the Standard German word “Kuß” (*kiss*) and by the Austrian German word “Busserl” (WBÖ: 1,1515)<sup>5</sup>, as can be seen in Figure 1 below. WWM uses also the Standard German word “Kuss” for expressing a meaning of the entry “Bussal”, as can be seen in Figure 2. Our assumption is thus that linking entries in distinct dialectal dictionaries can be implemented on the base of meanings that are expressed by similar means across the dictionaries.

In this paper we first briefly describe the two dialectal dictionaries we have been considering for our experiments. We then depict the processes we applied to the entries of the dictionaries for extracting and analyzing the expressions that express their meanings, and their encoding in the representation languages RDF<sup>6</sup>, SKOS-XL<sup>7</sup> and *lemon*<sup>8</sup> for supporting their publication in the LOD.

The ultimate goal of our work is not only to be able to cross-link the lexical resources described in this paper, but also to link them in the Linked Data cloud with available data sets for highly-resourced languages and to elevate this way our dialectal and historical lexical resources to the same “digital dignity” as the mainstream languages have already gained.

We show also how this encoding allows enriching our lexical data with additional lexical information, mainly senses and multilingual variants.

## 2 The selected dialectal Dictionaries WBÖ and WWM

We describe in this section briefly the main characteristics of the two dialectal dictionaries we selected for conducting our experiments on cross-linking.

### 2.1 The Architecture of the selected dialectal Dictionaries in a Nutshell

The chosen dialectal dictionaries, WBÖ as well as WWM, are scientific dictionaries. Each dictionary offers references and example sentences for illustrating contexts of use for the entries. Whereas the documentation and interpretation in WBÖ is exhaustive, WWM is much shorter and comprehensive. WBÖ is a dialectal dictionary for all Bavarian dialects in the former Austrian Hungarian Monarchy (status: about 1915), whereas WWM is a dictionary for the dialect of the city and county of Vienna.

---

5 See ÖWB 141: Bussel das, -s/-[n] (ugs.): Busserl; [ugs.] = colloquial language

6 See <http://www.w3.org/RDF/>

7 See <http://www.w3.org/TR/skos-reference/skos-xl.html>

8 See (McCrae & al., 2012).

- grammar: Every entry informs about grammatical properties of the word
- etymology: Every entry contains information about the etymology of the word.
- definition(s): Definitions are a central position in both (onomasiological) dictionaries. Complementing the definitions, WBÖ presents a lot of examples of spoken and written dialect, phrases, songs and poems. Due to the fact, that approximately 10% of the material consists of excerpts of written texts and that the main aim in the beginning was to document the development of a word from its beginning to the actual dialect (see Arbeitsplan 1912) the emphasis on written texts is very high. Furthermore, WBÖ definitions often include a lot of encyclopaedic information about rural traditions and traditional customs.

WWM presents this type of semantic information in a much more concise way.

- meanings, as a core part of definitions: Although conceived as monolingual reference work, many dialectal dictionaries share with bilingual dictionaries the fact that they express the meanings of their entries in a different language. The meta-language for expressing the meanings of entries in both WBÖ and WWM is Standard German, sometimes accompanied by Austrian German. So for example the meaning of the WBÖ entry “Puss” is expressed by the Standard German word “Kuß” (*kiss*) and by the Austrian German word “Busserl” (WBÖ 1,1515)<sup>9</sup>, as can be seen also in Figure 1. As the reader will see later, we take advantage of this property of dialectal dictionaries concerning the expression of meanings of entries: Linking entries in distinct dialectal dictionaries can be implemented on the base of the Standard German expressions of meanings that are shared across those dictionaries.
- references to dictionaries of adjacent German dialects: WBÖ puts the information presented into a whole dialectal language area in quoting the neighbouring German dialectal dictionaries. WMM does not include this position due to not being embedded into the same / similar methodological background.
- phonetics: Phonetics played an important role in the so called “Junggrammatische Schule”<sup>10</sup>, which is the methodological background for the WBÖ. WWM offers headwords that are transliteration based on phonetics.
- compounds: Compounds are treated within the base word entry, e.g. “Nuss)puss” (the famous creamy hazelnut truffle “Nussbusserl”). They might be dealt with in the position ‘Komp.’ (*Compounds*) within the main entry (as this is the case for the entries *Puss* (WBÖ) or *Bussal* (WWM)).
- cross references to derivations and related words: There is a position, where cross references to derivations and related words are stored, e.g. *Syn.* → (*Fotzen*)*pemperer* (WBÖ); (*Syn.*: *Schmåtss*

---

9 See ÖWB 141: Bussel das, -s/-[n] (ugs.): Busserl; [ugs.] = colloquial language.

10 See <http://en.wikipedia.org/wiki/Neogrammarian> for more details.

(WWM). Including articles of derivations or related words, especially those with “less information value” due to a rationalisation concept for WBÖ (see Straffungskonzept 1998: §§ 1.2.1-1.2.3).

- editor: Finally, each WBÖ entry closes with the initial of the author, e.g. *W.B.* = *Werner Bauer*; in the WWM similar signing is not existing.

In the partial entry for the word “Puss” in WBÖ shown in Figure 1, the reader can see in the right column the details for two selected meanings expressed in Standard German: 1) “Kuss” (*Kiss*) and 2) “Kl. süßes Gebäck” (*small sweet pastry*).

<p><b>Puss, Puss(e)lein</b>  M. (jedoch meist neutr.Dem.), Kuß („Busserl“), Gebäck, PflN s-,mbair. m. SI, Egerl. nur als → (<i>Zwick[er]</i>)-, Simmersdf. Igl.; Schallw., vgl. KLUGE<sup>20</sup> 114; frühhd. <i>buß</i> M. Kuß GÖTZE Frühhd.Gl. 44; s.a. KRANZMAYER Kennw. 10; entl. ins Magy. als <i>puszi</i> Kuß u. <i>puszedli</i> Gebäck KOBILAROV-GÖTZE 355f., ins Slow. als <i>púšek</i> Kuß PLETERŠNIK 2,366 u. ins Kä.Slow. als <i>pushei</i> Kuß GUTSMANN Dt.-Wind.Wb. 261. — Bayer.Wb. 1,295, Schwäb. Wb. 1,1558.</p>	<p>Bed.: 1. Kuß im gesamten Verbr.Geb. (meist als 1. od. 2. Dem.), Syn. → (<i>Fotz</i>)<i>pemperer</i>,  2. Kl. süßes Gebäck m. flacher kreisförmiger Unterseite u. gewölbter Oberseite ugs. (meist 2., seltener 1. Dem.), s.a. EBNER<sup>2</sup> 51; rundes Nußgebäck auf Kirchtagen Gott.Wb. 1,91 (2.Dem.);</p>
---	---

Figure 1: WBÖ 3,1515f.: Puss, Puss(e)lein.

In the second example, taken from the WWM and displayed in Figure 2, the reader can see that very similar meanings are provided for the entry “Bussal, Bussi, Bussl”. While the first meaning is expressed by using exactly the same Standard German word “Kuss” in both cases, the second meaning (*small sweet pastry*) is expressed in each dictionary by using variants: “Kl. süßes Gebäck” vs. “kleines Süßgebäck”.

<p><b>Bussal, Bussi, Bussl</b>, <i>das</i>, 1) <i>Kuss</i> (Syn.: <i>Schmâtss</i>); 2) <i>kleines Süßgebäck</i>;  Pl. <i>Bussaln</i>; viele Komp. wie <i>Nussbussal</i> usw. –  Etym.: bair.-österreich. Schallwort <i>Puss</i> <i>Kuss</i>.</p>
--

Figure 2: WWM 199: Bussal, Bussi, Bussl.



## 2.2 Access Structure

The main access structure, for both WBÖ and WWM, is the macrostructure, namely the headword.<sup>11</sup>

- WBÖ has chosen due to etymologic-historic considerations a sophisticated, artificial headword, which is difficult to be used as access structure by scientists and in particular causes problems for laypersons. As an example, the German headword “deutsch” (*german*) is represented in WBÖ as “te-ütsch”; the Standard German headword “Pflaumenbaum” and the Standard Austrian word “Zwetschkenbaum” (*plum tree*) are represented in the WBÖ as “Zwëtschken)päum”. And a subentry of the main entry “Päum” (*tree*), the WBÖ headword “Busserl”, lacks a standard German representation.
- WWM chooses a transliterated headword, based on phonetics.

So that a cross-referencing and interlinking of dialectal dictionaries, even within the same language area (here: Bavarian variants), does not work without the development of mapping rules. Furthermore, such a mapping would offer just a flat and non-hierarchical interlinking.

This situation motivated the main approach of the work presented here, which consists in investigating if individual word senses, the meanings of entries expressed in Standard and Austrian German words, can serve as an access point for the cross-linking of our two selected dialectal dictionaries, as well as reference point for linking to lexical resources and other knowledge sources in the Web. The following section describes the processes we applied to the entries of the dialectal dictionaries for extracting and analyzing the expressions that express their meanings

## 3 Extraction and Linguistic Analysis of Expressions introducing the Meanings of Entries

Our first task consisted in detecting and extracting automatically from both dictionaries the strings expressing the core meanings for each entry. Fortunately both dictionaries have been made available to us in an electronic version: WBÖ in a proprietary XML schema and WWM in Microsoft Word. We used the TEI “OxGarage”<sup>12</sup> service to convert the WWM Word document into a TEI compliant XML representation. In both XML representations it was straightforward to describe in Perl scripts the patterns for extracting the meanings of the entries expressed in Standard or Austrian German.

But as mentioned at the end of section 2.1, there are discrepancies in the use of Standard or Austrian German word forms across the dictionaries, so that it is often not possible to establish a relation bet-

---

11 Other important positions, navigation and access structures within (dialectal) dictionaries are – or could be – space and time. Geo-referencing with time-stamp stored within a GIS offers possibilities for spatio-temporal visualisation as well as analysis (Wandl-Vogt 2010) and exploratory visually conducted analysis (Theron & Wandl-Vogt 2014).

12 See <http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient/>

ween words expressing meanings across the dialectal dictionaries. Since pure string matching cannot provide accurate comparisons between those expressions, there is the need to apply basic natural language processing to the expressions and to reduce those to their lemmatized form and to mark them up with part-of-speech and morphological information. The comparison of expressions standing for meanings in both dictionaries is then made on the base of such linguistic information associated to the strings. We provided an automatic linguistic analysis of those extracted meanings, using lexical and syntactic analysis grammars written within the SCHUG tools (Declerck 2002). This included tokenization, lemmatisation, Part-of-Speech (POS) tagging and constituency as well as dependency analysis. The strings marking in both dictionaries the “small sweet pastry” meaning are enriched with the following linguistic features:

- (1) WBÖ: (NP süßes (ADJ, lemma = süß, MOD) Gebäck (N, lemma = Gebäck, HEAD)) - *sweet pastry*
- (2) WWM: (NP (kleines (ADJ, lemma = klein, MOD) Süßgebäck (N, compound: süß (ADJ, lemma = süß, MOD) + Gebäck (N, lemma = Gebäck, HEAD)), HEAD)) - *small sweet pastry*

In the examples (1) and (2), we can see the distinct serializations of similar concepts in German. The second example uses a compound noun (“Süßgebäck”), which has the same meaning as the simple nominal phrase in the first example (“süßes Gebäck”). In order to automatically establish this similarity, it is necessary to first perform a morphological decomposition of the head noun in the second example. And we need the lemma of the adjective in the first example, in order to be compared with the first element of the compound noun in the second example.

The fact, that both linguistically analyzed strings expressing the meanings share the same lemmas for adjectival modifiers and head nouns is the base for cross-linking the entries. As we want to formally express this relation, we need to use an appropriate representation language, opting here for Semantic Web standards (e.g. compatible to RDF), also in order to be able to publish our data in the Linked Data cloud.

## 4 Porting the dictionary data into the Linked Open Data framework

To mark linguistically analyzed meanings as related, it is requested to use semantic web representation languages, like those developed in the context of W3C<sup>13</sup> standardization activities: RDF<sup>14</sup>, SKOS, SKOS-XL<sup>15</sup>, *lemon*<sup>16</sup>. With this step we want to benefit from the inherent linking (and merging) possibilities offered by Semantic Web languages used in the Linked Data framework, and more specifically

---

13 See <http://www.w3.org/>.

14 See [http://de.wikipedia.org/wiki/Resource\\_Description\\_Framework](http://de.wikipedia.org/wiki/Resource_Description_Framework).

15 See <http://www.w3.org/2004/02/skos/> and <http://www.w3.org/TR/skos-reference/skos-xl.html> respectively.

16 See (McCrae & al., 2012).

we aim at contributing to the emerging Linguistic Linked Open Data cloud<sup>17</sup>, integrating dialectal language data into this framework.

## 4.1 Porting the dictionaries into SKOS

Based on the Resource Description Framework (RDF), SKOS (Simple Knowledge Organization System): provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary.<sup>18</sup>

Our experiment with SKOS is thus kind of novel, since we apply it to dictionaries, although one can for sure consider dictionaries as being very close to thesauri and in our approach we encode elements of entries (basically the meanings) of the dictionaries as concepts being part of a conceptual scheme. We chose this representation language, since:

- SKOS concepts can be “semantically related to each other in informal hierarchies and association networks”<sup>19</sup>
- “the SKOS vocabulary itself can be extended to suit the needs of particular communities of practice”<sup>20</sup>
- SKOS “can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools.”<sup>21</sup>

With the use of SKOS (and RDF), we are also in the position to make our dictionary resources compatible with other language resource available in the LOD cloud. Examples of such resources are the actual DBpedia instantiation of Wiktionary<sup>22</sup> or version 2.0 of BabelNet<sup>23</sup>.

We decided in the most recent version of our model to encode the strings standing for introducing each entry of a dictionary as a `skos:Concept` being a member of a `skos:Collection`, while each associated sense is encoded as `skos:Concept` that is part of a `concept:Scheme`. In the first case we deal with a flat list of elements, while in the second case we can model (hierarchical) relations between the meanings (also called “senses”). In the following pages of this section, we present some examples of our model applied to WBÖ, using the so-called turtle serialization.

---

17 See <http://linguistics.okfn.org/resources/lod/>

18 <http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/>

19 Ibid.

20 Ibid.

21 Ibid.

22 See <http://dbpedia.org/Wiktionary>. There, *lemon* is also used for the description of certain lexical properties.

23 <http://babelnet.org/>

(3) icltt:Dictionary

```
    rdf:type owl:Class ;
    rdfs:comment "Modeling the ICLTT dictionaries"@en ;
    rdfs:label „Wörterbuch“@de , „Dictionary“@en ;
    rdfs:subClassOf owl:Thing .
```

We first introduce (ex:3) a “Dictionary” class, of which the WBÖ dictionary is an instance of (ex:4).

(4) icltt:wboe

```
    rdf:type icltt:Dictionary , skos:Collection ;
    rdfs:comment "OEAW Dictionary for Bavarian"@en ;
    rdfs:label „Wörterbuch der bairischen Mundarten in Österreich“@de , „Bavarian
    dialects of Austria“@en ;
    icltt:hasLanguage icltt:bar ;
    skos:member icltt:concept_puss .
```

As the dialectal dictionaries are encoded as skos:Collection, entries of the dictionaries are modelled as being a skos:member of such collections. The WBÖ entry “Puss” (encoded in ex:4 as the icltt:concept\_puss) is therefore listed as a member (for reason of space we display here only one member of the collection). The icltt:concept\_puss is introduced in our model as an instance of the class icltt:Entry (ex:5):

(5) icltt:concept\_puss

```
    rdf:type icltt:Entry ;
    rdfs:label “puss”^^xsd:string .
```

(6) icltt:Entry

```
    rdf:type owl:Class ;
    rdfs:label “Entry”^^xsd:string ;
    rdfs:subClassOf skos:Concept .
```

In doing this we have introduced entries of the WBÖ as a concept being a member of a collection. We still need to introduce in our model the concrete information about the entries, and we describe this step in the next section.

## 4.2 Representing the Headwords and the Meanings in SKOS-XL and lemon

Contrary to most knowledge objects described in the LOD, we do not consider strings (encoding in WBÖ lemma and word forms as part of the language) as being just literals, but as being also knowledge objects. We considered therefore the use of SKOS-XL and of the *lemon* model<sup>24</sup> for representing the string used for headwords and senses. SKOS-XL has proven to be adequate for encoding strings as complex knowledge objects, but not for representing the linguistically analyzed expressions used for marking the meanings. For this we opted for the *lemon* model, which is compatible to SKOS.

(7) `icltt:concept_puss`

```

  rdf:type icltt:Entry ;
  rdfs:label "puss"^^xsd:string ;
  skosxl:prefLabel icltt:entry_puss .

```

In the displayed code in ex:7, the reader can see now the complete representation of the “`icltt:concept_puss`” object (extending ex:5): the concrete headword in the dictionary is pointed to by the means of the `skos-xl` property “`prefLabel`”, which contrary to both the `rdfs:label` and `skos:label` properties is not having a literal as the possible value of the range of the property, but which is having an object (“`icltt:entry_puss`”, as shown in ex:7) as a value in its range:

(8) `icltt:entry_puss`

```

  rdf:type icltt:Lemma ;
  icltt:hasPos icltt:noun ;
  lemon:sense icltt:gebäck , icltt:kuss , icltt:süßes_gebäck ;
  skosxl:literalForm "Puss"@bar .

```

As the reader can observe in ex:8, we can encode the fact that “`puss`” is a lemma and that it is a noun. More importantly, we can include the senses associated to the WBÖ entry. In fact we are adding a new sense, “`icltt:gebäck`” (*pastry*). This is a direct consequence of the linguistic analysis of the expression “`süßes Gebäck`” we described in section 3 (ex:1 and ex:2): since the word “`Gebäck`” is the head noun in this nominal phrase, we can assume that this head noun is also a meaning (or sense) to be associated to the entry. In doing so we introduce a hierarchical organization of the meanings associated to an entry: “`süßes Gebäck`” is a specialization of “`Gebäck`” (using the “`skos:broader`” relation, as shown in ex:16). As mentioned above, this is the reason why we use the `skos:ConceptScheme` construct in order to encode the senses associated to the entries (ex:9-ex:11):

---

<sup>24</sup> *lemon* is also available as an ontology: <http://lemon-model.net/>

(9) icltt:Senses\_ICLTT

```
    rdf:type skos:ConceptScheme ;
    rdfs:comment "Senses that are used in ICLTT dictionaries"@en ;
    rdfs:label „Senses“@en , "Bedeutungen"@de .
```

(10) icltt:Sense

```
    rdf:type owl:Class ;
    rdfs:label "Sense"@en ;
    rdfs:subClassOf skos:Concept ;
    owl:equivalentClass lemon:LexicalSense .
```

(11) icltt:kuss

```
    rdf:type icltt:Sense ;
    rdfs:label "kiss"@en , "Kuss"@de ;
    skos:inScheme icltt:Senses_ICLTT ;
    skosxl:prefLabel icltt:sense_kuss .
```

Lemmas of the expressions used in the original dictionaries for marking meanings are indirectly linked to the class Sense, and are directly instances of the class icltt:Lemma (ex:12-ex:14):

(12) icltt:sense\_gebäck

```
    rdf:type icltt:Lemma ;
    rdfs:label "Gebäck"@de ;
    icltt:hasPos icltt:noun ;
    skosxl:literalForm "Gebäck"@de .
```

(13) icltt:sense\_kuss

```
    rdf:type icltt:Lemma ;
    rdfs:label "Kuss"@de ;
    icltt:hasPos icltt:noun ;
    skosxl:literalForm "Kuss"@de .
```

(14) icltt:sense\_süß

```
    rdf:type icltt:Lemma ;
    rdfs:label "süß"@de ;
    icltt:hasPos icltt:adj ;
    skosxl:literalForm "süß"@de .
```

We introduce in ex:14 the class “CompoundSense” that allows us to mark the fact that the sense(s) can be resulting from a compound term or a phrase used to express the meaning of an entry.

(15) icltt:CompoundSense

```
    rdf:type owl:Class ;  
    rdfs:label “Composition of Sense”@en ;  
    rdfs:subClassOf icltt:Sense .
```

An instance of such a class is displayed in ex:16, in which the reader can see how we model for the time being the hierarchical relation between the sense “gebäck” and “süßes Gebäck” (using the “skos:broader” relation). We can also encode the fact that the sense of the entry “süß\_gebäck” is composed of two senses, but the model needs to be further developed, since it is clear that the sense “süß” cannot be considered only as a sub-sense of “süß\_gebäck”, but more as a “modifying” sense. We are currently working on representing with the help of *lemon* such cases of linguistic dependencies.

(16) icltt:süß\_gebäck

```
    rdf:type icltt:CompoundSense , lemon:LexicalSense ;  
    rdfs:label „sweet pastry”@en , „süßes Gebäck”@de ;  
    lemon:subsense icltt:süß , icltt:gebäck ;  
    skos:broader icltt:gebäck ;  
    skos:inScheme icltt:Senses_ICLTT .
```

In ex:16 we can see the advantage of using a representation model that can encode linguistic properties. In this case, it is for example necessary to tokenize the string representing the meaning of the entry “Puss”: the first token can then be lemmatized to “süß” (*sweet*), while for the second token the lemma is identical to the written form used. We represent the tokenization information using the *lemon* property “decomposition”, as can be seen in ex:17:

(17) lemon:decomposition

```
    rdfs:domain lemon:LexicalSense ;  
    rdfs:range icltt:Sense .
```

For the time being we introduce in our model an explicit listing of components as subclasses of icltt:CompoundSense (see ex:18 and ex:19). The way this encoding is used is shown in ex:21.

(18) icltt:Component1

```
    rdf:type owl:Class ;  
    rdfs:label “”^^xsd:string ;  
    rdfs:subClassOf icltt:CompoundSense .
```

(19) icltt:Component2

```

rdf:type owl:Class ;
rdfs:label ""^^xsd:string ;
rdfs:subClassOf icltt:CompoundSense .

```

### 4.3 Linking to Resources available in the LOD

As the reader can see in the examples (20) and (21) further below, we decided to use the DBpedia instantiation of Wiktionary as a reference for the senses (meanings) of the entries of the dictionary, pointing thus to linguistic and knowledge objects that are already in the LOD. To be more precise, the link to DBpedia/Wiktionary is applied for each token of the expressed meanings. In the case of “süßes Gebäck”, we can thus point to two URLs in DBpedia/Wiktionary, each representing the adequate senses for the actual token. In the name of the URLs used for pointing to DBpedia/Wiktionary we have implicitly also the information about the language and the PoS of the entry. But one could point to the RDF version of ISO data categories<sup>25</sup> for making this information explicit in our model.

Additionally the linking to the appropriate senses in DBpedia/Wiktionary allows accessing all the corresponding multilingual lemmas associated in this resource with a sense. Looking for example at <http://wiktionary.dbpedia.org/page/sweet-English-Adjective-1en> (corresponding to the URL for the German word, we use in ex:20), we get more than 70 expressions in more than 60 languages.

And the URL <http://wiktionary.dbpedia.org/page/pastry-English-Noun-1en> refers to ca. 30 expressions in about 25 languages. Having one unique URL for the “sweet” sense of “süß” allows to link all the corresponding entries to a unique reference point, and so to improve comparability of dictionary resources, also at the semantic level.

(20) icltt:süß

```

rdf:type lemon:LexicalSense , icltt:Component1 ;
rdfs:label „sweet“@en , „süß“@de ;
skos:exactMatch <http://wiktionary.dbpedia.org/resource/süß-German-Adjective-1de> ;
skos:inScheme icltt:Senses_ICLTT .

```

(21) icltt:gebäck

```

rdf:type lemon:LexicalSense , icltt:Component2 ;
rdfs:label „Gebäck“@de , „pastry“@en ;
skos:exactMatch <http://wiktionary.dbpedia.org/resource/Gebäck-German-Noun-1de> ;
skos:inScheme icltt:Senses_ICLTT ;
skos:narrower icltt:süß_gebäck ;
skosxl:prefLabel icltt:sense_gebäck .

```

---

25 See <http://www.isocat.org/>



In the two examples just above, the reader can see how we can link the senses of the entries to existing sources in the LOD. We use for this the `skos:exactMatch` property (although we could also use *lemon* properties for this). But our model also makes clear that the DBpedia/Wiktionary URL we use for each token is valid only in the context of the compound term we are dealing with. The word “süß” has in DBpedia/Wiktionary more senses, but in the context of “süßes Gebäck” only the one sense that refers to “sweet” is adequate. Using the lemon model allows us thus to disambiguate senses associated to the components of complex terms used in the dictionaries for expressing a meaning.

## 5 Cross-referencing of Dictionary Entries through shared Meanings

The establishment of a relation between the entry “Puss” in WBÖ and the entry “Bussal” in WWM is made possible on the base of the successful mapping of both the adjectival modifier “süß” and the head noun “Gebäck”, which are present in both the definitions in WBÖ and WWM, but used in the context of textual variants, as can be seen in the examples (1) and (2). Interesting is also the fact that a user searching the electronic version of the dictionaries could give the High German form “Gebäck” and would get from both dictionaries all the entries which have this word in their definition, also if the word is used in a compound form. The same for the High German adjectival form “süß”, also irrespectively if this form is inflected or part of a compound word. Our work is thus also addressing in the longer term the semantic access to dialectal dictionaries.

## 6 Conclusion

We presented an approach consisting in extracting meanings associated to entries in two dialectal dictionaries. Comparison of the expressions used to mark those meaning can be done only after applying basic natural language processing to those expressions. Expressions that are judged as being similar are cross-linked. Furthermore we encode those meanings in Semantic Web representation languages and can so link to lexical and knowledge resources available in the Linked Data Framework. This step in supporting potentially the semantically base cross-linking of our lexical entries to other dialectal dictionaries published in the web.

Current work is dedicated in improving the model for an adequate representation of more complex linguistic phenomena, and also in investigating how our approach could be applied for linking our dictionaries not only to DBpedia/Wiktionary but also to other lexical resources. We think here in particular to portals that already offer a network of dictionaries, like the Trier Wörterbuchnetz<sup>26</sup>, which

---

26 <http://woerterbuchnetz.de/>

contains a lot of dialectal dictionaries, also offering cross-links between entries. A next step will consist in published the content of our dictionaries in the Web, so that references from other sources in the LOD to our dictionaries can be implemented.

## 7 References

- Arbeitsplan (1912)*. Accessed at <http://www.oeaw.ac.at/icltt/dinamlex-archiv/Arbeitsplan.pdf> [10/04/2014]
- Declerck, T. (202). A set of tools for integrating linguistic and non-linguistic information. In: *Proceedings of SAAKM (ECAI Workshop)*.
- Declerck, T., Lendvai, P., Mörth, K. (2013). Collaborative Tools: From Wiktionary to LMF, for Synchronic and Diachronic Language Data. In Francopoulo, G. (ed) *LMF Lexical Markup Framework*. Wiley 2013.
- Francopoulo, G. (2013) *LMF -- Lexical Markup Framework*. Wiley.
- Gennari, J., Ferguson, R., Grosso, W. E., Crubezy, M., Eriksson, H., Noy, N. F., Tu, S. W., Musen, M. A. (2002). The Evolution of Protégé: An Environment for Knowledge-Based Systems Development
- Hornung, M., Grüner, S. (2002) *Wörterbuch der Wiener Mundart*; Neubearbeitung. öbvht, Wien.
- McCrae, J., Aguado-de-Cea, G., Buitelaar P., Cimiano P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. In: *Language Resources and Evaluation*. Vol. 46, Issue 4, Springer:701-719.
- Miles, A., Matthews, B., Wilson, M. D., Brickley, D. (2005). SKOS Core: Simple Knowledge Organisation for the Web. In *Proc. International Conference on Dublin Core and Metadata Applications*, Madrid, Spain,
- Moulin, C. (2010). Dialect dictionaries - traditional and modern. In: Auer, P., Schmidt, J.E. (2010) (eds) *Language and Space. An International Handbook of Linguistic Variation. Volume 1: Theories and Methods*. Berlin / New York. pp: 592-612. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science / Manuels de linguistique et des sciences de communication 30.1). Österreichisches Wörterbuch, 42. Auflage. 2012.
- Romary, L. (2009). Questions & Answers for TEI Newcomers. *Jahrbuch für Computerphilologie 10*. Mentis Verlag.
- Schreibman, S. (2009). The Text Encoding Initiative: An Interchange Format Once Again. *Jahrbuch für Computerphilologie 10*. Mentis Verlag.
- Straffungskonzept für das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1998)*. Accessed at [http://www.oeaw.ac.at/icltt/dinamlex-archiv/Straffungskonzept\\_1998.pdf](http://www.oeaw.ac.at/icltt/dinamlex-archiv/Straffungskonzept_1998.pdf) [10.04.2014].
- Theron, R., Wandl-Vogt, E. (2014). The fun of exploration: how to access a non-standard language corpus visually. In: *VisLR: Visualization as added value in the development, use and evaluation of LRs. Workshop-Proceedings of LREC2014*.
- Wandl-Vogt, E. (2005). From paper slips to the electronic archive. Cross-linking potential in 90 years of lexicographic work at the Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In: *Complex 2005. Papers in computational lexicography*. Budapest: 243-254.
- Wandl-Vogt, E. (2008). ..wie man ein Jahrhundertprojekt zeitgemäß halt: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX). In: Ernst, P. (ed) 2008, *Bausteine zur Wissenschaftsgeschichte von Dialektologie / germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der internationalen Gesellschaft für Dialektologie des Deutschen*, Wien: 93-112.
- Wandl-Vogt, E. (2010). Point and find: the intuitive user experience in accessing spatially structured dialect dictionaries. *Slavia Centralis* 3 (2010): 35-53.
- Wandl-Vogt, E., Declerck, T. (2013). Mapping a Traditional Dialectal Dictionary with Linked Open Data. In: Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. *Electronic lexicography in*

*the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.* Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

*Wiktionary RDF extraction.* Accessed at: <http://dbpedia.org/Wiktionary> [10/04/2014]

*Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1963-).* Verlag der Österreichischen Akademie der Wissenschaften. Wien. Accessed at: <http://hw.oeaw.ac.at/cl?frames=yes> [10/04/2014].

### **Acknowledgements**

The DFKI part of this work is supported by the LIDER Project. LIDER: “Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe” is a FP7 project with reference number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies.



# Nutzung des DWDS-Wortprofils beim Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen

Jörg Didakowski, Nadja Radtke  
Berlin-Brandenburgische Akademie der Wissenschaften,  
Technische Universität Dortmund  
didakowski@bbaw.de, nadja.radtke@tu-dortmund.de

## Abstract

Im Hinblick auf eine Zuarbeit für die lexikographische Arbeit an einem lexikalischen Informationssystem zu deutschen Stützverbgefügen wird das DWDS-Wortprofil vorgeschlagen. Mithilfe dieses Werkzeugs kann eine zeitintensive und mühsame Rechercharbeit über eine Textsuchmaschine vermieden werden, indem auf Basis eines ausgewogenen Korpus potenzielle Stützverbgefüge bereitgestellt werden. Des Weiteren wird die Einbeziehung von Assoziationsmaßen vorgeschlagen, um die Menge des zu sichtenden Materials für die lexikographische Arbeit weiter reduzieren zu können.

**Keywords:** Stützverbgefüge; Assoziationsmaße; computerlinguistische Verfahren

## 1 Einleitung

Neben den traditionellen Textkorpora stehen den Lexikographen heutzutage auch digitale Textkorpora zur Verfügung, die über Abfrage- und Analysewerkzeuge die Möglichkeiten der lexikographischen Arbeit stark erweitern (Engelberg & Lemnitzer 2009). Eines dieser Werkzeuge stellt das DWDS-Wortprofil dar, welches beim Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen (das SVG-Wiki) hilft bzw. den Aufbau von diesem gar erst ermöglicht. Dieses Werkzeug ist Teil des Digitalen Wörterbuchs der deutschen Sprache (DWDS), eines Projekts der Berlin-Brandenburgischen Akademie der Wissenschaften. Ziel des vorliegenden Beitrags ist es, die Potenziale der Nutzung des DWDS-Wortprofils beim Aufbau des SVG-Wikis aufzuzeigen.

Im 2. Kapitel des Beitrags stellen wir zunächst das SVG-Wiki vor, dabei legen wir den Gegenstand fest, bestimmen die Zielgruppe, führen die Komponenten des SVG-Wikis ein und heben abschließend die korpusbasierte Erarbeitung des SVG-Wikis hervor. Im 3. Kapitel beschreiben wir das DWDS-Wortprofil, das für die Datenerhebung des SVG-Wikis grundlegend ist. Anschließend gehen wir im 4. Kapitel des Beitrags genauer auf die Nutzung des DWDS-Wortprofils beim Aufbau des SVG-Wikis ein.

## 2 Das SVG-Wiki: Ein lexikalisches Informationssystem zu deutschen Stützverbgefügen

Den Gegenstand des SVG-Wikis bilden Stützverbgefüge (SVG, engl. *support verb constructions*) des Deutschen wie z.B. *Anwendung finden*, *zur Anwendung kommen* und *Kritik üben*, die aus einem prädikativen Nomen (z.B. *Anwendung*) und einem semantisch blassen Stützverb (z.B. *finden* oder *kommen*) konstruiert werden<sup>1</sup>. Abgegrenzt werden SVG von freien Konstruktionen (z.B. *zur Party kommen*) sowie von Idiomen (z.B. *zu Potte kommen*); beschrieben werden sie in Bezug auf ihre Systematik und hinsichtlich ihrer Leistungen (z.B. Pottelberge van 2001; Seifert 2004; Heine & Wotjak 2005; Storrer 2007).

SVG haben bereits Eingang in die Grammatiken, Wörterbücher und Lehrwerke des DaF-Unterrichts gefunden. Wie sie am besten eingeführt und behandelt werden, wird in der Forschungsliteratur intensiv diskutiert. Vermittelt werden sie bei der Arbeit mit dem Wortschatz sowie im Grammatikunterricht, indem einerseits ausgewählte SVG aufgeführt und andererseits ausgewählte grammatische und stilistische Eigenschaften der SVG beschrieben werden. Wie man dabei zu einer adäquaten Auswahl der SVG gelangt und wie Vollständigkeit und Systematik hinsichtlich der grammatischen und stilistischen Eigenschaften der SVG erreicht werden können, bleibt jedoch offen. Die zu vermitteln den SVG auszuwählen, um dann diese in Hinblick auf ihre Systematik behandeln zu können, fällt hier nach wie vor in das Aufgabenfeld der DaF-Lehrenden. Eine vor diesem Hintergrund hilfreiche Ressource zu deutschen SVG für die DaF-Lehrenden gibt es noch nicht. Das SVG-Wiki schließt diese Lücke, indem es die DaF-Lehrenden als zukünftige Nutzer vorsieht und aufbauend auf den bereits ausgereiften Vorschlägen für die Printlexikographie (Heine 2006; Heine 2008) eine Wörterbuchkomponente (Spezialwörterbuch) mit einer Grammatikkomponente verbindet. Das SVG-Wiki ist als ein digitales wikibasiertes Informationssystem realisiert<sup>2</sup>, das den Nutzern im Internet zur freien Verfügung stehen wird und kontinuierlich erweitert werden soll (siehe Abbildung 1).

---

1 In der deutschen linguistischen Fachliteratur findet man unterschiedliche Termini wie *Nominalisierungsverbgefüge*, *Funktionsverbgefüge* und *Streckverbgefüge*, die ebenfalls unterschiedlich begrifflich gefasst sind. Bei der Bestimmung des Gegenstandes des SVG-Wikis bedienen wir uns einem möglichst weiten Begriff der SVG: Das prädikative Nomen steht dabei im Akkusativ oder kommt als Präpositionalphrase vor, es ist abstrakt, wird deverbale oder deadjektivisch gebildet, kann ein Fremdexem sein und idiomatisch verwendet werden.

2 Das Wikisystem (hier: MediaWiki) ermöglicht u.a., die jeweiligen Komponenten des SVG-Wikis kollaborativ auszubauen. So können z.B. die Nutzer die noch nicht im SVG-Wiki berücksichtigten Stützverben zur Erweiterung des SVG-Wikis auf den dafür vorgesehenen Seiten des SVG-Wikis vorschlagen.



Abbildung 1: Hauptseite des SVG-Wikis.

Ziel des SVG-Wikis ist es, die SVG in ihrem Grundbestand festzuhalten, die DaF-Lehrenden bei der Auswahl der SVG zu unterstützen und die SVG in Bezug auf ihre grammatischen und stilistischen Eigenschaften zu beschreiben.

Das Besondere an dem SVG-Wiki ist, dass die jeweiligen Komponenten korpusbasiert erarbeitet werden. Ausgegangen wird hierbei von einer Lemmaliste der Stützverben, die aus 23 ausgewählten Grammatiken und unter Berücksichtigung der Forschungsliteratur erstellt wurde; sie enthält ca. hundert Stützverben mit den für sie charakteristischen und für die Datenerhebung relevanten Merkmalen. So wird z.B. bei dem Stützverb *finden* eingetragen, dass dieses mit einem prädikativen Nomen im Akkusativ vorkommt. Anhand dieser Lemmaliste werden zunächst die prädikativen Nomina der jeweiligen Stützverben<sup>3</sup> mithilfe des DWDS-Wortprofils ermittelt; von den ermittelten prädikativen Nomina ausgehend, besteht im weiteren Schritt der Erarbeitung die Möglichkeit, die im SVG-Wiki noch nicht berücksichtigten Stützverben ebenfalls mithilfe des DWDS-Wortprofils zu entdecken. Anschließend findet die lexikographische Beschreibung der Stützverben, der prädikativen Nomina und der SVG anhand des DWDS-Kernkorpus statt. Parallel dazu wird die Grammatikkomponente des SVG-Wikis erarbeitet. Das DWDS-Kernkorpus stellt somit die primäre Quelle für die Wörterbuchbasis der Wörterbuchkomponente dar und bildet gleichzeitig die Datengrundlage für die Grammatikkomponente des SVG-Wikis.

3 Kamber (2008) geht ebenfalls in seiner korpusbasierten Untersuchung zu den nominalen Prädikaten des Deutschen von den jeweiligen Verben aus.

### 3 Das DWDS-Wortprofil

Das DWDS-Wortprofil (Didakowski & Geyken 2013) stellt Kookkurrenzpaare für verschiedene grammatische Relationen wie z.B. Akkusativ-/Dativobjekt, Genitivattribut, Adjektivattribut und präpositionales Komplement/Modifizierer bereit. Die Kookkurrenzpaare werden mithilfe von computerlinguistischen Verfahren automatisch extrahiert. In Kilgarrieff et al. (2004) wird für die automatische Extraktion grammatischer Kookkurrenzpaare die flache *Sketch-Grammar* vorgeschlagen, mit der über reguläre Ausdrücke Kookkurrenzpaare für bestimmte grammatische Relationen extrahiert werden können. Ivanova et al. (2008) zeigen jedoch, dass es für das Deutsche sinnvoll ist, um gute Ergebnisse zu erzielen, auf eine reichhaltigere linguistische Analyse zurückzugreifen. Beim DWDS-Wortprofil wird für eine reichhaltigere Analyse die TAGH-Morphologie (Geyken & Hanneforth 2006) und der regelbasierte Parser SynCoP (Syntactic Constraint Parser, Didakowski 2008) verwendet. So kann die relativ reichhaltige Morphologie und freie Wortstellung im Deutschen angemessen behandelt und die Kookkurrenzen mit gewünschter Qualität extrahiert werden.

Im DWDS-Wortprofil sind die Kookkurrenzpaare mit Werten verschiedener Assoziationsmaße versehen. Derzeit werden drei verschiedene Assoziationsmaße berechnet: 1) die reine Frequenz, 2) das auf dem Dice-Koeffizienten basierende logDice-Maß (Rychlý 2008) und 3) das auf Mutual-Information basierende MI-log-Freq-Maß (Kilgarrieff & Tugwell 2002). Mithilfe dieser Maße können Kookkurrenzpaare nach Verbindungsstärke bzw. Anziehungskraft sortiert werden. Hierbei wird das Assoziationsmaß in der Regel so gewählt, dass die entsprechende Sortierung für eine bestimmte Aufgabe am geeignetsten ist (Evert 2008).

Die Korpusgrundlage für das DWDS-Wortprofil bilden das DWDS-Kernkorpus und verschiedene verbreitete Zeitungen (*Süddeutsche Zeitung, DIE ZEIT, Berliner Zeitung, DIE WELT, Der Tagesspiegel, Bild*). Das DWDS-Kernkorpus ist ein Referenzkorpus der *deutschen* Sprache des 20. Jahrhunderts und ist ausgeglichen bezüglich verschiedener Textsorten, die zudem gleichmäßig über das 20. Jahrhundert verteilt sind. Es umfasst über 100 Millionen laufende Wortformen (Tokens) und hat damit eine vergleichbare Größe wie das British National Corpus (Geyken 2007). Das DWDS-Kernkorpus stellt somit als ausgeglichenes Referenzkorpus das Herzstück der Korpusbasis des DWDS-Wortprofils dar und nimmt damit eine besondere Stellung ein. Die gesamte Korpusgrundlage des DWDS-Wortprofils umfasst ca. 1,7 Milliarden laufende Wortformen (Tokens) und reicht zeitlich vom Anfang des 20. Jahrhunderts bis heute. Die Kookkurrenzpaare sind hierbei für die gesamte Korpusbasis und auch für die einzelnen Subkorpora berechnet. So können Kookkurrenzpaare z.B. ausschließlich auf Basis des DWDS-Kernkorpus abgefragt werden.

Das DWDS-Wortprofil ist einerseits über die DWDS-Webseite und andererseits innerhalb der CLARIN-Infrastruktur über WebLicht (Hinrichs et al. 2010) zugänglich, wo es in Verarbeitungsketten integriert werden kann.

Ein Beispiel für eine DWDS-Wortprofil-Abfrage auf der DWDS-Webseite ist in Abbildung 2 zu sehen. Hier wurde das Verb *finden* für die Akkusativ-/Dativobjekt-Relation unter Verwendung des MI-log-



Freq-Maßes auf der Basis des DWDS-Kernkorpus abgefragt. Die relevanten Kookkurrenzpartner zu dem Verb *finden* werden als Wortwolke dargestellt. Je größer der Wert des Assoziationsmaßes eines Kookkurrenzpaares ist, desto größer wird der Kookkurrenzpartner in der Wolke dargestellt. Eine alternative Darstellungsform zu dieser Wortwolke ist die Tabellenansicht, in der die Kookkurrenzpartner nach dem Assoziationsmaß sortiert aufgelistet und genauere Informationen zu Wortkategorien und Assoziationswerten aufgeführt sind.

Hervorzuheben ist dabei, dass im DWDS-Wortprofil über die einzelnen Kookkurrenzpartner direkt auf die entsprechenden Korpusbelege zugegriffen werden kann. Erst dadurch wird eine sinnvolle lexikographische Arbeit möglich. In Abbildung 3 sind die Belege für das Kookkurrenzpaar *Anwendung finden* aufgeführt.

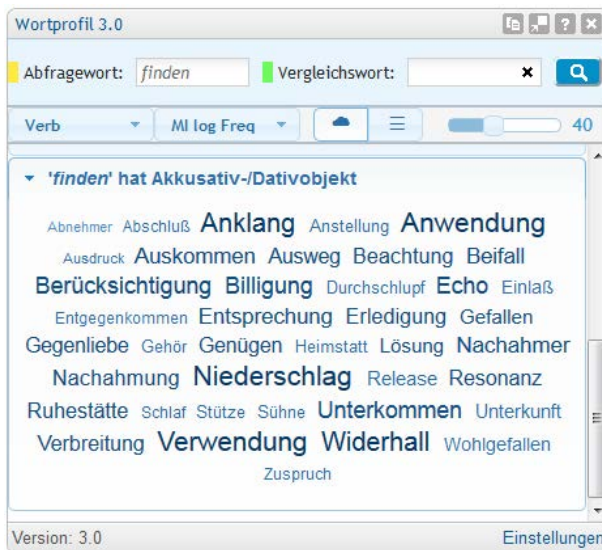


Abbildung 2: DWDS-Wortprofil-Wortwolke.

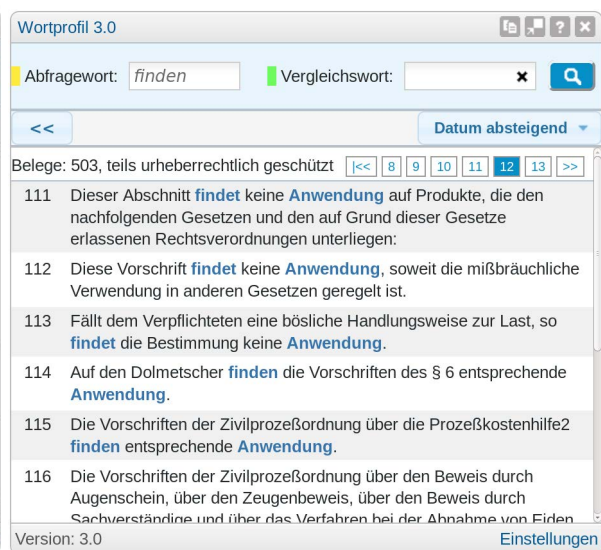


Abbildung 3: DWDS-Wortprofil-Belege.

Mit dem DWDS-Wortprofil ist es also möglich, auf strukturierte Weise Kookkurrenzpaare mit den dazugehörigen Korpusbelegen zu ermitteln. Verschiedene Assoziationsmaße können dazu verwendet werden, bestimmte Kookkurrenzpaare aus der Menge der Kookkurrenzpaare hervorzuheben. Die große Korpusbasis zusammen mit dem DWDS-Kernkorpus als ausgeglichenem Bestandteil gewährleistet hierbei ein breites Spektrum an Kookkurrenzen und repräsentative Ergebnisse.

## 4 Nutzung des DWDS-Wortprofils beim Aufbau des SVG-Wikis

Im Folgenden begründen wir die Wahl der Korpusbasis für die Ermittlung der prädikativen Nomina beim Aufbau des SVG-Wikis und gehen kurz auf verschiedene Möglichkeiten zur Ermittlung der prädikativen Nomina ein. Hierbei heben wir die Potenziale des DWDS-Wortprofils hervor. Im Weiteren

erläutern wir, ob die Assoziationsmaße dabei helfen können, die Menge der Kookkurrenzpaare für mögliche prädikative Nomina zu verkleinern, sodass weniger Kookkurrenzpaare gesichtet werden müssen und dabei trotzdem der Grundbestand der SVG festgehalten werden kann.

#### 4.1 Ermittlung der prädikativen Nomina

Bei der Ermittlung der prädikativen Nomina wird die Korpusbasis auf das DWDS-Kernkorpus eingeschränkt, damit mit Blick auf die Zielsetzung des SVG-Wikis die Dekaden des 20. Jahrhunderts sowie verschiedene Textsorten gleichermaßen vertreten sind.

Zur Ermittlung der prädikativen Nomina kann einerseits die DWDS-Suchmaschine und andererseits das DWDS-Wortprofil genutzt werden. Bei der Nutzung der DWDS-Suchmaschine erhält man z.B. für die Abfrage zu dem Verb *finden* eine Liste mit 82.864 Treffern, in der nach den prädikativen Nomina manuell gesucht werden muss. Bei der Nutzung des DWDS-Wortprofils reduziert man bereits durch die Wahl einer grammatischen Relation die Menge der Treffer. So erhält man bei dem Verb *finden* durch die Wahl der Akkusativ-/Dativobjekt-Relation eine Liste mit 779 Kookkurrenzpaaren, die dann durch das Zugreifen auf einzelne Korpusbelege manuell nach prädikativen Nomina klassifiziert werden können. Hierbei beträgt die durchschnittliche Anzahl an Kookkurrenzpaaren für die 31 in den Grammatiken am häufigsten genannten Stützverben pro Verb mit dem Nomen im Akkusativ ca. 437 und pro Verb mit dem Nomen als Präpositionalphrase ca. 843. Hervorzuheben ist im Weiteren, dass dem DWDS-Wortprofil eine reichhaltigere linguistische Analyse (siehe Kapitel 3) zugrunde liegt und somit bestimmte Fälle, die die Suche mit der DWDS-Suchmaschine zusätzlich erschweren, vermieden werden. Zu solchen Fällen gehören Verben mit einem abtrennbaren Präfix (z.B. *ausüben*) oder Verben, die bezüglich einer Wortform homograph zu einem anderen Verb sind (z.B. *geraten* und *raten*). Somit ermöglicht das DWDS-Wortprofil, die Ermittlung der prädikativen Nomina überhaupt in einem realistischen Zeitrahmen bewältigen zu können.

#### 4.2 Verwendung von Assoziationsmaßen bei der Ermittlung der prädikativen Nomina

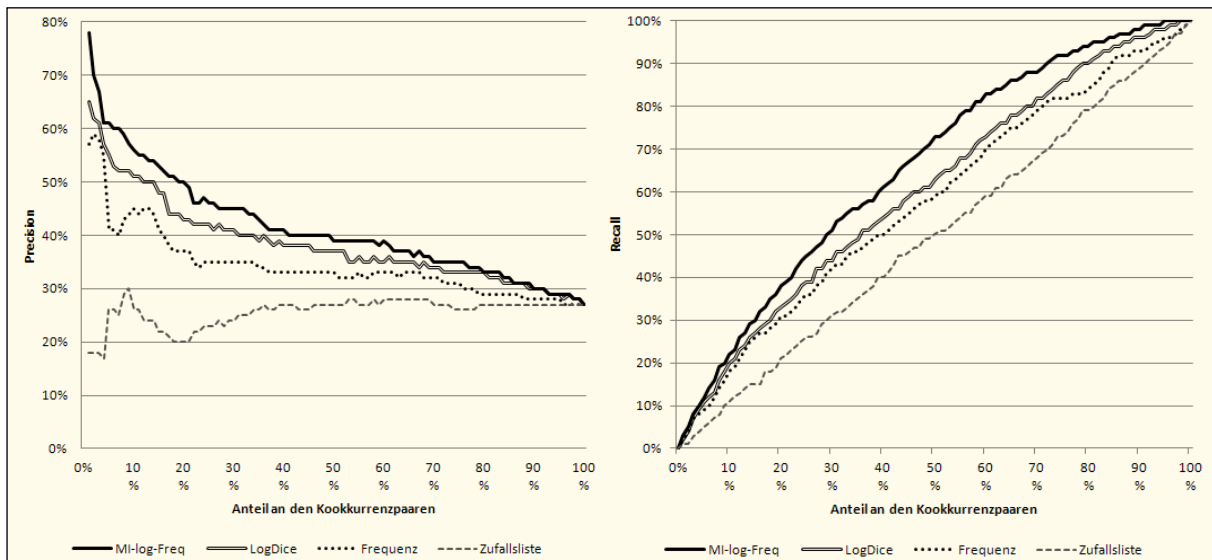
Über das DWDS-Wortprofil ist es möglich, die abgefragten Kookkurrenzpaarlisten nach verschiedenen Assoziationsmaßen zu sortieren (siehe Kapitel 3). Hierbei ist die Frage, ob die Assoziationsmaße in der Lage sind, die Kookkurrenzpaarlisten so zu sortieren, dass am Anfang der Listen die Dichte der prädikativen Nomina sehr hoch ist und am Ende nur wenige prädikative Nomina vorkommen. So könnten die Kookkurrenzpaarlisten verkleinert werden, ohne dass zu viele prädikative Nomina verloren gehen. Auf diese Weise kann Recherchearbeit eingespart werden.<sup>4</sup>

---

4 Langer (2009) versucht, Funktionsverbgefüge vollautomatisch aus Korpora zu gewinnen. Dies will er hauptsächlich über Assoziationsmaße realisieren. Er zeigt, dass die Maße für so eine Aufgabenstellung

Um in Bezug zu der oben genannten Fragestellung eine Bewertung der Assoziationsmaße durchzuführen, folgen wir der Methodik in Evert et al. (2000) und beurteilen die Maße anhand von Precision und Recall. Hierzu wurden vier Verben herangezogen: *bringen*, *finden*, *kommen* und *üben*. Zu diesen Verben wurden mithilfe des DWDS-Wortprofils unter Berücksichtigung der jeweiligen grammatischen Relationen und durch die Wahl des DWDS-Kernkorpus als Korpusbasis Kookkurrenzpaarlisten ermittelt, wobei eine Minimalfrequenz für die Kookkurrenzpaare auf fünf festgelegt wurde. Die Listen wurden dann manuell vollständig gesichtet und nach prädikativen Nomina klassifiziert. Insgesamt wurden 859 prädikative Nomina aus 9.166 Kookkurrenzpaaren identifiziert. Die prädikativen Nomina machen demnach 27% der Gesamtmenge aus. Über die so erstellte Referenzmenge können nun Precision und Recall ermittelt werden. Zur Bewertung des Nutzens der Assoziationsmaße wird hier zusätzlich eine Zufallsortierung der Kookkurrenzlisten hinzugezogen. Die Zufallsliste deckt den Fall ab, dass kein Assoziationsmaß zur Sortierung verwendet wird.

Der Verlauf von Precision und Recall zu den einzelnen Assoziationsmaßen und der Zufallsliste ist in den Diagrammen in Abbildung 4 und 5 zu sehen. Auf den X-Achsen ist der Anteil an Kookkurrenzpaaren, der durch das Verkürzen der Kookkurrenzpaarlisten entsteht, in Prozent aufgetragen. Precision und Recall zu den einzelnen Anteilen sind jeweils auf den Y-Achsen in Prozent ablesbar.



**Abbildung 4: Precision-Kurven.**

**Abbildung 5: Recall-Kurven.**

Der Verlauf der Precision-Kurven in Abbildung 4 zeigt, dass über die Sortierung nach dem MI-log-Freq-Maß die besten Precision-Werte erreicht werden. Hier liegt der Anteil der prädikativen Nomina sogar bei 78%, wenn 1% der Kookkurrenzpaare herangezogen wird. Mit anwachsendem Anteil an Kookkurrenzpaaren flachen die Precision-Kurven der Assoziationsmaße anfangs ab und fallen stetig auf das Grundniveau von 27%, welches durch den Anteil an prädikativen Nomina gesetzt ist. Bei der Zu-

---

nicht ausreichend sind. Die vollautomatische Extraktion, die Langer (2009) im Sinn hat, wird in unserer Vorgehensweise jedoch nicht verfolgt.

fallsortierung hingegen bewegt sich die Precision lediglich nahe am Grundniveau und sogar darunter. Der Verlauf der Recall-Kurven in Abbildung 5 zeigt ergänzend dazu, dass unter Verwendung des MI-log-Freq-Maßes nur die Hälfte der Kookkurrenzpaare gesichtet werden muss, um bereits 72% der prädikativen Nomina zu ermitteln. Sind die Kookkurrenzpaare nach dem Zufall sortiert, bekommt man hingegen lediglich ca. 50% der prädikativen Nomina.

Hier wird deutlich, dass die Assoziationsmaße hilfreich sind, wenn man bei der Ermittlung der prädikativen Nomina den Umfang und den damit verbundenen zeitlichen Aufwand reduzieren möchte und gleichzeitig möglichst viele prädikative Nomina als Grundbestand ermitteln will. Hierbei hat sich das MI-log-Freq-Maß als am geeignetsten herausgestellt.

## 5 Zusammenfassung

Das DWDS-Wortprofil hat das Potential, den Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen entscheidend zu erleichtern. Die Aufgabe des DWDS-Wortprofils liegt bei dem Aufbau des SVG-Wikis darin, mögliche prädikative Nomina für Stützverben bereitzustellen. Dadurch wird eine zeitintensive und mühsame Recherche über eine Textsuchmaschine vermieden. Unter dem Aspekt zeitlicher Restriktionen ist das DWDS-Wortprofil sogar unabdingbar. Über eine zusätzliche Bewertung nach Assoziationsmaßen kann zudem weitere Recherchezeit eingespart werden.

## 6 Literaturhinweise

- Das digitale Wörterbuch der deutschen Sprache (DWDS). Accessed at: [www.dwds.de](http://www.dwds.de) [11/04/2014].
- Didakowski, J. (2008). Local Syntactic Tagging of Large Corpora Using Weighted Finite State Transducers. In A. Storrer, A. Geyken et al. (eds.) *Text Resources and Lexical Knowledge. Selected Papers from the 9th Conference on Natural Language Processing, KONVENS 2008*. Berlin et al.: Mouton de Gruyter, pp. 65-78.
- Didakowski, J., Geyken, A. (2013). From DWDS corpora to a German Word Profile – methodological problems and solutions. In *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information, 2nd Work Report of the Academic Network “Internet Lexicography” (OPAL - Online publizierte Arbeiten zur Linguistik X/2012)*. Mannheim: Institut für Deutsche Sprache, pp. 43-52.
- Engelberg, S., Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.
- Evert, S., Heid, U. et al. (2000). Methoden zum qualitativen Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In W. Zühlke, Ernst G. Schukat-Talamazzini (eds.) *Sprachkommunikation, KONVENS 2000*. Berlin et al.: VDE, pp. 215-220.
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling, M. Kytö (eds.) *Corpus Linguistics. An International Handbook of the Science of Language and Society*. Berlin: Mouton de Gruyter, pp. 1212-1248.
- Geyken, A., Hanneforth, T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In A. Yli-Jyrä, L. Karttunen et al. (eds.) *Finite State Methods and Natural Language Processing*. Berlin et al.: Springer, pp. 55-66.

- Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In Ch. Fellbaum (eds.) *Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies*. London et al.: Continuum, pp. 23-41.
- Heine, A., Wotjak, B. (2005). Zur Abgrenzung und Beschreibung verbonominaler Wortverbindungen (Wortidiome, Funktionsverbgefüge, Kollokationen). In *Deutsch als Fremdsprache. Zeitschrift für Theorie und Praxis des Deutschunterrichts für Ausländer*, 42(3), pp. 143-153.
- Heine, A. (2006). Funktionsverbgefüge in System, Text und korpusbasierter (Lerner-) Lexikographie. Frankfurt a. M. et al.: Peter Lang.
- Heine, A. (2008). Funktionsverbgefüge richtig verstehen und verwenden. Ein korpusbasierter Leitfaden mit finnischen Äquivalenten. Frankfurt a. M. et al.: Peter Lang.
- Hinrichs, E., Hinrichs, M. et al. (2010). WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010, System Demonstrations (ACLDemos '10), Association for Computational Linguistics*. Stroudsburg, PA (USA), pp. 25-29.
- Ivanova, K., Heid, U. et al. (2008). Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. In *Proceedings of the 6th International Conference on Language Resources and Evaluation, 2008*. Marrakech (Morocco), pp. 2101-2107.
- Kilgarriff, A., Tugwell, D. (2002). Sketching Words. In M.-H. Corréard (eds.) *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins, EURALEX*, pp. 125-137.
- Kilgarriff, A., Rychlý, P. et al. (2004). The Sketch Engine. In *Proceedings of EURALEX 2004*. Lorient (France), pp. 105-116.
- Kamber, A. (2008). Funktionsverbgefüge – empirisch. Eine korpusbasierte Untersuchung zu den nominalen Prädikaten des Deutschen. Tübingen: Max Niemeyer.
- Langer, S. (2009). Funktionsverbgefüge und automatische Sprachverarbeitung. München: LINCOM. *MediaWiki*. Accessed at: [www.wikipedia.org](http://www.wikipedia.org) [11/04/2014].
- Pottelberge van, J. (2001). Verbonominale Konstruktionen, Funktionsverbgefüge. Vom Sinn und Unsinn eines Untersuchungsgegenstandes. Heidelberg: Universitätsverlag C. Winter.
- Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka, A. Horák (eds.): *Proceedings of Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008*. Brno: Masaryk University, pp. 6-9.
- Seifert, J. (2004). Funktionsverbgefüge in der deutschen Gesetzessprache (18. – 20. Jahrhundert). Hildesheim et al.: Georg Olms.
- Storrer, A. (2007). Corpus-based investigations on German support verb constructions. In Ch. Fellbaum (eds.) *Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies*. London et al.: Continuum, pp. 164-187.
- WebLicht*. Accessed at: <https://weblicht.sfs.uni-tuebingen.de> [11/04/2014].



# Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies

Iztok Kosem<sup>1</sup>, Polona Gantar<sup>2</sup>, Nataša Logar<sup>3</sup>, Simon Krek<sup>4</sup>

<sup>1</sup>Trojina, Institute for Applied Slovene Studies

<sup>2</sup>Fran Ramovš Institute for the Slovene Language, ZRC SAZU, Ljubljana, Slovenia

<sup>3</sup>Faculty of Social Sciences, University of Ljubljana

<sup>4</sup>Jožef Stefan Institute, Ljubljana

iztok.kosem@trojina.si, apolonija.gantar@guest.arnes.si,

natasa.logar@fdv.uni-lj.si, simon.krek@guest.arnes.si

## Abstract

Due to increasingly large amounts of authentic data to analyse, lexicographers are nowadays looking to language technologies to provide them with not only the tools to analyse the data, but also with tools and methods that ease and speed up the data analysis. One of the most promising avenues of research has been the automation of early stages of the corpus data analysis, with the aim to summarize, and consequently reduce, the amount of corpus data that the lexicographers need to examine. However, most of this research deals with general lexicography; terminology is yet to extensively test these methods. This paper attempts to address this gap by presenting two separate Slovene research projects, one lexicographic (Slovene Lexical Database) and the other terminological (Termis), that used the same method of automatic extraction of corpus data (presented in Kosem et al. 2013). After describing the projects and the corpora use, similarities and differences in the parameter settings and the quality of extracted data in the two projects are presented. We conclude with discussing the further potential of automation in both general and specialised lexicography.

**Keywords:** data extraction; terminology; general language; collocations; dictionary; GDEX

## 1 Introduction

In recent years, lexicography has witnessed several projects where automation of different aspects of lexicographer's work has been successfully implemented, such as detection of new words or meanings (Cook et al. 2013) or initial data extraction (Kosem et al. 2013). This trend of increasing the role of a computer in the dictionary-making process follows Rundell and Kilgarriff's (2011) vision of focusing lexicographer's tasks towards validating and completing the data extracted by a computer.

The calls for automation originate mainly from general lexicography where lexicographers are faced with increasingly larger corpora that they need to analyze. But what about using automation in the making of dictionaries, such as terminological dictionaries, where much smaller and more specialized corpora are used? To what extent can automation methods used in general lexicography be trans-

ferred to specialized lexicography or terminology? This paper attempts to provide some answers to these questions by describing and discussing two Slovenian projects, one lexicographic (Slovene Lexical Database) and the other terminological (Termis), that tested the use of automation in database compilation.

We first briefly describe both projects, and the corpora used for automatic extraction of data. This is followed by the description of the automatic process, and an overview of the settings used in the two projects. Then, the findings are presented, focusing on the differences as well as similarities identified between the results of automatic data extraction in the two projects. We conclude with some thoughts on the further potential of automation in both general and specialised lexicography, and outline our plans for the future.

## 2 Slovene Lexical Database

The Slovene Lexical Database (SLD; Gantar & Krek 2011) is one of the results of the Communication in Slovene<sup>1</sup> project that has developed language data resources, natural language processing tools and resources, and language description resources for Slovene. The Slovene Lexical Database has a twofold goal: it is intended as the basis for the future compilation of different dictionaries of Slovene, both monolingual and bilingual, and as such its concept is biased towards lexicography. Secondly, it will be used for the enhancement of natural language processing tools for Slovene. The database is conceptualized as a network of interrelated lexico-grammatical information on six hierarchical levels with the semantic level functioning as the organizing level for the subordinate ones. The six levels are:

- lemma or the headword,
- senses and subsenses (labelled with semantic indicators and in many cases described with semantic frames),
- multi-word expressions,
- syntactic structures (representing a formalization of typical patterns on the clause and phrasal level),
- collocations, and
- corpus examples.

---

1 The operation is partly financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia. The operation is being carried out within the operational program Human Resources Development for the period 2007–2013, developmental priorities: improvement of the quality and efficiency of educational and training systems 2007–2013. Project web page: <http://eng.slovenscina.eu/>.



### 3 Terminological database Termis

Applied research project Termis<sup>2</sup> took place between 2011 and 2013. The aim of the project was the compilation of an online dictionary-like terminological database for the discipline of public relations. The basis of the project was KoRP,<sup>3</sup> a corpus of public relations texts (Logar 2013). It has been envisaged from the beginning that the entries in the terminological database would contain English translations of headwords, explanations, syntactic and collocational information, and corpus examples. The database is now completed and is freely accessible online at <http://www.termania.net>. It comprises 2000 entries that also offer links to the KoRP corpus and the Gigafida corpus, a reference corpus of Slovene.

### 4 Using automatic data extraction in the two projects

The decision to use automatic extraction of lexical information from the corpus in both projects comes from the need to reduce time and cost connected with the production of dictionaries, by utilizing new possibilities offered by state-of-the-art tools for corpus analysis. The main idea behind using automatic extraction of corpus data is to reduce the amount of time spent by lexicographers on examining corpus data, especially on browsing through plethora of corpus examples. Lexicographic analysis remains corpus-based (or driven); however, the initial selection of corpus data to be analysed is left to the computer. The lexicographer then examines, validates, and completes the information and shapes it into the final dictionary entry.

The automatic method used in the two projects relies heavily on Word Sketch (Kilgarriff & Tugwell 2002) and GDEX (Good Dictionary Examples; Kilgarriff et al. 2008), two functions that are part of the Sketch Engine corpus tool. The method requires a lemma list, sketch grammar for the building of word sketches, GDEX configuration(s), and settings that set thresholds for data extraction. An API script is then used to extract from the corpus collocates under grammatical relations, defined in the sketch grammar, and examples of their use. The method is described in more detail in Kosem et al. (2013), thus the next sections focus on the main differences in the automatic method used by the two projects.

#### 4.1 Corpora

The basis for the extraction of lexical information for the Slovene Lexical Database was the Gigafida corpus<sup>4</sup> (Logar Berginc et al. 2012), containing 1.18 billion words or 39,427 texts created between 1990

---

2 <http://www.termis.fdv.uni-lj.si/>

3 [http://nl.ijs.si/noske/sl-spec.cgi/first\\_form?corpname=korp\\_sl](http://nl.ijs.si/noske/sl-spec.cgi/first_form?corpname=korp_sl)

4 <http://www.gigafida.net/>

and 2011 with printed texts representing 84.35% and internet texts 15.65%. Printed part contains fiction (2%), non-fiction and textbooks (4%), and periodicals such as daily newspapers (56%) and magazines (21%). Text originating from the web were published on news portals, pages of large Slovene companies and more important governmental, educational, research, cultural and similar institutions. Automatic extraction of lexical information for the Termis project was conducted on a much smaller, specialised corpus – the KoRP corpus – containing 1.8 million words. The texts in the KoRP corpus were selected according to carefully designed criteria (Logar 2007) that make the corpus representative of a public relations field in Slovenia. It is important to note that the two corpora were lemmatised and morphosyntactically tagged with the same statistical tagger (Grčar, Krek & Dobrovoljc 2012), enabling comparisons of extracted data.

## 4.2 List of lemmas

The two projects used completely different approaches to devising a list of lemmas for automatic extraction. For the Slovene Lexical Database, a more homogenous group of lemmas was used, mainly comprising of not too frequent lemmas that were either monosemous or less polysemous according to sloWNet, a Slovene version of Wordnet (Fišer, 2009). Less polysemous nature of lemmas also enabled a better comparison of data extraction with the Termis project, given that the terms in Termis were mainly monosemous. An additional criterion for selection, which was preferred but not mandatory, was the absence of the lemma in the Dictionary of Standard Slovenian (SSKJ). The final selection included 515 nouns, 260 verbs, 275 adjectives and 117 adverbs and was dominated by lemmas with frequency between 1000 (0.85 per million words) and 10,000 (8.5 per million words).

In the Termis project, the lemma list was in fact a headword list and was built using a term extraction tool (Vintar 2010)<sup>5</sup>. The list contained 2127 items: 941 nouns, 199 verbs and 987 multi-word terms. Single- and multi-word term candidates have been extracted using morphosyntactic patterns and term weights, calculated by comparing their frequencies in the KoRP corpus and in a reference corpus of Slovene called FidaPLUS (Arhar Holdt & Gorjanc 2007), as well as phraseological stability of the extracted terminological unit. Each term candidate was carefully examined in its natural environment – the texts in the KoRP corpus – by a terminologist and experts in the field of public relations.

## 4.3 GDEX configurations

The GDEX tool (Kilgarriff et al. 2008) ranks corpus examples according to their quality, using measurable parameters such as example length, whole sentence form, syntax, and presence/absence of rare words, etc. The majority of work associated with devising GDEX configurations for automatic extraction was done during the SLD project; drawing on the experience in developing the first version of

---

5 <http://lojze.lugos.si/cgittest/extract.cgi>

GDEX for Slovene (Kosem et al., 2011), four different configurations were designed, one for each word class in the SLD (noun, verb, adjective, adverb), the process involving several iterations of evaluation and comparison of results produced by the last two versions of configuration (Kosem et al. 2013). A good indication of the difference between the first version of GDEX for Slovene and the version for automatic extraction is that the former was designed to provide at least three good examples among the ten offered, while the latter aimed to have the top three examples meet the criteria of a good example. The Termis project's point of departure was using the final GDEX configurations used by the SLD project, evaluating them on a sample of lemmas and making adjustments to the heuristics, which proved to be minor, until the results were satisfactory. In the end, two different GDEX configurations were used, one for nouns and multi-word units, and one for verbs.

#### 4.4 Settings for extraction

This part of the automatic extraction introduced the greatest number of differences between the two projects. Preparation of settings for extraction included providing values for the following six parameters:

- number of examples per collocate
- number of collocates per grammatical relation
- minimum frequency of a collocate
- minimum frequency of a grammatical relation
- minimum salience of a collocate
- minimum salience of a grammatical relation.

For the SLD project, three examples per collocate were extracted, and for the Termis project two examples per collocate. Both projects used a limit of maximum 25 collocates per grammatical relation. The values of the remaining four parameters had to be obtained with statistical and manual analysis of the word sketches of a sample of lemmas used in automatic extraction. Namely, initial tests during the SLD project showed that the same values could not be used for all grammatical relations and collocates; for example, more salient and frequent relations of word classes (e.g. *adjective + noun* for adjectives) required higher thresholds due to a large number of collocates. Also, corpus frequency of the lemma played a vital role in setting the values; more frequent lemmas had more extensive word sketches and required higher thresholds, whereas rarer lemmas required lower thresholds or no thresholds at all. Consequently, both projects divided lemma lists into different frequency groups, with different settings used for each group. The SLD project used three frequency groups for each word class, with different frequency ranges for different word classes. On the other hand, the Termis project used three frequency groups for verbs, four frequency groups for nouns, and three frequency groups for multi-word units. Each category in the Termis project contained one group, the so-called 0 group, that included low frequency lemmas for which all the data available in the word sketches was extracted.

The only values that were shared by the two projects were values for minimum collocation salience for nouns and values for minimum gramrel salience for verbs; all other values were (much) lower for the Termis project than for the SLD project. This was a direct result of the difference in the sizes of the corpora used for automatic extraction of data.

#### 4.5 Extracted lexical information: general language vs. specialized language

It is worth noting that a term as a name for a concept in a certain discipline is more difficult to specify than it is presented and argued in the general theory on terminology (Wüster 1931; Felber 1984) – at least if terms are observed and identified in the context (Pearson 1998, as well as other perspectives, e.g. Cabré Castelví 2003). Such complexity of terms has been adequately summarized by Sager (1998/99) who argued that terms are merely words with a specific function, or in other words, terms are formally not very different from other words. This fact causes great difficulties to terminographers during preparatory stages, i.e. while preparing the headword list; on the other hand, this similarity between terms and other words is an advantage during the extraction of lexical context, as terminography can utilize lexicographic knowledge and tools for the analysis and description of a general language.

So far, we have compared grammatical relations/syntactic structures found in both Slovene Lexical Database and Termis, using a smaller number of noun entries that have a higher frequency per million words in the KoRP corpus than in the Gigafida corpus. The analysis showed that a large percentage of words acquire the specialised meaning only at a context level, especially with compounds or when we are dealing with polysemous words that have one of their meanings used also in a specialised domain or have developed their own specialised meaning.

The comparative analysis also focused on identifying syntactic structures common to both the general corpus (Gigafida) and the specialised corpus (KoRP), more specific to one of the corpora, or exclusive to one of the corpora. Similar comparison was made for collocations in both vocabularies. The sketch grammar contains 258 grammatical relations functioning as syntactic structures, and the automatically extracted data for noun entries showed that there were 69 (27%) attested syntactic structures, i.e. structures with identified collocates, in both corpora, 188 (73%) syntactic structures were found only in the Gigafida corpus, while one syntactic structure was found only in the KoRP corpus. These findings confirm that terminology does not differ from general language on a syntactic level, i.e. does not form terminology-specific syntactic structures. There are exceptions, however they are specific to particular lexical items; thus, a syntactic structure can be found in the language, but is not typical for a specific verb, noun, adjective etc. as used in the general language. An example of this is the structure VERB + NOUN<sub>4</sub> for the collocation *communicate message*, which is typical for the field of public relations, but not for general Slovene where the pattern *communicate + about + NOUN<sub>5</sub>* is more commonly used.

## 5 Discussion

The automatic extraction approach proved successful in both projects, in terms of providing good enough data for devising database entries and saving a great deal of lexicographer's/terminologist's time spent on more routine tasks. One of the important findings is that the steps used in the general language project (SLD) could be replicated in the terminological project (Termis), with some elements requiring little change (e.g. GDEX configurations) or no change at all (e.g. sketch grammar). Also, the evaluation of extracted corpus sentences in both projects reported good quality of the examples. The comparison clearly shows that the main work on any future project adopting this methodology would be dedicated to determining the settings for data extraction. Namely, this step exhibited the greatest differences between the projects, mainly on account of a significant difference in the size of the corpora used for automatic extraction.

Different nature of projects also enabled us to evaluate and test the approach on different lemmas in terms of corpus frequency and consequently in the amount of corpus data available. In SLD, the minimum frequency of a lemma was 600 occurrences (0.5 times per million words)<sup>6</sup> in the Gigafida corpus, whereas the threshold in Termis was determined by terminological potential of the word rather than its frequency (for example, some terms had only two or three occurrences in the KoRP corpus<sup>7</sup>). For high frequency lemmas, more work on settings for extraction was required in order to find the right balance between exporting enough data and excluding irrelevant grammatical relations and/or collocates. For very rare lemmas, i.e. for those in groups 0 in the Termis project<sup>8</sup>, it was established that the value of the automatic approach is mainly in saving lexicographer's time by directly exporting all the data for each lemma and importing it into the dictionary-writing system, thus changing the lexicographer's task from analysis-selection-copying to validation-deletion.

The automatic extraction of data for multi-word units was conducted only for Termis, as the project was conducted after the conclusion of the SLD project when a new feature called Multi-word links had already been implemented in the Sketch Engine. The automatic extraction of lexical information was only possible for two-word patterns such as *adjective + noun* and *noun + noun*, and not for others (e.g. *noun + preposition + noun*). It is therefore not possible to make comparisons of the projects as far as automatic extraction of data for multi-word units is concerned. Nonetheless, we can report that the data obtained in the Termis project was found to be of similar quality as the data for single-word terms, with the main difference being in the GDEX configuration and settings used.

What is left for lexicographers to do are tasks such as sense division, definition writing, distributing and cleaning the automatically extracted information etc.; and as shown by studies such as Kosem et al. (2013), some of those tasks can be left to non-lexicographers, e.g. by using crowd-sourcing. Further-

---

6 Majority of lemmas had frequency between 1000 (0.85 per million words) and 10,000 (8.5 per million words) in the Gigafida corpus.

7 This still meant that these terms had a higher frequency per million words (1.1) than the least frequent lemmas in the SLD project.

8 Groups 0 contained 889 terms in total.

more, the feedback from the terminologists devising entries in the Termis project showed that many extracted examples already contained definitions of terms or at least the information needed to devise them, indicating further avenues for the implementation of automation. It is noteworthy that the entries in the Termis database contain (encyclopaedic) definitions that are short (one sentence), medium-length (multi-sentence paragraph) or even longer (several paragraphs); in general they are longer than definitions (or semantic frames) in the Slovene Lexical Database.

## 6 Conclusion

Technological advances gave rise to corpora, enabling lexicographers to describe language more accurately and in greater detail than ever before, but ironically, corpora have now become a problem for lexicographers due to the increasingly larger amounts of data they contain. Consequently, it seems inevitable that more and more lexicographic tasks will become automated. There is simply too much data to analyse and not enough time to do it in – in addition, users want quick(er) access to up-to-date information. Initial experience on a Slovene lexicographic project has showed promising results, but it is even more encouraging that the automatic approach appears to be suitable also for terminological purposes.

The automatic method by Kosem et al. (2013) has the most potential for projects where a dictionary or a database is devised from scratch,<sup>9</sup> but it can also be useful for existing dictionaries. For example, periodical automatic extraction of regularly updated corpus data could facilitate quicker detection and description of new meanings and usages of the words. This remains one of the avenues of future research; namely, how to automatically extract and include in the database only the new information on the use of a particular word or phrase. By this we do not mean only new words and meanings, but also new uses of existing meanings.

Future plans as far as the Slovene Lexical Database is concerned include a more in-depth evaluation of entries devised with automatically extracted data, as well as their comparison with manually devised entries. We also aim to test automatic extraction on more frequent lemmas, where we expect much more work with setting parameters for extraction. Further use of the automatic approach is planned on the terminological side, possibly by testing its usefulness in a few other domains. Finally, we aim to explore automatic extraction of information not covered by the existing automatic method. One of such areas is definition extraction; for example, future plans with the Termis database include conducting an experiment on automatic definition extraction from the KoRP corpus, using the recently-developed methodology, specially adapted for Slovene (Pollak 2014).

---

9 For example, the automatic data extraction method is an integral part of a proposal for a new dictionary of contemporary Slovene (Krek et al., 2013).

## 7 References

- Arhar Holdt, Š., Gorjanc, V. (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovnstvo*, 52(2), pp. 95-110.
- Cabré Castellví, M. T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology* 9(2), pp. 163-199.
- Felber, H. (1984). *Terminology Manual*. Paris: Infoterm.
- Fišer, D. (2009). SloWNet – slovenski semantični leksikon. In M. Stabej (ed.) *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 145-149.
- Cook, P., Lau, J. H., Rundell, M., McCarthy, D., Baldwin, T. (2013) A lexicographic appraisal of an automatic approach for detecting new word senses In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 49-65.
- Gantar, P., Krek, S. (2011). Slovene lexical database. In D. Majchraková, R. Garabík (eds.) *Natural language processing, multilinguality: sixth international conference, Modra, Slovaška, 20-21 October 2011*, pp. 72-80.
- Grčar, M., Krek, S., Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In T. Erjavec, J. Žganec Gros (eds.) *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 89-94.
- Logar, N. (2007). Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah: doktorska disertacija. Ljubljana: Filozofska fakulteta.
- Logar, N. (2013). *Korpusna terminografija: primer odnosov z javnostmi*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.
- Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š. & Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the 13th EURALEX international congress*. Barcelona: Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, pp. 425-432.
- Kilgarriff, A., Tugwell, D. (2002). Sketching words. In H. Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*. Euralex, pp. 125-137.
- Kosem, I., Gantar, P., Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 32-48.
- Kosem, I., Husák, M., McCarthy, D. (2011). GDEX for Slovene. In I. Kosem, K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New applications for new users*, Proceedings of eLex 2011, Bled, 10-12 November 2011. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 151-159.
- Krek, S., Kosem, I., Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*, v1.1. Available at: [http://www.sssj.si/datoteke/Predlog\\_SSSJ\\_v1.1.pdf](http://www.sssj.si/datoteke/Predlog_SSSJ_v1.1.pdf)
- Pollak, S. (2014). *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov/Semi-automatic domain modeling from multilingual corpora (Semi-automatic Domain Modeling from Multilingual Corpora)*. PhD thesis. Ljubljana: University of Ljubljana, Faculty of Arts, Department of Translation. Accessed at: [http://kt.ijs.si/theses/phd\\_senja\\_pollak.pdf](http://kt.ijs.si/theses/phd_senja_pollak.pdf). [25/03/2014]
- Pearson, J. (1998). *Terms in context*. Amsterdam, Philadelphia: John Benjamins.

- Rundell, M., Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds.). *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Amsterdam: Benjamins, pp. 257-281.
- Sager, J. C. (1998/99). In Search of a Foundation: Towards the Theory of the Term. *Terminology*, 5(1), pp. 41-57.
- Vintar, Š. (2010). Bilingual term recognition revisited: the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2), pp. 141-158.
- Wüster, E. (1931). *Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik*. Berlin: VDJ.



# A Corpus-based Dictionary of Polish Sign Language (PJM)

Jadwiga Linde-Usiekniewicz, Małgorzata Czajkowska-Kisil, Joanna Łacheta, Paweł Rutkowski  
University of Warsaw  
jlinde@uw.edu.pl, maczajkis@wp.pl,  
j.lacheta@uw.edu.pl, p.rutkowski@uw.edu.pl

## Abstract

The aim of this paper is to give a general overview of an on-going lexicographic project devoted to Polish Sign Language (PJM), a natural language used by the Deaf community in Poland. The project in question will result in the first-ever PJM dictionary based on extensive corpus data (encompassing more than 300 hours of video material recently collected by the Section for Sign Linguistics of the University of Warsaw). The present article discusses the most important assumptions and methodological foundations of the current PJM dictionary project and confronts them with previous work on PJM and Signed Polish glossaries, as well as with international standards in contemporary sign language lexicography. The design of the new PJM dictionary is discussed in detail, including the most problematic issues, such as lemmatization, sense division and sense description principles. Sample entries are given as illustration. It is important to note that apart from filling an important gap in the availability of sign language teaching and learning materials in Poland, the PJM dictionary outlined in this paper is also likely to further the recognition of PJM as a full-fledged natural language.

**Keywords:** Polish Sign Language (PJM); corpus-based dictionary; sign language lexicography

## 1 Introduction

Polish Sign Language (*polski język migowy*, usually abbreviated as PJM) is a natural visual-spatial language used by the Deaf community in Poland. It emerged around 1817, with the foundation of the first school for the deaf in Warsaw, and has been continually in use since then. The current number of PJM users is estimated to exceed 50,000. Despite being one of the largest minority languages in Poland, PJM has not – until recently – attracted much attention from the hearing linguistic community. The first-ever academic unit specializing in research on the grammatical and lexical properties of PJM was created at the University of Warsaw in 2010 (the Section for Sign Linguistics, SSL). The present paper is devoted to a large-scale research project that is currently being developed by the SSL team with the aim of creating a corpus-based dictionary of PJM.

## 2 Previous PJM lexicography

As happened with most sign languages worldwide (Zwitserlood 2010 and the references therein), early dictionaries of PJM were relatively simple glossaries or wordlists. The first one, published in 1879 and reprinted as Hollak et al. 2011, featured spoken Polish words as headwords, while the body of the entry featured descriptions of how the sign was actually produced or, if the sign language equivalent of a Polish word consisted of a combination of signs, the component signs (described elsewhere in the dictionary) were listed. Thus the entries read as follows (translation ours): “ALLOW – move the hands down in front of oneself and nod seriously”; “WIT – the sign *quick thinking* followed by the sign *wisdom*” (Hollak et al. 2011). The dictionary was designed for families of deaf people and for educators. Despite its simplicity, it is a very important source of knowledge on the history of the PJM lexicon.

Again, as in most countries, sign language was banned from Deaf education in Poland for most of the 20<sup>th</sup> century. The departure from strict oralism in Deaf education and the reintroduction of signs was followed by the creation and establishment of a hybrid artificial code, called the Language-Sign System (*system językowo-migowy*, usually abbreviated as SJM), to be used in Deaf schools as the official language of instruction. SJM was a combination of a set of PJM signs, other artificially created signs and Polish grammar, and is, therefore, often referred to as Signed Polish. Importantly, natural sign language signs were supplemented with signs created as word-formation calques from Polish; the syntax was that of spoken Polish, and in its most elaborate form, it involved finger-spelled Polish morphological affixes.

At that time another dictionary, or rather glossary, was published (Hendzel 1986/2006), in which signs were represented by black-and-white photos of people producing them, with arrows representing movement. The Polish wordlist was based on Bartnicka & Sinielnikoff’s (1978) learner’s dictionary of Polish. For each sign, spoken Polish equivalents were added, e.g. for a sign referring to friendship the equivalents are Polish lexemes meaning ‘friendship’, ‘friendly’, ‘friendlily’ ‘to be friends’, ‘male friend’, ‘female friend’. The dictionary presents both true PJM signs and signs that are used in SJM exclusively (cf. Ruta & Wrześniewska-Pietrzak 2013).

The latest general lexicographical publication (Kosiba & Grenda 2011) follows roughly the same principle, though the pictures are in color. Both Hendzel (1986) and Kosiba & Grenda (2011) contain over 2000 signs. Both publications arrange their entries by alphabetical order of Polish words. In the latter publication the headword list was compiled on the basis of various sources, including specific purpose glossaries and phrasebooks, based on both SJM (e.g. Szczepankowski 2000) and PJM (Grzesiak 2008, 2010a,b,c).

In general, for all the lexicographic work mentioned above (Grzesiak’s phrasebooks being a notable exception), the lexicographic procedure for eliciting appropriate signs was based on Polish language lexemes, with all the possible dangers and actual errors this procedure may entail (Zwitserlood 2010: 455).

### 3 Sign language lexicography world-wide

In the meantime, sign language lexicography world-wide has taken a radical turn from glossaries to descriptive or explanatory dictionaries. The explanatory character of such dictionaries has consisted in the meaning of each sign being given not as a spoken language equivalent, but as a sense definition. As such, entries in sign language dictionaries started to conceptually resemble those of monolingual dictionaries of spoken language, though in contrast to true monolingual dictionaries the lexicographic description was not provided in the same language as the headword: thus in the American Sign Language Dictionary (Costello 1998) the sign referring to ‘bear’ (animal) is identified as a noun and provided with an English definition “[...] a large, heavy mammal with thick, rough fur [...]” (Zwitserslood 2010: 447). This model was also adopted for the Auslan dictionary (Johnston 1989), where the English-language definitions were adopted from the Collins-Cobuild dictionary (Johnston, p.c.). Thus the two explanatory dictionaries were not in fact truly monolingual; a true monolingual sign language dictionary should have the definitions signed (Johnston, p.c.). Yet there is no actual dictionary of a sign language that uses the sign language as the lexicographic metalanguage (Kristoffersen & Troelsgard 2012, Zwitserslood, Kristoffersen & Troelsgard 2013). Both ASL and Auslan dictionaries were originally published as books, with signs represented by drawings. In both publications the headsigns (or lemmas) were ordered not by alphabetical order of English-language equivalents, but by formal features of signs: handshape, hand orientation, sign location, direction of movement (for more detailed discussion of sign representation systems both in printed and on-line dictionaries see Zwitserslood 2010; Kristoffersen & Troelsgard 2010, 2012; Zwitserslood, Kristoffersen & Troelsgard 2013). Nowadays, sign language dictionaries are generally fully electronic, with videos representing headsigns and even sign language examples, as can be seen for instance in the Danish Sign Language Dictionary and Finnish Sign Language Dictionary. Moreover, signs can be accessed not only by formal features but also by topics, and the dictionaries can be used as thesauri as well.

### 4 Research into PJM – grammar and dictionary

The changes in sign language lexicography are part of a general advance in sign linguistics, brought about by modern video and IT technology that allowed for compilation and analysis of sign language corpora (Crasborn *et al* 2008). A sign language corpus is also being compiled in Poland. The PJM corpus project was launched in 2010 and its first phase will conclude in 2014. The underlying idea is to compile a collection of video clips showing Deaf people (native signers) using PJM in a variety of different contexts. As of early 2014, more than 80 people have already been filmed. When the project is completed, approximately 300 hours of footage will be available for research purposes. The PJM corpus is diversified geographically, covering more than 10 Polish cities with significant Deaf populations. The

group of signers participating in the project is well balanced in terms of age, gender, as well as for social and educational background (respective sociological metadata is an integral part of the corpus). Recording sessions always involve two signers and a Deaf moderator. The procedure of data collection is based on an extensive list of tasks to be performed by the two informants. Typically, the signers are asked to react to certain visual stimuli, e.g. by describing a scene, naming an object, (re-)telling a story, or explaining something to their partner.

The elicitation materials include pictures, videos, graphs, comic strips, etc., with as little reference to written Polish as possible. The participants are also requested to discuss a number of topics pertaining to the Deaf. Additionally, they are given some time for free conversation (they are aware of being filmed but no specific task is assigned to them). The latter two parts of the recording session scenario are aimed at collecting spontaneous and naturalistic data.

The raw material obtained in the recording sessions is further tokenized, lemmatized, annotated, glossed and translated using the iLex software developed at the University of Hamburg (Hanke & Storz 2008). The annotation conventions employed have been designed especially for the purposes of PJM (cf. Rutkowski, Łozińska, Filipczak, Łacheta & Mostowski 2014).

The two basic outcomes of the project are the compilation of a PJM dictionary and a grammar of PJM. The two have to be compatible in terms of underlying theoretical assumptions about sign language linguistics; moreover, it has been decided that the lexicography part of the project has to suit the grammatical description and not vice versa.

The methodology used to produce the PJM dictionary is based on methodologies established for compiling corpus-based spoken language dictionaries. The lexicographers working on the project are native signers, either hearing bilinguals (children of Deaf parents) or Deaf signers with near-native fluency in spoken/written Polish, all of them trained in monolingual lexicography. Thanks to iLex, they have access to all tokens (occurrences) of a particular sign and on the basis of this usage data they establish the meaning of each sign. It is a challenging procedure that needs to take into account various issues such as homonymy versus polysemy, division into syntactic categories, sense division, and adequate sense description, described below.

## **5 The design of the PJM dictionary**

### **5.1 Target audience and general purpose**

Whenever any dictionary is compiled, whether of sign or spoken language, the basic question that has to be addressed is what kind of purpose it is meant to serve and, in consequence, what kind of user group it is aimed at. In terms of purpose, dictionaries can be divided into two types: descriptive or explanatory, i.e. monolingual and bilingual. Importantly, in lexicographic tradition it is the expla-

natory dictionaries that have always been kept in higher regard: having an explanatory dictionary devoted to it gives a language stature and recognition. This socio-cultural and sociolinguistic aspect of lexicography has led us to design our dictionary as primarily explanatory. However, since it is not the Polish Deaf community that needs to be made to recognize PJM as a separate language of a linguistic minority, but the general society, compiling a truly monolingual explanatory dictionary of PJM would not only be technically and methodologically feasible, but also would defeat the very purpose of giving PJM its due recognition. That is why we opted for a hybrid explanatory dictionary, i.e. one in which the lexicographic description is given in spoken Polish, as was done in the ASL and Auslan dictionaries mentioned above. In spite of Zwitterlood's (2010: 463-464) criticism of the apparently similar procedure adopted by the publishers of the Dutch Sign Language dictionary (Schermer & Koolhof 2009), we believe that it offers several important advantages.

First of all, the meanings of signs are explicitly described and not left to be inferred from a set of Polish equivalents, which would have been the case had we followed the bilingual-like model that has recently gained popularity in sign language lexicography. There is a general risk of wrong inferences being drawn about the meaning of the source language item from the translation equivalents (Linde-Usiekniewicz & Olko 2006) if semantic equivalents are interpreted as translation equivalents or vice-versa (see Piotrowski 1989 for the distinction, and Piotrowski 1994: 104-155 for more detailed discussion of equivalence in general). While in the Danish dictionary this risk is minimized by linking Danish equivalents to their corresponding entries in a general Danish explanatory dictionary (Kristoffersen, p.c.), this option was not available to us.

Explicit semantic description of sign meanings makes the dictionary a useful source for research into the PJM lexicon and PJM semantics, independently of the researchers' actual proficiency in PJM. It is also a valuable source material for hearing people (native speakers of spoken Polish) learning PJM.

Paradoxically, while the design of the dictionary is such that it gives grounds for the sociolinguistic recognition of PJM, it might appear as if it were of less practical value to the Deaf community. Yet this is not the case. The majority of Polish Deaf are in fact bilingual and use spoken Polish (through lip-reading for speech and in written form). For more technical and academic texts they use Polish language dictionaries, which, in their great majority, are meant for native Polish users. Thus the PJM dictionary in its present form is a kind of a bridge dictionary (Williams 2008), which offers training in the way word senses are being defined in Polish lexicography. Yet since the original bridge dictionaries of English have English lemmas and English Collins-Cobuild definitions translated into the native language of the learners, the PJM dictionary, seen in that light, is actually the reverse combination: sign language lemmas and Polish definitions.

## 5.2 Directionality and looking-up options

The dictionary is conceived in principle as unidirectional, with sign language signs as lemmas. Nevertheless, when appropriate, for each defined sense, a Polish equivalent or equivalents are provided, and all equivalents are listed alphabetically, so the reverse direction is also available. If several entries share the same Polish equivalent, it is cross-referred to all of them. In the reverse looking-up mode, the definitions provide cues to the appropriate sense of a polysemous or even homonymous Polish word.

PJM signs can be identified by the aforementioned features of handshape, orientation, location and movement, as it is done in all sign language dictionaries.

## 5.3 Lemmatization: homonymy vs. polysemy

There are two ways in which lemmatization can be carried out in sign lexicography (and in spoken language lexicography as well): purely on formal grounds and on the basis of a combination of both formal and semantic criteria. For reasons of convenience, we opted for purely formal criteria, with no homonymous entries, i.e. with homonyms described in a single entry. Nevertheless, at the entry level homonymy-like phenomena are differentiated from polysemy. In some cases the homonymy is obvious, as the attested meanings fall into discrete bundles of related senses. For example there is a sign that may refer to an ache, pain and related phenomena, as well as to the function of director, directorship, etc. However, since there is a strong tendency for iconicity to be the motivating factor for a sign's signifier, a less obvious example consists of a single sign referring to a crown, and by the same token to a monarch, ruling, etc., as well as to someone having a spherical object on their head (Figure 1). It could be argued that the two groups of senses share the same etymology or motivation so no homonymy is involved, yet on the other hand the semantic relation is nothing but tenuous. In the case of homonymy-like phenomena the entry is split into several sections devoted to macrosenses, i.e. bundles of related senses, as shown in Figure 1. The decision as to whether a given entry should be split or not lies with the lexicographers, who nevertheless often consult other native signers to verify their intuitions about relevant groups of senses being interconnected or not.

## 5.4 Syntactic issues

Another challenge lies in classifying senses and usage-types as belonging to a syntactic category. Sign languages, including PJM, have almost no inflection and parts of speech (POS) cannot be established on morphological grounds. Syntactic criteria are not functional enough, as research into PJM to date has tended to show that the same sequence of signs can be interpreted as corresponding to a complex nominalization or to a full clause. Moreover, in contrast to spoken languages for which there are established grammars, the grammar of PJM is being investigated in parallel, on the basis of the same cor-

pus. However, identifying the syntactic character of a sign is necessary in order to provide the most adequate semantic definition of a given sense in Polish. At the same time, the way a given definition is formulated in Polish, i.e. using Polish verbs, nouns or adjectives as key elements, could erroneously suggest the POS of the sign being defined: using infinitives in definitions suggests a uniquely verbal character, using nouns suggests a nominal character, etc. In order to match senses with definitions we decided to eschew POS identification for signs and provide usage-type information instead. The usage-type information reflects the syntactic category identified on a semantic basis (Wierzbicka 2000; Schwager & Zeshan 2008).

Thus the syntactic properties of a sign are introduced in the dictionary as different types of usage (see Figure 1 and 2). The main usage types are named with terms based on the traditional Polish POS system. More importantly, irrespectively of their correspondence or non-correspondence to Polish POS, they are seen as falling into two largely distinct groups: one corresponding to words with predominantly conceptual meaning and the other corresponding to words with predominantly procedural meaning (Wilson 2011). Signs belonging to the first group are described as having: verb-like usage (i.e. referring to an act, an activity, a process or an event or a situation); noun-like usage (i.e. denoting a person, an animal or an object); adjective-like usage (i.e. denoting an entity's feature or property); adverb-like usage (i.e. denoting a feature of an activity or a process); and numeral-like usage, corresponding to cardinal and ordinal numerals. The second group contains pronoun-like usage, preposition-like usage and conjunction-like usage.

Not all signs and sign senses of PJM have a corresponding POS in spoken Polish. Sign languages are known for classifier signs (Zwitserlood 2012) both of nominal and verbal character. While nominal classifier signs mainly serve to identify referents, verbal classifiers usually involve incorporating both the path of movement or direction (if applicable) and the shape of the object the sign refers to. One example of such a classifier sign was the macrosense 'having a spherical object on one's head', the second macrosense of the sign also associated with king, reign and crown (see below).

## 5.5 Sense division and sense description

For all the senses associated with conceptual meaning the appropriate definition is not presented as semantic equivalence, as has been done in Costello (1998) and in Johnston (1989), but as a description of the sign denotation or reference. Thus, the appropriate entry fragment for the 'crown/king/rule...' sign is presented in Figure 1:

[MACROSENSE] **I. REIGN.**  
 [SYNTACTIC CATEGORY] **A. in nominal usage:**  
     [MICROSENSE] **1.** denotes a person, usually of noble blood, who rules a country  
     [MICROSENSE] **2.** denotes a headgear, usually made of precious metal and precious stones,  
     that a ruler wears for ceremonial purposes.  
 [SYNTACTIC CATEGORY] **B. in verbal usage:**  
     [MICROSENSE] refers to a situation in which a person, usually of noble blood, rules a country  
 [MACROSENSE] **II. CLASSIFIER**  
 [SYNTACTIC CATEGORY] **A. in verbal usage:**  
     [MICROSENSE] refers to a situation in which somebody wears or carries something on their  
     head

**Figure 1: The entry for king/crown/rule (translated into English).**

By contrast, signs with procedural meaning (or procedural senses of a given sign) are described in terms of their function; thus the sign corresponding to the conjunction ‘or’ is described as shown in Figure 2:

[SYNTACTIC CATEGORY] *used as conjunction*  
     [MICROSENSE] **1.** is used to connect two expressions or sentences, at least one of which is true  
     [MICROSENSE] **2.** is used to connect two expressions referring to two possibilities, from which a  
     choice needs to be made.

**Figure 2: The entry for conjunction ‘or’ (translated into English).**

The sense definitions tend to be highly detailed, and at the same time, formulated in a simple language, following the spirit, but not the letter of the Collins-Cobuild project (Sinclair 1987, and particularly Hanks 1987), and its Polish counterpart, Bańko 2000. To underline the fact that the dictionary definitions describe signs and do not provide complete semantic equivalents, the definitions are formulated in terms of specific frames, already mentioned above: thus, signs in noun-like usage *denote* entities further specified in the definition; signs in adjective-like usage *describe* entities as having some property further specified in the definition and signs in verb-like usage *refer* to a situation further described in the definition. For verb-like usages the definition contextually specifies the “subject-like” and the “object-like” syntactic arguments. Directional verbs are specified as such in the syntactic category.

Sense definitions tend to be extended and comprehensive, as we are trying not to use examples either to further specify the range of meaning or to supplement an over-general definition, but to illustrate the sign use. Specifically we will be using examples to show nouns used as modifiers, or being modified by other nouns, since this is a syntactic property of sign languages, including PJM, that is absent in Polish. Adjectives will be illustrated either by attributive or predicative use, and specifically to show if they tend to be used pre- or post-nominally (Rutkowski, Łozińska, Łacheta & Czajkowska-Kisil 2013, Rutkowski, Czajkowska-Kisil, Łacheta & Kuder 2013).



In order to arrive at the actual sense division for each sign the lexicographer consults the recorded corpus, using the iLex software (Hanke & Storz 2008). Since many tokens of the same sign are repeated in recordings of different participants performing the same task (questionnaires, recounting pictures and movies) the lexicographers are not required to check all instances of use in the same context. (This is one of the advantages of working with an elicited corpus, as opposed to spoken language corpora, where corpus search brings numerous repeated instances of the word's use). However, the lexicographers are supposed to check all the instances in which the sign appears in free discourse. Since the lexicographers are themselves native signers with fully developed vocabulary (teachers of PJM and PJM interpreters) they may apply their own knowledge of the sign, confirmed by other deaf consultants, to further develop the entry, thus providing senses not attested in the corpus. On the basis of both corpus-attested usage and not corpus-attested but nevertheless confirmed usage, they establish sense division. In so doing, they are not supposed to be guided by Polish equivalents of the sign. Thus, having a single Polish equivalent does not constitute evidence for a sign having a single sense. On the other hand, having two different translation equivalents in Polish does not constitute evidence for there being two different senses, as is also the case for two spoken languages as represented in bilingual dictionaries (Bogusławski 1995, Linde-Usiekniewicz & Olko 2006, Linde-Usiekniewicz 2011, Lew 2013). The general tendency is for “lumping” as opposed to splitting, i.e. establishing senses conceptually and not by reference or denotation.

## 6 Video materials

Video clips for lemmas will be recorded in order to provide non-native users with a neat example of the sign production. Variant realizations, usually differing from the basic form by one parameter alone, will also be recorded. Within the body of the entry, some examples of actual use will be provided as clips. These would be chosen from corpus recordings to complement the information explicitly provided in the entry. For example signs with nominal usage and sense defined in the entry will be illustrated in their attributive use (i.e. where they would be translated into Polish as adjectives); ordinary verbs will be exemplified by their use with standard arguments and also in patterns where they would be translated by nominalizations. Examples will also be provided for directional verbs and for classifiers. Though the examples will be initially chosen from the corpus material they will be reproduced and re-recorded in controlled conditions, for greater clarity. Examples will be glossed and accompanied by Polish translations.

## 7 Other features

The entries will also feature information about geographical restrictions in sign use (if applicable), since some of the signs tend to be used only in some geographically restricted areas (i.e. they are regionalisms). Another feature, meant mainly for non-native signers, is that of cross-reference and comparison. The 'compare' feature will direct the user either to a sign that is produced in a similar way, and therefore may be confused, or to a sign that differs in form but has the same Polish equivalent.

## 8 Size and coverage

As to the dictionary size, in order to be as comprehensive as possible we plan to include all signs which are represented by more than 5 tokens in the corpus. However, the headsign list will have to be complemented by signs taken from other sources, since the corpus frequency is influenced by the nature of the corpus: it is an elicited corpus, based largely on specific visual stimuli, with signs corresponding to these stimuli largely overrepresented.

## 9 Concluding remarks

Overall, the main objective of the PJM dictionary project described here is to fill an important gap in the availability of sign language teaching and learning materials in Poland, by providing a dictionary of groundbreaking functionality. Moreover, it is our expectation that the resulting dictionary is likely to further the recognition of PJM as a full-fledged natural language. As such, we have offered some justification herein for the methodological assumptions and choices made in developing this project, as being appropriate for this particular set of circumstances.

## 10 References

- Bańko, M. (ed.) (2000). *Inny słownik języka polskiego*. Warsaw: Wydawnictwo Naukowe PWN.
- Bartnicka, B. & Sinielnikoff, R. (1978). *Słownik podstawowy języka polskiego dla cudzoziemców*. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.
- Bogusławski, A. (1995). Bilingual general purpose dictionary. A draft instruction with commentaries. In J. Wawrzyńczyk (ed.) *Bilingual Lexicography in Poland: Theory and Practice*. Warsaw: Katedra Lingwistyki Formalnej Uniwersytetu Warszawskiego, pp. 15-55.
- Costello, E. (1998). *Random House Webster's American Sign Language Dictionary*. New York: Gallaudet University Press.

- Crasborn, O., Hanke, T., Efthimiou, E., Zwitserlood, I., & Thoutenhoofd, E. (eds.) (2008). Construction and Exploitation of Sign Language Corpora. *3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris.
- Grzesiak, I. (ed.). (2008). Mini-rozmówki migowo (PJM)-polskie, polsko-migowe (PJM) ze słowniczkiem. Znaki migowe i przykładowe dialogi przydatne w placówkach administracji publicznej. Olsztyn: Fundacja na Rzecz Głuchych i Języka Migowego.
- Grzesiak, I. (ed.) (2010a). Mini-rozmówki migowo (PJM)-polskie, polsko-migowe (PJM) ze słowniczkiem. Znaki migowe i przykładowe dialogi przydatne w placówkach opieki zdrowotnej. Olsztyn: Fundacja na Rzecz Głuchych i Języka Migowego.
- Grzesiak, I. (ed.) (2010b). Mini-rozmówki migowo (PJM)-polskie, polsko-migowe (PJM) ze słowniczkiem. Znaki migowe i przykładowe dialogi z zakresu rozwoju zawodowego. Olsztyn: Fundacja na Rzecz Głuchych i Języka Migowego.
- Grzesiak, I. (ed.) (2010c). *Piłka nożna. Słowniczek migowo (PJM)-polski, polsko-migowy (PJM)*. Olsztyn: Fundacja na Rzecz Głuchych i Języka Migowego.
- Hanke, T. & Storz, J., (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood, E. Thoutenhoofd, (eds.) *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. Paris: ELDA.
- Hanks, P. (1987). Definitions and explanations. In J. Sinclair (ed.) *Looking up*. London: HarperCollins Publishers Limited, pp. 116-136.
- Hendzel, J. (1986/2006). *Słownik polskiego języka migowego*. Olsztyn: Rakiel.
- Hollak J., Jagodziński, T., Świdorski, T., Twardowska, E., Turkowska, K. & Dutkiewicz D. (2011). *Słownik mimiczny dla głuchoniemych i osób z nimi styczność mających*. Łódź: Polski Związek Głuchoniemych
- Johnston, T. (1989). *Auslan Dictionary: a Dictionary of Australian Sign Language (Auslan)*. Adelaide: TAFE National Centre for Research and Development.
- Kosiba, O & Grenda, P. (2011). *Leksykon języka migowego*. Bogatynia: Silentium.
- Kristoffersen, J. & Troelsgard, T. (2010). The Danish Sign Language Dictionary. In A. Dykstra & T. Schoonheim, (eds.): *Proceedings of the XIV EURALEX International Congress*. Leeuwarden: Fryske Akademy
- Kristoffersen, J. & Troelsgard, T. (2012). The Electronic Lexicographical Treatment of Sign Languages: The Danish Sign Language. In S. Granger & M. Paquot, (eds.): *Electronic Lexicography*. Oxford: Oxford University Press, pp. 293-315.
- Lew, R. (2013). Identifying, ordering and defining senses, In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London: Bloomsbury Publishing, pp. 284-302
- Linde-Usiekniewicz, J. (2011). Polszczyzna w leksykografii dwujęzycznej – dylematy i postulaty. In W. Gruszczyński & L. Polkowska (eds.). *Problemy leksykografii. Historia – metodologia – praktyka*, Kraków: Wydawnictwo Lexis, pp. 105-122.
- Linde-Usiekniewicz, J. & Olko, M. (2006). Multilingual dictionaries on-line: reality and perspectives. In V. Koseka-Toszeńska & R. Roszko (eds.) *Semantyka a konfrontacja językowa*, Vol. 3. Warsaw: SOW, pp. 43-59
- Piotrowski, T. (1989). The bilingual dictionary – a manual of translation or a description of lexical semantics. In Z. Saloni (ed.) *Studia z polskiej leksykografii współczesnej III*, Białystok: Dział Wydawnictw Filii UW, pp. 41-52.
- Piotrowski T. (1994). *Problems in Bilingual Lexicography*. , Wrocław: Wydawnictwa Uniwersytetu Wrocławskiego.
- Ruta K. & Wrześniewska-Pietrzak M. (2013). Rzecz o nieobecnych. O słownikach polskiego języka migowego (presentation). *IV Glosa do leksykografii polskiej*, Warsaw, September 22 -23, 2013
- Rutkowski, P., Czajkowska-Kisil, M., Łacheta, J. & Kuder A. (2013). The Internal Structure of Nominals in Polish Sign Language (PJM): A Corpus-based Study (poster) *Theoretical Issues in Sign Language Research 11 – TISLR 11* London, July 11.

- Rutkowski, P., Łozińska, S., Filipczak, J., Łacheta J., & Mostowski, P. (2014). Jak powstaje korpus polskiego języka migowego (PJM), *Polonica* 33.
- Rutkowski, P., Łozińska, S., Łacheta, J. & Czajkowska-Kisil M., (2013). Constituent order in Polish Sign Language (PJM), *Theoretical Issues in Sign Language Research 11 – TISLR 11* London, July 11.
- Schermer G.M., & Koolhof C. (eds.) (2009). *Van Dale Basiswoordenboek Nederlandse Gebarentaal*. Utrecht: Van Dale.
- Schwager, W., & Zeshan, U. (2008), Word classes in sign languages. Criteria and classifications. *Studies in Language*, 32 (3), pp. 509–545.
- Szczepankowski, B. (2000). *Słownik liturgiczny języka migowego*. Warsaw: Wydawnictwo św. Jacka.
- Sinclair, J. (ed.). (1987). Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary. London: HarperCollins Publishers Limited.
- Tennant, R. (1998). *The American Sign Language Handshape Dictionary*, Gallaudet University Press
- Williams, G.. (2008). A Multilingual Matter: Sinclair and the Bilingual Dictionary, *International Journal of Lexicography*, Vol. 21 , No. 3, pp. 255-266.
- Wilson, D. (2011). The conceptual-procedural distinction: Past, present and future. In V. Escandell-Vidal, M. Leonetti & A. Ahern (eds.) *Procedural meaning: Problems and perspectives* Bingley: Emerald Group Publishing Limited, pp. 3-31.
- Wierzbicka, A. (2000). Lexical prototypes as a universal basis for cross-linguistic identification of parts of speech. In P. M. Vogel & B. Comrie (eds.), *Approaches to the Typology of Word Classes*. Berlin: Mouton de Gruyter, pp. 285–317.
- Zwitsersloot, I. (2010), Sign language lexicography in the early 21st century and a recently published Dictionary of Sign Language of the Netherlands, *International Journal of Lexicography*, Vol. 23 No. 4, pp. 443–476
- Zwitsersloot, I. (2012)., Classifiers. In R. Pfau, M. Steinbach & B. Woll (eds.) *Sign Language: An International Handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science: 37), Amsterdam: De Gruyter Mouton, pp. 158-186.
- Zwitsersloot, I., Kristoffersen, J. & Troelsgard, T. (2013). Issues in Sign Language Lexicography. In H. Jackson (ed.) *The Bloomsbury Companion to Lexicography*. London: Bloomsbury Publishing, pp. 259-283.

### **Acknowledgements**

This research has been supported financially by Poland's National Science Center (*Narodowe Centrum Nauki*) within the project *Iconicity in the grammar and lexicon of Polish Sign Language (PJM)* (grant number: 2011/01/M/HS2/03661). The authors are also grateful to Daniel J. Sax for his help with the preparation of this paper.

# Laying the Foundations for a Diachronic Dictionary of Tunis Arabic: a First Glance at an Evolving New Language Resource

Karlheinz Mörth<sup>1</sup>, Stephan Procházka<sup>2</sup>, Ines Dallaji<sup>2</sup>

<sup>1</sup>Institute of Corpus Linguistics and Text Technology (Austrian Academy of Sciences)

<sup>2</sup>Department of Oriental Studies (University of Vienna)

Karlheinz.Moerth@oeaw.ac.at, stephan.prochazka@univie.ac.at, ines.dallaji@univie.ac.at

## Abstract

Arabic lexicography has a long tradition. However, at the time of writing this report, there exist only a very few digital products, let alone products documenting Arabic dialects. Our paper presents the TUNICO project (Linguistic Dynamics in the Greater Tunis Area) and digital language resources which are being produced as part of the project. The TUNICO working group is working on a digital diachronic dictionary of Tunis Arabic which is being compiled as part of a larger linguistic endeavour to document the variety of the Tunisian capital. One of the interesting features of the project is that it draws on a number of heterogeneous sources: text books, grammatical descriptions and a corpus of spoken youth language which is currently being compiled. In this project, the dictionary is used as an analytical tool, as a research instrument, by integrating the various sources into one new coherent language resource thus allowing researchers to gain unprecedented insights in material that partly has been available for quite some time.

**Keywords:** eLexicography; diachronic lexicography; lexicography tools

## 1 Introduction

The compilation of the diachronic dictionary of Tunis Arabic has been started as part of a larger project investigating the linguistic dynamics caused by recent demographic and social changes in the metropolitan area of Tunis (hence Tunis Arabic and not Tunisian Arabic). The TUNICO project (funded by a three year grant of the *Austrian Science Fund*<sup>1</sup>) will produce two digital language resources: (a) a corpus of unmonitored speech (dialogues as well as narratives) and (b) a dictionary based on this corpus and on other historical sources published in print form.

TUNICO in turn has grown out of an ongoing cooperative project which goes by the name *Vienna Corpus of Arabic Varieties* (VICAV). VICAV has been already started several years ago and is being run with a twofold perspective in mind: proceeding from linguistic research questions VICAV has been designed

---

1 Project number P 25706-G23

as a forum for collecting, producing and making available digital language resources of a wide range of spoken Arabic varieties. In addition, the project also pursues text technological interests investigating relevant standards, developing tools and workflows. At the heart of the VICAV collection there are so-called language profiles. This type of text consists in concise sketches of spoken linguistic varieties. The intention has been to proceed in a complementary manner to similar endeavours (such as the *Encyclopedia of Arabic Language and Literature* which is a standard reference work in the field). For the time being, the concept does not foresee the production of detailed grammatical descriptions. The focus is rather put on general information, research histories, relevant literature and sample texts. Another language resource represented in the collection are lists of salient grammatical features. The working group has attempted to identify particular linguistic items that are repeatedly used elsewhere in comparative Arabic investigations and make them comparable in example sentences that are the same across the various linguistic varieties. There are also digital dictionaries, texts, bibliographies, descriptions of relevant workflows and best-practises such as thorough encoding guidelines that can be reused for similar purposes in other projects. VICAV is intended as a cross-disciplinary platform for researchers in the field enabling them to exchange data, to collaborate effectively on new digital resources and to publish their findings, tools and data.

Both projects, TUNICO and VICAV, are joint initiatives of the University of Vienna (Department of Oriental Studies) and the Austrian Academy of Sciences (Institute for Corpus Linguistics and Text Technology). They are typical examples of a new brand of research that understands itself as digital humanities pursuing research with innovative methods and in accordance with new paradigms such as collaborative work, transdisciplinarity and open humanities.

## 2 A New Digital Dictionary

In the history of lexicography, dictionaries documenting Arabic dialects are a rather recent phenomenon. While the situation with respect to print dictionaries has improved for many areas, there are almost no digital Arabic dictionaries available so far, let alone dictionaries that come in a digitally reusable form, live up to modern standards or cover varieties other than Modern Standard Arabic.

With respect to Tunis, the situation is no different. There exists no comprehensive dictionary of the Arabic dialect of Tunis. Nicolas 1911 can be regarded as a good basis for diachronic research. However, it is – by and large – outdated. Other sources for lexicographic data are the works of Beaussier/Lentin (2006, a fusion of the 1958 edition and the 1959 supplement) which also include data on Tunis, Quéménéur (1961 and 1962) who provided useful lists of lexicographical items and Abdellatif 2010 who produced a quite useful amateur glossary. The eight volume glossary compiled by Marçais/Guîga 1958-61 covering the vocabulary of the village of Takroûna (ca.100 km south-east of Tunis) is still of unmatched value for the documentation of the lexicon of the Arabic vernacular of the Tunis area. However, it reflects mainly rural speech and is based on material from the 1920s.

Our project was designed to create up-to-date and easily accessible lexical information on Tunis Arabic, taking into account historical as well as contemporary data, by compiling a small, micro-diachronic and machine-readable dictionary of the variety. One of the many advantages of such a machine-readable dictionary is that queries in both directions (in our case Tunis Arabic – English/German/French and vice versa) are possible. All hitherto published dictionaries except the outdated work by Nicolas (1911) are unidirectional Arabic – French.

### 3 Heterogeneous Sources

One innovative aspect of the project lies in the fact that it is not only drawing on contemporary data taken from a digital corpus. However, it will also incorporate various sources reaching back as far as the 19<sup>th</sup> century (Stumme 1893/1896a-b). By integrating both corpus data and historical sources, we will create a new language resource, a new dictionary. Technically, the intention has been to keep each bit of information added to the dictionary traceable to its origin, thus allowing coming generations of researchers to interpret the data in accordance with their particular needs.

The basis of the dictionary was laid by data taken from didactic materials that were compiled by participants in the project for university classes. The glossaries of this course of spoken Tunisian Arabic could be easily recycled for the purpose and transformed into digital dictionary entries. In the next phases of the project, this data will be enriched from three main additional sources: the newly created corpus, interviews with first language speakers, and historical publications on the linguistic variety under investigation.

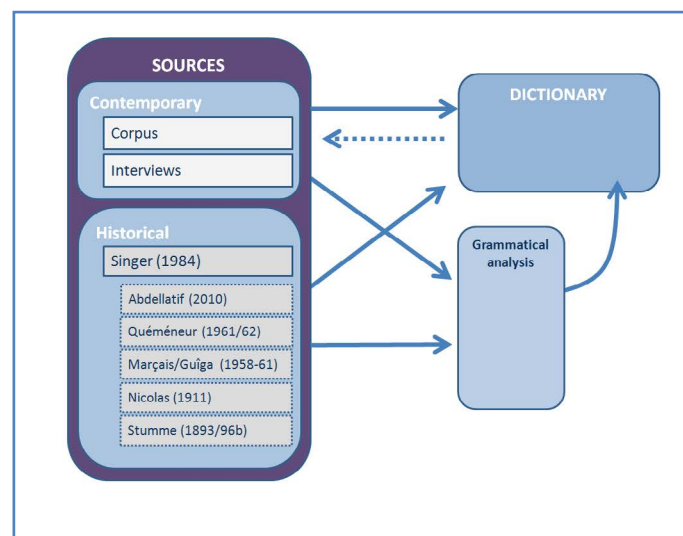


Figure 1: Basic chart of used sources.

### 3.1 Corpus of Spoken Youth Language

As can be seen in figure 1, the contemporary sources consist in a corpus of youth language by which we understand language produced by people under the age of 25. In Arabic dialectology, focussing on the language of the younger generation is a rather uncommon approach as traditionally linguistic interests were usually directed towards the past. Dialectological investigations often been focussed on the language of the older generation attempting to gain knowledge on older more conservative “pure” linguistic varieties. This has led to a situation (and not only in Arabic studies) where we often only know older forms of particular varieties and comparatively little about contemporary language. TUNICO (and also other language resources of the VICAV collection) has been designed to remedy this shortfall focussing on modern language, contemporary usage and lexical neologisms which, in the case of Tunis, are most often not of Arabic but French (or English) origin.

A first set of data was collected in August and September 2013 and is currently being transcribed. The field workers recorded some 33 hours of Tunis Arabic, the recordings contain data from approximately 90 different interviewees. The final version of the dictionary will contain the most frequent lemmas represented in the transcribed corpus, which will constitute the foundation of the dictionary’s contemporary layer.

### 3.2 Additional Interviews

The second contemporary source, the interviews, will be created at a later stage of the project. As we expect numerous lacunae in the lemma list and constituents of the dictionary entries, the data gained from the corpus and also the historical sources will have to be completed with information gained from additional interviews conducted with Tunisian informants. This will be done during the third and the last field campaigns. The interviews will differ from those done in the first two campaigns as they will be semi-structured interviews that aim at the elicitation of lexical data absent in the corpus. Having collected plenty of dialogues, it is planned to also go for narratives in this phase of the project.

### 3.3 Historical Data

The historical aspect will be introduced by way of lexicographic items excerpted from print publications, especially the very rich lexical material contained in Singer’s monumental grammar (1984; almost 800 pages) of the Medina of Tunis, which hitherto has been difficult to use which is mainly due to a lack of an index<sup>2</sup>. Singer’s data will be evaluated systematically and integrated into the dictionary (the material will, of course, be indicated by reference to his book; however on account of the unclear

---

2 It is important to note that Singer’s study is based on fieldwork carried out in the early 1960s (Singer 1984: VIII); the texts and the glossary which were advertised in the foreword (p.X) have never been published.



copyright status it is not planned to create a digital version of the book itself). Additional resources (Nicolas 1911, Marçais/Guîga 1958-61, Quéméneur 1962, Abdellatif 2010) will also be consulted in order to verify and complete the collected data. The diachronic dimension will help to better understand processes in the development of the lexicon.

The rich material gathered from young people whose parents are often not natives of the city of Tunis but have migrated to the capital from rural regions or other cities of Tunisia will hopefully enable us to analyse recent developments in the lexicon, particularly semantic changes including semantic shift, semantic reduction, and semantic extension of lexical items. Our diachronic approach will also make possible to determine the influence of other dialectal varieties of Arabic which in many cases are a result of the pan-Arabic satellite channels. One focus of the interviews carried out during our fieldwork is to gain newly coined vocabulary that appeared during and after the revolution of 2011, which had an immense impact on Tunisian society and hence also on Tunisian language. This vocabulary is, however, different from real youth slang that is often only used in in-group conversation. Studies on this particular field of the lexicon of Arabic dialects are extremely rare (the best publication on this topic is Caubet 2004 dealing mostly with Morocco).

As we are not mainly interested in the “pure and real” dialect we will also pay attention to the incorporation of foreign elements into the Arabic language as spoken in Tunis both with regard to semantics and morphology. The latter is characterized by a high degree of integration into the morphological frame of Arabic. Particularly verbs of foreign origin have to be adapted to Arabic patterns for the sake of inflection. A similar development is often found with pluralisation of nouns and adjectives.

The direct connection of corpus and dictionary (see figure 1) will facilitate research on phrases, idioms and collocations. Apart from some very well-studied varieties of Arabic (especially Egyptian and Moroccan Arabic) phraseology and related subjects such as collocation have been largely neglected, mainly because of a lack of text corpora sizeable enough for these purposes. The linkage of dictionary and corpus will enable users to investigate in which way given lexical items are connected to one another.

## 4 Modelling the Dictionary Entries

In the world of digital dictionary production, a considerable number of competing formats co-exist. We are far from any real standardisation in the field and our paper will not resume the discussion as to which format is best suited for which task (cf. Budin et al. 2012). Let it suffice to state, that using the TEI dictionary module to encode digitized print dictionaries has become a fairly common standard procedure in digital humanities. However, it has been shown that the TEI dictionary module is also usable for NLP purposes (Budin 2012). Data modelling for our project has been undertaken with two perspectives in mind: (a) to achieve a high degree of interoperability with comparable dictionaries of

other varieties of Arabic (already existing at the same department) and (b) to stay as compliant as possible with the ISO standard LMF (Romary 2013).

A major issue in this endeavour is how to represent and how to harmonise the diverging systems of transcription and transliteration found in the historical sources. As in comparable other projects, researchers in our project try to reduce the rich inventory of combinations of diacritics and characters by applying a basically phonemic transcription. Following a widespread convention in Arabic dialectology, the data is presented in a broad phonological transcription that does not usually indicate allophones. Basically, the set of characters used in the dictionary follows widely used conventions in Arabic studies. It is by and large the system used in standard reference works such as the *Encyclopedia of Arabic Language and Linguistics* (Leiden: Brill, 2006-2009). In the future, it is planned to provide the data in IPA-transcription too.

As can be seen in figure 2 we indicate the so-called root for each lexical item in the dictionary. The root is an intrinsic feature of Arabic word formation. In all layers of Arabic the bulk of the vocabulary is built on the principle of root and pattern. To express certain semantic terms, i.e. words, a purely consonantal root carrying the basic semantic information is combined with a limited set of patterns utilizing a fixed sequence of consonants, vowels, and optional prefixes and suffixes. To make comparative cross-dialectal search possible we have decided to indicate the root in a strictly etymological way. This means, each root reflects the corresponding root of Standard Arabic wherever possible. We are convinced that this approach does not reduce the usability of the on-line dictionary because, for users familiar with Arabic morphology, it is easy to detect the dialectal root in question. The main advantage of this approach lies in the possibility to find the reflexes of a certain Standard Arabic root in all dictionaries simultaneously.

## 4.1 Basic Schema

The schema applied in the compilation of the new dictionary has been used before in other projects for various languages and serves as the structural foundation of the dictionary entries. It imposes a number of very strict structural constraints on the TEI elements to ensure a high degree of interoperability with other components of the existing dictionary infrastructure at the ICLTT (Budin 2012). These constraints are defined by means of an XML Schema which only allows the use of a small subset of TEI elements and only a very few combinations thereof. The typical, slightly simplified basic structure of an entry taken from the Tunis dictionary is shown below. The entry begins with the lemma, this is followed by morphological forms and grammatical information. The system provides for translations in several languages. In the Tunis dictionary, we intend to offer German and English translations. Resources permitting, we will also add French.

```

<entry xml:id="ktaab_001">
  <form type="lemma">
    <orth xml:lang="ar-aeb-x-tunis-vicav">ktāb</orth>
    <orth xml:lang="ar-aeb-x-tunis-arabic">كتاب</orth>
  </form>

  <form type="inflected" ana="#n_pl">
    <orth xml:lang="ar-aeb-x-tunis-vicav">ktub</orth>
    <orth xml:lang="ar-aeb-x-tunis-arabic">كتب</orth>
  </form>

  <gramGrp>
    <gram type="pos">noun</gram>
    <gram type="gender">masculine</gram>
    <gram type="root" xml:lang="ar-aeb-x-tunis-vicav">ktb</gram>
  </gramGrp>

  <sense>
    <cit type="translation" xml:lang="en">
      <quote>book</quote>
    </cit>

    <cit type="translation" xml:lang="de">
      <quote>Buch</quote>
    </cit>

    <cit type="translation" xml:lang="fr">
      <quote>livre</quote>
    </cit>
  </sense>
</entry>

```

Figure 2: Basic encoding of a typical dictionary entry.

As can be seen in the example above, *sense* elements can have multiple translations. In a similar manner, every *entry* can contain an unspecified number of *form* elements. These can represent different morphological forms, variants such as for instance competing plurals or varying phonological representations. All these cases are treated similarly. Hierarchies are avoided, all *form* elements are placed directly inside the *entry* element.

```

...
  <form type="inflected" ana="#n_pl">
    <orth xml:lang="ar-aeb-x-tunis-vicav">xdim</orth>
    <bibl>
      <author>Ritt-Benmimoun</author>
      <date>2012/2013</date>
    </bibl>
  </form>

  <form type="inflected" ana="#n_pl">
    <orth xml:lang="ar-aeb-x-tunis-vicav">xidmāt</orth>
  </form>
...

```

Figure 3: Overabundance in plural forms.

The two inflected forms represent both common plurals. In cases of a clear distribution across registers, labels can be used to assign information regarding the register of the particular form. However, for the time being such forms are merely collected without adding information concerning the for-

mality scale. Once the corpus is available, we intend to add frequency data to the lemmas and the inflected forms. As to the encoding of the frequency information, discussions concerning data modelling are still ongoing.<sup>3</sup>

### Modelling Diachrony

Diachrony, or as some might insist micro-diachrony (as we are only talking about a time-span of roughly a century), is represented in the dictionary by indicating the source from which the data was taken. To this end, we make use of the *bibl* (bibliographic citation) element. This is a loosely-structured element the sub-components of which may or may not be explicitly tagged (TEI Guidelines 2013).

```

...
<bibl>
  <author>Ritt-Benmimoun</author>
  <date>2012/2013</date>
</bibl>
...
<bibl>
  <author>Singer</author>
  <date>1958</date>
  <biblScope unit="page">56</biblScope>
</bibl>
...

```

Figure 4: Bibliographic citations in TEI (P5).

Diachrony is established by adding these *bibl* elements to *form* and/or *sense/cit* elements. As stated before, any entry can have multiple forms and also can have multiple instances of the same morphological form. The absence of a *bibl* element indicates that the form has been entered from contemporary sources. In this manner, each element can be historically classified. In the following example, the *entry*, a noun, has two plural forms. By contrast to the example above which displays synchronous data (*xdim* vs. *xidmât* are both still in use) the second form here represents evidence of a historic form. An analogous contemporary form could so far not be verified.

```

...
<form type="inflected" ana="#n_pl">
  <orth xml:lang="ar-aeb-x-tunis-vicav">ktub</orth>
</form>

<form type="inflected" ana="#n_pl">
  <orth xml:lang="ar-aeb-x-tunis-vicav">uktba</orth>
  <bibl>
    <author>Singer</author>
    <date>1958</date>
    <biblScope unit="page">594</biblScope>
  </bibl>
</form>
...

```

Figure 5: *ktub* vs. *uktba*.

3 Details of these discussions were presented in the TEI members Meeting 2014 (Rome) and have been submitted for publication in jTEI 8 (papers from the 2013 conference).

## 5 Tools

The dictionary entries are compiled making use of the *Viennese Lexicographic Editor* (VLE), a general purpose XML editor providing a number of functionalities typically needed in compiling lexicographic data. It allows to collaboratively work on lexicographic data. From the very beginning of its development, it was designed to process standard-based lexicographic and terminological data such as LMF, TBX, RDF or TEI. VLE can automate many editing procedures. Most of these functions can be applied both to single and/or multiple entries. VLE has been implemented as a standalone desktop application (for Windows). VLE is the client of a server-client architecture. In order to realise a working environment, a web-server is needed. The server-side scripts (*php + mysql*) are also freely available and easy to setup. The program can check the structural integrity (well-formedness) of input on the fly and can validate the data against XML Schemas.

One of the particular features of VLE is a special module optimising access to external language resources such as corpora, other dictionaries, word lists etc. which makes it particularly well suited for deployment in our project with its various digital resources.<sup>4</sup> The fact that VLE is a product of our institute and constantly being updated will ease the implementation of necessary interfaces between the corpus and the dictionary infrastructures.

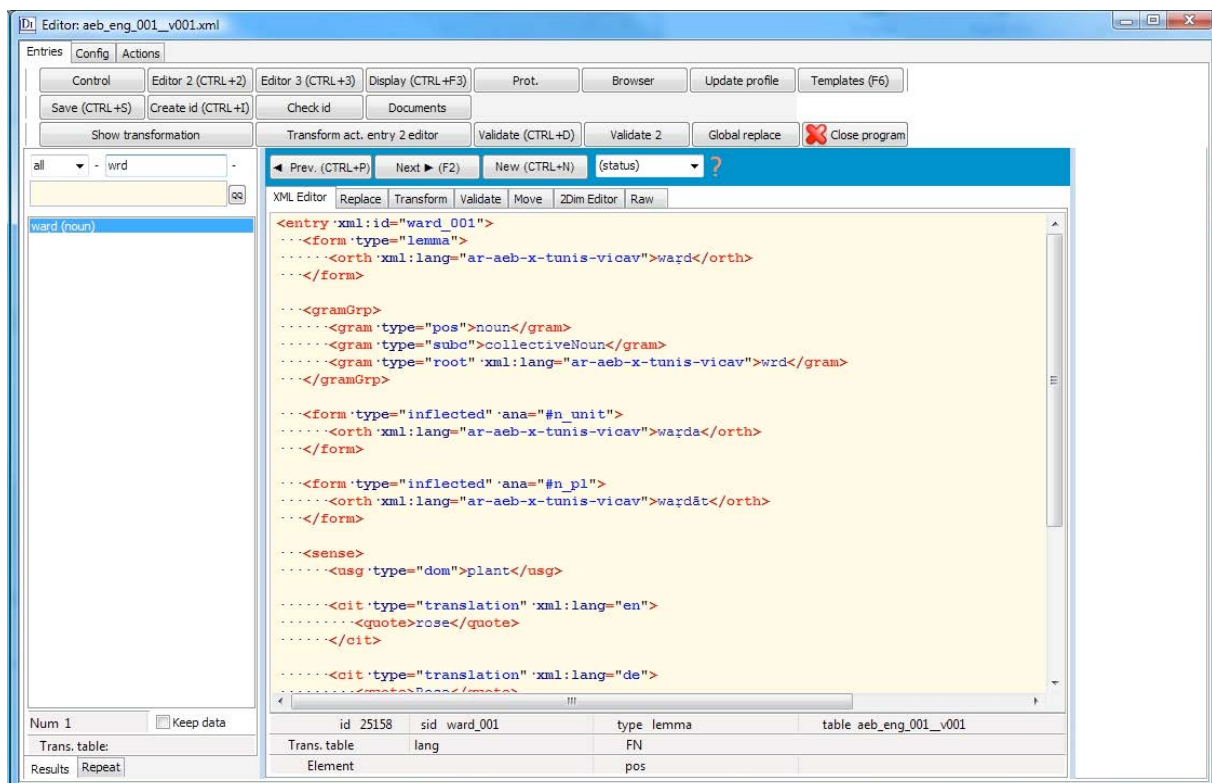


Figure 6: The dictionary editor.

4 The tools can be downloaded from the *Language Resources Portal* (CLARIN Centre Vienna): <http://clarin.oew.ac.at/ccv/vle>.

The online publication of both the corpus and the dictionary will be undertaken by means of *corpus\_shell*, a modular framework of reusable software components which has also been developed at the ICLTT over the past couple of years. It is used to access and publish heterogeneous and distributed language resources such as language corpora, dictionaries, encyclopaedic databases, prosopographic databases, bibliographies, metadata, and schemata. Its core functionality is encapsulated in self-contained components exposing well-defined interfaces based on acknowledged standards. The principle idea behind the architecture is to decouple the modules serving data from the user-interface components.<sup>5</sup> This software was used in several other projects before, it is the backbone of the *Language Resources Portal*<sup>6</sup> which is run at the Austrian Academy of Sciences.

## 6 Status and Outlook

The project is being conducted in the context of CLARIN-AT, the local branch of the European infrastructure consortium CLARIN-ERIC (Common Language Resources and Technology Infrastructure). Both the corpus and the dictionary were planned as in-kind contributions to the CLARIN network for the years to come. The build-up of the corpus, the compilation of the dictionary and the development of software are being undertaken in the spirit of open-access and open-source. So far, no binding decision has been made as to the licence under which the particular language resources will be available. However, there is a strong case for a Creative Commons licence, CC-BY being the favoured option, which has been used for comparable other projects of the department. Discussions with interested researchers and other stakeholders have shown that the permission to create derivative works is usually regarded as an important prerequisite in order to ensure reuse of data. The tools, workflows and specifications created in this project can potentially also be used for other languages and many other applications.

At the time of writing this paper, the dictionary already contained roughly five thousand raw entries and several hundred edited entries. We are planning to make data available as soon as data from the corpus and historical sources have been added to the basic entries, i.e. already during the production phase which is meant to allow and to encourage other researchers in the field to contribute to this work. To our knowledge, this lexicographic undertaking is not only the first scholarly attempt to make available a digital dictionary of a spoken Arabic variety, but also the first attempt at creating a digital dictionary presenting diachronic data of a spoken Arabic variety.

---

5 More details at [clarin.oeaw.ac.at/ccv/corpus\\_shell](http://clarin.oeaw.ac.at/ccv/corpus_shell).

6 [clarin.oeaw.ac.at/ccv/](http://clarin.oeaw.ac.at/ccv/)



## 7 Selected References

- Abdellatif, K. (2010). Dictionnaire «de Karmous» du Tunisien. Accessed at: <http://www.fichier-pdf.fr/2010/08/31/m14401m/> [06/04/2014].
- Baccouche, T., Mejri, S. (2000). L'Atlas Linguistique de Tunisie: problématique phonologique. In *Revue Tunisienne de Sciences Sociales* 120, pp. 151-156.
- Banski, P., Wójtowicz, B. (2009). FreeDict: an Open Source repository of TEI-encoded bilingual dictionaries. In TEI-MM, Ann Arbor. Accessed at: <http://www.tei-c.org/Vault/MembersMeetings/2009/files/Banski+Wojtowicz-TEIMM-presentation.pdf> [06/04/2014].
- Beaussier, M. (2006). *Dictionnaire pratique arabe-français: (arabe maghrébin); constitué du «Dictionnaire pratique arabe-français» de Marcelin Beaussier dans l'édition de Mohamed Ben Cheneb & de son «Supplément» par Albert Lentin*, Paris: Ibis Press.
- Bel, N., Calzolari, N. & Monachini, M. (eds.) (1995). Common Specifications and notation for lexicon encoding and preliminary proposal for the tagsets. MULTTEXT Deliverable D1.6.1B. Pisa.
- Budin, G., Majewski, S. & Mörth, K. (2012). Creating Lexical Resources in TEI P5: A Schema for Multi-purpose Digital Dictionaries. In *Journal of the Text Encoding Initiative 3 (Special issue on TEI and linguistics)*.
- Hass, U. (ed.) (2005). Grundfragen der elektronischen Lexikographie: Elexiko, das Online-Informationssystem zum deutschen Wortschatz. Berlin, New York: W. de Gruyter.
- Ide, N., Kilgariff, A. & Romary, L. (2000). A Formal Model of Dictionary Structure and Content. In *Euralex 2000 Proceedings*, pp. 113-126.
- Marçais, W. (1925). Textes arabes de Takroûna. I. Textes, Transcription et Traduction annotée. Paris.
- Marçais, W., Guîga, A. (1958-61). *Textes arabes de Takroûna. II: Glossaire*. 8 vol. Paris.
- Mörth, K., Budin, G. (2011). Hooking up to the corpus: the Viennese Lexicographic Editor's corpus interface. In *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011*. Bled (Slovenia), pp. 52-59.
- Nicolas, A. (1911). Dictionnaire français-arabe: idiome tunisien and Dictionnaire arabe-français. Tunis.
- Quéméneur, J. (1962). Glossaire de dialectal. In *IBLA* (1962), pp. 325-67.
- Romary, L., Salmon-Alt, S. & Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. In *Workshop on enhancing and using electronic dictionaries*. Coling 2004, Geneva.
- Romary, L., Wegstein, W. (2012). Consistent Modelling of Heterogeneous Lexical Structures. In *Journal of the Text Encoding Initiative 3 (Special issue on TEI and linguistics)*.
- Romary, L. (2013). TEI and LMF crosswalks. In *Stefan Gradmann and Felix Sasaki (eds.), Digital Humanities: Wissenschaft vom Verstehen*. Humboldt Universität zu Berlin. Accessed at: <http://hal.inria.fr/hal-00762664> [08/03/2014].
- Singer, H. (1984). Grammatik der Arabischen Mundart der Medina von Tunis. Berlin-New York.
- Sperberg-McQueen, C.M., Burnard L. & Bauman S. (eds.) (2010). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlottesville, Nancy. Accessed at: <http://www.tei-c.org/Guidelines/P5/> [08/03/2014].
- Stumme, H. (1893). Tunisische Märchen und Gedichte. Band I: Transcribierte Texte nebst Einleitung; Band II: Übersetzung. Leipzig.
- Stumme, H. (1896a). Grammatik des tunsischen Arabisch nebst Glossar. Leipzig.
- Stumme, H. (1896b). Neue tunsische Sammlungen. (Kinderlieder, Strassenlieder, Auszählreime, Rätsel, 'Arôbi's, Geschichtchen u.s.w.). Berlin (ZAOS II).
- Stumme, H. (1898). Märchen und Gedichte aus der Stadt Tripolis in Nordafrika. Eine Sammlung prosaischer und poetischer Stücke im arabischen Dialekt der Stadt Tripolis, nebst Übersetzung, Skizze des Dialekts und Glossar. Leipzig.
- Versteegh, K., Eid, M., Elgibali, A., Woidich, M & Zaborski, A. (eds.) (2005-2009). *Encyclopedia of Arabic Language and Linguistics*. 5 vols. Leiden, Boston: Brill





# BabelNet meets Lexicography: the Case of an Automatically-built Multilingual Encyclopedic Dictionary

Roberto Navigli  
Sapienza University of Rome  
navigli@di.uniroma1.it

## Abstract

In this paper we provide a first study of the lexicographic quality of BabelNet, a very large automatically-created multilingual encyclopedic dictionary. BabelNet 2.0, available online at <http://babelnet.org>, covers 50 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech. It is obtained from the automatic integration of several language resources, namely: WordNet, Open Multilingual WordNet, Wikipedia and OmegaWiki. Here we present a first analysis of the dictionary entries in terms of their coverage of English and Italian word tokens in a large corpus and in comparison to existing, well-established dictionaries, namely the Oxford Dictionary of English and the Treccani Italian dictionary. We observe that BabelNet contains most meanings of the frequent words under analysis and provides additional, often domain-specific meanings and their textual definitions unavailable in traditional dictionaries, as well as encyclopaedic coverage for those words.

**Keywords:** Multilinguality; Encyclopedic dictionaries; Quality evaluation of automatically-created dictionaries

## 1 Introduction

The textual content that is available on the Web is becoming ever increasingly multilingual, providing an additional wealth of valuable information. Most of this information, however, remains inaccessible to the majority of users because of language barriers. Consequently, both humans and automatic systems need tools which will enable them to enjoy the beauty and the usefulness of this varied multilingual world.

The wide majority of bilingual paper dictionaries, however, focus on a given language pair, which are the languages on which the lexicographers, and authors of the dictionary, are expert in. As a result, the sense inventories of dictionaries for different language pairs are different, even if the dictionaries are printed by the same publisher. Integrating these inventories, thereby enabling the creation of a multilingual dictionary, is therefore a very arduous task.

MultiJEDI (Multilingual Joint word sense Disambiguation, <http://multijedi.org>) is a major project under way in the Linguistic Computing Laboratory at the Sapienza University of Rome. MultiJEDI is a 5-year Starting Independent Research Grant funded by the European Research Council (ERC) that started in February 2011. The project aims to investigate new, groundbreaking directions in the field of Word Sense Disambiguation (WSD), the task of computationally determining the meaning of words in context (Navigli, 2009; 2012). The key intuition underlying the project is that we now have the capabilities to transform multilinguality from an obstacle to Natural Language Understanding into a powerful catalyst for the task. As a core tool for enabling multilinguality the project aims to create a very large automatically-created multilingual encyclopedic dictionary, called BabelNet, made available online at <http://babelnet.org>. BabelNet is a novel language resource in several respects, including: being a multilingual dictionary which covers tens of languages; providing both encyclopaedic and lexicographic coverage; including information which is usually not available within dictionaries, such as images, fine-grained category information, multiple textual definitions for the same entry, hyperlinks to other entries, and much more.

Since integrating dictionaries of different kinds and nature, especially on a multilingual scale, is admittedly a hard, ambitious task, in this paper we analyze the lexicographic quality of BabelNet, especially in terms of the user perspective, and compare it against manually created dictionaries, so as to determine the added value of an automatic dictionary integration process. Our analysis is performed both at the corpus level, by studying the coverage provided by BabelNet of word occurrences within text (on a portion of the American National Corpus - ANC), and at the inventory level, i.e. by comparing the BabelNet sense inventory with that of other well-established resources, such as the Oxford Dictionary of English and the Treccani dictionary of Italian. Our analysis shows that the richness and amount of information available in BabelNet largely exceeds that of manually created lexicographic resources.

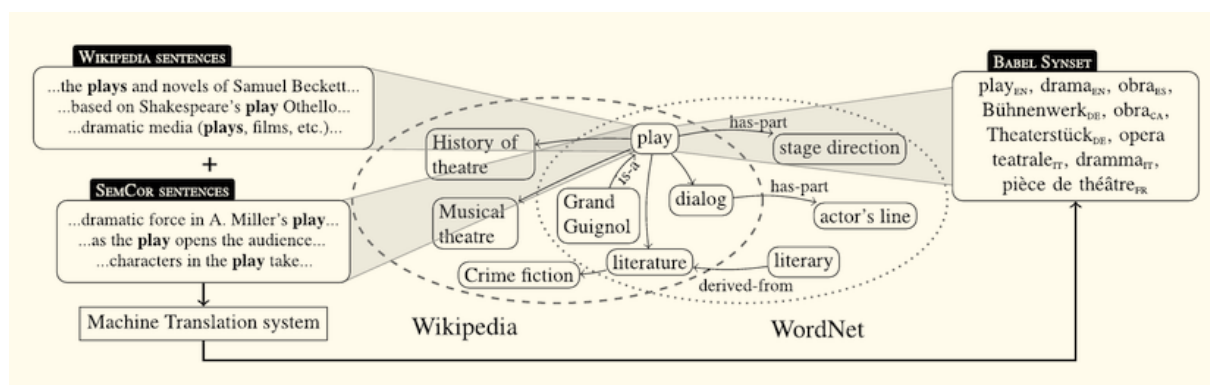


Figure 1: The BabelNet structure.

## 2 BabelNet 2.0

BabelNet is based on the key idea that different language resources, such as WordNet (Fellbaum, 1998), i.e., the largest machine-readable computational lexicon of English, and Wikipedia (<http://www.wikipedia.org>), i.e., the most popular multilingual encyclopedia, provide complementary knowledge that can be integrated into a single unified multilingual semantic network covering as many languages as possible. BabelNet, available online at <http://babelnet.org>, is therefore a large-scale “encyclopedic dictionary”. BabelNet encodes knowledge as a labeled directed graph  $G = (V, E)$  where  $V$  is the set of nodes – i.e., concepts such as *play* and named entities such as *Shakespeare* – and  $E \subseteq V \times R \times V$  is the set of edges connecting pairs of concepts (e.g., *play* is-a *dramatic composition*). Each edge is labeled with a semantic relation from  $R$ , e.g., [is-a, part-of, ...,  $\epsilon$ ], where  $\epsilon$  denotes an unspecified semantic relation. Importantly, each node  $v \in V$  contains a set of lexicalizations of the concept for different languages, e.g., { *play*<sub>EN</sub>, *Theaterstück*<sub>DE</sub>, *dramma*<sub>IT</sub>, *obra*<sub>ES</sub>, ..., *piece de theatre*<sub>FR</sub> }. We call such multilingually lexicalized concepts Babel synsets. Concepts and relations in BabelNet are harvested from the largest available semantic lexicon of English, WordNet, and a wide-coverage collaboratively-edited encyclopedia, Wikipedia. In order to construct the BabelNet graph, we extract at different stages:

- from WordNet, all available word senses (as concepts) and all the lexical and semantic pointers between synsets (as relations);
- from Wikipedia, all the Wikipages (i.e., Wikipages, as concepts) and semantically unspecified relations from their hyperlinks.

A graphical overview of BabelNet is given in Figure 1. As can be seen, WordNet and Wikipedia overlap both in terms of concepts and relations: this overlap makes the merging between the two resources possible, enabling the creation of a unified knowledge resource. In order to enable multilinguality, we collect the lexical realizations of the available concepts in different languages. Finally, we connect the multilingual Babel synsets by establishing semantic relations between them. Thus, our methodology consists of three main steps:

- (1) We integrate WordNet and Wikipedia by automatically creating a mapping between WordNet senses and Wikipages. This avoids duplicate concepts and allows their inventories of concepts to complement each other.
- (2) We collect multilingual lexicalizations of the newly-created concepts (i.e., Babel synsets) by using (a) the human-generated translations provided by Wikipedia (i.e., the inter-language links), as well as (b) a machine translation system to translate occurrences of the concepts within sense-tagged corpora.
- (3) We create relations between Babel synsets by harvesting all the relations in WordNet and in the wikipedias in the languages of interest.

Its current version, i.e., BabelNet 2.0, covers 50 languages and provides both lexicographic and encyclopedic knowledge for all the open-class parts of speech. It is obtained from the automatic integration of the following resources:

- WordNet, a popular computational lexicon of English (<http://wordnet.princeton.edu>, version 3.0);
- Open Multilingual WordNet (<http://www.casta-net.jp/~kuribayashi/multi/>), a collection of word-nets available in different languages;
- Wikipedia, the largest collaborative multilingual Web encyclopedia (<http://wikipedia.org>);
- OmegaWiki, a large collaborative multilingual dictionary (<http://omegawiki.org>).

The number of lemmas for each language ranges between more than 8 million (English) and almost 100,000 (Latvian), with a dozen languages having more than 1 million lemmas. The number of polysemous terms ranges between almost 250,000 in English to only a few thousand for languages such as Galician, Latvian and Esperanto, with most languages having several tens of thousands of polysemous terms. BabelNet 2.0 contains about 9.3 million concepts, i.e., Babel synsets, and above 50 million word senses (regardless of their language). It also contains about 7.7 million images and almost 18 million textual definitions, i.e., glosses, for its Babel synsets. The synsets are linked to each other by a total of about 262 million semantic relations (mostly from Wikipedia). Language distribution of lemmas, synsets and senses is graphically shown in Figure 2. It can be seen that the top 9 languages cover approximately half of the language resource in all respects.

Details on the automatic construction procedure can be found in (Navigli and Ponzetto, 2012) and in (Navigli, 2014), where many applications to Word Sense Disambiguation, Open Information Extraction and Linked Open Data are also reported.

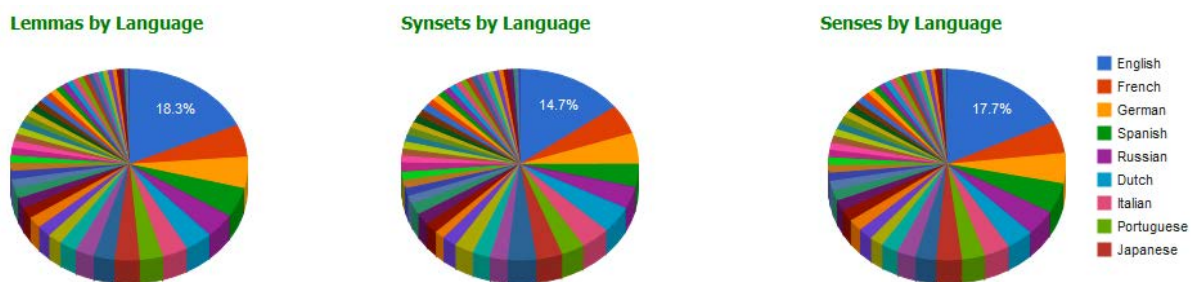


Figure 2: Statistics on the number of lemmas, synsets and senses for the main languages in BabelNet.

### 3 Corpus coverage in English

To determine corpus coverage, we used the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008) which consists of parts of the American National Corpus (<http://www.anc.org>) covering a wide range

of genres of written and spoken textual data amounting to over 500k words. This project aims at organizing and addressing the problems arising against the creation of a resource with multiple annotations. The corpus is available in different formats such as GrAF, in-line XML, token/part of speech sequences, RDF encoding and CoNLL format. The key feature of this corpus is the availability within a single resource of many different linguistic annotations; to date, it contains 17 different types of linguistic annotation, such as sentence boundary, part of speech and syntactic dependency among others. These annotations are the result of a semi-automatic effort in which automatic systems have been coupled with an iterative process of manual evaluations and annotations for retraining the automatic approaches and fine-tuning annotator guidelines to improve inter-annotator agreement. Moreover, the fact that it is freely available (<http://www.anc.org/data/masc/>) makes it an invaluable resource for both industry and academic communities in order to produce and improve cutting-edge language technologies.

For our statistics, we considered the set of open-class words in MASC 3.0, totaling 233115 open-class word tokens, and determined, first, the percentage of word tokens for which BabelNet contains an entry for the corresponding lemma and part of speech tag and, second, the percentage of word tokens for which BabelNet contains either a single-word entry or a multiword expression which covers two or more word token in the given sentence. We calculated that 95.15% of open-class word tokens in MASC are covered in BabelNet in the first case, while if we also consider multiword expressions, our coverage increases to 95.53%. We performed the same calculations using the lexicon of the Oxford Dictionary of English (ODE, Soanes & Stevenson, 2003), obtaining 83.91% of single-word tokens covered and 84.03% of tokens covered by any multiword or single-word expression. This shows higher lexicographic coverage (+10%) in BabelNet than in the ODE for the English language. We note that, for many of the uncovered word tokens, the problem is a wrong part-of-speech tag assigned to them (e.g., *achievable* tagged as a noun, *calculus* as an adjective, etc.).

In the future we plan to obtain similar statistics for other languages. However, we note that this requires part-of-speech tagging systems in order to find the appropriate lemma within the dictionary.

## 4 Dictionary comparison

We performed a comparison of BabelNet against important dictionaries for two different languages, namely: the Oxford Dictionary of English for the English language and the Treccani dictionary for the Italian language.

### 4.1 English dictionary comparison

As regards English, we compared the lexicographic entries in BabelNet against those of the Oxford Dictionary of English (<http://www.oxforddictionaries.com/>) for ten of the 1000 most frequent English

lemmas, namely: *work, time, country, head, room* (nouns), *remember, wait, close, write, contain* (verbs). An analysis of the definitions in the two resources resulted in the following findings:

- **Sense coverage:** in general, the two dictionaries share most of the senses, with additional senses on both sides. However, BabelNet provides a considerably higher number of senses, especially domain-specific ones for nouns and more fine-grained verb sense distinctions. Examples include: a specific thermodynamics sense of *work*, the computer system sense of *time* as well as its representation in ISO time format, *country* in the music style sense, several meanings of *head*, among which: the tip of an abscess, the front a military formation, a difficult juncture and many others; *write* in the sense of coding a computer program. The ODE also includes a few senses which are not covered in BabelNet. For instance, *work* as the operative part of a clock or a defensive structure and *write* in the sense of underwrite (an insurance policy). Finally, we note that BabelNet covers all the most important encyclopaedic meanings of the nominal lemmas, e.g., *head* as the linux program, several films, companies, albums and songs named *Work, Time, Country* and so on.
- **Quality of sense definitions:** the quality of the sense definitions in the Oxford Dictionary of English is generally higher, with carefully selected usage examples. BabelNet, however, has the advantage of providing several synonyms for the same word sense (e.g., *caput, mind, brain, psyche, chief, head word* etc. for different meanings of *head*, *piece of work, employment, study, mechanical work* for *work*, etc.)
- **Quantity of sense definitions:** The number of definitions per sense is considerably higher in BabelNet, thanks to its integration of different language resources. We show statistics in Table 1 (left). It can be seen that, for nouns, BabelNet provides five times the number of definitions per lemma on average while, for verbs, this difference drops to less than 3 times, which is still very high. Interestingly, for nouns BabelNet provides several multiple definitions for the same sense.

## 4.2 Italian dictionary comparison

We then compared the quality of ten of the 1000 most frequent Italian lemmas in BabelNet against the Treccani Italian dictionary (<http://www.treccani.it/vocabolario>), namely: *lavoro, tempo, paese, testa, sala* (nouns), *ricordare, aspettare, chiudere, scrivere, contenere* (verbs). An analysis of the definitions resulted in the following findings:

- **Sense coverage:** in general, the two dictionaries share most of the senses, with additional senses on both sides. Like for English, BabelNet provides coverage for very domain-specific nominal senses, such as *work* in project management, *work* in applied sciences, the linguistic sense of *tempo*, *testa* as the word in a grammatical constituent; the Treccani dictionary, instead, tends to encode all the traditional, regional or historical lexicographic sense distinctions of our words, including some which – due to lack of translations into Italian – are unavailable in BabelNet. Examples include: *sala* in the sense of the complex of acts by which a change of ownership was made in Ger-

manic law; *testa* in the regional Apulian use denoting a species of fish, i.e., *Trigla*; an uncommon usage of *paese* as painted landscape (as in *pittore di paesi*). As regards verbs, we did not find relevant differences between the two dictionaries. Finally, we note that BabelNet covers all the most important encyclopaedic meanings of the nominal lemmas, including a town in Italy called *Paese*, a magazine and a company producing tissues called *Tempo*, several towns and a necropolis called *Sala*, a surname and a novel called *Testa*, etc.

- **Quality of sense definitions:** the quality of the sense definitions in the Treccani dictionary is generally higher, with carefully selected usage examples. However, BabelNet has the big advantage of providing several synonyms for the same word sense (e.g. *opera* for the piece of work sense of *lavoro*; *collocamento*, *impiego* and *occupazione* for its employment sense, *compito*, *faccenda*, *incarico* and *incombenza* for its undertaking sense, etc.).
- **Quantity of sense definitions:** The number of definitions per sense is considerably higher in BabelNet, thanks to its integration of different language resources. We show statistics in Table 1 (right). It can be seen that we have a considerably lower number of sense definitions in BabelNet. This is due to the fact that many of the lexical resources integrated, while providing much lexicographic coverage, do not provide textual definitions for the senses they encode. This is particularly true for verbs (and adjectives and adverbs), to which resources like Wikipedia cannot contribute. Interestingly, however, BabelNet provides a higher number, more than twice overall, of senses than the Treccani dictionary, thanks to its integration of several different language resources contributing to its lexical richness also in non-English languages.

		English		Italian	
		BabelNet	ODE	BabelNet	Treccani
Nouns	Total (average) # of senses	79 (15.8)	29 (5.8)	82 (16.4)	30 (6.0)
	Total (average) # of definitions	126 (25.2)	29 (5.8)	37 (7.4)	93 (18.6)
Verbs	Total (average) # of senses	45 (9.0)	17 (3.4)	35 (7.0)	19 (3.8)
	Total (average) # of definitions	50 (10.0)	17 (3.4)	3 (0.6)	44 (8.8)
Total	Total (average) # of senses	124 (12.4)	46 (4.6)	117 (11.7)	49 (4.9)
	Total (average) # of definitions	176 (17.6)	46 (4.6)	40 (4.0)	137 (13.7)

**Table 1: Statistics of our ten frequent words for English (left) and Italian (right) using two different dictionaries. Only lexicographic entries are considered (BabelNet encyclopaedic synsets are excluded from these statistics).**

### 4.3 Validation of lexicographic entries with Video Games with a Purpose

As BabelNet is the output of an automatic mapping algorithm (Navigli and Ponzetto, 2012), some of the entries which contain information from several resources, e.g. both WordNet and Wikipedia, might have been merged incorrectly starting from two different senses of the same word. Moreover,



the automatic translation system used to increase the set of multilingual lexicalizations of our Babel synsets might produce wrong translations.

We therefore proposed validating BabelNet using video games with a purpose (Vannella et al. 2014). The annotation tasks are transformed into elements of a video game where players perform their task by playing the game, rather than by performing a more traditional annotation task. While prior efforts in Natural Language Processing have incorporated games for performing the annotation and validation task (Siorpaes and Hepp, 2008; Herdagdelen and Baroni, 2012; Poesio et al., 2013), these games have largely been text-based. In contrast, this year we proposed two video games with graphical 2D gameplay, whose fun nature provides an intrinsic motivation for players to keep playing, thereby increasing the quality of their work and keep the cost per annotation low. The first game, *Infection*, validates concept-concept relations, and the second, *The Knowledge Towers*, validates image-concept relations. In experiments involving online players, we demonstrated that, first, players do not need financial incentives to increase the quality of their annotations, second, in a comparison with crowdsourcing, we demonstrated that video game-based annotations consistently generated higher-quality annotations and, third, we found that video game-based annotation can be more cost-effective than crowdsourcing or annotation tasks with game-like features. However, these games did not focus on the validation of the lexicographic entry itself, but on hyperlinks between entries and concept-associated images in BabelNet.

In the future we plan to develop video games that will enable the addition, integration and validation of textual definitions, as well as the validation and addition of senses in arbitrary languages.

#### **4.4 General remarks**

Our objective was not to show that BabelNet is better than a traditional dictionary, especially for resource-rich languages such as Italian and English. However, our first analysis shows that, thanks to its integration of several online resources, a multilingual dictionary such as BabelNet provides adequate coverage of lexicographic entries while at the same time containing several synonyms, multiple definitions, hyperlinks to other senses in the dictionary, encyclopedic coverage, which is inherently impossible to achieve in a traditional dictionary, and, last but not least, multilingual interlinking across senses.

In our evaluation we have not taken into account many other features of BabelNet, such as its semantic network structure, which can be explored by humans to better understand the semantics of a concept and exploited by machines to perform automatic tasks such as Word Sense Disambiguation and Entity Linking (Moro et al., 2014), and its availability as a Linked Open Data (LOD) thanks to a Lemon-RDF encoding of the network (Ehrmann et al., 2014).



## 5 Conclusion

In this paper we first presented BabelNet, a multilingual encyclopaedic dictionary automatically constructed from online language resources, and then performed a first qualitative analysis of the BabelNet inventory. Our analysis was performed both in terms of coverage of a large English corpus, i.e., MASC, a subset of the American National Corpus, also in comparison with the Oxford Dictionary of English (ODE), and in terms of coverage and quality of the entries when compared to the ODE for English and the Treccani dictionary of Italian on a random sample of 10 frequent words for the two languages.

## 6 References

- Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., Navigli, R. (2014) Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. *Proc. of the 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 26-31 May, 2014
- Fellbaum, C. editor. (1998). *WordNet: An Electronic Database*. MIT Press.
- Herdagdalen, A. & Baroni, M. (2012). Bootstrapping a game with a purpose for common sense collection. *ACM Transactions on Intelligent Systems and Technology*, 3(4):1-24.
- Ide, N., Baker, C., Fellbaum, C., Fillmore, C. & Passonneau, R. (2008). MASC: the Manually Annotated Sub-Corpus of American English. In *Proceedings of LREC 2008*.
- Moro, A., Raganato, A. & Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*.
- Navigli, R. (2009). Word Sense Disambiguation: a Survey. *ACM Computing Surveys*, 41(2), ACM Press, 2009, pp. 1-69.
- Navigli, R. (2012). A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In *Proc. of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM 2012)*, Spindleruv Mlyn, Czech Republic, January 21-27th, 2012, pp. 115-129.
- Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, Elsevier, 2012, pp. 217-250.
- Navigli, R. (2014) BabelNet and Friends: A manifesto for multilingual semantic processing. *Intelligenza Artificiale*, 7(2), pp. 165-181, IOS Press, 2013.
- Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems*, 3(1):3:1-3:44, April.
- Soanes, C. & A. Stevenson, editors (2003). *Oxford Dictionary of English*. Oxford University Press.
- Siorpaes, K. & Hepp, M. (2008) Ontogame: Weaving the Semantic Web by online games. In Sean Bechhofer, Manfred Hauswirth, Jrg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications*, vol. 5021 of *Lecture Notes in Computer Science*, pp. 751-766. Springer Berlin, Heidelberg.
- Vannella, D., Jurgens, D., Scarfini, D., Toscani, D., Navigli, R. (2014) Validating and Extending Semantic Knowledge Bases using Video Games with a Purpose. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, USA, June 22-27, 2014.

### Acknowledgements

The author gratefully acknowledges the support of the ERC Starting Grant MultiJEDI No. 259234. Additional thanks go to Daniele Vannella and Andrea Moro for their help with calculating the BabelNet statistics.



# A Simple Platform for Defining Idiom Variation Matching Rules

Koichi Takeuchi<sup>#</sup>, Ulrich Apel<sup>§</sup>, Ray Miyata<sup>¶</sup>, Ryo Murayama<sup>¶</sup>,  
Ryoko Adachi<sup>¶</sup>, Wolfgang Fandler<sup>§</sup>, Iris Vogel<sup>§</sup>, Kyo Kageura<sup>¶</sup>

<sup>#</sup>Okayama University, Japan,

<sup>§</sup>Tübingen Eberhard Karls University, Germany,

<sup>¶</sup>University of Tokyo, Japan

koichi@cl.cs.okayama-u.ac.jp, ulrich.apel@uni-tuebingen.de, ray.miyata@gmail.com,  
utatanenohibi@gmail.com, littlemarmalade@gmail.com, fandler@japanologie.uni-tuebingen.de,  
iris@fruchtfledermaus.de, kyo@p.u-tokyo.ac.jp

## Abstract

In this demonstration, we present a system which enables people who are learning languages to define idiom variation matching rules, with special reference to variations created by insertion. The system is fully operational, currently providing Japanese idiom entries taken from Japanese-English and Japanese-German dictionaries, and is used by both graduate and undergraduate university students who are studying Japanese and Japanese native speakers.

**Keywords:** Idiom variations; Language learning; Japanese-German dictionary

## 1 Introduction

Matching idiom occurrences in texts to dictionary entry forms is critical for developing a satisfactory automatic dictionary lookup system. There are important studies of idioms in linguistics (Čermák, 2001; Fraser, 1970; Moon, 1998; Nicolas, 1995; Numberg et al., 1994) but they do not provide tractable variation rules in different languages. In the field of computational linguistics, some important contributions have been made in idiom variation matching and related issues since the mid 1990s (Breidt et al., 1996; Breidt and Feldweb, 1997; Carl and Rascu, 2006; Michiels, 2000; Proszeky and Kis, 2002; Takeuchi et al., 2007). Nevertheless, most available dictionary lookup systems and MT systems do not incorporate flexible idiom matching functions. Given this situation, we developed a system which enables language learners and practitioners to define flexible idiom matching rules.

## 2 Idiom variations

Major idiom variations can be categorised into three types (Kageura and Toyoshima, 2006), namely (i) insertion (e.g. “make unholy allowance for” as a variation of “make allowance for”); (ii) change of or-

der (e.g. “the bucket is kicked” as a variation of “kick the bucket”); and (iii) paradigmatic replacement (e.g. “head screwed on wrong” as a variation of “head screwed on right”). We focus on variation by insertion in our platform, because (a) this is the most frequently observed variation, (b) simple change of order can mostly be dealt with straightforwardly and such complex variations as combinations of change of order with insertion are relatively rare, and (c) dealing with paradigmatic replacement is a problem to be solved not by defining syntagmatic rules but by lexical resources such as thesauri. We may extend the target classes to include change of order variations in the future.

Note that while linguists are likely to argue that “you cannot passivise ‘kick the bucket’ and say ‘the bucket is kicked,’” these kinds of variations do occur, albeit rarely, in real-world texts, and as such it is important for language practitioners and learners to be able to retrieve the underlying idiom from the variation.

### **3 System for Defining Idiom Variation Matching Rules**

#### **3.1 Access**

The system can be accessed at <http://edu.ecom.trans-aid.jp>.

#### **3.2 System Concepts**

The basic policies we adopted are as follows:

(a) We took a restrictive approach rather than a generative approach; we assume that gapped matching of constituent elements is carried out by the base lookup algorithm, and that the rules defined in the platform are to be used to filter out false positives. This has two practical merits. First, the idiom variation rules can be added to dictionary lookup systems as a separate module. Second, if we assume that the matching rules will be used in a translation-aid environment, overmatching (as long as it is not excessive) is less harmful than misses.

(b) We only assume morphological analysers and/or POS-taggers for preprocessing; we do not use parsers. This is because (i) morphological analysers and POS-taggers are available for a wider range of languages than parsers, (ii) we found that there is no difference in performance all in all in a test in English, and (iii) as overmatching is less critical than misses, the merits of using parsers are less important in the application we assume.

(c) We assume that the rules will be defined not only by trained linguists but by ordinary speakers, practitioners and learners of that language. To facilitate this, we restricted the range of variations that can be specified in one rule, by prohibiting the combinations of AND and OR choices. For instance, one cannot write: N (adj|adv|N)+(postp) V in a single rule.

### 3.3 System Constitution

The system requires dictionary entries consisting of ordinary words. In addition, it requires a separate list of idioms. They should be registered to the system in advance. Currently, a Japanese-English dictionary and a Japanese-German dictionary are registered, through which Japanese idiom variation rules can be defined and validated. The entries are morphologically analysed, and indexes are made for the entry forms as well as the constituent elements of entries. The base lookup module consists of (a) matching individual entries and (b) gapped matching, with up to eight intervening elements, of the constituent elements of idioms. Variation matching rules are defined by users of the system through the Web interface. Currently, we assume that the target idioms for which variation rules are defined are determined by users. To develop variation rules systematically, it would be better for the system to provide users with idioms. This is to be incorporated at the next developmental stage. The rules are used as filters for the gapped matching of idioms. They can be downloaded as a separate file, which can be used as an add-on providing filtering rules for lookup systems.

#### Interface and Usage

The initial screen consists of a search box into which text (a sentence) that contains an idiom can be input. Figure 1 shows the system output when a user inputs the sentence “このままでは足がすぐ出る。” (kono mama deha ashi ga sugu deru = we will run short of money soon). The system outputs an idiom entry matched to this input, together with word-level matching information. Note that the idiom entry “足が出る” is retrieved through gapped matching.



Figure 1: Initial system output for “このままでは足がすぐ出る”.

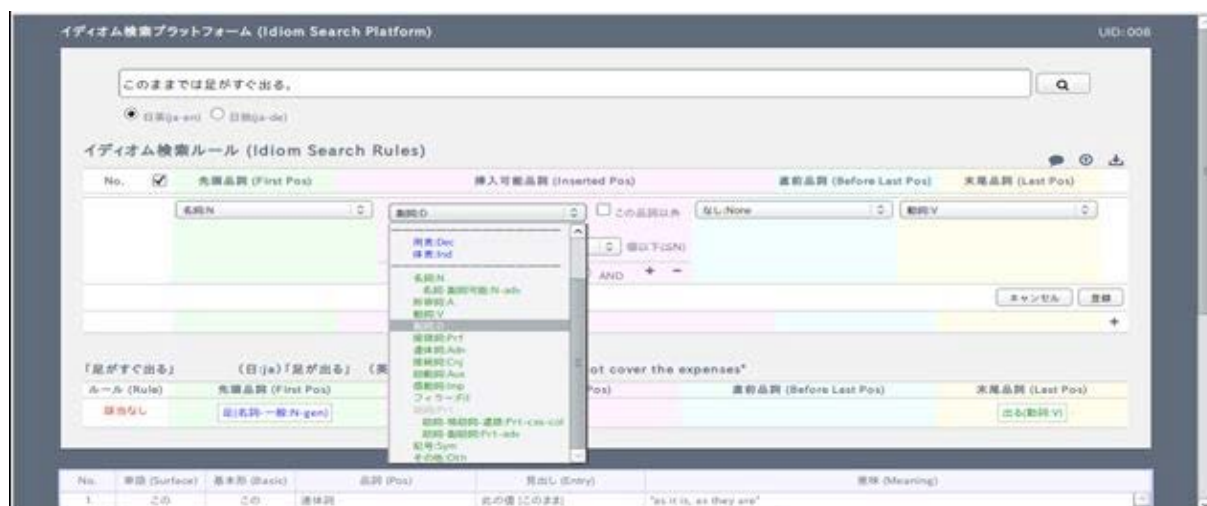


Figure 2: Defining an idiom variation matching rule.



Figure 3: Validating the rule.

This example shows that an adverb can be inserted between “足が” (ashi ga) and “出る” (deru), which enables us to define a rule allowing for the insertion of (an) adverb(s) in this position. The definition of a rule can be done by choosing POS patterns from the pull-down menu for the two constituent elements of the idioms and possible POSes which can be inserted in between (Figure 2). Once the rule is registered, the matching is carried out by applying the rule, which is shown by the rule number given in the result in Figure 3. Rules are defined as POS-based filtering patterns, so rules can be applied to other idioms which were not used for defining the pattern. The rule set can be downloaded as a text file.

## 4 Prospects

The system is deliberately simple for two reasons: (a) so that non-computationally oriented language practitioners and learners can use it and (b) so that the resultant variation rules can be exported to a variety of dictionary lookup systems. The variation matching rules defined in this platform can be straightforwardly incorporated into the dictionary lookup routine of the translation aid system we maintain (Utiyama, et al. 2009) (and in other systems if only mapping of POS sets is made). Currently, the system assumes that individual users define rules independently. While adding a collective mode is technically not difficult, whether that would be effective in defining rules is not yet clear. We are currently discussing this issue with a limited number of users of the system, while testing the system by providing users with a set of idioms and variation examples, and asking them to individually construct rules. The rules thus created will be unified and adjusted. After this cycle, we will be able to determine whether adopting collective coordination from the start would be more useful or not.

## 5 References

- Breidt, E., Segond, F., & Valetto, G. (1996). Formal description of multi-word lexemes with the finite-state formalism IDAREX. In *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark, pp. 1036-1040.
- Breidt, E., & Feldweg, H. (1997). Accessing foreign languages with COMPASS. In *Machine Translation*, 12, pp. 153-174.
- Carl, M., & Rascu, E. (2006). A dictionary lookup strategy for translating discontinuous Phrases. In *Proceedings of the European Association for Machine Translation 2006*, Oslo, Norway, pp. 49-58.
- Čermák, F. (2001). Substance of idioms: Perennial problems, lack of data, or theory? In *International Journal of Lexicography*, 14(1), pp. 1-20.
- Fraser, B. (1970). Idioms within a transformational grammar. In *Foundations of Language*, 6, pp. 22-42.
- Kageura, K., & Toyoshima, M. (2006). Analysis of idiom variations in English for the enhanced automatic look-up of idiom entries in dictionaries. In *Proceedings of the 12th EURALEX International Congress*, Torino, Italy, pp. 989-995.
- Michiels, A. (2000). New developments in the DEFI matcher. In *International Journal of Lexicography*, 13(3), pp. 151-167.
- Moon, R. (1998). *Fixed Expressions and Idioms in English*. Oxford, United Kingdom: Clarendon Press.
- Nicolas, T. (1995). Semantics of idiom modification. Everaert, M. et al. (eds.) *Idioms: Structural and Psychological Perspectives*. Hillsdale: Lawrence Erlbaum, pp. 233-252.
- Numberg, G., Sag, A. I., & Wasow, T. (1994). Idioms. In *Language*, 70(3), pp. 491-538.
- Proszeky, G., & Kis, B. (2002). Context-sensitive electronic dictionaries. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, pp. 1-5.
- Takeuchi, K., Kanehira, T., Hilao, K., Abekawa, T., & Kageura, K. (2007). Flexible automatic look-up of English idiom entries in dictionaries. In *Proceedings of the Machine Translation Summit XI*, Copenhagen, Denmark, pp. 451-458.
- Utiyama, et al. (2009) Minna no Hon'yaku: A website for hosting, archiving, and promoting translations. In *Translating and the Computer* 31, London.

**Acknowledgements**

This work is supported by the JSPS-DAAD bilateral collaboration project “Flexible matching of Japanese collocations in a translation environment with Japanese as the source language”.



# Tracing Sentiments: Syntactic and Semantic Features in a Subjectivity Lexicon

Jana Šindlerová, Kateřina Veselovská, Jan Hajič jr.  
Charles University in Prague, Faculty of Mathematics and Physics  
{sindlerova,veselovska,hajicj}@ufal.mff.cuni.cz

## Abstract

In this paper we present a syntactic and semantic analysis of verbal entries in the Czech subjectivity lexicon Czech SubLex 1.0 concerning their semantic and valency properties with respect to the roots and degree of subjectivity and evaluativeness. We demonstrate that evaluative verbs share certain abstract syntactic patterns with valency positions encoding the position of the source and target of evaluative stance. These patterns then roughly correspond to semantic properties of the verbs. For example, verbs propagating sentiments to the Actor position usually describe events of destruction (negative sentiments), or progress (positive sentiments), or events of direct experiencing emotional states, whereas verbs propagating sentiments to the Addressee or Patient position usually describe events of taking and communicating a stance, stopping or eliminating, or praising. The analysis represents the first step towards enhancing the lexical database with syntactic and semantic features in order to suit the lexicon for the task of the detection of targets (and sources) within evaluative stances.

**Keywords:** subjectivity lexicon; valency; sentiment analysis

## 1 Introduction

Sentiment analysis is a subfield of natural language processing searching for, extracting and classifying opinionated segments of text. One of its main goals is the identification of a positive or negative polarity, or neutrality of a sentence (or, more broadly, a text). Usually, this takes place by means of detecting the polarity items, i.e. words or phrases inherently bearing a positive or negative value. The polarity items are usually collected in the so-called subjectivity lexicons. The implementation of polarity items from the subjectivity lexicon into the data is the first step towards sentiment analysis.

There are more ways to build a subjectivity lexicon. A popular method for languages with sparse resources is providing a small amount of manually selected seed words and using bootstrapping algorithms to grow the list of word candidates (Banea et al. 2008). Polarity items can also be found by training probabilistic models on manually annotated data. Other approaches use translations of existing subjectivity lexicons (Milhacea et al. 2007), sometimes enhanced with triangulation methods (Steinberger et al. 2012).

There is a number of papers dealing with the topic of building subjectivity lexicons for particular languages, see e.g. (Bakliwal et al. 2012), (De Smedt and Daelemans 2012), (Jijkoun and Hofmann 2009) or (Perez-Rosas et al. 2012). The ongoing research on sentiment analysis in Czech language (Habernal et al. 2013),

(Veselovská et al. 2012) manifested the need for compiling a subjectivity lexicon for Czech. In 2013, the Czech Sublex 1.0, a subjectivity lexicon for Czech, was made publicly available.<sup>1</sup>

The experiments with incorporating the subjectivity lexicon into sentiment classifiers (Veselovská et al. 2013) hint that unfortunately, plain lexical databases do not suffice. The performance of classifiers suffers from the lack of verb sense disambiguation stage. Moreover, it is suggested that sentiments should be approached in a compositional way (Neviarouskaya et al. 2009), combining lexical, semantic and syntactic information. Incorporating syntactic and semantic information into lexicons has already become an established lexicographic praxis, a variety of valency lexicons have been edited so far and semantic class annotation has become a popular method of enhancing lexicographic annotation of verbs. Since verbs are considered the core of the sentence, naming the events and linking the participants into a coherent situation, it is of importance to capture their properties, syntactic and semantic, in complexity.

In this paper we present a preliminary linguistic analysis of verb entries in a Czech subjectivity lexicon Czech SubLex 1.0 concerning their semantic and valency properties with respect to the roots and degree of subjectivity and evaluativeness. On the basis of the analysis, we suggest enhancing the lexicon data with pointers to a Czech valency lexicon and to a semantic class database.

## 2 Methodology and Data

The presented analysis is based on the entries from the Czech SubLex 1.0. The core of the Czech subjectivity lexicon has been gained by automatic translation of a freely available English subjectivity lexicon,<sup>2</sup> see (Milhacea et al. 2007), which is a part of the OpinionFinder system (Wilson et al. 2005) for automatic subjectivity detection in English. The clues in this lexicon were collected from a number of both manually and automatically identified sources (Riloff and Wiebe 2003). The lexicon data have been translated into Czech via the parallel corpus CzEng 1.0 (Bojar et al. 2012) containing 15 million parallel sentences (233 million English and 206 million Czech tokens) from seven different types of sources automatically annotated at surface and deep layers of syntactic representation.

Czech sentence	English translation	Syntactic pattern
Topolánek není důvěryhodný.	Topolánek is not trustworthy.	ACT <sub>TARGET</sub> PRED <sub>COPULA</sub> PAT <sub>EVAL</sub>
Považuji tento film za špatný.	I consider the film poor.	ACT <sub>SOURCE</sub> PRED <sub>PSYCH</sub> PAT <sub>TARGET</sub> EFF <sub>EVAL</sub>
Zeman si se vyjádřil o Klausovi kriticky.	Zeman spoke critically of Klaus.	ACT <sub>SOURCE</sub> PRED <sub>COMM</sub> PAT <sub>TARGET</sub> EFF MANN <sub>EVAL</sub>
Jiří Paroubek udělal chybu.	Jiří Paroubek has made a mistake.	ACT <sub>TARGET</sub> PRED <sub>EMPTY</sub> CPHR <sub>EVAL</sub>

**Table 1: Examples of syntactic patterns for non-evaluative verbs in evaluative stances.**

1 <http://hdl.handle.net/11858/00-097C-0000-0022-FF60-B>

2 Available at [http://mpqa.cs.pitt.edu/lexicons/subj\\_lexicon](http://mpqa.cs.pitt.edu/lexicons/subj_lexicon)

Czech sentence	English translation	Syntactic pattern
Líbí se mi to jméno.	I like the name	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>
Duchovní láska člověka obohacuje.	Spiritual love enriches the man.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT EFF
Nový ministr zdravotnictví dráždí novináře.	The new health minister irritates journalists.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT <sub>SOURCE</sub>
Novináři kritizují nového ministra zdravotnictví.	Journalists criticize the new health minister.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>

**Table 2: Examples of syntactic patterns for evaluative verbs.**

The output of the translation were 7228 items – evaluative expressions. These items have been further manually inspected for reliability. During the refinement process, items have been removed from the lexicon when considered errors or unreliable items. The reasons for excluding an item from the list were manifold, reaching from the “lost in translation” phenomenon, through predictable errors of the system (e.g. in translating negated items) to a significantly different degree of evaluativeness of the Czech word. The final set consists of 4625 evaluative expressions, of which 1549 are verbs. Within the analysis, each verbal item of the lexicon has been considered separately in order to decide which valency argument of the verb corresponds to the target of the sentiment propagated by the verbal evaluative semantics.

The current version of Czech Sublex 1.0 contains information about word lemma, part of speech, polarity orientation and source-language equivalent. Our intention to the future is to add information about valency and semantic class characteristics to relevant entries, possibly by means of pointers to existing resources, such as VALLEX 2.5<sup>3</sup> (Lopatková et al. 2007), FrameNet (Ruppenhofer et al. 2006), or VerbNet (Schuler 2005).

### 3 From Lexicon to Sentence: Preparing the Grounds for Compositional Approach

There are several reasons why enhancing a subjectivity lexicon with information about verbal valency is valuable. First, different valency frames serve as unique identifiers of verb senses. It is a common phenomenon that individual senses of a lemma differ with respect to the presence (or absence), degree and orientation of polarity. Disambiguating different senses of a verb lemma allows us to identify sentiments more precisely. For example, in case of the verb “abdicate”, we are able to differentiate between the intransitive pattern “to leave a position” which does not constitute evaluative meaning directly, and the transitive pattern (“abdicate one’s responsibilities”), meaning “to fail” and creating an evaluation stance with opinion target at the position of the Actor. There is a whole group of verbs in

3 <http://hdl.handle.net/11858/00-097C-0000-0001-4908-9>

the original English lexicon sharing the non-evaluative semantics of “action under a physical disorder”, which in their second sense describe an evaluative stance (“hobble”, “jolt”, “stammer” etc.). By feeding them into the translation process without verb sense disambiguation we risk gaining a considerable number of inappropriate lexical units which may later spoil the polarity tracking results. In a larger perspective, such a disambiguation process represents a decision between real subjective sentiments and the so-called “good” or “bad news” (objectively presented positive and negative content), a task recognized as important e.g. in the sentiment analysis of the news. (Balahur et al. 2010).

Valency is expected to be helpful in the task of the identification of the target of the evaluation as well. Traditionally, the subjectivity analyses distinguish three components of an evaluative private state<sup>4</sup> that need to be distinguished (Wiebe et al. 2004): the source, i. e. the entity expressing the private state, the target, i.e. the evaluated entity, and the evaluation, i.e. polar elements, words or phrases bearing positive or negative value.

Sublex verb	Czech sentence	English translation
abdikovat (abdicate)	Císař Vilém II. byl přinucen abdikovat.	The emperor Wilhelm II was forced to abdicate.
amputovat (amputate)	Lékaři mu museli amputovat chodidlo.	The doctors had to amputate his foot.
degenerovat (degenerate)	Schopnost naučit se mluvit u člověka degeneruje.	The ability of learning to speak degenerates in humans.
dovádět (frolic)	Tanečnice na parketu dovádí jako malé děti.	The dancers frolicked on the dance floor like little children.
hladovět (starve)	Přiberu pět kilo, pak zase hladovím.	I put on five kilos, then I starve again.

**Table 3: Examples of verbs not propagating sentiments to any of its arguments.**

From the corpora of evaluative texts we are able to extract typical abstract syntactic (and semantic) patterns for expressing evaluative meaning.<sup>5</sup> Some verbs only serve as syntactic hints for evaluative words (evaluative nouns, adjectives, or adverbs), typically, this is the case of copular verbs, “psych” verbs (verbs describing mental action), communication eliciting verbs, or light verbs marking complex predication (phrasal verbs etc.), see table 1.

Other verbs function as bearers of the evaluation themselves, these are verbs which can be found in a subjectivity lexicon. In a typical verb-centered evaluative stance, evaluation as such is carried by the verb, while the source and the target of the evaluation occupy the positions of verb arguments. The

4 A subjective state, i.e. a state not open to objective observation or verification (see e.g. Ruppenhoffer et al. 2008).

5 The patterns presented in the tables are constructed in accordance with Functional Generative Description valency theory labelling standards, see (Sgall et al. 1986) based on the tectogrammatic layer of description. In this theoretical approach, the tectogrammatic layer represents deep syntactic relations between words, with certain extension into the area of semantic relations.

verbs in the lexicon then differ with respect to the question of sentiments propagation to individual arguments. Examples of syntactic patterns for evaluative verbs can be found in table 2.

A number of verbs which appear in the lexicon do not propagate sentiments to any of its arguments. These are most probably candidates for what we call “good/bad news” items. We describe good/bad news items as terms designating generally positive or negative situations or facts (like “war”, “disaster” “luck” etc.). The good/bad news verbs (in their primary meaning) do not evoke a positive or negative attitude to an entity/situation/fact occupying any of the valency positions, rather they function at the same time both as the polar word and the target of the sentiment. Examples of such verbal items are listed in table 3.

Due to the fact that none of the good news/bad news verbs propagates the sentiment to any of its valency participants, it is necessary to mark them as a separate category in the lexicon. Still, it is beneficial to keep them in the lexicon because they provably, though indirectly, influence emotions of the reader.

Table 4 contains verbs propagating sentiments to the Actor position. They usually describe events of destruction (negative sentiments), or progress (positive sentiments), or events of direct experiencing emotional states.

The interesting thing with verbs propagating sentiments to the Actor position is that they are usually verbs allowing the Abstract Cause-Subject alternation (Levin 1993), i.e. an alternation of valency participants of the type “Mike distorted the wonderful moment with a scream” and “Mike’s scream distorted the wonderful moment”. Different aspects of the semantic shift between the two alternations are widely discussed (Alexiadou and Schäfer 2006) and the shift of the sentiment focus can be seen as significant in this respect.

As can be seen from table 5, verbs propagating sentiments to an Addressee or Patient position usually describe events of taking and communicating a stance (both polarities), stopping or eliminating (negative), or praising (positive).

Another pattern is evident from the data: the target of the evaluation is the centre of the evaluative stance. The way the source of evaluation is expressed is dependent on the verb’s semantic choice of the target argument. If the target is expressed by a PAT argument, the source occupies the ACT position. If the target is selected at the ACT position, the source must be expressed external to the clausal structure (e.g. by means of “in my opinion” etc.).

The issue of propagating sentiments is more than complicated. There are of course more argument types than we have suggested so far to which sentiments can be propagated. The sentiments may be propagated to more than one argument in a structure. For example, in a sentence “John criticized Mary for her not coming,” the negative sentiment affects not only “Mary” as the patient, but also “her not coming” as the cause of critique. The same may apply to verbs allowing the Abstract Cause-Subject alternation, where the sentiments may affect secondarily not only the Actor position, but also the position of the Abstract Cause if present overtly.

Sublex verb	Czech sentence	English translation	Syntactic pattern
bavit (amuse)	Hoteliérství mě baví ze všeho nejvíc.	I most enjoy being a hotel owner.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT <sub>SOURCE</sub>
děsit (freak)	Nekonečná samota tě děsí.	The neverending solitude freaks you out.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT <sub>SOURCE</sub>
kazit (spoil)	Nedovolím ti kazit mi život.	I won't allow you spoil my life.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT
narušit (distort)	Nádhernou chvíli narušil výkřik.	The wonderful moment was distorted by a scream.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT
naštvat (upset)	Rozhodčí naštvál domácího borce.	The referee upset a guy from the home team.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT <sub>SOURCE</sub>
ohrozit (endanger)	Těžba ohrozí existence jejich domovů.	Mining will endanger the existence of their homes.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT
zachránit (rescue)	Dobré jméno vlády zachránil ministr Bursík.	The Government's credit was saved by minister Bursík.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> PAT EFF
zlepšit se (improve)	Zlepšila se jí pleť a rozjasnily oči.	Her skin improved and her eyes brightened.	ACT <sub>TARGET</sub> PRED <sub>EVAL</sub> ORIG PAT

**Table 4: Example of verbs propagating sentiments to the Actor position.**

Deciding the sentiment propagation direction also may be a nontrivial question – which of the two affected arguments receives the sentiment primarily and which acquires it on the basis of some semantic transfer. To make the issue even more fuzzy, there is more than one kind of sentiment. We must distinguish between the sentiments that affect the “source of evaluation” in the text and sentiments which affect the “perceptor” of the text. Thus for example, in the sentence “John ignored Mary without reason,” Mary is the target of negative sentiments of John, but John may also be a target for negative sentiments held by the reader. Similar transfer of sentiments from the “inner” sentential structure (textual source of sentiments) to the “external” reader’s perception appears with many verbs propagating sentiment to a non-actor position (cf. table 5). In a valid lexicon for opinion target extraction, we must keep the information about which of the two sentiments (reader-oriented or source-oriented) we want to trace.

Though the overall situation is quite complex, it can be seen from the analysis that verbs propagating sentiments to the same arguments usually belong to the same semantic classes, or at least share the same semantic components. The clusters of semantically similar verbs arising in the analysis are well traceable in common semantic class databases, like VerbNet or FrameNet. Thus, exploring the semantic affiliation of the verbs and recording it in the lexicon may be beneficial for further lexicon bootstrapping tasks.

## 4 Conclusions and Future Work

We have described a newly emerged Czech subjectivity lexicon SubLex 1.0. The method of automatic translation from a source to a target language is on one hand quick and easy, on the other hand demands a further refinement processes, which may be costly, and brings certain challenging consequences: the target lexicon has different properties (part-of-speech distribution, degree of evaluativeness of individual words, possibly even polarity orientation of the individual words) than the source one.

Subjectivity lexicon capturing the information about prior polarities of words is a useful and needed resource for sentiment analysis of textual data. Nevertheless, it does not suffice for sentiment analysis tasks on its own. For a successful analysis of sentiment, syntactic and semantic patterns must be also employed, in order to prevent mistakes and handle the data appropriately.

We have offered here a brief analysis of the subjectivity lexicon data both in the contrastive perspective to the original items and in the mutual relations of the lexical items in the paradigm of valency and semantic class characteristics. This analysis is a first step towards a more thorough research into the linguistic properties of expressing evaluation and towards implementing the theoretical knowledge into sentiment analysis experiments.

The methodology described above is expected to ease the process of identification of the source and target of the evaluation, which would not be possible with a simple plain text with no semantic features annotated. In the near future, we would like to accomplish the extended annotation of the Czech SubLex with labels designating the typical deep syntactic pattern of the evaluative stance and verify our findings by a series of experiments. In the first step, we would like to map the verbs from SubLex also into the PDT-Vallex (Urešová 2009), a valency lexicon which is interlinked with the dependency tree-bank, and try to automatically extract the sources and targets of the evaluation on the syntactically annotated data of the PDT, where both syntactic and semantic roles are manually annotated.

verb	Czech sentence	English translation	Syntactic pattern
bát se (fear)	Bojím se, že přijdu o všechny své peníze.	I fear losing all my money.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>
degradovat (degrade)	Tento přístup degraduje ženy na pouhé sexuální objekty.	This approach degrades women to mere sex objects.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>
doporučit (recommend)	Studium lingvistiky bych doporučil každému studentovi.	I would recommend studying linguistics to any student.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> ADDR PAT <sub>TARGET</sub>
důvěřovat (trust)	Tvému úsudku plně důvěřuji.	I fully trust your opinion.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>
eliminovat (eliminate)	Je potřeba eliminovat falešná doznání.	It is necessary to eliminate false confessions.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>

verb	Czech sentence	English translation	Syntactic pattern
kárat (reproach)	Vedoucí káral nevkusně oděného účetního.	The manager reproached the tastelessly dressed accountant.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>
odmítnout (reject)	Odmítnul nabídku členství v KSČ.	He rejected the offer of becoming a member of the communist party.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>
oslavovat (praise)	Švýcaři oslavují nového šampióna ve sjezdovém lyžování.	The Swiss praise the new champion in alpine skiing.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub>
prosazovat (advocate)	Rychlé přijetí evropské měny prosazuje Jan Švejnar.	Jan Švejnar advocates prompt adoption of the euro.	ACT <sub>SOURCE</sub> PRED <sub>EVAL</sub> PAT <sub>TARGET</sub> EFF

**Table 5: Example of verbs propagating evaluation to the position of Addressee or Patient.**

## 5 References

- Alexiadou, A. & Schäfer, F. (2006). Instrument Subjects Are Agents or Causers. In *Proceedings of WCCFL* (Vol. 25, No. 40-48).
- Bakliwal, A., Piyush, A. & Varma, V. (2012). Hindi Subjective Lexicon: A Lexical Resource for Hindi Adjective Polarity Classification. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, pp. 1189-1196.
- Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belayeva, J. (2012). Sentiment Analysis in the News. *arXiv preprint arXiv:1309.6202*.
- Banea, C., Mihalcea, R. & Wiebe, J. (2008). A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources. In *The Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 2764-2767.
- Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M. & Tamchyna, A. (2012). The Joy of Parallelism with CzEng 1.0. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012)*, pp 3921-3928.
- De Smedt, T. and Daelemans, V. (2012). Vreselijk mooi! (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives. In *Proceedings of the 8<sup>th</sup> Language Resources and Evaluation Conference (LREC 2012)*, pp. 3568-3572.
- Habernal, I., Ptáček, T., & Steinberger, J. (2013). Sentiment Analysis in Czech Social Media Using Supervised Machine Learning. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2013)*.
- Jijkoun, V. & Hofmann, K. (2009). Generating a Non-English Subjectivity Lexicon: Relations That Matter. In *Proceeding of: EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics*.
- Levin, B. (1993). English verb classes and alternations: A preliminary investigation. University of Chicago press.
- Lopatková, M. et al. (2007). VALLEX 2.5 – Valency Lexicon of Czech Verbs, version 2.5, *Software prototype*, 16: 1.
- Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning Multilingual Subjective Language via Cross-lingual Projections. In *Annual Meeting – Association for Computational Linguistics*, (Vol. 45., No. 1, p. 976).
- Neviarouskaya, A., Prendiger, H. & Ishizuka, M. (2009). Semantically distinct verb classes involved in sentiment analysis. In *IADIS AC (1)*, pp. 27-35.



- Perez-Rosas, V., Banea, C. & R. Mihalcea (2012) Learning Sentiment Lexicons in Spanish. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC2012)*, pp 3077-3081.
- Riloff, E. & Wiebe, J. (2003) Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R. & Scheffczyk, J. (2006) FrameNet II: Extended theory and practice. Accessed at: <http://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf> [04/12/2013].
- Ruppenhofer, J., Somasundaran, S. & Wiebe, J. (2008). Finding the Sources and Targets of Subjective Expressions. In *The Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 2781-2788.
- Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. PhD thesis. University of Pennsylvania, US.
- Sgall, P., Hajičová, E. & Panevová, J. (1986). The meaning of the sentence in its semantic and pragmatic aspects. Springer.
- Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S. & Zavarella, V. (2012). Creating Sentiment Dictionaries via Triangulation. In *Decision Support Systems* 53(4), pp. 689-694.
- Urešová Z. (2009). Building the PDT-VALLEX valency lexicon. In: *On-line Proceedings of the fifth Corpus Linguistics Conference*, University of Liverpool, UK.
- Veselovská, K., Hajič Jr, J., & Šindlerová, J. (2012). Creating Annotated Resources for Polarity Classification in Czech. In *Proceedings of KONVENS*, pp. 296-304.
- Veselovská, K., Hajič Jr, J. & Šindlerová, J. (2013). Subjectivity Lexicon for Czech: Implementation and Improvements. To appear in *Proceedings of KONVENS*. 2013.
- Wiebe, J., Wilson, T., Bruce, R., Bell, M. & Martin, M. (2004). Learning subjective language. In *Computational linguistics* 30(3), pp. 277-308.
- Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E. & Patwardhan, S. (2005). OpinionFinder: A System for Subjectivity Analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, Association for Computational Linguistics, pp. 34-35.

## Acknowledgement

This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). This research has been partially supported by SVV project number 260 104. Participation in the conference was supported by Foundation of Vilem Mathesius.



# **Lexicography and Corpus Linguistics**



# Compatible Sketch Grammars for Comparable Corpora

Vladimír Benko

Slovak Academy of Sciences, L. Štúr Institute of Linguistics

Comenius University in Bratislava, UNESCO Chair in Translation Studies

vladob@juls.savba.sk

## Abstract

Our paper describes an on-going experiment aimed at creating a family of billion-token web corpora that could to a large extent deserve the designation “comparable”: corpora are of the same size, data gathered by crawling the web at (approximately) the same time, containing similar web-specific domains, genres and registers, further pre-processed, filtered and deduplicated by the same tools, morphologically annotated by (possibly) the same tagger and made available via Sketch Engine. To overcome the problem of great differences in the existing sketch grammars for the respective languages, a set of “compatible” sketch grammars have been written that will aid contrastive linguistic research and bilingual lexicographic projects. The sketch grammars use a uniform set of rules for all word categories (parts of speech) and the resulting set of tables is displayed in a fixed order in all languages.

**Keywords:** comparable web corpora; sketch grammars; bilingual lexicography

## 1 Introduction

Ten years after its introduction to the lexicographic community at the Lorient Euralex Congress (Kilgarriff et al., 2004), Sketch Engine (*SkE*) has become a standard tool in numerous lexicographic projects, as well as in various areas of corpus-based linguistic research. Sketch grammars for corpora in many languages have been written (cf. References). Recently published open-source tools for efficient web crawling (Suchomel a Pomikálek, 2012) stimulate the building of very large web corpora, the analysis of which is hardly imaginable without a powerful summarisation machine such as *SkE*. Newly implemented *SkE* support for parallel and comparable corpora (Kovář, 2013) facilitate its use in the area of bilingual lexicography and contrastive linguistic research.

In bilingual and multilingual linguistic work with *SkE*, we often encounter the problem of sketch grammars defining the collocational profiles of a headword and its translation equivalent for the respective languages. Those sketch grammars have often been created for different purposes, having in mind different user requirements, with resulting word sketches being rather disparate, making its use for contrastive linguistics problematic. Our paper suggests an alternative approach to the creation of sketch grammars, within the framework of which the respective grammars can be made compatible for (almost) all languages.

## 2 The Aranea project

### 2.1 Why new corpora?

Besides our interest in testing the new corpus-building tools, the motive for starting a new corpus project was the lack of suitable corpora that could be used by students of foreign languages and translation studies at our university. The existing web corpora families that are available for download do not cover all the languages needed. As for corpora stored at the *SkE* web site<sup>1</sup>, they (1) are not available for download, (2) are mostly too large for classroom use<sup>2</sup>, and (3) have too different sketch grammars, which makes them difficult to use in a mixed-language classroom.

We expect that a set of corpora for several languages of equal size and built by a standardized methodology can not only be used for teaching purposes, but also in other areas of linguistic research (contrastive studies) and in lexicography (both mono- and bilingual).

### 2.2 The names

For our corpora, we have decided to use “language-neutral” Latin names denoting the language of the texts and the corpus size. The whole corpus family is called *Aranea*, and the respective members bear the appropriate language name, e.g. *Araneum Anglicum*, *Araneum Francogallicum*, *Araneum Russicum* for English, French, and Russian, respectively, etc.

Each corpus exists in several editions, differing by their sizes. The basic (medium-sized) version, *Maius* (“greater”), contains approximately 1.2 billion tokens (i.e., over 1 billion words). This size can be reached relatively quickly for all participating languages, and for the “large” ones with plenty of web data available, it usually takes just one or two days of download time. The 10% random sample of *Maius*, called *Minus* (“smaller”), is to be used for teaching purposes (e.g. for lessons in the framework of Corpus Linguistics programmes for students of foreign languages and translation studies). A 1% sample, *Minimum* (“minimal”), is not intended to be used directly by the end users, and is utilized in debugging the processing pipelines and tuning the sketch grammars. And lastly, the largest *Maximum* (“maximal”) edition will contain as much data as can be downloaded from the web for the particular language, and its size is mostly determined by the configuration of the server.

---

1 <http://www.sketchengine.co.uk/>

2 According to our experience, the ideal corpus for teaching corpus linguistics is about BNC-sized, i.e. it contains some 100 million tokens. As it is not easy to prevent students from invoking search operations taking several minutes to evaluate, billion-plus token corpora proved to be quite unsuitable for teaching purposes.

## 2.3 Web crawling

The source data acquisition is being performed by means of *SpiderLing*<sup>3</sup>, a web crawler optimized for collecting textual data from the web. The system contains an integrated character encoding (*chared.py*) and language recognition (*trigrams.py*) module, as well as a tool for boilerplate removal (*justext*). The input seed URLs (some 1,000 for each language) have initially been harvested by BootCAT<sup>4</sup> (Baroni and Bernardini; 2004).

Several input parameters of the crawling process are to be set by the user, most notably the language name, a file containing sample text in the respective language (to produce a model for language recognition), a language similarity threshold (a value between 0 and 1; default 0.5), the number of parallel processes, and the crawling time.

In our processing, we usually crawled the web in 24-hour slots (the process could then be re-started) with all other values set to defaults. The only exception was crawling for Slovak and Czech, where we used 7-day slots, as the process was much slower here. The language similarity threshold also had to be changed in case of Slovak and Czech. As these languages are fairly similar, the trigram method did not seem to be able to distinguish between them sufficiently. We have therefore increased the similarity threshold value to 0.65 (saving many “good” documents, and causing many “wrong” ones to pass the filter) and removed the unwanted texts by subsequent filtration based on character frequencies .

## 2.4 Post-download processing

Besides the basic filtration aimed to remove texts with incorrect or misinterpreted character encoding, missing diacritics and texts with non-standard proportion of punctuation and uppercase characters, the main processing operation in this phase is tokenization. As the tokenization strategy has to be compatible with that of the corpus used to train the tagger, we decided to use the tokenizers supplied with Tree Tagger and TaKIPI for the respective languages. In the future, we would like to make use of the *unitok.py* universal tokenizing program developed at Masaryk University in Brno (Jakubiček; 2014).

## 2.5 Deduplication

The whole procedure (Benko; 2013) consists of three stages. The first stage detects near-duplicate documents by means of the Onion (Pomikálek; 2012) utility (similarity threshold 0.95), and the duplicate documents are deleted. The second stage deduplicates the remaining text at the paragraph level using the same procedure and settings. The tokens of the duplicate paragraphs, however, are not deleted but rather they are marked to make them “invisible” during corpus searches, while they can be displayed

---

3 <http://nlp.fi.muni.cz/trac/spiderling>

4 <http://bootcat.sslmit.unibo.it/>

as context at the boundary of non-duplicate and duplicate text. In the last stage, we make use of our own tool based on the fingerprint method (with ignoring punctuation, special graphics characters and digits) to deduplicate the text at the sentence level. The tokens of duplicate sentences are marked similarly to the previous stage. This last step can “clean up” the duplicities among the short segments that fail to be detected as duplicates by Onion.

As deduplication is beyond the scope of our paper, we only mention here that the process has typically removed some 20–45% of tokens in the *Maius* versions of our corpora

## 2.6 Morpho-syntactic annotation

For languages covered by the parameter files of Tree Tagger (Schmid; 1994), this tagger has been used to annotate the respective corpora. For Polish, the TaKIPI (Piasecki; 2007), and for Czech, the Morče (Hajič; 2004) taggers were used, respectively. The question of tools for tagging Hungarian and Ukrainian data has not been resolved yet.

## 2.7 Tagging-related filtration

To improve the precision of tag assignments, a series of pre- and post-tagging filters have been developed that fix issues introduced by Unicode encoding of the source text<sup>5</sup>. The filtration fixes known tagger issues for the respective languages, namely the misassigned tags for many punctuation and special graphic characters (that are often tagged as nouns, adjectives, or abbreviations, and sometimes even as verbs with subcategories). For some languages, an additional tag with masked subcategories for gender and number is created, that is later used by some rules within the respective sketch grammars.

## 2.8 Current state of the project

At present, eight language versions of the *Aranea* corpus family have been created, containing both *Maius* and *Minus* editions as follows (in chronological order): *Araneum Russicum* (Russian), *Araneum Francogallicum* (French), *Araneum Germanicum* (German), *Araneum Hispanicum* (Spanish), *Araneum Polonicum* (Polish), *Araneum Anglicum* (English), *Araneum Nederlandicum* (Dutch), and *Araneum Slovaccum* (Slovak).

---

5 As an example we can point out the problem of the “apostrophe” character in French texts. As much as 8 different Unicode characters representing the apostrophe (with just two of them being “canonical”) can be found in the texts collected from the web. As the Tree Tagger French parameter file originated in the pre-Unicode era, even one of the two “canonical” representations would not be processed (i.e., tokenized and lemmatized) properly without special measures, and tokens like “l” and “d”, that belong to the most frequent ones, would be mistagged.



The crawling has also been done for *Araneum Bohemicum* (Czech). This data is now being pre-processed to be ready for annotation that will be performed by the Institute of Theoretical and Computational Linguistics at the Faculty of Arts of Charles University in Prague.<sup>6</sup>

The first stage of our project will be completed by *Araneum Hungaricum* (Hungarian), *Araneum Italicum* (Italian), and *Araneum Ucrainicum* (Ukrainian). With the exception of the last mentioned, we expect to complete the whole venture by the end of 2014.

For all of the languages mentioned, sketch grammars have been written and at least two rounds of testing have been performed for each corpus. The procedure involved is described in the following section.

### 3 Sketch grammars

A sketch grammar<sup>7</sup> is a set of rules based on the CQL (Corpus Query Language<sup>8</sup>) used by the Sketch Engine to generate the respective collocation profiles (“word sketches”) for all lexical units (lemmas) in a corpus. The word sketches are pre-computed in advance, which makes the system user-friendly and very fast.

A sketch grammar rule consists of (1) an optional comment indicated by hash “#” character, (2) the rule type marked by an asterisk “\*”, (3) the rule name preceded by the equal sign “=”, and (4) a list of CQL expressions. For example, a rule describing the relationship between two nouns (in English using the Penn Treebank tagset) might look as follows:

```
# noun followed by another noun
*DUAL
=modifier/modified
      2: [tag="NN.*"] 1: [tag="NN.*"]
```

The “1:” label denotes the “keyword”, i.e. the lemma the word sketch is created for, and the “2:” label marks the lemma of the collocate. The “\*DUAL” keyword indicates that the rule is to be used twice, the second time with swapped labels, i.e. exchanging the positions of the keyword and the collocate. The text following the slash “/” character will be used as a name for the second use of the rule.

In reality, the rules usually look slightly more complex to indicate that “intermediate” words may be present between a keyword and a collocate, or in the vicinity of them.

---

6 Besides Ukrainian, Czech is the only language within the *Aranea* project with no free tagging tool available.

7 <https://www.sketchengine.co.uk/documentation/wiki/SkE/GrammarWriting>

8 <https://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying>

### 3.1 What's in a name

Unlike Juliet Capulet<sup>9</sup>, we believe that the name is often really important, and the sketch grammar rule name is exactly such a case. On one hand, it is the only component of the sketch grammar that is not predetermined, and thus can be “virtually anything”. On the other hand, the name is the only clue for the user about the contents of the respective word sketch tables, and therefore should be as informative as possible. It has, however, to be very short as the name is displayed in the heading of the respective word sketch table within a only a limited space available. Rule names longer than 10–12 characters would increase the table widths, and the resulting word sketches could possibly not fit the screen.

Most sketch grammars used for corpora available at the *SkE* site follow the naming conventions introduced by A. Kilgarriff in the first English and French sketch grammars. These rule names are motivated syntactically, i.e. they denote the syntactic function of the collocate, with that of the keyword being implied. For example the rule name:

=modifier/modified

is representing two rule names with readings as follows: “collocate is a modifier of the keyword”, and “collocate is modified by the keyword”, respectively.

The syntactically motivated rules are transparent and user-friendly for description of basic relationships between subjects, object, modifiers/attributes, and verbs/predicates, but in more complex cases this strategy is not easily applicable. The nature of the problems can be observed in the Czech sketch grammar written by P. Smrž (Kilgarriff et al.; 2004). Some examples of rule names are as follows:

is\_subj\_of/has\_subj  
 is\_obj4\_of/has\_obj4  
 prec\_prep  
 gen1/gen2

As it can be seen, it is not really easy for the user the figure out “who is who” in the keyword – collocate – syntactic function “puzzle”. Moreover, rule names like “prec\_verb” do not denote any syntactic functions but rather just describe collocational relationships.

There are two notable deviations from the “traditional” rule name conventions in the sketch grammars. In the grammar for the Slovenian FidaPLUS corpus<sup>10</sup>, S. Krek (Krek; 2006) uses rule names containing (among other features) Slovenian “case questions”. For example, the “*koga-česa*” name means

9 Juliet: “*What’s in a name? that which we call a rose / By any other name would smell as sweet*” (William Shakespeare: Romeo and Juliet, Act II, Scene 2).

10 <http://www.sketchengine.co.uk/documentation/wiki/Corpora/FidaPLUS>

that only collocates of the keyword that are in genitive case are displayed, with the syntactic function of the collocate being implied.

The second notable exception is the sketch grammar written by P. Whilelock (2010) for the Oxford English Corpus<sup>11</sup> (OEC) where the rule names not only name the syntactic function, but also the PoS of the keyword and the collocate and their mutual position within the collocation. For example, the “V\* ADJ” rule name stands for verb modified by an adjective, with asterisk indicating the keyword.

### 3.2 Sketch grammar for Slovak corpora

In our Institute, the *SkE* has been extensively used since autumn 2007 with several Slovak and Czech corpora. These corpora serve as a source of lexical evidence for our monolingual and bilingual lexicographic projects, as well as for other linguistic research activities.

The sketch grammar used in our *SkE* installation has been optimized for a lexicographic use, and differs from most “traditional” grammars for corpora stored at the *SkE* web site in several aspects:

- The rule names are not motivated syntactically (i.e., they do not indicate the syntactic relationship between the keyword and the collocate) but rather collocationally
- The right-hand or left-hand position of the collocate towards the keyword is indicated explicitly in the rule name
- The keyword’s PoS in the rule is not specified, i.e., it covers any PoS
- Recall is preferred over precision
- The number of rules and the order of resulting tables is fixed
- The object names within the rules are governed by the following rules:
- The keyword is denoted by the X symbol
- The keyword’s grammatical attributes (mostly in unary rules) are indicated by lowercase abbreviation, e.g., gen(X) indicates the genitive case of keyword
- The collocate’s PoS is indicated by an abbreviation with a leading capital letter, e.g., Aj X indicates a left-hand adjective collocate
- Y indicates a collocate that is from any PoS category
- Z indicates a collocate from any PoS category not covered by the other “explicit” rules

### 3.3 Rule name summary

The core of our grammar consists of rules covering four basic autosemantic word classes. Taking into account our experience with early versions of the grammar, the rules for verbs (Vb X/X Vb) and adverbs (Av X/X Av) do not distinguish the left and right position of the respective collocate.

---

11 <http://www.sketchengine.co.uk/documentation/wiki/Corpora/OEC>

For nouns, two separate rules take into account the position of the collocate (Sb X, X Sb). Similar situations can be found with adjectives (Aj X; X Aj), prepositions (Pp X; X Pp) and for immediate autosemantic collocates (Y X; X Y). The “catch all” rules for the remaining word classes (Z X; X Z) cover mostly numerals and pronouns, as well as some synsemantic word classes.

The remaining two binary (symmetric) rules map the relationship of coordination, i.e., the situation where a keyword and a collocate with compatible morphological tags are separated by a comma (X/Y, X/Y) or a conjunction (X/Y Cj X/Y).

The four trinary rules cover relationships among a keyword, collocate, and preposition in different positions (Pp Y X, Pp X Y, Y Pp X, and X Pp Y).

Our set of rules is complemented by unary rules showing the frequency distribution of the keyword’s forms according to grammatical categories and subcategories..

The compatible grammars

In creating sketch grammars for a group of languages, it is convenient not to use the “native” tagsets for the respective languages, but rather to use a common symbolic notation. This can be done, e.g., by means of a macro processor (such as m4). We have, however, decided to adopt a different approach and to create a simple universal tagset (“*Araneum Universal Tagset*” – AUT) similar to that of the *Universal PoS Tagset*<sup>12</sup> (UPT; Petrov et al., 2011), and to map all the respective tagsets into this tagset at the source vertical data level, i.e. to create a new layer of annotation. The AUT contains 11 tags for “traditional” part of speech categories, 7 additional tags for other elements, and one tag to indicate errors in the mapping process.

aTag	PoS	aTag	PoS	aTag	PoS
Dt	determiner/article	Pp	preposition	Xx	other (content word)
Nn	noun	Cj	conjunction	Xy	other other (function word)
Aj	adjective	Ij	interjection	Yy	unknown/foreign/alien
Pn	pronoun	Pt	particle	Zz	punctuation
Nm	numeral	Ab	abbreviation/acronym	Er	mapping error
Vb	verb	Sy	symbol		
Av	adverb	Nb	number		

**Table 1: Araneum Universal Tagset (AUT).**

The compatible sketch grammar using AUT consists of three sections. The first part (AUT-based) contains unary rules showing PoS category distribution for a particular lemma. The second part is

12 The AUT PoS tags for the eleven „traditional“ word classes directly correspond with those of UPT, with the difference being just in the names as we wanted to keep the names of the PoS categories identical with those used in the sketch grammar rule names introduced before the UPT tagset has been published. The additional 7 categories accommodate information provided by the respective “native” tagsets that is being ignored by UPT. For example, the “Xx” (other: content word) tag is assigned to participles in Slovak that have a category of their own in the SNK Slovak tagset.

tagset-dependent and contains unary rules showing PoS subcategories provided by the respective tagset. Due to differences in the depth of the morpho-syntactic annotation, the number of subcategories varies among the languages. With verbs, e.g., we have just 5 subcategories for Spanish, while more than 20 for Polish. The final third section (*AUT*-based) covers the collocational relationships of the respective keyword by means of binary, symmetric and trinary rules.

The compatible sketch grammar is basically identical for all the languages with one important exception: the number of intermediate tokens between a keyword and a collocate is increased by one for languages having articles in their language system.

## 4 Discussion and conclusion

A collocationally-based sketch grammar has (against a traditional one) several advantages. It can symmetrically cover all relationships between keywords and collocates of all word classes (parts of speech). As the PoS category is not tested for the keyword, a word sketch can be created even in cases of incorrectly assigned tags. If the same (compatible) sketch grammar is used with corpora for two or more languages, the resulting word sketches can be conveniently used in contrastive linguistic research, as well as within bilingual lexicographic projects.

The disadvantage of our approach is that not all tables for some words represent linguistically relevant relationships, and they may contain a lot of noise. We believe, however, that having a fixed number of tables gives the user a clear overview, and he or she can easily ignore the irrelevant data.

In the Appendix, we present the word sketches for the lemma “without” created by means of compatible sketch grammars from four Aranea web corpora..

## 5 Further work

In the near future, we plan to carry out activities within several tracks. Firstly, we would like to improve the quality of the Aranea corpus data itself (by means of better filtration, normalization and deduplication), as well as its morpho-syntactic annotation by means of long-term evaluation of the resulting word sketches. Secondly, we want to include new languages into our Aranea corpus family and to write the respective corpus grammars, at least for the languages taught at Slovak universities. And finally, we plan to tune the global parameters of the compatible sketch grammars, as well as provide language-specific improvements so that the bilingual word sketches provide more relevant results.

## 6 References

- Baroni, M. – Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In: *Proceedings of LREC'2004*. Lisbon: ELRA, 2004.
- Benko, V. (2013). Data Deduplication in Slovak Corpora. In: *Slovko 2013: Natural Language Processing, Corpus Linguistics, E-learning*. Katarína Gajdošová (Ed.), Adriána Žáková. Lüdenscheid: RAM-Verlag, 2013, pp. 27–39.
- Hong, J. F. – Huang, C. R. (2007). Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research. URL: <http://www.ling.sinica.edu.tw/eip/FILES/publish/2007.7.18.93102662.8243902.pdf>.
- Jakubíček, M. (2014). Personal communication.
- Kilgarriff, A. et al. (2004). The Sketch Engine. In: G. Williams and S. Vessier (eds.), *Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6–10, 2004*. Lorient: Université de Bretagne-Sud, pp. 105–116.
- Khokhlova, M. (2010). Building Russian Word Sketches as Models of Phrases. In: *Proc. EURALEX 2010, Leeuwarden*, July 2010.
- Kovář, V. (2013). New features in the Sketch Engine interface. Part 1. In: *SKEW-4 Workshop*, Tallinn, October 2013. URL: [https://www.sketchengine.co.uk/documentation/raw-attachment/wiki/SKEW-4/Program/ske\\_interface\\_part1.pdf](https://www.sketchengine.co.uk/documentation/raw-attachment/wiki/SKEW-4/Program/ske_interface_part1.pdf)
- Krek, S. – Kilgarriff, A. (2006). Slovene Word Sketches. In: *Proceedings of 5th Slovenian and 1st international Language Technologies Conference 2006*. October 9th – 10th 2006. Jožef Stefan Institute, Ljubljana, Slovenia
- Macoveiciuc, M. – Kilgarriff, A. (2010). The RoWaC Corpus and Romanian Word Sketches. In: *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*. Edited by Dan Tufis and Corina Forascu. Romanian Academy.
- Petrov, S. – Das, D. – McDonald, R. (2012). A Universal Part-of-Speech Tagset. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: ELRA, May 2012.
- Piasecki, M. (2007). Polish tagger TaKIPI: Rule based construction and optimisation. *TASK QUARTERLY* 11, No 1–2, 151–167
- Radziszewski, A. – Kilgarriff, A. – Lew, R. (2011). Polish Word Sketches In: Zygmunt Vetulani (ed.) *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the 5th Language & Technology Conference*. Poznań : Fundacja Uniwersytetu im. A. Mickiewicza.
- Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
- Srdanović, E. I. – Erjavec, T. – Kilgarriff, A. (2008). A web corpus and word-sketches for Japanese. In: *Journal of Natural Language Processing* 15/2, 137– 159. (reprinted in *Information and Media Technologies* 3/3, 2008, 529– 551)
- Suchomel, V. – Pomikálek, J. (2012). Efficient Web Crawling for Large Text Corpora. In: *7th Web as Corpus Workshop (WAC-7)*, Lyon, 2012.
- Tiberius, C. – Kilgarriff, A. (2009). The Sketch Engine for Dutch with the ANW corpus. In: *Fons Verborum, Festschrift for Fons Moerdijk*. Instituut voor Nederlandse Lexicologie, Leiden, The Netherlands.
- Whitelock, P. (2010). Personal communication.

### Appendix

To demonstrate the compatible word sketches, we present screen shots for the preposition “without” in four languages (English, French, German, and Russian). Prepositions belong to word classes that are usually either not covered by the respective traditional sketch grammars at all, or that produce a limited number of output word sketch tables only.



For all languages involved, we can observe the typical binary collocations with noun and verbs. The collocations with adjectives usually form a multi-word expression that is not fully displayed in the word sketches, but many of those can be easily recognized even without going into the actual concordances.

Note: Due to the longer adjectives in Russian, the interesting table with verbal collocates did not fit onto the screen.

**without** (*non-verb*) Araneum Anglicum Maius (En Web 1.2.01) 1,20 G freq = **413459** (344.5 per million) Click on collocates in boldface to get multi word sketches.

<b>X</b>	<b>X/Y, X/Y</b> 31449 -0.0	<b>X/Y Cj X/Y</b> 23389 -0.0	<b>YX</b> 291638 -0.1	<b>XY</b> 379129 -0.1
Pp(X) 413459 -0.5	<b>except</b> 141 3.61	<b>within</b> 753 3.35	<b>complete</b> 1130 4.76	<b>limitation</b> 4001 7.46
	whereas 36 2.92	<b>with</b> 7024 2.62	<b>reproduce</b> 355 4.63	<b>doubt</b> 4084 7.08
	<b>unless</b> 111 2.83	<b>between</b> 237 1.13	<b>viagra</b> 318 4.54	<b>notice</b> 3171 6.71
	albeit 11 2.65	<b>under</b> 111 0.41	<b>survive</b> 662 4.45	<b>hesitation</b> 1133 6.47
	although 97 1.76	beyond 28 0.39	<b>incomplete</b> 190 4.08	<b>regard</b> 2352 6.21
	<b>though</b> 160 1.64	against 80 0.17	<b>function</b> 323 4.05	<b>prescription</b> 1401 6.16
	despite 42 1.57	except 12 0.09	<b>live</b> 2861 4.0	<b>express</b> 1024 6.13
	<b>upon</b> 118 1.25	whether 44 -0.0	<b>impossible</b> 385 4.0	<b>prior</b> 2708 6.1
	<b>like</b> 402 1.21	<b>from</b> 687 -0.06	<b>exist</b> 1153 3.98	<b>permission</b> 1676 6.09
	via 45 1.2	<b>at</b> 662 -0.08	<b>possible</b> 1451 3.88	<b>Borders</b> 837 6.08
	<b>because</b> 352 1.2	toward 17 -0.13	<b>cialis</b> 164 3.8	<b>compromise</b> 1176 6.06
	till 14 1.17	upon 45 -0.13	<b>detention</b> 184 3.79	<b>delay</b> 1265 5.99
	<b>within</b> 166 1.17	behind 25 -0.14	<b>proceed</b> 288 3.75	<b>prejudice</b> 920 5.93
	unto 16 1.16	<b>through</b> 145 -0.17	<b>reporter</b> 213 3.58	<b>fear</b> 2381 5.87
	<b>while</b> 253 1.11	<b>into</b> 210 -0.18	<b>die</b> 672 3.39	<b>consent</b> 1235 5.83
	<b>under</b> 169 1.01	<b>for</b> 1374 -0.23	<b>even</b> 2829 3.38	<b>exception</b> 1332 5.7

<b>Nn X</b> 292904 -0.1	<b>X Nn</b> 401566 -0.1	<b>Aj X</b> 75340 -0.0	<b>X Aj</b> 110090 -0.1	<b>Vb X/X Vb</b> 704673 -0.1
<b>viagra</b> 467 5.09	<b>consent</b> 4488 7.65	<b>incomplete</b> 195 5.4	<b>express</b> 1335 7.7	<b>compromise</b> 1273 5.53
<b>cialis</b> 255 4.43	<b>permission</b> 4964 7.61	<b>impossible</b> 737 5.4	<b>written</b> 2518 7.36	<b>ca</b> 4366 5.46
<b>detention</b> 223 4.07	<b>limitation</b> 4115 7.46	<b>complete</b> 1228 5.13	<b>prior</b> 3214 6.69	<b>sacrifice</b> 1016 5.21
<b>reporter</b> 231 3.7	<b>notice</b> 4362 7.14	meaningless 82 4.38	<b>undue</b> 371 6.38	<b>leave</b> 5606 4.96
<b>life</b> 3435 3.61	<b>doubt</b> 4205 7.09	<b>generic</b> 158 4.33	<b>parental</b> 419 5.95	<b>worry</b> 1337 4.91
<b>buy</b> 129 3.56	<b>hesitation</b> 1212 6.49	<b>possible</b> 1826 4.31	<b>proper</b> 1536 5.95	<b>could</b> 10736 4.89
<b>sentence</b> 259 3.4	<b>delay</b> 1674 6.34	<b>useless</b> 106 4.24	<b>adequate</b> 731 5.95	<b>resort</b> 701 4.83
<b>prescription</b> 177 3.38	<b>regard</b> 2478 6.26	worthless 65 4.15	<b>explicit</b> 451 5.81	<b>can</b> 29062 4.82
<b>taxation</b> 116 3.35	<b>prescription</b> 1463 6.17	indefinite 56 4.11	<b>further</b> 2508 5.2	<b>survive</b> 1130 4.72
<b>imprisonment</b> 111 3.29	<b>Borders</b> 818 5.97	lonely 58 3.58	<b>foregoing</b> 174 5.16	<b>allow</b> 4815 4.66
<b>sex</b> 362 3.28	<b>authorization</b> 989 5.94	consecutive 68 3.51	<b>slight</b> 357 5.06	<b>depart</b> 694 4.63
Viagra 91 3.26	<b>fear</b> 2517 5.92	usable 45 3.49	<b>formal</b> 483 4.85	<b>live</b> 4724 4.62
<b>party</b> 829 3.24	<b>prejudice</b> 947 5.91	<b>able</b> 1160 3.46	<b>added</b> 238 4.83	<b>imagine</b> 1196 4.57
<b>nothing</b> 650 3.18	<b>exception</b> 1407 5.74	<b>difficult</b> 466 3.39	<b>excessive</b> 208 4.63	<b>lose</b> 2519 4.55
<b>anything</b> 655 3.14	<b>ado</b> 628 5.65	unthinkable 30 3.37	<b>conscious</b> 294 4.62	<b>go</b> 12313 4.53
<b>day</b> 2643 3.11	<b>warning</b> 1087 5.63	inconceivable 28 3.33	<b>additional</b> 1159 4.61	<b>would</b> 13830 4.45

<b>Av X/X Av</b> 112453 -0.1	<b>ZX</b> 226890 -0.0	<b>XZ</b> 304244 -0.1	<b>Pp X</b> 92448 -0.0	<b>X Pp</b> 99334 -0.0
<b>whatsoever</b> 340 5.61	<b>those</b> 2787 4.04	<b>any</b> 27935 6.89	<b>albeit</b> 112 4.95	<b>into</b> 2483 3.37
<b>ever</b> 3020 5.13	<b>any</b> 3658 3.97	<b>them</b> 5997 4.18	<b>onto</b> 173 3.62	<b>except</b> 139 3.35
<b>indefinitely</b> 128 4.83	<b>themselves</b> 630 3.63	<b>a</b> 49871 3.78	<b>into</b> 2579 3.42	<b>whether</b> 379 3.03
<b>freely</b> 262 4.82	<b>off</b> 1294 3.44	<b>him</b> 2262 3.62	<b>through</b> 1679 3.35	<b>of</b> 38204 2.97
<b>necessarily</b> 454 4.82	<b>i.e.</b> 147 3.42	<b>an</b> 7055 3.48	<b>although</b> 240 2.98	<b>from</b> 5422 2.92
<b>overly</b> 155 4.72	<b>another</b> 911 3.38	<b>some</b> 2749 3.32	<b>than</b> 1688 2.82	<b>unless</b> 119 2.75
<b>even</b> 6980 4.71	<b>them</b> 3308 3.33	<b>it</b> 13245 3.23	<b>across</b> 305 2.73	<b>through</b> 1091 2.72
<b>anywhere</b> 360 4.45	<b>itself</b> 387 3.05	<b>their</b> 6214 3.21	<b>because</b> 971 2.64	<b>upon</b> 330 2.67
<b>too</b> 2634 4.42	<b>yourself</b> 311 3.0	<b>its</b> 3091 3.17	<b>like</b> 1071 2.61	<b>about</b> 2781 2.66
<b>properly</b> 346 4.36	<b>it</b> 11199 2.99	<b>yourself</b> 354 3.1	<b>from</b> 4267 2.58	<b>like</b> 1093 2.63
<b>actually</b> 1420 4.2	<b>these</b> 1877 2.96	<b>the</b> 79951 3.08	except 77 2.52	<b>for</b> 9121 2.5
<b>explicitly</b> 142 4.17	<b>this</b> 6489 2.93	<b>your</b> 5936 3.08	<b>on</b> 6479 2.5	<b>until</b> 310 2.44
<b>easily</b> 607 4.11	<b>himself</b> 318 2.88	<b>her</b> 2447 3.0	<b>for</b> 9024 2.48	<b>behind</b> 140 2.23
<b>overboard</b> 67 4.08	<b>him</b> 1310 2.85	<b>either</b> 373 2.93	<b>at</b> 3684 2.39	<b>on</b> 5230 2.19
<b>completely</b> 592 4.08	<b>no</b> 1735 2.84	<b>whom</b> 248 2.85	<b>upon</b> 269 2.39	<b>at</b> 3079 2.13
<b>physically</b> 161 4.06	<b>to</b> 40462 2.79	<b>these</b> 1714 2.82	whereas 33 2.32	<b>till</b> 35 2.08

**ohne** Araneum Germanicum Maius (De Web 1.2.01) 1,20 G freq = **592115** (493.4 per million) Click on collocates in boldface to get multi word sketches.

X	X/Y, X/Y	34121 0.2	X/YC X/Y	41201 0.2	YX	323276 0.3	XY	483395 0.4
Pp(X) 505192 3.0	<b>dass</b>	1892 4.32	einschließlich	57 3.19	<b>Kredit</b>	2342 6.4	<b>vorherig</b>	4624 7.62
Cj(X) 86923 0.9	<b>ob</b>	270 3.32	seitens	40 2.8	<b>jederzeit</b>	1222 5.36	<b>Schufa</b>	3481 7.57
	<b>oder</b>	1064 3.25	<b>mit</b>	8640 2.61	<b>Reporter</b>	414 5.2	<b>Zweifel</b>	3621 7.26
	d.h.	21 3.05	inklusive	34 2.33	<b>gänzlich</b>	475 5.11	<b>Rücksicht</b>	2860 7.24
	<b>wie</b>	659 2.85	<b>zwischen</b>	382 1.97	<b>ganz</b>	8310 4.75	<b>Weiteres</b>	1810 6.82
	<b>und</b>	5054 2.84	trotz	71 1.8	<b>Fahren</b>	306 4.63	<b>Einschränkung</b>	2002 6.49
	einschließlich	39 2.69	mittels	40 1.8	<b>völlig</b>	929 4.54	<b>Angabe</b>	4455 6.46
	bzw.	54 2.23	innen	12 1.79	<b>Fass</b>	236 4.34	<b>gesondert</b>	1607 6.35
	inklusive	30 2.2	außer	30 1.55	<b>Abnehmen</b>	236 4.32	<b>Zustimmung</b>	2145 6.3
	außer	45 2.17	samt	12 1.54	<b>Girokonto</b>	288 4.28	<b>Abzug</b>	1372 6.19
	weil	61 2.1	aufs	22 1.53	<b>Handy</b>	539 4.26	<b>Gewähr</b>	1324 6.19
	<b>als</b>	464 2.06	außerhalb	39 1.51	<b>Kreditkarte</b>	348 4.24	<b>Problem</b>	6810 6.14
	infolge	12 1.72	wider	11 1.47	<b>Rechnung</b>	486 4.13	<b>Umweg</b>	1192 6.11
	indem	12 1.6	ans	18 1.4	<b>Leben</b>	2437 4.12	<b>Unterbrechung</b>	1176 6.11
	aufs	20 1.43	ob	71 1.38	<b>allerdings</b>	1516 4.09	<b>ausdrücklich</b>	2164 6.1
	sondern	55 1.37	<b>gegen</b>	269 1.36	<b>Pfanne</b>	226 4.03	<b>Behinderung</b>	1596 5.93

Nn X	311082 0.3	X Nn	606564 0.5	Aj X	89134 0.2	X Aj	200309 0.5	Yb X/X Vb
<b>Kredit</b>	2584 6.56	<b>Zustimmung</b>	5620 7.5	<b>gedruckt</b>	973 7.25	<b>vorherig</b>	4938 8.37	<b>auskommen</b>
<b>Publikation</b>	1004 5.87	<b>Schufa</b>	3453 7.29	<b>gänzlich</b>	485 6.2	<b>gesondert</b>	1704 7.27	<b>gestatten stat</b>
<b>Reporter</b>	415 5.26	<b>Zweifel</b>	4051 7.21	<b>völlig</b>	1022 4.96	<b>ausdrücklich</b>	3061 7.06	<b>funktionieren</b>
<b>Fahren</b>	337 4.82	<b>Ankündigung</b>	3067 7.07	anwaltlichen	83 4.73	<b>nennenswert</b>	811 6.66	<b>verändern</b>
<b>Rechnung</b>	739 4.76	<b>Rücksicht</b>	3041 7.07	<b>gesamt</b>	1581 4.62	<b>schriftlich</b>	2599 6.55	<b>verlaufen</b>
<b>Handy</b>	708 4.67	<b>Aufwand</b>	3653 6.93	<b>erhoben</b>	133 4.52	<b>lästig</b>	639 6.18	<b>verlieren</b>
<b>Girokonto</b>	349 4.55	<b>Einschränkung</b>	2510 6.59	<b>Prepaid</b>	112 4.5	<b>erkennbar</b>	856 6.09	<b>leben</b>
<b>Kreditkarte</b>	402 4.48	<b>Genehmigung</b>	2522 6.55	<b>personenbezogen</b>	248 4.5	<b>störend</b>	448 5.8	<b>laufen</b>
<b>Abnehmen</b>	247 4.43	<b>Weiteres</b>	1818 6.52	<b>komplett</b>	820 4.38	<b>weit</b>	11902 5.58	<b>zögern</b>
<b>Fass</b>	239 4.41	<b>Angabe</b>	4874 6.49	<b>selbstverständlich</b>	433 4.38	<b>zusätzlich</b>	3407 5.58	<b>dürfen</b>
<b>Leben</b>	2895 4.37	<b>Grund Gründen</b>	3484 6.45	viagra	61 4.27	<b>möglich</b>	4803 5.48	<b>verlassen</b>
<b>Angebot</b>	1875 4.2	<b>Einwilligung</b>	1781 6.31	<b>vollkommen</b>	226 4.26	<b>ersichtlich</b>	345 5.34	<b>nachdenken</b>
<b>Tarif</b>	439 4.19	<b>Problem</b>	7956 6.31	undenkbar	64 4.24	<b>fremd</b>	827 5.32	<b>bleiben</b>
<b>Pfanne</b>	242 4.17	<b>Abzug</b>	1758 6.29	<b>kommerziell</b>	153 4.22	<b>unnötig</b>	449 5.28	<b>können</b>
<b>Abmahnung</b>	267 4.13	<b>Umweg</b>	1353 6.01	verlinkten	97 4.19	<b>aufwendig</b>	523 5.21	<b>kündigen</b>
<b>Baufinanzierung</b>	197 4.05	<b>Gewähr</b>	1384 5.99	berufsmäßig	52 4.19	<b>finanziell</b>	1189 5.19	<b>überstehen</b>

Av X/X Av	173826 0.4	Z X	285861 0.3	X Z	356121 0.4	Pp X	67488 0.1	X Pp	111231 0.3
<b>jemals</b>	724 6.31	<b>man</b>	8344 4.02	<b>jegliche</b>	4363 7.76	<b>dank</b>	140 3.13	<b>seitens</b>	198 4.65
<b>jederzeit</b>	1696 6.07	<b>14</b>	690 3.93	<b>irgendwelche</b>	1088 6.08	<b>pro</b>	258 2.87	<b>durchs</b>	80 3.87
<b>vorher</b>	1598 5.93	<b>diese</b>	1417 3.91	<b>dabei</b>	6094 5.38	<b>bei</b>	5283 2.86	<b>ans</b>	124 3.79
<b>jedoch</b>	4386 5.15	<b>nicht</b>	20169 3.85	<b>irgendeine</b>	556 4.95	einschließlich	51 2.85	<b>auf</b>	15175 3.71
<b>allerdings</b>	2735 5.01	<b>wer</b>	1280 3.83	<b>allzu</b>	476 4.74	<b>gegenüber</b>	221 2.71	<b>außer</b>	138 3.48
<b>ganz</b>	9471 4.96	<b>keine</b>	4041 3.74	<b>dafür</b>	1743 4.61	inklusive	49 2.7	<b>aufs</b>	105 3.43
<b>niemals</b>	434 4.86	<b>daher</b>	740 3.57	<b>jede</b>	5801 4.4	<b>trotz</b>	135 2.67	<b>von</b>	16501 3.42
<b>leider</b>	1386 4.66	<b>solche</b>	1093 3.54	<b>zu</b>	52715 4.36	<b>nach</b>	2766 2.58	<b>durch</b>	3718 3.34
<b>irgend</b>	201 4.65	<b>sie</b>	6700 3.53	<b>nichts</b>	1282 4.16	<b>innerhalb</b>	209 2.49	innen	56 3.32
<b>kaum</b>	1261 4.63	<b>niemand</b>	257 3.45	<b>keine</b>	4669 3.94	<b>für</b>	7039 2.48	<b>über</b>	3330 3.0
<b>freilich</b>	204 4.62	<b>eine</b>	37041 3.41	<b>solche</b>	1384 3.85	<b>wegen</b>	200 2.46	<b>an</b>	7594 2.92
<b>also</b>	3264 4.58	<b>jemand</b>	317 3.39	<b>jedwede</b>	159 3.75	<b>seit</b>	561 2.46	<b>innerhalb</b>	277 2.84
<b>bisher</b>	1131 4.56	<b>was</b>	2380 3.32	<b>darauf</b>	740 3.74	zeit	20 2.46	<b>per</b>	225 2.79
<b>auch</b>	35534 4.56	<b>er</b>	5165 3.25	<b>jemand</b>	424 3.71	<b>von</b>	8310 2.43	<b>in</b>	22140 2.69
<b>sofort</b>	1136 4.53	<b>nichts</b>	655 3.24	<b>viele</b>	4400 3.71	mittels	63 2.36	<b>mit</b>	8122 2.52
<b>meistens</b>	430 4.52	<b>es</b>	9643 3.23	<b>darüber</b>	676 3.69	innen	20 2.23	<b>gegen</b>	606 2.5



**SANS** Araneum Francogallicum Maius (Fr Web 1.2.02) 1,23 G freq = **1023016** (829.5 per million) Click on collocates in boldface to get multi word sketches.

<b>X</b>	<b>X/Y, X/Y</b> 165883 -0.2	<b>X/Y Cj X/Y</b> 89170 -0.1	<b>Y X</b> 638024 -0.1	<b>X Y</b> 878596 -0.1
Pp(X) <a href="#">1012928</a> -0.6	<b>sauf</b> <a href="#">248</a> 3.9	<b>avec</b> <a href="#">11553</a> 4.27	<b>reporter</b> <a href="#">975</a> 5.06	<b>doute</b> <a href="#">85421</a> 9.99
Cj(X) <a href="#">10088</a> -0.0	<b>jusque</b> <a href="#">599</a> 3.29	<b>hors</b> <a href="#">104</a> 2.71	<b>non</b> <a href="#">5532</a> 4.86	<b>cesse</b> <a href="#">20493</a> 9.02
	<b>malgré</b> <a href="#">217</a> 3.13	soit <a href="#">77</a> 2.17	<b>c'est-à-dire</b> <a href="#">804</a> 4.49	<b>oublier</b> <a href="#">15100</a> 7.48
	hormis <a href="#">51</a> 3.03	malgré <a href="#">78</a> 1.8	<b>fonctionner</b> <a href="#">1169</a> 4.48	<b>fil</b> <a href="#">7755</a> 7.07
	<b>hors</b> <a href="#">147</a> 3.01	parce <a href="#">36</a> 1.58	<b>organisme</b> <a href="#">1212</a> 4.2	<b>précédent</b> <a href="#">6972</a> 6.97
	<b>avec</b> <a href="#">4598</a> 2.94	<b>entre</b> <a href="#">519</a> 1.56	<b>répéter</b> <a href="#">637</a> 4.16	<b>préavis</b> <a href="#">3701</a> 6.96
	<b>depuis</b> <a href="#">869</a> 2.89	sauf <a href="#">41</a> 1.52	<b>rester</b> <a href="#">3613</a> 4.16	<b>autorisation</b> <a href="#">5545</a> 6.84
	<b>chez</b> <a href="#">541</a> 2.89	<b>en</b> <a href="#">5765</a> 1.48	<b>accepter</b> <a href="#">1380</a> 4.15	<b>autant</b> <a href="#">9143</a> 6.81
	<b>vers</b> <a href="#">626</a> 2.88	<b>sous</b> <a href="#">284</a> 1.45	<b>presque</b> <a href="#">971</a> 4.14	<b>relâche</b> <a href="#">3002</a> 6.72
	<b>pendant</b> <a href="#">439</a> 2.83	<b>pendant</b> <a href="#">162</a> 1.45	<b>vivre</b> <a href="#">2985</a> 4.14	<b>compter</b> <a href="#">12738</a> 6.67
	<b>sous</b> <a href="#">725</a> 2.77	<b>contre</b> <a href="#">250</a> 1.37	<b>consommer</b> <a href="#">669</a> 4.14	<b>frontière</b> <a href="#">4720</a> 6.61
	<b>dès</b> <a href="#">262</a> 2.63	à <a href="#">7766</a> 1.37	<b>interdire</b> <a href="#">864</a> 4.01	<b>conteste</b> <a href="#">2740</a> 6.6
	à <a href="#">18464</a> 2.62	dès <a href="#">98</a> 1.29	<b>mourir</b> <a href="#">943</a> 3.98	<b>faille</b> <a href="#">3165</a> 6.6
	<b>au</b> <a href="#">7379</a> 2.59	<b>du</b> <a href="#">9242</a> 1.28	<b>vétérinaire</b> <a href="#">364</a> 3.94	<b>consentement</b> <a href="#">2514</a> 6.28
	<b>devant</b> <a href="#">323</a> 2.58	<b>jusque</b> <a href="#">137</a> 1.22	<b>voiture</b> <a href="#">857</a> 3.87	<b>moindre</b> <a href="#">3757</a> 6.26
	<b>en</b> <a href="#">12377</a> 2.58	<b>de</b> <a href="#">23057</a> 1.17	<b>dérouler</b> <a href="#">816</a> 3.85	<b>limite</b> <a href="#">4816</a> 6.24

<b>Nn X</b> 526473 -0.1	<b>X Nn</b> 969635 -0.1	<b>Aj X</b> 118195 -0.1	<b>X Aj</b> 167461 -0.1	<b>Vb X/X</b> 1329224 -0.1
<b>reporter</b> <a href="#">920</a> 5.16	<b>doute</b> <a href="#">96148</a> 10.11	<b>vétérinaire</b> <a href="#">260</a> 5.06	<b>lucratif</b> <a href="#">2851</a> 8.2	<b>oublier</b> <a href="#">16163</a> 7.31
<b>organisme</b> <a href="#">1441</a> 4.53	<b>cesse</b> <a href="#">20510</a> 8.93	<b>impossible</b> <a href="#">736</a> 5.01	<b>préalable</b> <a href="#">4847</a> 7.94	<b>compter</b> <a href="#">14308</a> 6.61
<b>journal</b> <a href="#">1822</a> 4.3	<b>fil</b> <a href="#">7829</a> 7.02	<b>tierce</b> <a href="#">139</a> 4.55	<b>moindre</b> <a href="#">5688</a> 7.7	<b>parler</b> <a href="#">12885</a> 5.91
<b>voiture</b> <a href="#">1075</a> 4.28	<b>préavis</b> <a href="#">3832</a> 6.89	modifiable <a href="#">97</a> 4.54	<b>expresse</b> <a href="#">1296</a> 7.5	<b>soucier</b> <a href="#">2432</a> 5.71
<b>nuit</b> <a href="#">1242</a> 4.16	<b>autorisation</b> <a href="#">6013</a> 6.88	<b>gratuit</b> <a href="#">747</a> 4.34	<b>nul</b> <a href="#">2863</a> 7.15	<b>tarder</b> <a href="#">2551</a> 5.71
<b>Internet</b> <a href="#">1353</a> 4.07	<b>précédent</b> <a href="#">6605</a> 6.82	<b>accessible</b> <a href="#">517</a> 4.29	<b>apparent</b> <a href="#">1106</a> 6.93	<b>laisser</b> <a href="#">9961</a> 5.61
<b>réseau</b> <a href="#">2004</a> 4.07	<b>relâche</b> <a href="#">3004</a> 6.59	<b>possible</b> <a href="#">1898</a> 4.26	<b>frais</b> <a href="#">3179</a> 6.59	<b>attendre</b> <a href="#">7517</a> 5.61
<b>acceptation</b> <a href="#">367</a> 4.02	<b>faille</b> <a href="#">3200</a> 6.49	<b>utilisable</b> <a href="#">108</a> 4.08	<b>frontière</b> <a href="#">1676</a> 5.97	<b>perdre</b> <a href="#">5087</a> 5.31
<b>amour</b> <a href="#">1131</a> 3.91	<b>conteste</b> <a href="#">2760</a> 6.47	remboursable <a href="#">69</a> 3.93	<b>égal</b> <a href="#">1040</a> 5.64	<b>pouvoir</b> <a href="#">54251</a> 5.31
<b>jour</b> <a href="#">4495</a> 3.88	<b>limite</b> <a href="#">5331</a> 6.32	<b>sexuel</b> <a href="#">320</a> 3.89	<b>réel</b> <a href="#">2466</a> 5.61	<b>passer</b> <a href="#">13149</a> 5.21
<b>médecin</b> <a href="#">739</a> 3.88	<b>surprise</b> <a href="#">3308</a> 6.29	<b>correct</b> <a href="#">133</a> 3.85	<b>gras</b> <a href="#">748</a> 5.6	<b>hésiter</b> <a href="#">2779</a> 5.11
<b>licencement</b> <a href="#">300</a> 3.86	<b>consentement</b> <a href="#">2647</a> 6.24	<b>partiel</b> <a href="#">173</a> 3.8	<b>fixe</b> <a href="#">640</a> 5.5	<b>savoir</b> <a href="#">12005</a> 5.11
<b>connexion</b> <a href="#">400</a> 3.79	<b>souci</b> <a href="#">3214</a> 6.03	réalisable <a href="#">63</a> 3.71	<b>excessif</b> <a href="#">420</a> 5.44	<b>rester</b> <a href="#">7627</a> 5.01
<b>monde</b> <a href="#">3438</a> 3.77	<b>hésitation</b> <a href="#">2068</a> 6.02	estre <a href="#">81</a> 3.7	<b>supplémentaire</b> <a href="#">1195</a> 5.42	<b>vivre</b> <a href="#">6437</a> 5.01
<b>sexe</b> <a href="#">522</a> 3.74	<b>arrêt</b> <a href="#">3353</a> 6.0	<b>immédiat</b> <a href="#">197</a> 3.67	<b>valable</b> <a href="#">526</a> 5.37	<b>bouger</b> <a href="#">1713</a> 5.01
<b>gens</b> <a href="#">1270</a> 3.74	<b>frontière</b> <a href="#">3229</a> 5.98	inconcevable <a href="#">53</a> 3.66	<b>explicite</b> <a href="#">298</a> 5.17	<b>regarder</b> <a href="#">3542</a> 5.01

<b>Av X/X Av</b> 287079 -0.1	<b>Z X</b> 526005 -0.1	<b>X Z</b> 554614 -0.1	<b>Pp X</b> 258101 -0.1	<b>X Pp</b> 349901 -0.1
<b>autant</b> <a href="#">10797</a> 7.41	<b>ceci</b> <a href="#">1004</a> 4.96	<b>aucun</b> <a href="#">36125</a> 8.34	<b>devant</b> <a href="#">691</a> 3.61	<b>jusque</b> <a href="#">1966</a> 4.88
<b>toutefois</b> <a href="#">3317</a> 6.99	<b>cela</b> <a href="#">6226</a> 4.58	<b>quoi</b> <a href="#">3522</a> 5.73	<b>chez</b> <a href="#">870</a> 3.52	<b>pour</b> <a href="#">27339</a> 4.31
<b>jamais</b> <a href="#">11860</a> 6.7	<b>la le</b> <a href="#">9272</a> 4.47	<b>OGM_</b> <a href="#">963</a> 5.49	<b>vers</b> <a href="#">992</a> 3.49	<b>avec</b> <a href="#">9295</a> 3.94
<b>trop</b> <a href="#">11522</a> 6.58	<b>la</b> <a href="#">2222</a> 4.32	<b>la le</b> <a href="#">10965</a> 4.71	<b>du</b> <a href="#">42461</a> 3.48	<b>pendant</b> <a href="#">1000</a> 3.89
<b>forcément</b> <a href="#">1325</a> 5.84	<b>se</b> <a href="#">32382</a> 4.28	<b>la</b> <a href="#">2817</a> 4.65	à <a href="#">31897</a> 3.41	<b>sur</b> <a href="#">13933</a> 3.83
<b>nécessairement</b> <a href="#">855</a> 5.74	<b>un</b> <a href="#">77428</a> 4.26	<b>toi</b> <a href="#">1128</a> 4.64	<b>de</b> <a href="#">107821</a> 3.4	à <a href="#">42038</a> 3.8
<b>rien</b> <a href="#">6754</a> 5.71	<b>on</b> <a href="#">10142</a> 4.24	<b>eux</b> <a href="#">1697</a> 4.46	<b>pendant</b> <a href="#">677</a> 3.39	<b>depuis</b> <a href="#">1684</a> 3.78
<b>vraiment</b> <a href="#">3859</a> 5.5	<b>elle</b> <a href="#">9775</a> 4.24	<b>lequel</b> <a href="#">3094</a> 4.41	<b>après</b> <a href="#">1187</a> 3.35	<b>de</b> <a href="#">137573</a> 3.75
<b>non</b> <a href="#">7636</a> 5.4	<b>il</b> <a href="#">25892</a> 4.1	<b>lui</b> <a href="#">6975</a> 4.33	<b>dans</b> <a href="#">11797</a> 3.18	<b>au</b> <a href="#">15714</a> 3.68
<b>priori</b> <a href="#">564</a> 5.33	<b>votre</b> <a href="#">5454</a> 4.07	<b>y</b> <a href="#">7918</a> 4.32	<b>pour</b> <a href="#">11453</a> 3.06	<b>par</b> <a href="#">11444</a> 3.67
<b>réellement</b> <a href="#">926</a> 5.33	<b>vous</b> <a href="#">12695</a> 4.07	<b>leur</b> <a href="#">7251</a> 4.22	<b>avec</b> <a href="#">4790</a> 2.99	<b>devant</b> <a href="#">715</a> 3.59
<b>c'est-à-dire</b> <a href="#">989</a> 5.31	<b>tu</b> <a href="#">2004</a> 3.99	<b>se</b> <a href="#">30956</a> 4.21	<b>sur</b> <a href="#">7429</a> 2.92	<b>dans</b> <a href="#">15547</a> 3.58
<b>exprès</b> <a href="#">387</a> 5.23	<b>ce</b> <a href="#">38418</a> 3.91	<b>me</b> <a href="#">6916</a> 4.14	<b>au</b> <a href="#">9114</a> 2.9	<b>vers</b> <a href="#">1043</a> 3.52
<b>presque</b> <a href="#">1466</a> 5.05	<b>y</b> <a href="#">5689</a> 3.85	<b>son</b> <a href="#">19423</a> 4.14	<b>derrière</b> <a href="#">177</a> 2.85	<b>contre</b> <a href="#">1189</a> 3.51
<b>préalablement</b> <a href="#">416</a> 5.02	<b>leur</b> <a href="#">5436</a> 3.81	<b>que</b> <a href="#">19955</a> 4.11	<b>contre</b> <a href="#">726</a> 2.83	<b>envers</b> <a href="#">242</a> 3.43
<b>apparemment</b> <a href="#">419</a> 4.92	<b>qui</b> <a href="#">24786</a> 3.8	<b>votre</b> <a href="#">5045</a> 3.96	<b>en</b> <a href="#">14810</a> 2.83	<b>quant</b> <a href="#">240</a> 3.42

X	X/Y, X/Y	16565	-0.2	X/Y Cj	9119	-0.1	YX	424569	-0.3	XY	620883	-0.3		
Pp(X)	749749	-2.3	безо	15	3.46	вне	76	2.84	обойтись	12297	8.55	исключение	9151	7.69
nom(X)	6477	-0.0	вне	90	3.05	у	526	0.2	обойтись	5964	8.15	сомнение	8080	7.63
gen(X)	737804	-1.3	кроме	230	2.46	помимо	12	-0.07	остаться	14808	6.92	весть	4171	7.49
dat(X)	1640	-0.0	ради	54	1.86	вместо	16	-0.28	пропасть	2590	6.68	попечения	2636	7.02
acc(X)	1003	-0.0	посредством	18	1.21	против	44	-0.28	невозможный	2621	6.53	преувеличение	2256	6.76
loc(X)	2266	-0.0	вместо	39	0.99	вокруг	14	-0.74	оставить	4941	6.24	особый	9784	6.71
ins(X)	559	-0.0	путем	24	0.98	для	445	-0.81	невозможно	2174	6.07	труд	8844	6.61
			помимо	24	0.91	со	90	-0.84	немыслимый	954	5.93	лишний	4748	6.58
			среди	74	0.65	до	163	-0.86	оставлять	1678	5.37	учет	6474	6.55
			выше	16	0.6	из-за	22	-0.91	пропавшими	485	5.2	малое	1960	6.51
			возле	11	0.56	от	294	-1.01	практически	2233	5.14	предварительный	2943	6.33
			от	782	0.4	после	78	-1.19	оставаться	3850	5.12	малейший	2032	6.25
			для	989	0.35	среди	18	-1.39	прожить	741	4.98	присмотр	1582	6.25
			из	811	0.33	из	237	-1.44	жить	4890	4.89	посредник	1786	6.21
			у	552	0.27	с	296	-2.38	вообще	2127	4.79	согласие	2584	6.19
			внутри	22	0.27				почти	1985	4.67	разбор	1567	6.16

Nn X	364089	-0.3	X Nn	844256	-0.4	Aj X	88968	-0.2	X Aj	194940	-0.3	Vb Xj
оставление	423	5.12	сомнение	9447	7.64	немыслимый	1008	7.55	малейший	2061	7.24	обойтись
кредит	1386	4.88	исключение	9375	7.52	пропавшими	471	7.31	предварительный	3438	7.19	обойтись
репортер	312	4.68	весть	4173	7.12	невозможный	2767	7.25	видимый	1687	7.12	остаток
дым	377	4.51	труд	10846	6.8	пропавшим	176	5.94	лишний	4940	7.08	оставление
участок	1432	4.4	согласие	4377	6.69	худой	324	5.78	особый	10017	6.96	пропуск
кофе	432	4.26	попечения	2737	6.66	мыслимый	187	5.55	посторонний	1570	6.9	оставление
чай	565	4.17	преувеличение	2693	6.56	предпринимательский	249	5.14	уважительный	761	6.6	прожорливый
лицо	2825	4.14	ведомо	2464	6.49	минеральный	263	4.66	должный	1241	6.53	жизнь
секс	396	4.02	учет	6771	6.48	неполный	156	4.63	невозможный	1714	6.32	оставление
отпуск	412	4.01	разрешение	4377	6.35	исковый	124	4.43	дополнительный	4464	6.13	предварительный
наличные	219	3.98	ограничение	3996	6.31	ровный	205	4.27	излишний	669	5.86	смоченный
квартира	1366	3.95	усилие	4131	6.25	гладкий	146	4.12	надлежащий	563	5.67	мысленный
бой	586	3.93	малое	1968	6.12	земельный	290	3.92	немыслимый	397	5.53	повседневный
жизнь	5441	3.92	потеря	3707	6.11	апелляционный	83	3.91	хирургический	475	5.48	мочевыделительный
столбик	187	3.91	ущерб	2715	6.09	длительный	397	3.85	единый	2288	5.27	справочный
заработок	387	3.84	вмешательство	2359	6.03	Послеоперационный	39	3.82	специальный	3111	5.13	выжженный

Av X/X Av	100363	-0.2	Z X	231029	-0.2	X Z	226728	-0.3	Pp X	70963	-0.1	X Pp	79547	-0.1
невозможно	7192	8.32	куда	1227	5.54	всякий	21391	8.49	ко	271	3.49	со	2113	3.69
немыслимо	439	6.99	тут	1299	4.65	какой-либо	10149	8.07	сверх	23	2.68	над	724	3.57
нельзя	4436	6.65	даже	4526	4.49	никуда	887	6.39	за	2747	2.45	для	6323	3.02
извне	309	5.98	ж	324	4.4	никак	1459	6.18	на	12506	2.44	на	16431	2.83
практически	3089	5.85	никак	412	4.35	таковой	1174	6.03	обо	39	2.41	к	4975	2.75
трудно	1261	5.83	поэтому	1540	4.28	оное	252	5.04	до	1499	2.34	с	8158	2.4
сложно	797	5.79	ведь	1216	4.15	онный	281	4.87	во	994	2.33	от	3131	2.4
вообще	3670	5.78	теперь	1167	4.12	чей-либо	190	4.68	посреди	19	2.29	о	2583	2.28
скучно	252	5.77	здесь	1347	4.05	какой	2738	3.89	в	23004	2.25	из-за	206	2.24
вовсе	983	5.53	не	40489	3.96	ничто	1527	3.82	через	483	2.21	об	527	2.21
бесплатно	507	5.37	бы	4480	3.82	то	13929	3.79	при	1481	2.2	свыше	40	2.21
можно	13945	5.32	весь	8436	3.81	они	13906	3.63	к	3292	2.16	около	223	2.2
почти	2600	5.25	тоже	1295	3.73	она	9102	3.56	из	2797	2.11	до	1265	2.09
желательно	325	4.96	туда	277	3.73	он	15472	3.55	об	486	2.09	пред	27	2.07
возможно	826	4.93	как-то	277	3.62	пять	539	3.45	сквозь	29	2.09	между	375	2.04
тяжело	302	4.91	просто	1910	3.61	ваш	2980	3.43	под	609	2.06	в	19486	2.01



# From GLÀFF to PsychoGLÀFF: a Large Psycholinguistics-oriented French Lexical Resource

Basilio Calderone, Nabil Hathout, Franck Sajous  
CLLE-ERSS (CNRS&Université de Toulouse-Le Mirail)  
{basilio.calderone, nabil.hathout, franck.sajous}@univ-tlse2.fr

## Abstract

In this paper, we present two French lexical resources, GLÀFF and PsychoGLÀFF. The former, automatically extracted from the collaborative online dictionary Wiktionary, is a large-scale versatile lexicon exploitable in Natural Language Processing applications and linguistic studies. The latter, based on GLÀFF, is a lexicon specifically designed for psycholinguistic research.

GLÀFF, counting more than 1.4 million entries, features an unprecedented size. It reports lemmas, main syntactic categories, inflectional features and phonemic transcriptions. PsychoGLÀFF contains additional information related to formal aspects of the lexicon and its distribution. It contains about 340,000 entries (120,000 lemmas) that are corpora-attested. We explain how the resources have been created and compare them to other known resources in terms of coverage and quality. Regarding PsychoGLÀFF, the comparison shows that it has an exceptionally large repertoire while having a comparable quality.

**Keywords:** French lexicon; lexical resource for psycholinguistic studies; Wiktionary

## 1 Introduction

Lexical resources play an important role in psycholinguistics by providing researchers with a set of experimentally relevant corpus information concerning words and, to a lesser extent, their sub-lexical components. In particular, psycholinguists working on lexical access need to manipulate a set of formal properties of words, such as syllabification, phonemic transcription, lemmas, inflected forms or orthographic/phonological neighborhood (i.e., the number of words differing from the target word by only one character/phoneme). Word frequency is possibly the most crucial information to be accounted for in psycholinguistic studies, and it is generally provided for either wordforms, lemmas, or both. The most well-known resource for English, German and Dutch is probably CELEX (Baayen et al. 1995). Many other languages, including French, lack a similar resource.

Some freely available French morphological lexicons, such as Lefff (Clément et al. 2004) and Morphalou (Romary et al. 2004), contain inflected forms, lemmas and morphosyntactic tags. These resources, designed in the first place for natural language processing (NLP) or lexicography do not include, ho-

wever, phonemic transcriptions that are necessary to set up psycholinguistics experiments, for extensive morphology and in the design of tools such as phonetizers. One noticeable exception is Lexique (New 2006), a free lexicon quite popular in psycholinguistics. This lexicon includes phonemic transcriptions, word frequencies and various features relevant to this field. However it has a limited coverage, especially in terms of inflected forms. All other resources that have both exploitable coverage and phonemic transcriptions, such as BDLex (Pérennou and de Calmès 1987), ILPho (Boula De Mareuil et al. 2000) or GlobalPhone (Schultz et al. 2013) are not free. Besides their cost, derivative works cannot be redistributed, which constitutes an impediment for collaborative research. As of today, no French lexicon meets all following requirements: free license, wide coverage, phonemic transcriptions and word frequencies.

In this article, we present a psycholinguistics-oriented resource based on *Wiktionnaire*,<sup>1</sup> the French edition of Wiktionary. In a previous work (Sajous et al. 2013a), we automatically extracted GLÀFF, “*a Large Versatile French Lexicon*”. This large-scale resource contains, for each entry, inflectional and phonemic information. PsychoGLÀFF is a new step leveraging Wiktionnaire’s content.<sup>2</sup> Grounded on GLÀFF, PsychoGLÀFF is a lexicon that contains additional features specifically designed for meeting psycholinguistic needs.

The paper is organized as follows: Section 2 gives an overview of Wiktionnaire and its features relevant to lexical resources building. GLÀFF is then described in Section 3. Finally, we present in Section 4 PsychoGLÀFF, a lexicon designed for psycholinguistics use and compare it to Lexique in terms of coverage and word frequency. Conclusions and future directions of work are discussed in Section 5.

## 2 Wiktionnaire as a source of lexical knowledge

Wiktionary is a free multilingual dictionary available online. As its mother project Wikipedia, Wiktionary is based on the wiki paradigm: every internet user may contribute by adding content or modifying existing one. Launched in 2003, the Wiktionary project boasts, ten years later, more than two million entries for its French language edition, the Wiktionnaire. The impressive size of its headword list has to be tempered: inflected forms, discussion pages and, more surprisingly, “*pages describing in French words from other languages*” are counted as regular entries. However, once these latter entries excluded, Wiktionnaire still accounts for 1.4 million entries (186,000 lemmas).

While Wikipedia has been extensively used in various disciplines, its lexicographic counterpart seems to have received less attention from the scientific community. Wiktionary was first used in NLP by Zesch et al. (2008) to compute semantic relatedness. Its potential as an electronic lexicon was studied for the first time by Navarro et al. (2009) for French and English synonymy mining. Along the same line of research, Anton Perez et al. (2011) realized the integration of the Portuguese edition of

1 <http://fr.wiktionary.org>

2 GLÀFF and PsychoGLÀFF are freely available from the REDAC website: <http://redac.univ-tlse2.fr/lexicons/>

Wiktionary in the ontology Onto.PT (Gonçalo Oliveira and Gomes 2010). Serasset (2012) designed Db-nary, an open-source resource containing “easily extractable entries”. The French subpart of this resource contains 260,467 entries. Works led by Meyer and Gurevych (2012) and Gurevych et al. (2012) resulted in German and English ontologies based on Wiktionary. Sajous et al. (2010) has made available a structured XML version of this lexicon for French and English, called WiktionaryX.<sup>3</sup>

Although Wiktionnaire presents interesting features (unprecedented coverage, definitions, phonemic transcriptions, semantic relations, translations, free license),<sup>4</sup> the information it contains is difficult to extract. This probably explains the relatively small number of works using it. Wiktionnaire, as other Wiktionary’s language editions, is released as an “XML dump”, where XML only marks the macrostructure. The microstructure is encoded in a format called *wikicode*, inherent in the content management system *MediaWiki*. This format has no formally defined syntax, evolves over time, and is not stable from one language edition to another. This underspecified syntax makes therefore the automatic information extraction from the collaborative dictionary uneasy: multiple deviations from a “prototypical article” should be expected, as well as missing information, redundancy and inconsistency.

Figure 1 shows the entry *affluent* (adjective and noun ‘affluent’, and two inflection forms of the verb *affluer* ‘to flow into/to pour in’) as it is visible in Wiktionnaire. The corresponding wikicode of this article is shown in Figure 2. Inflected forms may appear in the article related to their lemma (as it is the case in Figure 1). They may also have a dedicated page (cf. Figures 3 and 4).

**affluent**

**Adjectif**

**affluent**

- (Géographie) Qui se jette dans un autre en parlant d'un cours d'eau.
- (Médecine) Qui afflue, qui se portent en abondance vers quelque partie du corps.

	Singulier	Pluriel
Masculin	affluent <i>/a.fly.ø/</i>	affluents <i>/a.fly.ø/</i>
Féminin	affluente <i>/a.fly.øt/</i>	affluentes <i>/a.fly.øt/</i>

**Nom commun**

**affluent** */a.fly.ø/* masculin

- (Géographie) Cours d'eau qui se jette dans un autre.

	Singulier	Pluriel
	affluent	affluents
	<i>/a.fly.ø/</i>	

**Forme de verbe**

**affluent** */a.fly/*

- Troisième personne du pluriel de l'indicatif présent de affluer.
- Troisième personne du pluriel du subjonctif présent de affluer.

Conjugaison du verbe affluer		
INDICATIF	Présent	ils/elles affluent
SUBJONCTIF	Présent	qu'ils/elles affluent

Figure 1: Article of the word ‘*affluent*’ in Wiktionnaire.

3 WiktionaryX, is an XML version of the English and French editions of Wiktionary, freely available at <http://redac.univ-tlse2.fr/lexiques/wiktionaryx.html>

4 For a more comprehensive description of the Wiktionnaire, see (Navarro et al. 2009) and (Sajous et al. 2010; 2013b).

The table of the adjective inflected forms (top-right in Figure 1) is not explicitly present in the wikicode, but is generated by the template `{{fr-accord-cons|a.fly.ã|t}}` (cf. Figure 2). There are hundreds of similar patterns in the wikicode. An example of the non-systematic wikicode's format and resulting article's layout can be seen in Figure 3: unlike the template `{{f}}` that defines the feminine gender of the form, there is no template specifying the grammatical number. The number can only be extracted by parsing the definition "*Féminin singulier*". The heterogeneity of the wikicode also concerns the phonemic transcriptions: they occur sometimes in the *Ligne de forme* (the line following the part of speech heading), as in Figure 3 for *'affluente'*, and sometimes, on the contrary, they are specified in a separate "*Prononciation*" section as in Figure 4 for *'affluentes'*.

To build GLÀFF and PsychoGLÀFF, we automatically extracted the inflected forms and lemmas in their dedicated pages, and detected the inflection templates. We also identified the phonemic transcriptions wherever they occur. We finally parsed the conjugation tables (cf. Figure 5). We thus collected as much (possibly redundant) information as possible and applied some heuristics to automatically detect major inconsistencies.

```

{{-adj-|fr}}
{{fr-accord-cons|a.fly.ã|t}}
'''affluent'''
# {{géographie|fr}} Qui se [[jeter|jette]] [[dans]] un [[autre]] en [[parlant]]
d'un [[cours]] d'eau.

{{-nom-|fr}}
{{fr-rég|a.fly.ã}}

{{-flex-verb-|fr}}
{{fr-verbe-flexion|affluer|ind.p.3p=oui|sub.p.3p=oui}}
'''affluent''' {{pron|a.fly|fr}}
# ''3ème pers. du pluriel de l'indicatif présent de'' [[affluer]].
# ''3ème pers. du pluriel du subjonctif présent de'' [[affluer]].

{{-pron-}}
{| class="wikitable"
| Adjectif et nom commun
* {{pron-rég|France|ã.n.a.fly.ã|titre=un affluent}}
|-
| Forme du verber affluer
* {{pron-rég|France (île-de-France)|a.fly}}

```

Figure 2: Wikicode of the article *'affluent'*.


<h1>affluente</h1> <hr/>  <b>Forme d'adjectif</b> <hr/> <p><b>affluente</b> <i>féminin</i> /a.fly.õt/</p> <ol style="list-style-type: none"><li>1. Féminin singulier de <b>affluent</b>.</li></ol>	<hr/> <pre>{{-flex-adj- fr}} '''affluente''' {{f}} {{pron a.fly.õt lang=fr}} #''Féminin singulier de'' [[affluent#fr-adj affluent]].</pre> <hr/>
---	--

Figure 3: Article and wikicode of 'affluente'.



<h1>affluentes</h1> <hr/>  <b>Forme d'adjectif</b> <hr/> <p><b>affluentes</b></p> <ol style="list-style-type: none"><li>1. Féminin pluriel d'<b>affluent</b>.</li></ol>  <b>Prononciation</b> <hr/> <ul style="list-style-type: none"><li>• /a.fly.õt/</li></ul>	<hr/> <pre>{{-flex-adj- fr}} '''affluentes''' # Féminin pluriel d''''[[affluent]]'''.</pre> <hr/> <pre>{{-pron-}} * {{pron a.fly.õt}}</pre> <hr/>
--	---

Figure 4: Article and wikicode of 'affluentes'.





**Wiktionnaire**  
Le dictionnaire libre

Page d'accueil  
Index alphabétique  
Portails thématiques  
Page au hasard  
Page au hasard par langue  
Poser une question

Contribuer  
Journal des contributeurs  
La Wikidémie  
Communauté  
Discuter sur IRC  
Modifications récentes  
Faire un don

Aide  
Outils  
Langues

## Annexe:Conjugaison en français/affluer

Conjugaison de **affluer**, verbe du 1<sup>er</sup> groupe, conjugué avec l'auxiliaire *avoir*.

### Modes impersonnels

Mode	Présent	Passé
<b>Infinitif</b>	affluer /a.flɥe/	avoir afflué /a.vwaʁ_a.flɥe/
<b>Gérondif</b>	en affluant /ɑ̃.n_a.flɥɑ̃/	en ayant afflué /ɑ̃.n_ɛ.jɑ̃.t_a.flɥe/
<b>Participe</b>	affluant /a.flɥɑ̃/	afflué /a.flɥe/

### Indicatif

Présent		Passé composé	
j'afflue	/ʒ_a.flɥ/	j'ai afflué	/ʒ_e a.flɥe/
tu afflues	/ty a.flɥ/	tu as afflué	/ty a.z_a.flɥe/
il/elle/on afflue	/[il/ɛl/ɔ̃] a.flɥ/	il/elle/on a afflué	/[i.l/ɛ.l/ɔ̃.n]_a.t_a.flɥe/
nous affluons	/nu.z_a.flɥɔ̃/	nous avons afflué	/nu.z_a.vɔ̃.z_a.flɥe/
vous affluez	/vu.z_a.flɥe/	vous avez afflué	/vu.z_a.ve.z_a.flɥe/
ils/elles affluent	/[il/ɛl].z_a.flɥ/	ils/elles ont afflué	/[i/ɛl].z_ɔ̃.t_a.flɥe/
Imparfait		Plus-que-parfait	
j'affluais	/ʒ_a.flɥɛ/	j'avais afflué	/ʒ_a.ve.z_a.flɥe/
tu affluais	/ty a.flɥɛ/	tu avais afflué	/ty a.ve.z_a.flɥe/
il/elle/on affluait	/[il/ɛl/ɔ̃] a.flɥɛ/	il/elle/on avait afflué	/[i.l/ɛ.l]_a.ve.t_a.flɥe/
nous affluions	/nu.z_a.flɥjɔ̃/	nous avions afflué	/nu.z_a.vjɔ̃.z_a.flɥe/
vous affluiez	/vu.z_a.flɥje/	vous aviez afflué	/vu.z_a.vje.z_a.flɥe/
ils/elles affluaient	/[il/ɛl].z_a.flɥɛ/	ils/elles avaient afflué	/[i/ɛl].z_a.ve.t_a.flɥe/
Passé simple		Passé antérieur	
j'affluai	/ʒ_a.flɥe/	j'eus afflué	/ʒ_y.z_a.flɥe/
tu affluas	/ty a.flɥa/	tu eus afflué	/ty y.z_a.flɥe/
il/elle/on afflua	/[il/ɛl/ɔ̃] a.flɥa/	il/elle/on eut afflué	/[i.l/ɛ.l/ɔ̃.n]_y.t_a.flɥe/
nous affluâmes	/nu.z_a.flɥam/	nous eûmes afflué	/nu.z_ym.z_a.flɥe/
vous affluâtes	/vu.z_a.flɥat/	vous eûtes afflué	/vu.z_yt.z_a.flɥe/
ils/elles affluèrent	/[il/ɛl].z_a.flɥɛʁ/	ils/elles eurent afflué	/[i/ɛl].z_yʁ.t_a.flɥe/
Futur simple		Futur antérieur	
j'affluerai	/ʒ_a.flɥ.ʁe/	j'aurai afflué	/ʒ_o.ʁe a.flɥe/
tu afflueras	/ty a.flɥ.ʁa/	tu auras afflué	/ty o.ʁa.z_a.flɥe/
il/elle/on affluera	/[il/ɛl/ɔ̃] a.flɥ.ʁa/	il/elle/on aura afflué	/[i.l/ɛ.l/ɔ̃.n]_o.ʁa a.flɥe/
nous affluerons	/nu.z_a.flɥ.ʁɔ̃/	nous aurons afflué	/nu.z_o.ʁɔ̃.z_a.flɥe/
vous affluez	/vu.z_a.flɥ.ʁe/	vous aurez afflué	/vu.z_o.ʁe.z_a.flɥe/
ils/elles afflueront	/[il/ɛl].z_a.flɥ.ʁɔ̃/	ils/elles auront afflué	/[i/ɛl].z_o.ʁɔ̃.t_a.flɥe/

### Subjonctif

Présent		Passé	
que j'afflue	/kə_ʒ_a.flɥ/	que j'aie afflué	/kə_ʒ_ɛ a.flɥe/
que tu afflues	/kə ty a.flɥ/	que tu aies afflué	/kə ty_ɛ.z_a.flɥe/
qu'il/elle/on afflue	/k_[il/ɛl/ɔ̃] a.flɥ/	qu'il/elle/on ait afflué	/k_[i.l/ɛ.l/ɔ̃.n]_ɛ.t_a.flɥe/
que nous affluions	/kə nu.z_a.flɥjɔ̃/	que nous ayons afflué	/kə nu.z_ɛ.jɔ̃.z_a.flɥe/
que vous affluiez	/kə vu.z_a.flɥje/	que vous ayez afflué	/kə vu.z_ɛ.je.z_a.flɥe/
qu'ils/elles affluent	/[il/ɛl].z_a.flɥ/	qu'ils/elles aient afflué	/k_[i/ɛl].z_ɛ.t_a.flɥe/

Figure 5: Conjugation table of the verb affluer (extract).



### 3 GLÀFF

In this section, we summarize some relevant characteristics of GLÀFF, first introduced in (Sajous et al. 2013a), from which PsychoGLÀFF is derived. The latest version of GLÀFF includes nouns, verbs, adjectives, adverbs, and function words. As can be seen in Figure 6, GLÀFF specifies for each entry:

- the wordform;
- the lemma;
- the part of speech and morphosyntactic features in GRACE format (Rajman et al. 1997);
- the phonological transcription(s) (when specified in Wiktionnaire) in IPA and in SAMPA with syllable boundaries.

```

affluent | Ncms | affluent | a.fly.ɑ̃ | a.fly.A~
affluent | Afpms | affluent | a.fly.ɑ̃ | a.fly.A~
affluents | Afpmp | affluent | a.fly.ɑ̃ | a.fly.A~
affluents | Ncmp | affluent | a.fly.ɑ̃ | a.fly.A~
affluent | Vmip3p- | affluer | a.fly | a.fly
affluent | Vmsp3p- | affluer | a.fly | a.fly
    
```

Figure 6: Extract of GLÀFF.

#### 3.1 Coverage

GLÀFF differs from the lexicons currently used in NLP and psycholinguistics by its exceptional size. Table 1 shows the number of inflected forms and lemmas for simple words (only letters) and non-simple words (containing spaces, dashes or digits) in five different French lexicons. GLÀFF contains from 3 to 4 times more tokens and from 3 to 9 times more inflected forms than the other lexicons.

	Categorized inflected forms			Categorized lemmas		
	Simple	Non simple	Total	Simple	Non simple	Total
Lexique	147,912	4,696	152,608	46,649	3,770	50,419
BDLex	431,992	4,360	436,352	47,314	1,792	49,106
Lefff	466,668	3,829	470,497	54,214	2,303	56,517
Morpha- lou	524,179	49	524,228	65,170	7	65,177
GLÀFF	1,401,578	24,270	1,425,848	172,616	13,466	186,082

Table 1: Size of five French lexicons (counting only nouns, verbs, adjectives and adverbs).

Table 2 reports the intersection of GLÀFF with the other lexicons. We observe that the magnitude of the intersection depends on the size of the lexicons: the bigger a lexicon, the larger its intersection

with the other ones. Three groupings emerge: Lexique has the smallest coverage, only containing 9% of GLÀFF and 22% to 26% of the entries of the other lexicons. BDLex, Lefff and Morphalou cover 76% to 80% of Lexique and about 30% of GLÀFF. Finally GLÀFF is clearly on top with coverage of 85% to 93%. In total, its coverage is 5% to 65% higher than the other lexicons.

	<b>Lexique</b>	<b>BDLex</b>	<b>Lefff</b>	<b>Morphalou</b>	<b>GLÀFF</b>
Lexique	-	26.03	25.20	22.46	8.95
BDLex	76.02	-	79.87	70.40	28.75
Lefff	79.50	86.28	-	72.32	30.04
Morphalou	79.58	85.43	81.24	-	32.03
GLÀFF	84.83	93.26	90.23	85.66	-

**Table 2: Intersection of five French lexicons (% of the categorized inflected forms).**

Size is a crucial aspect of the lexicons used for research in derivational and inflectional morphology or, more generally, in the development of NLP tools such as morphosyntactic taggers and parsers. In order to assess that GLÀFF’s largest size is actually useful, we compared the five lexicons with the vocabulary of four corpora of various types. Frantext 20<sup>e</sup> is constituted by 515 novels of 20th century French literature containing 30 million words. LM10 is a 200 million word corpus made up of the archives of the newspaper Le Monde from 1991 to 2000. The third corpus, containing 260 million words, consists of the articles from the French Wikipedia. Finally, FrWaC (Baroni et al. 2009) is a 1.6 billion words corpus of French web pages. Table 3 shows the coverage of the five lexicons with respect to the four corpora.

Threshold: frequency ≥		1	2	5	10	100	1000
<b>Frantext</b>	#forms	145,437	95,189	61,813	43,919	10,767	1,376
	Lexique	66.76	84.35	94.00	96.91	99.15	99.27
	BDLex	70.86	84.69	92.47	95.74	99.12	99.20
	Lefff	71.89	85.63	93.21	96.21	99.08	98.90
	Morphalou	73.93	86.66	93.29	96.00	98.48	97.09
	GLÀFF	76.92	88.57	94.54	96.72	98.77	98.76
<b>LM10</b>	#forms	300,606	172,036	106,470	77,936	29,388	83.21
	Lexique	29.59	47.28	65.23	76.31	93.81	98.58
	BDLex	37.77	55.79	71.76	80.93	95.53	98.69
	Lefff	39.64	58.22	74.33	83.20	95.99	98.90
	Morphalou	39.06	56.82	71.92	80.32	93.27	97.48
	GLÀFF	45.24	63.83	78.63	86.23	96.46	98.68
<b>Wikipedia</b>	#forms	953,920	435,031	216,210	136,531	35,621	7,956
	Lexique	9.13	18.27	31.52	43.03	78.58	95.72
	BDLex	12.29	22.89	36.80	48.04	79.39	95.33
	Lefff	12.88	23.94	38.26	49.65	80.57	95.71
	Morphalou	13.05	23.96	37.87	48.87	78.74	94.16
	GLÀFF	16.42	29.00	44.13	55.45	83.21	96.10
<b>FrWaC</b>	#forms	1,624,620	846,019	410,382	255,718	74,745	22,100
	Lexique	5.83	10.85	20.84	30.81	66.00	89.47
	BDLex	9.36	15.85	27.28	37.48	69.61	90.03
	Lefff	9.85	16.67	28.57	39.16	71.61	91.16
	Morphalou	10.09	16.89	28.53	38.68	69.36	88.51
	GLÀFF	13.13	21.13	34.29	45.35	76.39	92.76

**Table 3: Lexicon/corpus coverage (% of non-categorized inflected forms).**

The vocabulary is restricted to the forms that occur at least once, 2, 5, 10, 100 and 1000 times. The ranking of the corpora by coverage is the same for the five lexicons. Although their size affects the order, their nature is also crucial. For example, FrWaC being a collection of web pages, it contains a large number of “noisy” forms (foreign words, missing or extra spaces, missing diacritics, random spelling, etc.). Again, we see the division of lexicons into three groups. BDLex, Lefff and Morphalou have a quite close coverage. Except for Frantext 20<sup>e</sup>, Lexique has the smallest coverage. GLÀFF has the largest coverage for all corpora, except for LM10 at the 1000 threshold where it is surpassed by Lefff by 0.2%. The best coverage of Lexique for the Frantext 20<sup>e</sup> corpus, above the 10 threshold, while it has the weakest coverage in all other cases, is explained by the design of its vocabulary, extracted from this corpus. For the other corpora and up to the 100 threshold, the size of GLÀFF explains its larger coverage with re-

spect to the other lexicons (at the threshold 1, 14% to 53% larger for LM10 and 30% to 120% larger for FrWaC; at the threshold 10, 4% to 16% for LM10 and 15% to 47% for FrWaC). NLP tools that integrate GLÀFF should therefore offer an improved performance in the treatment of these corpora. In a qualitative study described in (Sajous et al. 2014), we observed that GLÀFF specific entries contains not only rare neologisms, but also very common words such as *attractivité* ‘attractivity’, *brevetabilité* ‘patentability’, *diabolisation* ‘demonization’, *employabilité* ‘employability’, *homophobie* ‘homophobia’, *hébergeur* ‘host’, *fatwa*, *institutionnellement* ‘institutionally’, *anticorruption* ‘anti-corruption’, etc. missing from the other lexicons.

### 3.2 Phonemic transcriptions

GLÀFF provides a phonemic transcription for about 90% of the entries. We evaluated the consistency of these transcriptions with respect to those of BDLex and Lexique (after conversion into IPA encoding).

Tables 4a to 4c report the top ten variations between pairs from the three lexicons. We only considered one phoneme differences, ignoring syllabification. The differences in transcriptions between GLÀFF and the other two lexicons are comparable to the differences observed between BDLex and Lexique. In particular, these differences are mostly due to the distinctions between the mid vowels, i.e. the front-mid vowels: [e] (close-mid) vs. [ɛ] (open-mid) and the back-mid vowels: [o] (close-mid) vs. [ɔ] (open-mid). This alternation is a well-known aspect of French phonology resulting from diatopic variations (North vs. South), as described in (Detey et al. 2010). Such expected oppositions account for about 91% of the divergences between BDLex and Lexique. Table 5 reports the percentage of identical phonological transcriptions shared by the lexicons and the percentage of the ‘comparable’ phonological transcriptions, i.e. disregarding the distinction between close-mid and open-mid vowels. GLÀFF and Lexique give identical transcriptions for 79.5% of entries whereas the percentage between GLÀFF and BDLex is lower, at 61.7%. Table 5 also reports the results of the comparison of syllabification in the three lexicons (performed on the basis of identical transcriptions only). This comparison shows that the three lexicons are quite similar with respect to syllabification (98%).

Comparing GLÀFF with the major resources that contain the same type of information clearly shows that the overall quality of the lexicon is quite satisfactory and is in all respect comparable to those of these resources.

Oper.	Phonemes	%	Σ %	Oper.	Phonemes	%	Σ %	Oper.	Phonemes	%	Σ %
r	ε/e	48.18	48.18	r	ɔ/o	60.03	60.03	r	e/ε	66.46	66.46
r	ɔ/o	32.17	80.36	i	ə	14.18	74.21	r	ɔ/o	10.58	77.05
r	o/ɔ	11.02	91.37	r	e/ε	6.90	81.11	i	ə	5.90	82.96
r	y/ɥ	1.83	93.21	r	ε/e	4.98	86.09	r	o/ɔ	4.36	87.32
r	ə/ø	1.44	94.64	r	ɑ/a	4.92	91.01	r	ɑ/a	3.84	91.17
r	ə/œ	1.39	96.03	r	s/z	1.25	92.26	r	ɥ/y	1.61	92.78
r	u/w	0.84	96.87	r	ə/ø	0.91	93.17	r	œ/ə	1.09	93.88
r	b/p	0.73	97.61	r	œ/ø	0.47	93.64	r	ø/ə	0.86	94.74
r	s/z	0.51	98.12	i	i	0.42	94.06	i	i	0.84	95.58
d	j	0.25	98.37	r	o/ɔ	0.38	94.44	r	w/u	0.79	96.38
(a) BDLex/Lexique				(b) GLÀFF/Lexique				(c) GLÀFF/BDLex			

**Table 4: The 10 most frequent differences in phonemic transcriptions (Operations: r = replacement, i = insertion, d = deletion).**

Lexicons		Intersection	Phonemic transcriptions		Syllabification
			Identical	Comparable	Identical
BDLex	Lexique	112,439	58.31	96.88	98.92
GLÀFF	Lexique	123,630	79.50	97.81	98.48
GLÀFF	BDLex	396,114	61.72	96.88	98.30

**Table 5: Inter-lexicon agreement: phonemic transcriptions and syllabification.**

## 4 From GLÀFF to PsychoGLÀFF

### 4.1 Overview

Our goal in creating PsychoGLÀFF is to provide psycholinguists with a set of features related to the formal aspects of the lexicon entries. For this purpose, we selected from GLÀFF only forms having non-zero frequency in at least one of the corpora mentioned in section 3.1. This means that PsychoGLÀFF only contains lexical entries attested in the corpora, amounting to about 340,000 forms for 120,000 lemmas.

In addition to GLÀFF's features, PsychoGLÀFF includes the following information for each entry:

- the absolute and relative frequencies of the wordform and of the lemma in the aforementioned French corpora (Frantext 20<sup>e</sup>, LM10 and FrWac);

- the length of the wordform (number of characters);
- the length of the phonological transcription(s) (number of phonemes);
- the syllabification and the CV structure of the wordform;
- the number of syllables;
- the geometric mean of the conditional character probabilities of bigrams, which calculates the probability of the bigram occurring given the preceding bigram;
- the geometric mean of the conditional character probabilities of trigrams, which calculates the probability of the trigram occurring given the preceding trigram;
- the geometric mean of the conditional character probabilities of 4-grams, which calculates the probability of the 4-gram occurring given the preceding 4-gram;
- the geometric mean of the conditional phoneme probabilities respectively calculated for bigrams, trigrams and 4-grams.
- the size of the orthographic neighbourhood, i.e. the number of words in the lexicon differing by one character (via deletion, insertion, or substitution);
- the size of the phonological neighborhood, i.e. the number of words differing from the phonological transcription by one phoneme (via deletion, insertion, or substitution);
- the size of the ratio between the number of consonants and syllables composing the phonological form. This score is meant to provide an estimate of the ‘syllabic complexity’ of the form.

The *n*-gram conditional probability represents a measure of phonotactic occurrence, defining the likelihood of occurrence of *n*-grams in French. This kind of measure is expected to be particularly helpful for the design of experimental stimuli in lexical access experiments (Storkel and Hoover 2011).

## 4.2 Comparison with Lexique

We compare hereafter, in terms of coverage and word frequencies, PsychoGLÀFF and Lexique, the most frequently used French lexicon in psycholinguistics.

Being directly extracted from GLÀFF, PsychoGLÀFF stands out with respect to the lexicons currently used in psycholinguistics mostly for its size, as it counts 337,572 entries. Table 6 shows the number of inflected forms and lemmas of Lexique and PsychoGLÀFF. The relative coverage of these lexicons is reported in Table 7.

	<b>Categorized inflected forms</b>	<b>Categorized lemmas</b>
Lexique	153,934	50,419
PsychoGLÀFF	337,572	121,021

**Table 6: Size of Lexicons (restricted to nouns, verbs, adjectives and adverbs)**

	<b>Lexique</b>	<b>PsychoGLÀFF</b>
Lexique		36.1 %
PsychoGLÀFF	78.9 %	

**Table 7: Lexicons relative coverage (% of categorized inflected forms)**

We observe that PsychoGLÀFF is more than twice larger than Lexique (for both inflected forms and lemmas) and has a total coverage of about 79% with respect to Lexique (which covers only 36.1% of the inflected forms of PsychoGLÀFF). PsychoGLÀFF reports the absolute and relative frequencies for its wordforms and lemmas. Frequencies are calculated on the basis of three stylistically different corpora of written French: the abovementioned Frantext20<sup>e</sup>, LM10 and FrWac (literature, newspaper and web corpora). Lexique reports word frequency estimates too. It exploits two smaller corpora: a) a written corpus made up of 218 books from Frantext 20<sup>e</sup>; b) a corpus of French subtitles for 9,474 movies and television series, assumed to be more representative of spoken French.

While GLÀFF and PsychoGLÀFF frequencies are exclusively based on written French, Lexique mixes together spoken-like and written resources for the calculation of wordform and lemma frequencies. Although the corpora used by the two lexicons have very different sizes, we attempted a comparison of the PsychoGLÀFF's frequencies with respect to the frequencies reported in Lexique, looking only at the intersection of the two lexicons.

Table 8 reports the correlation between the normalized frequencies of wordforms in PsychoGLÀFF (separately for Frantext, LM10 and FrWac) and Lexique (separately for books and movie subtitles). The data were normalized by one million words. It is not surprising that the correlation between Frantext's frequencies and Lexique's book frequencies is quite high (Pearson's coefficient  $\rho = .81$ ), the latter being a sub-corpus of Frantext 20<sup>e</sup>. Although PsychoGLÀFF frequencies are based exclusively on written corpora, we found a statistically significant correlation  $\rho = .68$  between Lexique's subtitle frequencies and the Frantext frequencies (the value slightly decreases for the subtitles/FrWac correlation). This seems to indicate that the lexical coverage of PsychoGLÀFF, though based on written sources, is comparable to a relevant extent to the coverage of corpora specifically devoted to spoken French.

		Lexique	
		Subtitles	Books
PsychoGLÀFF	Frantext	.68	.81
	LM10	.62	.59
	FrWac	.67	.62

**Table 8: PsychoGLÀFF/Lexique correlations with respect their normalized frequency values.**

An additional property of PsychoGLÀFF worth noting to is the presence of infrequent lexical items. This feature clearly derives from the nature of Wiktionnaire: being an online dictionary, it has not to conform to the same size constraints as printed ones. Bootstrapped by importation of articles from public domain dictionaries, it contains dated entries. Finally, being crowdsourced, it is regularly updated and contains general-domain neologisms, as well as subculture vocabulary and technical terms (Sajous et al. 2014). As a consequence, PsychoGLÀFF contains a large number of specific entries.

The bigger the corpus, the more low-frequency lexical items are likely to be included (while the size of the corpus is not likely to have a strong impact on the number of those words that are frequent or

very frequent in a language, representing to a certain extent the ‘essential lexicon’ of that language). The Figure 7 illustrates this point by showing the distribution of different frequency intervals for both Lexique’s and PsychoGLÀFF’s sub-corpora. A normalized frequency range of 10.01 or more corresponds to very high words frequency and is situated at the right edge of the graph. A frequency range of 0.01-0.1 corresponds to very low words frequency and is situated at the left edge of the figure.

Six intermediate ranges capture the frequency differences of the entire lexicon. The figure shows that the distribution of the frequency intervals is approximately the same for Lexique and PsychoGLÀFF, with the significant exception of the least frequent word class (< 0.2), for which the number of lexical items in PsychoGLÀFF is almost twice as large as that of Lexique. At the same time, PsychoGLÀFF contains many words of ordinary usage that are absent from Lexique, such as *acceptabilité* ‘acceptability’, *centralité* ‘centrality’, *Saturne* ‘Saturn’, etc. In this sense PsychoGLÀFF offers a much larger lexical repertoire not only in terms of tokens, but also in terms of types, which represents a particularly interesting feature for psycholinguistic studies and corpus investigations.

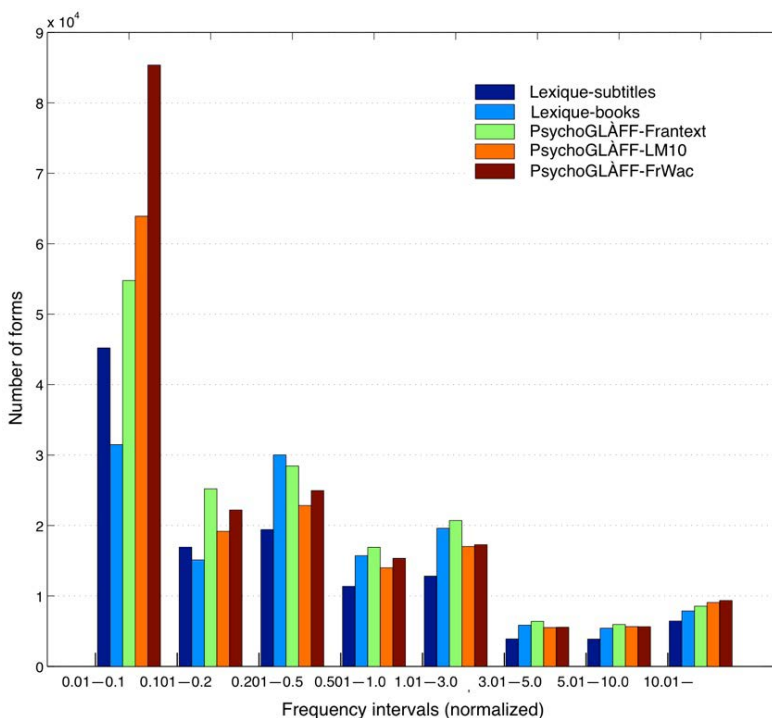


Figure 7: Distribution of forms with respect to their corpus frequency.

## 5 Conclusions and future directions

This paper presents a first version of PsychoGLÀFF, a large lexicon designed for psycholinguistic experimentation. PsychoGLÀFF was built on top of the inflectional and phonemic lexicon GLÀFF, itself acquired from Wiktionnaire, the French edition of the collaborative dictionary Wiktionary. In particular, PsychoGLÀFF contains the subset of GLÀFF’s corpora-attested entries. This resource complements



the inflectional and phonological information present in GLÀFF with features needed for experimental material calibration including frequency, lexical neighborhood, syllabic complexity and phonotactic likelihood.

Like GLÀFF, PsychoGLÀFF is characterized by an exceptional coverage, much higher than those of comparable resources as Lexique on one hand and Morphalou and Lefff on the other. We also show that the “primary” information (parts of speech, phonemic transcriptions, frequency) of PsychoGLÀFF and GLÀFF has a satisfactory quality. PsychoGLÀFF is a free resource distributed under a copy-lefted license and is available to all psycholinguistic researchers working on French. We hope that it will soon be adopted by this community whose feedback will allow us to improve the resource and appropriately respond to its needs.

In the near future, we plan to improve PsychoGLÀFF on several aspects. One of them will concern the description completeness and consistency of the lexicon. An online interface, comparable to GLÀFFOLI (the GLÀFF OnLine Interface) will also be developed, that will enable users to query the lexicon and develop experimental material interactively.

## 6 References

- Anton Pérez, L., Gonçalo Oliveira, H., and Gomes, P. (2011). Extracting Lexical-Semantic Knowledge from the Portuguese Wiktionary. In *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, pp. 703–717, Lisboa, Portugal.
- Baayen, R. H., Piepenbrock, R. and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Philadelphia: Linguistic Data Consortium.
- Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3): 209-226.
- Boula De Mareuil, P., Yvon, F., D’Alessandro, C., Aubergé, V., Vaissière, J., and Amelot, A. (2000). A French Phonetic Lexicon with variants for Speech and Language Processing. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 273–276, Athens, Greece.
- Clément, L., Lang, B. and Sagot, B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 1841–1844, Lisboa, Portugal.
- Detey, S., Durand, J., Laks, B., and Lyche, C. (2010). *Les variétés du français parlé dans l’espace francophone*. Paris: Ophrys.
- Gonçalo Oliveira, H. and Gomes, P. (2010). Onto.PT: Automatic Construction of a Lexical Ontology for Portuguese. In *Proceedings of 5th European Starting AI Researcher Symposium*, pp. 199–211. IOS Press.
- Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer C. M. and Wirth C. (2012). UBY - A Large-Scale Unified Lexical- Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, Avignon, France.
- Meyer, M. C. and Gurevych, I. (2012). *OntoWiktionary - Constructing an Ontology from the Collaborative Online Dictionary Wiktionary*. In *Semi-Automatic Ontology Development: Processes and Resources*, chapter 6, pp 131–161. IGI Global.
- New, B. (2006). Lexique 3: Une nouvelle base de données lexicales. *Actes de la 13<sup>e</sup> Conférence Traitement Automatique des Langues Naturelles (TALN 2006)*, Louvain-la-Neuve, Belgium.

- New, B., Brysbaert, M., Veronis, J. and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. In *Applied Psycholinguistics* (28): 661–677.
- Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P. and Huang, C. R. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the ACL Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*. ACL-IJCNLP 2009, Singapore.
- Pérennou, G. and De Calmès, M. (1987). BDLEX lexical data and knowledge base of spoken and written French. In *Proceedings of ECST 1987*, pp. 1393–1396, Edinburgh, UK.
- Rajman, M., Lecomte, J., and Paroubek, P. (1997). *Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique*. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.
- Romary, L., Salmon-Alt, S. and Francopoulo, G. (2004). Standards going concrete: from LMF to Morphalou. In *Workshop on Electronic Dictionaries*, Geneva, Swiss.
- Sajous, F., Navarro, E., Gaume, B., Prévot, L., et Chudy, Y. (2010). Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Loftsson, H., Rögnvaldsson, E., and Helgadóttir, S. (eds), *Advances in Natural Language Processing*, vol. 6233 of LNCS, 332–344. Springer.
- Sajous, F., Hathout, N. and Calderone, B. (2013a). *GLÀFF, un Gros Lexique À tout Faire du Français*. Actes de la 20<sup>e</sup> Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013), pp. 285-298, Les Sables d'Olonne, France.
- Sajous, F., Navarro, E., Gaume, B., Prévot, L. and Chudy, Y. (2013b). Semi-automatic enrichment of crowdsourced synonymy networks: the WISIGOTH system applied to Wiktionary. *Language Resources and Evaluation*, 47(1): 63-96.
- Sajous, F., Hathout, N. and Calderone, B. (2014). Ne jetons pas le Wiktionnaire avec l'oriipeau du Web ! Etudes et réalisations fondées sur le dictionnaire collaboratif. *Actes du 4<sup>e</sup> Congrès Mondial de Linguistique Française (CMLF 2014)*, Berlin, Germany.
- Schultz, T., Vu, N. T. and Schlippe, T. (2013). GlobalPhone: A multilingual text & speech database in 20 languages. In *Proceedings of the Conference on Acoustics, Speech, and Signal Processing*, pp. 8126–8130, Vancouver, Canada.
- Sérasset, G., (2012). Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.
- Storkel, H. L. and Hoover, J. R. (2011). The influence of part-word phonotactic probability/neighborhood density on word learning by preschool children varying in expressive vocabulary. In *Journal of Child Language*, 38, 628–643
- Zesch, T., Müller, C. and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

# RIDIRE. Corpus and Tools for the Acquisition of Italian L2

Alessandro Panunzi, Emanuela Cresti, Lorenzo Gregori  
University of Florence  
alessandro.panunzi@unifi.it, elicresti@unifi.it, lorenzo.gregori@unifi.it

## Abstract

This paper introduces the RIDIRE corpus, built by means of an open source tool (RIDIRE-CPI) for creating specifically designed web corpora through a targeted crawling strategy. The RIDIRE-CPI architecture combines existing open source tools with specifically developed modules, comprising a robust crawler, a user friendly web interface, several conversion and cleaning tools, an anti-duplicate filter, a language guesser, and a PoS-tagger. The RIDIRE corpus is a balanced Italian web corpus (1.5 billion tokens) designed for enhancing the study of Italian as a second language, while also being exploitable for lexicographic purposes. The targeted crawling was performed through content selection, metadata assignment, and validation procedures. These features allowed the construction of a large corpus with a specific design, covering a variety of language usage domains (News, Business, Administration and Legislation, Literature, Fiction, Design, Cookery, Sport, Tourism, Religion, Fine Arts, Cinema, Music). The RIDIRE query system allows research to be carried out on the whole corpus itself or on the sub-corpora. Specifically, available queries comprehend all the functions usually exploited in corpus-based lexicography: frequency lists, concordances and patterns, collocations, Sketches, and Sketch Differences.

**Keywords:** Corpus linguistics; Terminology; Collocations

## 1 Introduction

RIDIRE (acronym for *RIsorse Dinamiche dell'Italiano in Rete*, “Italian Dynamic Resources Online”; Monaglia & Paladini 2010) is a project which produced a large Italian language corpus, and an open-source tool for web corpora building and processing, named RIDIRE-CPI (Panunzi et al. 2012). The corpus - of 1.5 billion tokens - was built using web-crawling techniques and exploited the Italian content of the Internet. The corpus is now available online and is integrated with computational tools for the exploitation of vast corpora to enhance language usage in L2 Italian learners. RIDIRE is designed for use by both teachers and learners, who will be able to profit from access to a database of representative texts which characterize Italian culture. The database collects a massive amount of freely available content, covering a selection of domains which are relevant to Italian identity: law, religion, politics, literature, trade, administration, information, design, food, fashion. To reach this goal, a distributed

crawling infrastructure was created and a targeted crawling strategy pursued. This document will summarize the corpus design for the resource as well as the crawling techniques and processing tools used for deriving language corpora from the web. Also presented are examples of queries that are relevant for both learners and lexicographers.



Figure 1: The RIDIRE resource home page.

## 2 Corpus Design Strategy

Different kinds of projects have been carried out to exploit the language data populating the web (Kilgarriff & Greffentette 2003, Sharoff 2006). Among these, the WaCky initiative (Baroni et al. 2009) and the Italian web corpus ItWaC are important antecedents. More recently a new generation of web corpora have been created and processed with boilerplate cleaning and de-duplication tools and are available through Sketch Engine for a large number of languages (Kilgarriff et al. 2004); these are identified through their target size as the TenTen collection: 10 billion word corpora ( $10^{10}$ ). Such initiatives resulted in the development of dedicated software for crawling (Heritrix), text-processing, cleaning, and the large-scale use of existing technologies for morpho-syntactic annotation (TreeTagger) and online corpus querying (CQPweb). These technologies have been used in RIDIRE and adapted to its specific goals.

The RIDIRE project aimed to build an online database representative of a wide and significant Italian language universe which would have value for sourcing information on the use of Italian in various aspects of life and culture, for linguistic/lexicographic researches, and for didactic purposes. To build such a resource involved two corpus design requirements which did not characterize the web corpora collected in previous initiatives: a) the selection of linguistic resources which document the main domains of usage (life and culture); b) the enrichment of the resource with metadata which enables a perspicuous querying of the database in each specific domain.

The collection focuses on two sets of non-hierarchically structured domains, selected for their pragmatic relevance to the use of the Italian language. The first set is constituted by general non-semantic fields, in which language characterizes its function (up to 400 million words for each domain):

- News
- Business
- Administration and Legislation

The second consists of semantic fields in which Italian excellence is largely recognized (up to 100 million words for each domain):

- Literature
- Fiction
- Design
- Cookery
- Sport
- Tourism
- Religion
- Fine Arts
- Cinema
- Music

The possibility for learners to find specific information on the language usage characterizing a domain should enhance their ability to find the right expressions for it. From a lexicographic point of view, the presence of different domains allows the derivation of specific uses of a word and the description of its semantic variation across the different domains of language use. Table 1 and Figure 2 show the structure of the corpus and the quantitative measures for each domain.

DOMAINS	# WEBSITES	# PAGES	# TOKENS	# WORDS
Functional (total)	189	976,460	854,388,230	747,268,841
Information	27	550,169	216,431,868	186,577,769
Economics and Business	123	226,535	179,710,476	161,377,152
Administration and Law	39	199,756	458,245,886	399,313,920
Semantic (total)	816	907,374	660,243,564	566,229,119
Sport	49	138,235	98,172,470	82,695,548
Architecture and Design	142	136,725	93,822,675	81,235,939
Cooking	20	123,376	52,784,045	45,523,096
Cinema	25	122,850	51,466,145	44,370,692
Music	195	113,015	12,906,213	106,287,283
Fashion	103	74,584	24,645,980	21,690,140
Visual Arts	118	70,601	56,517,442	48,929,903
Religion	51	66,053	72,454,492	62,291,806
Literature and Theatre	113	61,935	85,474,102	73,204,712
<b>Total</b>	<b>2,010</b>	<b>3,767,668</b>	<b>1,514,631,794</b>	<b>1,313,497,960</b>

**Table 1: Number of crawled websites, pages, tokens and words per domain.**

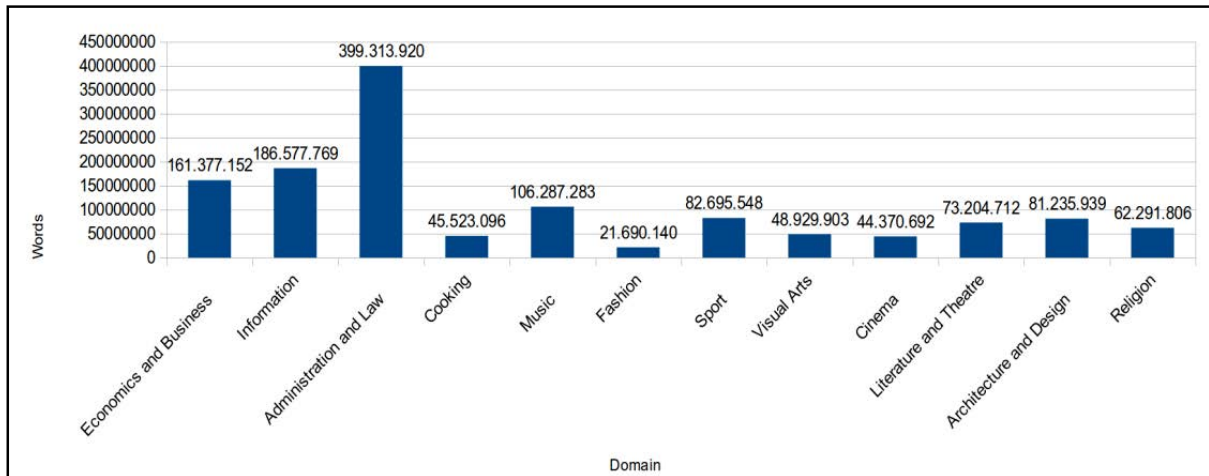


Figure 2: Words per Domain chart of RIDIRE Corpus.

### 3 The Crawling Infrastructure

The gathering of specific linguistic data for each sub-corpus requires a targeted crawling strategy performed by different teams of experts. The tool developed within the RIDIRE project for the crawling and the processing of the web resources (RIDIRE-CPI) is now open source and its user-friendly web interface is specifically intended to allow collaboration between users unskilled in web technology and text processing, working in a distributed environment. The application comprises:

- the crawling process
- the mapping of the resource in a MySQL database
- user interaction via web interface

RIDIRE-CPI has a modular architecture (see Figure 3), which is made up of:

- a web crawler
- a web interface for crawling management and validation
- conversion tools
- HTML cleaner tools
- anti-duplicate filters
- a language guesser
- a PoS-tagger

The crawling activity, as in the other cited web corpus initiatives, makes use of the Heritrix web crawler (version 3.1.1). However RIDIRE-CPI configures it via the web interface, making it suitable for use in a distributed environment. The crawling activity itself is structured into “jobs” (fully configured crawling sessions) in which the user determines three sets of parameters. First, the user selects the seed URLs from which the crawling activity starts. Then the formats of the resources that should be

downloaded are specified. In addition to HTML, RIDIRE-CPI is able to process TXT, RTF, DOC, and PDF documents. This feature is crucial, since many linguistically relevant resources from the web are not contained in web pages, but in documents of varying formats. The third set of parameters determine the strategy for the selection of content from websites. This step is important in downloading resources which comply with the representativeness requirement, since the reference unit for text on the web (when representing the language of a particular domain) is the web page rather than the website. As a matter of fact, only a subset of the web pages from a given site give information strictly concerning the specific domain to which the site belongs. Within the step, the user selects and/or discards the “resources” specifying

- which found URLs the crawler has to add to the queue (“URL to be navigated”);
- which resources the crawler has to download to the file system (“URL to be saved”)

Once all the parameters are defined by the user, the crawler starts from the first seed URL, which is put in the processing queue. The crawler accesses the web page relative to the first URL in the queue, extracts all the links that match the “URL to be navigated” rules and saves them in the queue; then, if the page is a “URL to be saved”, the crawler downloads the web page content and stores it on the file system. Finally, it goes back to the first step and proceeds recursively until the processing queue is empty.

To maximize the precision of the process, the user can decide to insert a list of complete URLs, to specify website areas with path substrings (any URL containing one of these strings) or to write a customized regular expression that matches desired page URLs. For instance, in Figure 4 the user decided to crawl the website <http://musica.atuttonet.it>, getting HTML pages only, and further navigating to any link found (this option is set with a regular expression in the *Pattern* field), downloading any pages that do not contain the word *varie* or *artisti* in the URL.

In this stage no technical competence is required, but a pre-analysis of the website(s) is necessary to ensure only relevant information is retrieved.

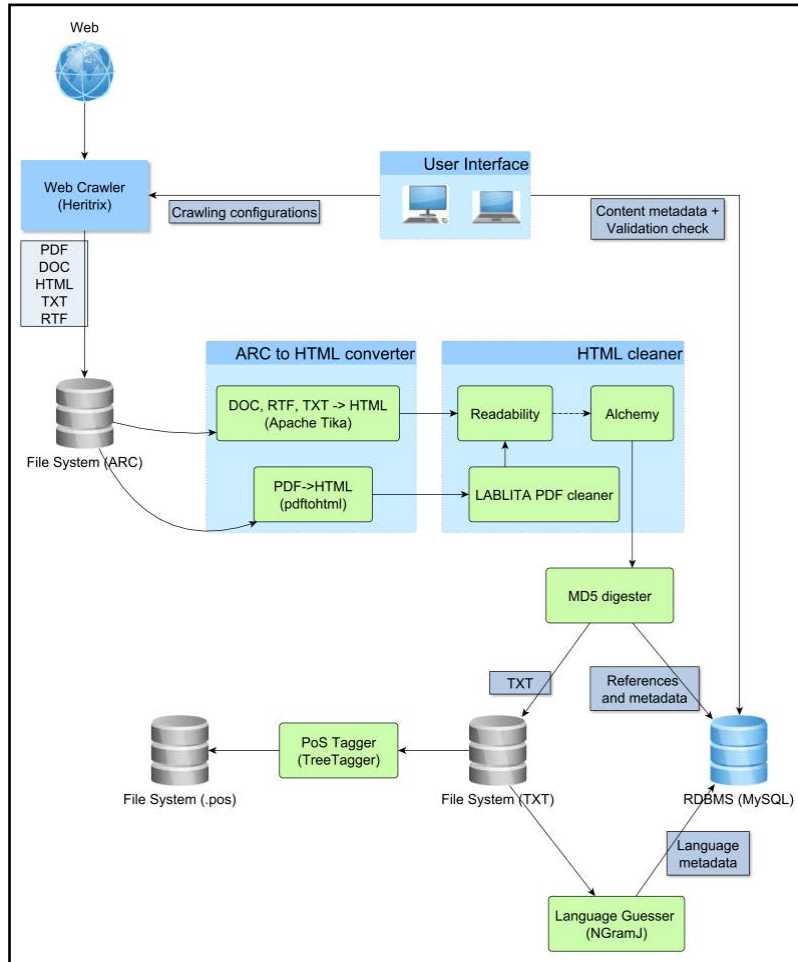


Figure 3: RIDIRE-CPI Architecture.

The screenshot shows the 'Job Creation' page. It includes the following fields and options:

- Job name:** atuttonet\_music
- Seeds:** http://musica.atuttonet.it/
- Formats:** A list of formats (DOC, HTML, PDF, RTF, TXT) with a 'HTML' button selected.
- URL to be navigated:**
  - Only URL containing these strings: (empty text area)
  - Pattern: ^.\*\$
  - Negative:
- URL to be saved:**
  - Only URL containing these strings: varie, artisti
  - Pattern: (empty text area)
  - Negative:

Figure 4: RIDIRE "Job Creation" page.



### 3.1 The Mapping Process

To be adequate for linguistic research, the crawled data needs to be processed by a procedure that includes text cleaning, duplicate removal, and PoS-tagging (Baroni et al. 2009). To this end, RIDIRE-CPI uses an automatic processing pipeline on the downloaded resources to extract the running text that will constitute the corpus itself. Web pages, as is well known, contain text that is not relevant for the constitution of a corpus e.g. advertising, navigation menus, disclaimers, credits, etc. (the so called “boilerplate”). Each terminated job is first converted into HTML, which involves several tools depending on the input format. After the conversion, the text cleaning is performed. The boilerplate is removed by means of two external tools freely available for research: Readability and Alchemy API. PDF files are more difficult to clean, so they are treated separately with a dedicated tool - PDF-Cleaner - that performs a deep filtering on the content.

Readability is the first option for the HTML cleaner, but if it won't yield results or outputs an error, the Alchemy API provides a second chance. The plain text documents output from the cleaning stage are then processed by a simple MD5 digester to get their signature, which acts as an anti-duplication system allowing the application to discard resources found with the same signature. The last phase of the mapping procedure is the part-of-speech tagging of the plain text resource. The PoS-tagging is performed by TreeTagger, which is run as an external executable by the main application. TreeTagger creates the PoS-tagged file in the correct file location directly.

### 3.2 Validation and Corpus Creation

RIDIRE-CPI integrates a validation interface dedicated to the evaluation of the crawled resources, which ensures that they belong to the specific domain they should represent. The validation procedure creates a random sample of the resources found and the user can check whether they are adequate with respect to the corpus design or content restrictions. A job can be considered “valid” if it contains non adequate resources under a given percentage (less than 10%, in principle). Since a manual revision is required for a high quality result, but checking the whole corpus is not an option due to its size, the validation process implemented in RIDIRE is a good trade-off between a clean corpus and a fast check. Figure 5 shows how the interface presents a random sampling of one crawled job, allowing direct access to a selection of pages whose adequacy in representing the given domain can be verified.

**The Job is valid**      **Validation data**

Validation  
 Threshold:  Percentage of non-valid resources beyond which the Job is considered non-valid.

Results: 1 - 10 of 34 results for page: 10

Show only resources to be validated

URL	Words	Lang.	MimeType	Valid
<a href="http://www.iperbole.bologna.it/quartiereporto/cons...">http://www.iperbole.bologna.it/quartiereporto/cons...</a>	968	it	application/pdf	Valid
<a href="http://www.iperbole.bologna.it/quartiereporto/cons...">http://www.iperbole.bologna.it/quartiereporto/cons...</a>	1293	it	application/pdf	Valid
<a href="http://www.iperbole.bologna.it/quartieresaragozza/...">http://www.iperbole.bologna.it/quartieresaragozza/...</a>	260	it	application/pdf	Valid
<a href="http://www.iperbole.bologna.it/quartierereno/piazz...">http://www.iperbole.bologna.it/quartierereno/piazz...</a>	174	it	application/pdf	Valid
<a href="http://www.iperbole.bologna.it/quartierereno/piazz...">http://www.iperbole.bologna.it/quartierereno/piazz...</a>	127	it	application/pdf	Valid
<a href="http://www.iperbole.bologna.it/quartierereno/piazz...">http://www.iperbole.bologna.it/quartierereno/piazz...</a>	7373	it	application/pdf	Valid
<a href="http://www.iperbole.bologna.it/quartierenavile/att...">http://www.iperbole.bologna.it/quartierenavile/att...</a>	443	it	application/pdf	Valid
<a href="http://www.comune.bologna.it/quartieresavena/qare...">http://www.comune.bologna.it/quartieresavena/qare...</a>	440	it	application/pdf	Valid
<a href="http://www.comune.bologna.it/quartierenavile/atti...">http://www.comune.bologna.it/quartierenavile/atti...</a>	264	it	application/pdf	Valid
<a href="http://www.comune.bologna.it/quartierenavile/atti...">http://www.comune.bologna.it/quartierenavile/atti...</a>	1503	it	application/pdf	Valid

Figure 5: Validation sampling.

Through the content selection, metadata assignment, and validation procedures, the RIDIRE-CPI allows the gathering of linguistic data from the web with a supervised strategy that allows a high level of control. The frequency lists of the various domains provide direct evidence that the crawling performed within expectations. The nouns (i.e. the referred entities) that ranked highly identify each domain (Religion, Fashion and Cookery) quite well, and are shown in Table 2.

## 4 Methods for the Extraction of Linguistic Information from Corpora in L2 Acquisition and Lexicography

Various experiences in trying to use corpora for second language acquisition purposes clearly show that both learners and teachers are scared by the complexities of techniques involved in corpus linguistics and that the resultant data is difficult to appreciate (Kilgarriff 2009). Concordances provide a large amount of fragmented information that is difficult to read, especially for second language learners. Despite the fact that corpora contain information that is needed and that the tools are pretty powerful (Sinclair 2004; Conrad 2006), the way to use these tools is undefined and the information retrieved is difficult to interpret, with the overall process being felt as time consuming. The challenge for corpus linguistics in the field of second language acquisition is to provide a simple way to link the actual needs of learners to corpus data.

Religion		Fashion		Cooking	
Lemma	Freq.	Lemma	Freq	Lemma	Freq
vita	210,420	collezione	56,685	ricetta	135,498
uomo	169,995	moda	50,381	iscritto	104,610
amore	110,831	anno	49,369	località	93,692
fedede	100,514	colore	32,777	acqua	82,492
mondo	98,913	abito	30,085	farina	81,695
pagina	95,462	mondo	28,816	volta	81,274
parola	92,532	donna	28,657	pasta	75,144
cuore	92,351	stile	26,815	zucchero	67,609
tempo	82,891	linea	26,026	minuto	66,579
giorno	76,190	pelle	20,962	impasto	65,074
figlio	70,231	capo	20,619	forno	61,672
persona	69,251	euro	19,199	olio	59,151
anno	69,054	modello	18,947	cucina	56,065
popolo	66,595	articolo	18,747	gr	55,079
modo	65,716	tempo	18,307	burro	52,101
preghiera	64,907	prodotto	17,365	uovo	49,057
cosa	57,020	marchio	16,968	cosa	48,276
santo	52,341	vita	16,388	tempo	47,712
fratello	51,370	accessorio	16,268	messaggio	47,453
famiglia	51,234	stilista	16,254	parte	46,829

**Table 2: The 20 most frequent nouns, taken from 3 different domains.**

The types of queries available in RIDIRE are inspired by those from the Sketch Engine and are available for both the general corpus and each sub-corpus:

- frequency lists
- concordances and patterns of words (ranked according to raw frequency)
- collocations (general and restricted to specific PoS)
- Sketches and Sketch Differences (between two words or domains) of collocates for the most relevant patterns of a word

The key strategy adopted in RIDIRE is to give a clear picture of the subset of problems that a learner can solve through corpora access, providing each problem area with a predetermined search path which leads to satisfactory results.

An extension of the concordances search function is the pattern search, where a user can view the concordances of a sequence of words (rather than a single one) specified by a form, lemma or PoS attribute; then, grouping the results together, he can see the more frequent usages of the sequence and

what the allowed syntactic structures are. In Figure 6 we searched the occurrences of the Italian verb *sperare* immediately followed by a preposition and we can see that there are five returned sequences (we excluded the rare occurrences): *sperare di* (68.37%), *sperare in* (13.88%), *sperare per* (4.24%), *sperare nel* (3.7%), *sperare nella* (3.26%). In this way a language learner can understand which prepositions may follow *sperare* and how they may be used by scrolling the occurrences list and looking at the different application contexts.

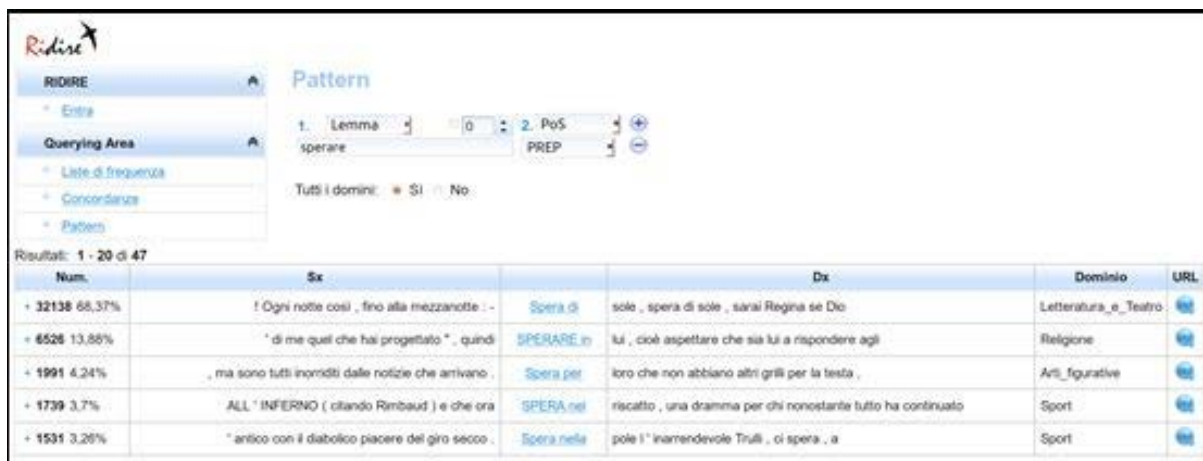


Figure 6: Pattern search grouped results.

RIDIRE is furthermore characterized by a set of sub-corpora representing Italian usage in different semantic and functional domains. The way in which a concept can be characterized in a given domain is partly a function of idiosyncratic usage conventions and corpus data can show this to the learner. In language this is reflected in particular by adjectives and adverbs, which show preferential meaning and associations and which vary across language usages. For instance, the variety of objects which are modified by the adjective *forte* (“strong”) vary when the context of usage is Religion or Cookery. The learner should wonder whether or not this adjective, learned in general, has specific meaning in a domain when applied to its particulars. Here, RIDIRE exploits its corpus variation. Corpus queries based on collocations demonstrate the possible choices, highlighting the adjective’s variation across domains.

The collocations in Figure 7 highlight the vastly different meanings conveyed by this adjective in each domain. In Religion, internal state is intensified (*fede*, “faith”; *tentazione*, “temptation”), while in Cookery flavours and smells are augmented. The meaning in one domain cannot automatically be extended to another.

Religion					Cookery				
Risultati: 1 - 20 di 4520					Risultati: 1 - 20 di 2049				
Parola	Punteggio	Freq. nel corpus	Freq. collocata	Freq. collocata attesa	Parola	Punteggio	Freq. nel corpus	Freq. collocata	Freq. collocata attesa
<a href="#">richiamo</a>	0,0077	47888	245	4.2191	<a href="#">farina</a>	0,169	8023	1254	0.2976
<a href="#">grido</a>	0,0071	20125	129	1.7731	<a href="#">sapore</a>	0,0138	49409	388	1.8329
<a href="#">tentazione</a>	0,0069	17810	116	1.5691	<a href="#">odore</a>	0,0098	16632	115	0.617
<a href="#">fede</a>	0,0065	163796	586	14.431	<a href="#">aroma</a>	0,0068	9707	56	0.3601
<a href="#">supplica</a>	0,0053	2467	49	0.2174	<a href="#">abbraccio</a>	0,0066	12520	64	0.4644
<a href="#">legame</a>	0,0051	47612	161	4.1948	<a href="#">drago</a>	0,0055	4226	30	0.1568
<a href="#">lacrima</a>	0,0037	18831	64	1.6591	<a href="#">gr</a>	0,0043	60518	149	2.245
<a href="#">vento</a>	0,0036	49803	117	4.3878	<a href="#">botte</a>	0,0043	7613	31	0.2824
<a href="#">impulso</a>	0,0036	19318	63	1.702	<a href="#">emicrania</a>	0,0043	644	16	0.0239
<a href="#">invito</a>	0,0032	66481	134	5.8572	<a href="#">ricotta</a>	0,0042	15348	46	0.5694

Figure 7: The first 10 collocations (lemmas) of the adjective *forte* in the Religion (left) and Cookery (right) domains.

Despite the versatility of the collocation extraction procedure and its implementation in linguistic applications, a basic knowledge of corpus querying techniques is required for correct usage. RIDIRE collocations across domains can also be extracted with the Sketches tool, which provides a more intuitive way to obtain linguistically relevant information. In other words, Sketches are more suitable for language learners that do not have high competence in corpus linguistics tools, as it provides them with an explicit language acquisition path.

A Sketch is a selection of relevant lemmas that co-occur with the key lemma in a specific syntactic pattern. The relevance of lemmas in each Sketch is determined by a lexical association measure (log-Dice in the RIDIRE implementation). Each Sketch corresponds to a precise grammatical relation<sup>1</sup>; for example, Figure 8 shows the *e\_o* Sketch for the adjective *forte* in all domains i.e. the first ten adjectives that co-occur with *forte*, linked to it by a copulative (*e*, “and”) or disjunctive (*o*, “or”) conjunction:

e_o	38731
<a href="#">deciso</a>	586 8,15
<a href="#">coraggioso</a>	381 7,73
<a href="#">chiaro</a>	1122 7,4
<a href="#">sano</a>	297 7,09
<a href="#">forte</a>	1148 6,92
<a href="#">debole</a>	313 6,87
<a href="#">potente</a>	266 6,71
<a href="#">incisivo</a>	159 6,61
<a href="#">competitivo</a>	200 6,46
<a href="#">intenso</a>	263 6,42
<a href="#">radicato</a>	118 6,41

Figure 8: Example of a Sketch.

1 RIDIRE Sketches (including both the lexical queries and the visualization layout) are realized with the rules of SketchEngine, that is considered the reference web application for corpus linguistics studies.

RIDIRE provides two extensions of the Sketch tool: Sketch Difference and Domain Sketch. The Sketch Difference tool shows the difference between the collocational behavior of two lemmas within the same syntactic pattern: we can see the words usable with the first lemma, with the second and with both of them.

In Figure 9 we see the difference between the Italian adjectives *forte* and *resistente* (“resistant”) in the Fashion domain; specifically, we select two important Sketches: *e\_o*, as in Figure 8, and *NofA*, which selects the nouns related to the adjective. From this example we can see that *forte* has a more varied usage in Fashion and is often related to the characterization of personality traits, while *resistente* is more specific and used for the technical specifications of clothing and accessories.

	NofA				e_o				
impatto	365	0	10,52	0	deciso	159	0	10,38	0
personalità	291	0	10,02	0	chiaro	59	0	8,31	0
legame	137	0	9,21	0	sicuro	38	0	8,14	0
tinta	169	0	9,18	0	indipendente	23	0	8,05	0
pezzo	254	0	9,15	0	determinato	17	0	7,93	0
carattere	122	0	8,87	0	sensuale	38	0	7,91	0
crescita	174	0	8,85	0	riconoscibile	16	0	7,7	0
identità	108	0	8,85	0	grintoso	19	0	7,67	0
appeal	73	0	8,36	0	sano	18	0	7,65	0
espansione	79	0	8,35	0	coraggioso	12	0	7,46	0
richiamo	71	0	8,26	0	emerso	10	0	7,44	0
presenza	99	0	8,24	0	vivace	19	0	7,37	0
colore	374	0	8,19	0	pratico	0	13	0	7,29
segnale	63	0	8,16	0	aerodinamico	0	3	0	7,31
emozione	70	0	8,12	0	protettivo	0	6	0	7,34
contrasto	94	0	8,09	0	funzionale	0	10	0	7,6
punto	152	0	7,96	0	duttile	0	4	0	7,76
messaggio	51	0	7,67	0	elastico	0	10	0	7,82
connotazione	39	0	7,61	0	leggero	0	56	0	7,94
impronta	40	0	7,57	0	comodo	0	22	0	7,94
contenuto	49	0	7,53	0	robusto	0	9	0	8,54
carica	38	0	7,43	0	capiente	0	13	0	8,63
polimero	0	2	0	7,73	impermeabile	0	24	0	8,96
guaina	0	3	0	8,14	flessibile	0	19	0	9,44
ultra	0	4	0	9,24	ultra	0	11	0	9,51

forte
-6.0
-4.0
-2.0
0.0
2.0
4.0
6.0
resistente

Figure 9: The Sketch Difference for the adjectives *forte* and *resistente* in the Fashion domain.





(O'Donovan & O'Neill 2008). In this respect, web corpora are particularly interesting, since the web can be nowadays considered as the main access to written language, both in comprehension and in production, for a large part of the population.

The dimension and the structure of the RIDIRE corpus make it particularly attractive for lexicographic purposes. For instance, its data have been explored by Carla Marellò for the study of Latin loanwords in Italian. The results showed that, in this respect, the corpus is richer than the modern dictionaries: all the Latinisms that are frequent in Italian monolingual dictionaries are frequent also in the corpus, but the corpus contains also various frequent Latinisms that are not reported in the dictionaries (but they probably should be).

The availability of very large corpora gave also a new perspective in the studies of collocations. Starting from these data, for example, it becomes possible to determine the input to which the learners are exposed while reading, and to select the collocations that should be considered during the compilation of monolingual and learner's dictionaries (Marellò 2013). The use of sketches, that are a sort of quick synopsis of the grammatical and collocational behavior of a word, makes available a wide range of usage pattern that should be considered during the dictionary creation process.

Moreover, Sketches are useful not only for the detection of collocations, but also to give a quick picture of the distinct meanings of a word, since different meanings often select different collocates (Kilgariff & Rundell 2002). It has to be noticed that the significance of this "extraction procedure" grows proportionally to the corpus dimension. If detecting meanings and collocations from very large corpora by means of concordance scanning could be very hard and time consuming, for the automatic collocation extraction procedures the bigger is the corpus, the better are the sketches (both in quantitative and in qualitative terms). Finally, the Sketch Differences tool is specifically interesting for comparing a word with its (near) synonyms and antonyms, in a pure lexicographic perspective.

## 5 Conclusions

Large scale corpora representing a language's domain of usage offer a unique source of data to both learners and lexicographers in accessing information about how the language is actually used. The computational tools now available, including those for web based infrastructures, allow the selection of the relevant information in a simple manner, overcoming significant difficulties encountered by corpus linguistics in meeting second language acquisition needs. Learners, teachers, and lexicographers, however, must be aware of the information required for a proper language acquisition that are up to usage conventions. On the basis of this understanding, corpus querying can be used to solve specific problems and be accepted as a modern method for use in the language acquisition process and in the dictionary creation.



## 6 References

- Alchemy API. Accessed at: <http://www.alchemyapi.com/> [06/04/2014].
- Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. In *Language Resources and Evaluation*, 43(3), pp. 209-226.
- Conrad, S. (2006). Challenges for English Corpus Linguistics in Second Language Acquisition Research. In Y. Kawaguchi, S. Zaima, T. Tackagaki, Y. Tsuruga, M. Usami (eds) *Linguistics Informatics and Spoken Language Corpora*. Amsterdam/Philadelphia: John Benjamins.
- CQPweb. Accessed at: <http://cwb.sourceforge.net/cqpweb.php> [06/04/2014].
- Heritrix. Accessed at: <http://crawler.archive.org/> [06/04/2014].
- Kilgarriff, A. (2009). Corpora in the classroom without scaring the students. In *Proceedings of 18th Internat. Symposium on English Teaching, Taipei*. Accessed at: <http://www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc> [06/04/2014].
- Kilgarriff, A. (2013). Using corpora as data sources for dictionaries. In H. Jackson (ed.), *The Bloomsbury Companion to Lexicography*. London: Bloomsbury, pp. 77-96.
- Kilgarriff, A., Greffentette, G. (2003). Introduction to the Special Issue on Web as Corpus. In *Computational Linguistics*, 29(3), pp. 1-15.
- Kilgarriff, A., Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications: A Case Study. In A. Braasch, C. Povlsen (eds), *Proceeding of the Tenth Euralex Conference, Copenhagen, 13-17 August 2002*. Copenhagen: University of Copenhagen, pp. 807-818.
- Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds) *Proceeding of the Eleventh Euralex Conference, Lorient (France), 6-10 July 2004*. Lorient: Université de Bretagne-Sud, pp. 105-116.
- Marello, C. (2013). Sembra che e subordinate soggettive. Primi sondaggi in italiano L2 scritto. In F. Geymonat (ed.) *Linguistica applicata con stile. In traccia di Bice Mortara Garavelli*. Alessandria: Edizioni dell'Orso, pp. 79-94.
- Moneglia, M., Paladini, S. (2010). Le risorse di rete dell'italiano. Presentazione del progetto "RIDIRE.it". In E. Cresti, I. Korzen (eds) *Language, Cognition and Identity*. Firenze: Firenze University Press, pp. 111-128.
- O'Donovan, R., O'Neill, M. (2008). A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In E. Bernal, J. DeCesaris (eds) *Proceeding of the Thirteenth Euralex Conference, Barcelona, 15-19 July 2008*. Barcelona: Universitat Pompeu Fabra, pp. 571-579.
- Panunzi, A., Fabbri, M., Moneglia, M., Gregori, L., Paladini, S. (2012). RIDIRE-CPI: an Open Source Crawling and Processing Infrastructure for Supervised Web-Corpora Building. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds) *Proceedings of Eighth Language Resources and Evaluation Conference (LREC 2012), Istanbul, 23-25 May 2012*. Paris: ELRA, pp. 2274-2279.
- Readability. Accessed at: <http://www.readability.com/> [06/04/2014].
- RIDIRE Corpus Online. Accessed at: <http://www.ridire.it> [06/04/2014].
- RIDIRE-CPI. <https://github.com/lablita/ridire-cpi> [06/04/2014].
- Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni, S. Bernardini (eds), *Wacky! Working papers on the Web as Corpus*. Bologna: Gedit, pp. 63-98.
- Sinclair, J. (ed.) 2004. *How to use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.
- Sketch Engine. Accessed at: <http://www.sketchengine.co.uk/> [06/04/2014].
- TreeTagger. Accessed at: <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/> [06/04/2014].
- WaCky. Accessed at: <http://wacky.sslmit.unibo.it/doku.php> [06/04/2014].

## **Acknowledgments**

The RIDIRE Project is funded by MIUR - FIRB 2007 and is promoted and maintained by SILFI (Società Internazionale di Linguistica e Filologia Italiana). The web application RIDIRE-CPI was developed by LABLITA and the corpus creation involved six Italian university departments: University of Florence (Dip. Italianistica and Dip. Sistemi e Informatica), University of Turin (Dip. Scienze Letterarie e Filologiche), University of Siena (Dip. Studi Aziendali e Sociali), University of Rome - Roma 3 (Dip. Italianistica), University of Naples - Federico II (Dip. Filologia Moderna).

# Empirical Approaches to German Paronyms

Petra Storjohann, Ulrich Schnörch  
Institut für Deutsche Sprache Mannheim  
storjohann@ids-mannheim.de, schnoerch@ids-mannheim.de

## Abstract

German lexical items with similar or related morphological roots and similar meaning potential are easily confused by native speakers and language learners. These include so-called paronyms such as *effektiv/effizient*, *sensitive/sensibel*, *formell/formal/förmlich*. Although these are generally not regarded as synonyms, empirical studies suggest that in some cases items of a paronym set have undergone meaning change and developed synonymous notions. In other cases, they remain similar in meaning, but show subtle differences in definition and restrictions of usage. Whereas the treatment of synonyms has received attention from corpus-linguists (cf. Partington 1998; Taylor 2003), the subject of paronyms has not been revisited with empirical, data-driven methods neither in terms of semantic theory nor in terms of practical lexicography. As a consequence, we also need to search for suitable corpus methods for detailed semantic investigation. Lexicographically, some German paronyms have been documented in printed dictionaries (e.g. Müller 1973; Pollmann & Wolk 2010). However, there is no corpus-assisted reference guide describing paronyms empirically and enabling readers to find the correct contemporary usage. Therefore, solutions to some lexicographic challenges are required.

**Keywords:** paronyms; synonyms; easily confused words; collocation profile

## 1 Introduction

This paper presents a new lexicographic project studying easily confusable words in language use and employing a data-driven approach to the investigation of German paronyms. Although there is a large spectrum of definitions, paronyms are generally referred to as lexical items with both related or similar morphological roots as well as slight morphological difference such as suffixes. But they are not only linked to another by similarity of form and/or sound but also have a similar semantic potential and are hence commonly confused for one another. Alternatively, it is one of the items of a paronymic pair that is commonly misused both by native speakers and learners, respectively. Examples of such items include for example *effektiv/effizient* (*effective/efficient*), *sensitive/sensibel* (*sensible/sensitive*), *Method/Methodik/Methodologie* (*method/methodology*), *formell/formal/förmlich* (*formal*). They have generally not been regarded as synonyms (cf. Lăzărescu 1995, 1999). However, first empirical studies suggest that in a number of cases items of a paronym set have undergone meaning change and developed synonymous notions (e.g. Storjohann 2013). It is therefore argued here that a sharp distinction between paro-

nymy and synonymy is not always justified. In other cases, they remain similar in meaning, but can show restrictions of usage and subtle differences in definition.

Whereas the treatment of synonyms has received attention from corpus-linguists (cf. Partington 1998, Taylor 2003), the subject of paronymy has not been revisited with empirical, data-driven methods neither in terms of semantic theory nor in terms of practical lexicography. Lexicographically, some German paronyms have been documented in printed dictionaries (Müller 1973; Pollmann & Wolk 2010), although not systematically. However, there is no corpus-guided reference guide describing paronym sets empirically enabling readers to find the correct usage of such lexical items. So overall, paronyms should to be addressed from new perspectives. Firstly, the phenomenon has not been accounted for comprehensively in linguistic theory. Secondly, from a corpus linguistic view, we need to search for suitable corpus methods for detailed semantic investigation. Vachková & Belica (2009) propose the comparison of collocation profiles. They suggest a data-driven method to analyse lexical usage of near-synonyms with self-organising feature maps (SOMs). In this paper, we argue that this might prove a suitable method for the treatment of paronyms too, as it provides insight into both, semantic overlap and differences and it provides instant access to contrastive patterns by examining the concrete collocational behaviour of two items under investigation. Finally, solutions to some lexicographic challenges are required.

## 2 Linguistic Treatment of Paronyms

As Hausmann (1990) points out the subject of paronymy has mainly been approached linguistically from typological, language contrastive perspectives, particularly in the area of translation studies. These focus on this lexical relation exclusively from a language learner's view. Depending on different parameters, paronyms have been defined differently, covering items such as false friends, homophones, homographs or even cognates (cf. Bußmann 2002). In this paper, paronymy is broadly understood as a lexical relation between two or more items within one language, which are semantically related, have a similar or identical root and which are similar in form and sound but show a slight morphological difference (see morphemic paronyms/paronyms proper in Bolshakov & Gelbukh 2003: 199). Some of the most stringent terminology and a classification model have been proposed by Lăzărescu (1995, 1999) treating paronyms exclusively from a L2-learner's perspective.

Generally, Lăzărescu (1995) distinguishes between phonetically- and orthographically-based lexical confusion (e.g. *Föhn/Fön*), lexicologically-related terms (e.g. *Schiffahrt/Schiffsfahrt*, *Kindbett/Kinderbett*) and grammatically-based items such as *wohnen/bewohnen* or *dort/dortig*. Another class of paronyms exhibits specific semantic-stylistic features, e.g. as between *Vatermord/Vatermörder*, *Etikett/Etikette*). All these share their potential to be commonly confused in context and then to cause confusion in text reception or production due to similar phonemic representation as well as similarities of form and meaning. Most of these examples are far less likely to be mixed-up by native speakers. The importan-

ce of paronyms is based on the assumption that these items play a vital role for users in the process of second language acquisition and foreign language communication in order to avoid misunderstandings (Làzàrescu 1999). Confusing paronyms is sometimes regarded a violation of semantic correctness. Prescriptive analysts favour semantic correction and the avoidance of such mishaps and argue that paronyms “are important for poorly educated native speakers and for foreigners” (Bolshakov & Gelbukh 2003: 199). Indeed, the alleged misuse of morphologically and semantically similar words also cause linguistic uncertainties for native speakers, as numerous language-related Internet blogs show.

Language learners and native speakers too share their concern of correct language use with the language community as demonstrated by threads and their opening questions such as “Was ist der genaue Unterschied zwischen *effektiv* und *effizient*?” (What is the exact difference between *effektiv* and *effizient*), (see for instance <http://www.gutefrage.net/frage/was-ist-der-genaue-unterschied-zwischen-effektiv-und-effizient>). The answers of the blog community are impressively diverse.

So far, there is no semantic account encompassing different perspectives on the phenomenon and no satisfactory lexical focus on this relation that goes beyond Làzàrescu’s categorisation system and that comprises semantic, diachronic and cognitive aspects. Effectively, there is a large interest in easily confused words from both learners and native users alike, but we lack an empirical treatment and full theoretical account of paronymy in general. So far, the question of what constitutes a relation of paronymy has not been satisfactorily been answered as besides lexical features it also involves cognitive aspects. Furthermore, we have no suitable, user-friendly, appropriate dictionaries documenting paronym behavior (see section 4). Hence, there are no widely tested methods that have proved suitable for semantic analysis of such words. To be able to derive conclusion and to develop hypotheses, it is suggested to work with corpus-driven procedures to examine paronyms closer. With the help of corpora and innovative tools it is possible revisit paronymy and also to open up new research questions. Since it is possible to analyse language use synchronically as well as diachronically and observe gradual meaning change even over a short period of time it is possible to detect slight semantic shifts or nuances and also to determine the degree of semantic overlap between similar lexical items both quantitatively and qualitatively. It is argued (see also Storjohann 2013) that in some cases there is no semantic violation when paronyms are being confused. The meanings of typically confused words are more freely exposed to semantic negotiation. Following a descriptive empirical view, the semantics of some paronymic lexical items have adopted new semantic aspects and undergone meaning changes that are observable as regular patterns in a corpus and not as single misused occurrences. Overall, corpus-driven research on paronymy demands a more differentiated look at the phenomenon than has previously been offered. Empirically-based investigations of paronyms can also provide valuable insights into cognitive aspects and the exact circumstances under which two items are being confused as well as possible principles of language change of conceptually associated terms.

### 3 Corpus-Linguistic Approaches to Paronyms

Currently, researchers face a range of techniques that can be incorporated into the analysis of texts, such as eliciting co-occurrences, extracting keyword lists, investigating concordance and analysing dispersion and frequency of words or patterns. Methodologically, it is advantageous to use corpus tools that are able to provide good access to patterns and structures of lexical use by exploring co-occurrences. Exploring collocational patterns and other syntagmatic patterns has become an established procedure in order to describe the contextual behaviour of a word in empirical lexical semantics and in lexicography. It has also become an established tradition within corpus-linguistics to lead researchers more towards difference-driven analysis (cf. Taylor 2013). However, the examination of paronym sets necessarily incorporates contrastive meaning analyses including the study of similarities too. Therefore, suitable methods should be capable of measuring semantic similarity or distance by contrasting collocation profiles pairwise to systematically uncover overlap and differences in terms of contextual behaviour.

Storjohann (2013) conducted a contrastive analysis of paronymic items on two sets of data. A semantic study of *effizient/effektiv* has been carried out on the basis of a large newspaper corpus, mainly by investigating their collocations. The data used in the analysis consists of 2.7 billion words (cf. Storjohann 2005) and it has been the basis of the lexicographic project *ellexiko* ([www.owid.de/wb/ellexiko/glossar/ellexiko-Korpus.html](http://www.owid.de/wb/ellexiko/glossar/ellexiko-Korpus.html)) at the Instiut für Deutsche Sprache. In addition, self-organising feature maps (SOMs) (cf. Kohonen 1990; Keibel & Belica 2007) have been employed which offer visual representation of topographic profiles of the involved lexical items and which complement the collocation analysis. This procedure is implemented into a corpus-linguistic research and development workbench called CCDB (Keibel & Belica 2007). The CCDB is a database containing numerous static lexical co-occurrence profiles. It has been used to extract topographic profiles to break down unstructured collocation patterns and hence complex semantic properties (see Figure 1 and 2). On the basis of collocation profiles, semantic structures are analysed, clustered, visualized in a two-dimensional lattice reflecting different degrees of similarity between various words. As emphasized before, the comparison of paronyms implies the analysis of difference as much as the analysis of overlap. These are systematically being identified between items with overlapping collocation profiles (cf. Vachková & Belica 2009). Self-organising features maps cluster all those items such that proximity of the grid reflects semantic similarity between their semantic profiles. The more similar the colours of two neighbouring groups, the more similar are their collocation profiles although a strict separation is not suggested, as SOMs imply a continuum of semantic shades. The more their colours differ the more semantic differences can be found with regard to their uses (see Figure 1 and 2).



© Cyril Belica: Modelling Semantic Proximity - Self-Organizing Map (version: 0.32, init tau: 0.04, dist: u, iter: 10000)

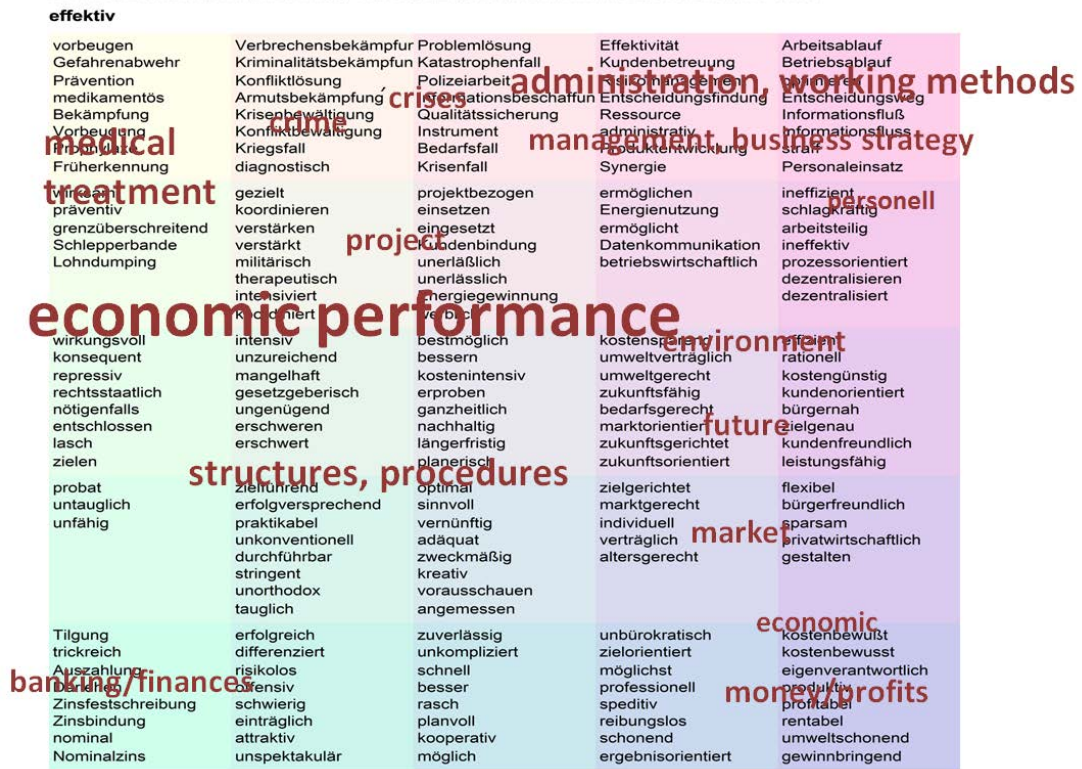


Figure 3: Topographic profiles of German *effektiv*.

© Cyril Belica: Modelling Semantic Proximity - Self-Organizing Map (version: 0.32, init tau: 0.04, dist: u, iter: 10000)

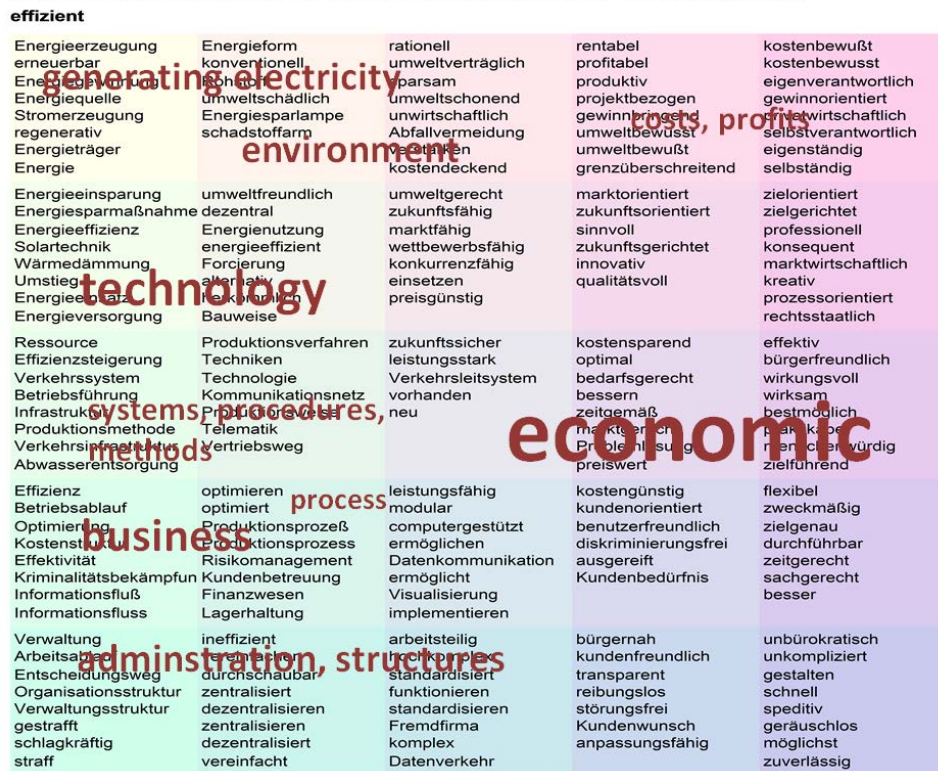


Figure 4: Topographic profiles of German *effizient*.

It is argued here, that the interpretation of such topographic feature maps could be the main entry point into data analysis of paronym behavior. It is a useful device for directing researchers to salient thematic domains associated with the individual terms. Through the process of abstraction a mental associations can be created by looking at individual squares and by moving from one square to the next (see Figure 1 and 2). As Vachková & Belica suggest:

Moving your focus forth and back, try to visualize the boundary where the initial sign eventually faded out, and where a notion of a new supersign entered your mind. Repeat for all corners and all directions. Try to assign each SOM square to at least one SOM supersign. (2009: 228-229)

Hence, a more abstract “supersigns” or superordinate concepts can be derived to categorise key semantic fields, clusters or domains. One major advantage of this procedure is the quick detection of thematic topics or domains in which the lexical items predominantly occur. For both lexical items, these thematic supersigns have been assigned and marked typographically. As a result, the interpretation of the findings as derived from supersigns or general concepts can then be directly compared, as summarised for example in Table 1.

<b>effizient</b>	<b>effektiv</b>
systems/procedures/structures administration costs/projects/economy	business/methods/work/management/personnel economic structures/performance
environment/generating electricity	fighting crime crisis management
technology	medical treatment/therapy
	environment/future
	banking & finance

**Table 3: Semantic contexts/domains.**

The domains in Table 1 are arranged according to their dominance in the underlying corpus. Semantic overlap between *effizient* and *effektiv* can be found in thematic domains where supersigns refer to the notion of business and administration and where both terms characterise methods, structures, procedures and issues of management. These are the discourse areas where ample evidence of synonymous contexts is provided in written German corpora (see examples 1 -3, for more examples see Storjohann 2013).

1. Zwar besteht auch in Hamburg die Vorschrift, dass ein Toter höchstens 36 Stunden in privaten Räumlichkeiten liegen darf und nach 14 Tagen begraben sein muss. „Aber Fakt ist, dass es eine weitere gesetzliche Reglementierung gar nicht gibt“, sagt Hillermann. Der sture Ablauf, der fast immer eingehalten wird, sei vielmehr das Ergebnis *effektiver* Arbeitsteilung von Spitälern, Bestattern und den Friedhofsbetreibern. Fast alle größeren Bestattungsunternehmen pflegen in



Deutschland eine *effiziente* Arbeitsteilung. (Die Zeit, 15.04.2004; Wie man in Deutschland begraben wird)

2. Im Jahr 2000 haben sich drei Fünftel der heimischen Gastbetriebe vom Dehoga abgespalten. [...] Der Hoga Rheinland, also der Hotel- und Gaststättenverband im nördlichen Rheinland-Pfalz, ist seinerzeit ausgetreten, weil er die Landes- und Bundesverbandsarbeit für nicht *effektiv* genug gehalten hat und die nicht üppig vorhandenen finanziellen Mittel lieber in regionale Arbeit und Projekte stecken wollte. Die Nordlichter warfen dem Dehoga zu wenig *effiziente* Arbeit vor. Wie sehen Sie das? (Rhein-Zeitung, 05.11.2004; Gastgeber wieder unter einem Dach)

Differences can be found for *effizient* being used as an attribute to characterise equipment, instruments and technological developments as well as being used to specify types and the use of energy resources. *Effektiv*, on the other hand, is more likely associated with the contextual domain of crime and crisis management. It is also attributed to characterise medical treatment and means of saving the climate. Furthermore, it is used within the context of banking and financing. Although, total semantic exclusion is not suggested, within these contexts both terms are far less likely being synonymously. As a result, the self-organising feature maps help us to structure the “unordered” semantics of a lexical item in use. It provides us with necessary details such as semantic dominance and contextual preferences in terms of referential domains and discursive foci.

In a further step, self-organisation maps can be used to contrast patterns of usage between two lexical items such as *effektiv* and *effizient* by comparing them with all those items with which they share parts of their collocation profiles.<sup>1</sup> This means that it is capable of measuring semantic similarity or distance by contrasting typical contextual behavior pairwise. The procedure is referred to as CNS-model (Contrasting Near-Synonyms, cf. Belica 2001 ff) and it is implemented into the workbench CCDB too. This procedure allows linguists to compare and contrast two words visually according to salient collocational contexts. *Effektiv* and *effizient* are being contrasted with each other and with those items with which they share parts of their co-occurrence profiles. The feature maps arrange specific usage aspects that the items in question share and those they do not have in common (see Figure 3).

---

1 Compare for example corpus software which facilitate the comparison of collocates visually, e.g. Sketch-Engine's Word Sketch Difference (<https://www.sketchengine.co.uk/>)

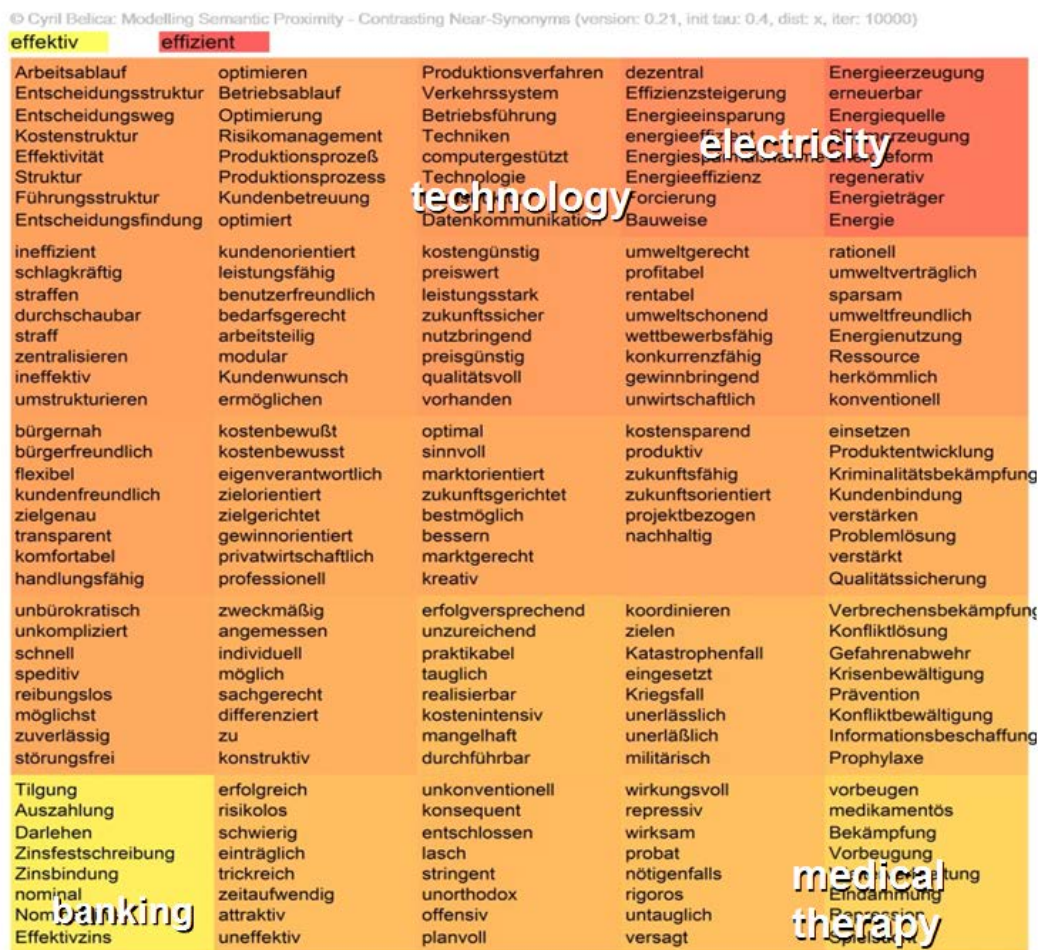


Figure 3: Contrasting German *effektiv* and *effizient* with SOM.

As indicated, the referential domains of ‘banking and financing’ (bottom left) and ‘medical treatment’ (bottom right) are preferably assigned to *effektiv*. Ways of generating electricity/energy and defining or specifying technology, however, are more likely a subject where the term *effizient* is preferably attested. In Figure 3, the subjective interpretation of the domains has been added to the map. Large parts of this feature map are shared by both terms. For the purpose being, these have not been further analysed also because a number of referential domains cannot always be clearly determined. Nevertheless, this feature map helps to validate the findings of conceptual similarities and differences as summarised in Table 1.

Generally, feature maps cannot serve as detailed lexicographic documentations to help users to be aware of “appropriate” or “false” usage. With respect to a paronym dictionary, more information as to concrete contextual usage is necessary. In the next stage, these domains or themes could, for example, be exemplified through statistically significant collocates extracted from an underlying corpus and assigned to corresponding discourses in order to illustrate specific preferences and restrictions (see Table 2). This process reveals for example, that one can modify procedures as processes (*Abläufe*, *Arbeitsabläufe*, *Arbeitsweise*, *Betriebsabläufe*), solutions (*Lösung*), structures (*Strukturen*), systems (*System*) and means/instruments (*Maßnahmen*) both as *effizient* and *effektiv* without implying much of a difference. It is within these contextual domains that similarity between the two terms is most evident

and examples of synonymous usage are being attested in the corpus (compare corpus examples 1 and 2). One does not typically characterise technology or the use of electricity/energy as *effektiv*. Here, the adjective *effizient* is being preferred. There are also efficient combustion engines (*Verbrennungsmotoren*) and efficient power stations (*Kraftwerke*). Other collocates on the topic of health and medical treatment include for example *therapy* (*Behandlungsmethoden, Therapie*) and *exercises for your back* (*Rückenübungen*) usually associated with *effective* but not with *efficient*. And terms referring to the banking and finance sector could be *interests, rates, return on capital* etc. These are modified in German by using *effektiv* (meaning real) exclusively, but never by *effizient*.

<b>effizient</b>		<b>effektiv</b>	
collocate	discourse domain	collocate	discourse domain
Abläufe, Betriebsabläufe, Arbeit, Arbeitsabläufe, Einsatz, Lösung, Strukturen, Verwaltungen, Maßnahmen, Arbeitsweise, Bewirtschaftung, Organisation, Wirtschaften, arbeiten, Aufgabenerfüllung, Bewirtschaftung, Methode, Abwicklung ...	systems procedures structures administration costs projects economy	arbeiten, gestalten, Maßnahme, Werbung, Methode, Maßnahme, Lösung, Strukturen, System, Arbeit, Verwaltung, Kommunikation, Kontrolle, Organisation, Arbeitsabläufe, Personaleinsatz, Zeitmanagement, organisieren, Controlling, Techniken, Strukturen ...	business methods work management personnel economic structures performance
Energie, Energieeinsatz, Energienutzung, Energieverwendung, Energieversorgung, Energieverbrauch, Heizen, Heizsysteme, Kohlekraftwerke, Stromnutzung, Stromerzeugung ...	environment generating electricity	bekämpfen, Mittel, Krisenmanagement, Strafverfolgung, schützen, Rechtsschutz, Selbstverteidigungstechniken, Verbrechensbekämpfung, Dopingbekämpfung, Polizeiarbeit, Überwachung ...	fighting crime crisis management
Motoren, Verbrennungsmotoren, Anlagen, Antriebstechnologie, Heizungen, Wärmepumpen, Kraftwerke, Geräte, Technologien, Wärmedämmung, ...	technology	Hochwasserschutz, Ressource, Klimaschutz, Wärmedämmung, Reduzierung	environment future
		Behandlungsmethoden, Ganzkörpertraining, Therapie, Rückenübungen, Prävention, Behandlung, Workout, Schmerztherapie, Beinmuskultraining, Behandlung ...	medical treatment therapy
		Jahreszins, Rendite, Zins, Zinssatz, Nominalzins	banking & finance

**Table 4: Collocates in corresponding discursive domains.**

As suggested by Vachková & Belica (2009), this approach to collocational patterning might be applicable to the lexicographic investigation of synonyms. It is argued that salient SOM features stimulate lexicographers' associative awareness and encourage guided mental imagery leading to valuable insights into both the word semantic structure and the process of discourse-based negotiation of lexical meaning (Vachková & Belica 2009: 239).

The notion of similarity has played a great role in lexicological areas, for instance in the corpus-linguistic investigation of sense relations by using collocational overlap to measure degrees of synonymy (cf. Partington 1998). Marková (2012), for example, puts forward examples of studies of German synonyms where she employed the CNS-model successfully. It is proposed here that consultations and interpretations of self-organisation feature maps might be a suitable approach to the analysis and semantic description of paronyms too where usage aspects that are shared and not shared can be uncovered. Feature maps can guide lexicographers to those contextual patterns where to look for further evidence for example through the analysis of collocations that can be attributed to specific thematic domains. Effectively, the chosen procedures result in a form of methodological triangulation comprising three different analytical stages: first, interpreting SOMs in order to associate domains, secondly, the CNS-Model to validate the previous interpretation and thirdly, collocation analysis to exemplify the given domains/topics contextually.

## 4 The Lexicographical Treatment

With regard to German, commonly confused words including some paronyms have been described in two printed reference books: Müller (1973) and Pollmann & Wolk (2010). Both are prescriptive documentations aiming at guiding users to the allegedly correct usage and describing a clear distinction between the items in question (see for example Figure 4).

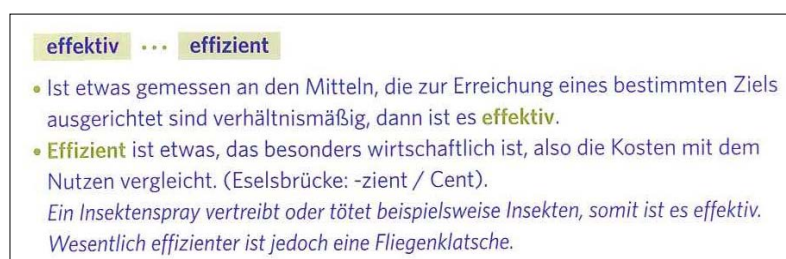
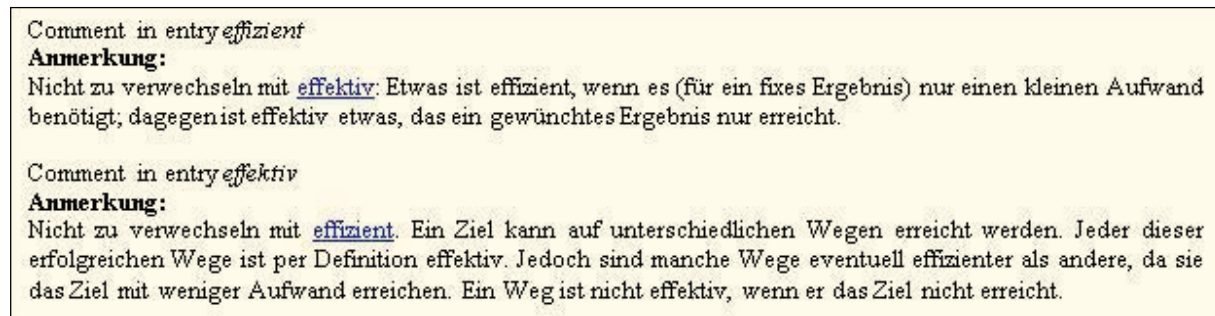


Figure 4: Dictionary entry *effektiv/effizient* in Pollmann & Wolk (2010).

The entries contain short meaning descriptions, occasional encyclopaedic comments and citations or examples. Moreover, some normative grammars and lexical studies concerning the didactics of normative language practice contain lists of some paronyms (e.g. Heringer 1989, 1995). Strictly normative language use is also propagated in wiktionary, a popular electronic resource which under an explicit

headline points out that confusion over the two words *effektiv* and *effizient* should be avoided (see Figure 5).



**Figure 5: Comments regarding the misuse of *effektiv* and *effizient* in wiktionary.**

From a lexicological point of view the remarks found there are questionable. However, they demonstrate that language users are aware of a potential conflict between *effektiv* and *effizient* and that misuse or confusion of this kind is a rather regular phenomenon that is to be avoided. Overall, in all reference guides findings are neither based on semantic examinations of current natural language in use nor on investigations of large data. Empirical corpus explorations open up the discrepancies to traditional descriptions. The usage restrictions that are documented in these reference books cannot be confirmed through corpus data. Entries lack collocational details referring to recurrent referential domains as for example illustrated in Figures 1 and 2. These provide essential information to users as to in which concrete contexts the corresponding adjectives might be more commonly found. As is the case for *effektiv/effizient*, strict usage lines cannot be sharply drawn which might have been expected intuitively. Conventional reference guides have so far focussed on the differences between commonly confused words. They entirely fail to explain existing similarities. In that respect the methodology is similar to most corpus-linguistic research. Corpus-assisted studies on semantically similar words have so far focussed on the differences between the individual items. As Pearce (2008: 21) points out there is a risk of “the privileging of differences over similarities. [...] the analyst is in danger of exaggerating the differences and overlooking similarities”. However, corpus studies also allow for the description of similarities which, on the one hand, might offer a deeper understanding why two words are regularly being confused and, on the other hand, it might indicate ongoing linguistic change worth documenting.

## 4.1 Challenges

A new project at the Institut für Deutsche Sprache will reopen the chapter on paronymy as a lexical as well as cognitive phenomenon. It will account for paronymy from a corpus-linguistic perspective and it will test methods that will hopefully prove suitable for semantic analysis of such words. It is hoped that data-driven investigations of paronyms can provide valuable insights into principles of



language change in semantically related lexical items. This could enable us to integrate paronymy into a wider theoretical framework. Part of this project is the compilation of the first corpus-assisted paronym dictionary which aims at guiding users descriptively through current usage and contexts. As an electronic resource it will provide adequate pairwise documentations combined with user-friendly navigation and search structures. In the near future, it will be an integral part of the German dictionary portal OWID (see: <http://www.owid.de/>).

From a lexicographic point of view, a number of challenges are encountered when documenting usage-based findings in a paronyms dictionary where users might demand definite answers for doubtful language situations. One central problem regards the interpretation and documentation of language change and normative restrictions. This is particularly relevant for pairs that are recorded as semantically distinct lexical items in traditional reference works and that have assimilated semantically over time due to common, allegedly “false” use. In some cases, corpus analyses signal tendencies that paronyms might have possibly turned into synonyms. Therefore, one of the major challenges of a corpus-based paronym dictionary is the interpretation of ambiguous data, especially paronym usage with a similar proportion between contexts with clear semantic difference between the terms and contexts exhibiting synonymous use. The lexicographic interpretation of such data requires a certain sensibility, as a specific conflict is expected to be encountered with corpus data. On the one hand, false language use caused by confusing paronyms needs prescriptive correction. On the other hand, gradual language change caused by frequent misuse of a certain lexical item needs descriptive documentation of contemporary language use. Above all, it should be able to explain semantic overlap and to sensitise users for continuous language change. Although this has not been studied, it is assumed that the expectations of users of such a dictionary have rather prescriptive notions due to their habits and their handling with other existing dictionaries. Therefore, the most challenging objective of this descriptive, usage-based dictionary certainly is to offer a reference guide that shows similarities and difference between paronymic items contrastively, including corpus samples, explanations and comments without neglecting to inform users about aspects of semantic overlap, gradual semantic changes, contextual vagueness, possible substitutability and at the same time still answering their look-up questions satisfactorily.

## 4.2 Presentation

As an e-dictionary the new German paronym guide can go far beyond the depth of information found in the two existing printed dictionaries. It will also have to consider different options with respect to navigation, visualisation, cross-referencing, linking and searching in order to exploit the possibilities of the electronic medium profitably and in order to create a user-friendly instrument. Traditional dictionary entries contain explanations of the formula “*effizient* is something that ...” or alternatively “if something is *effizient* than ...” (see Figure 4). The answer as to what exactly this something is can be found in contexts and collocates. Although at this point, no finite solution of the details of presentati-

on can be given, it is inarguably information on co-occurring patterns that most attention will be drawn to. Users can expect direct access to collocational patterns of two or more easily confused words together at the same time with their interpreted thematic domains in which both are likely to occur together with their preferences or their restrictions. Corpus samples will illustrate the information given. The depth of information could be realised as optional user-customised views. Overall, two aspects need further exploring. Firstly, the possibilities of the electronic medium still need to be examined and exploited to create a reliable and usable environment enabling users to make correct choices. Secondly, research on the users' needs and behaviour provides us with valuable insights (cf. Tarp 2008; Müller-Spitzer 2014) and these should be profitably incorporated into modes of presentation.

## 5 Conclusion

In this paper, a multifaceted approach to the study of regularly confused words in German has been discussed, suggesting a method of investigation that implies both differences and similarities between paronymic items. The phenomenon of paronymy has not been accounted for empirically. It has neither been reconsidered in recent linguistic theories and models nor in lexicographic practice. This gap will be closed by a new project which accounts for paronymy from a corpus-linguistic perspective where methods of investigation will be tested to find suitable tools for semantic analysis. With the help of the example *effective/effizient*, we have set out one possible way to implement different software-driven resources facilitating the search for similarity and difference. Corpus-assisted investigations of easily confused words and their usage over recent decades can provide valuable insight into principles of semantic shift. It is argued here, that such analyses might enable semanticists to integrate the phenomenon into a wider theoretical framework on the one hand and into appropriate lexicographic descriptions on the other hand.

## 6 References

- Belica, C. (1995). Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode. Mannheim: Institut für Deutsche Sprache.
- Belica, C. (2001ff). *Kookkurrenzdatenbank CCDB – V3.3. Eine korpuslinguistische Denk- und Experimentierplattform*. Mannheim: Institut für Deutsche Sprache. Accessed at <http://corpora.ids-mannheim.de/ccdb/> [27/3/2014].
- Belica, C. (2006). Modellierung semantischer Nähe: Kontrastierung von nahen Synonymen. Korpusanalytische Methode. Mannheim: Institut für Deutsche Sprache.
- Belica, C. (2011). Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In A. Abel, R. Zanin (eds.) *Korpora in Lehre und Forschung*. Freie Universität Bozen: Bozen University Press, pp. 155-178.

- Bolshakov, I.A., Gelbukh, A. (2003). Paronyms for accelerated correction of semantic errors. In *International Journal Information Theories & Applications*, 10, pp. 198-204.
- Bußman, H. (2002). *Lexikon der Sprachwissenschaft*. 3<sup>rd</sup> edition, Stuttgart: Kröner Verlag.
- ellexiko*. Accessed at <http://www.owid.de/wb/ellexiko/start.html> [27/3/2014].
- ellexiko-corpora*. Accessed at <http://www.owid.de/wb/ellexiko/glossar/ellexiko-Korpus.html> [27/3/2014].
- Hausmann, F.J. (1990). Das Wörterbuch der Homonyme, Homophone und Paronyme. In F.J. Hausmann, O. Reichmann, H.E. Wiegand (eds.) *Wörterbücher. Dictionaries. Dictionnaires. Vol. 2*. Berlin/New York: de Gruyter, pp. 1120-1125.
- Heringer, H.J. (1989). Lesen – lehren – lernen. Eine rezeptive Grammatik des Deutschen. Studienausgabe. Tübingen: Niemeyer.
- Heringer, H.J. (1995). Grammatik und Stil. Praktische Grammatik des Deutschen. Berlin: Cornelsen.
- Keibel, H., Cyril, B. (2007). CCDB. A Corpus-Linguistic Research and Development Workbench. In *Proceedings of the 4th Corpus Linguistics Conference, CL 2007, 27-30 July 2007*. University of Birmingham, UK.
- Kohonen, T. (1990). The Self-Organizing Map. New Concepts in Computer Science. In *Proceedings of the IEEE*, Vol. 78 (9), pp. 1464-1480.
- Lăzărescu, I. (1995). Deutsche Paronyme. In *Grazer Linguistische Studien* 43, pp. 85-93.
- Marková, V. (2012). Synonyme unter dem Mikroskop. Eine korpuslinguistische Studie. Tübingen: Gunter Narr.
- Müller, W. (1973). *Leicht verwechselbare Wörter. Duden Taschenwörterbücher*, Vol. 17. Mannheim: Bibliographisches Institut.
- Müller-Spitzer, C. (ed.) (2014). *Using Online Dictionaries*. Berlin/New York: de Gruyter. (Lexicographica: Series Maior 145).
- OWID Online-Wortschatz-Informationssystem Deutsch*. Accessed at: <http://www.owid.de/> [27/3/2014].
- Partington, A. (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.
- Pearce, M. (2008). Investigating the Collocational Behaviour of Man and Woman in the BNC using Sketch Engine. In *Corpora* 3(1), pp. 1-29.
- Pollmann, Ch., Wolk, U. (2010). *Wörterbuch der verwechselten Wörter*. Stuttgart: Pons.
- SketchEngine*. Accessed at: <https://www.sketchengine.co.uk/> [27/3/2014].
- Storjohann, P. (2005). Das ellexiko-Korpus. Aufbau und Zusammensetzung. In U. Haß (ed.) *Grundfragen der elektronischen Lexikographie. ellexiko – das Online-Informationssystem zum deutschen Wortschatz*. (Schriften des Instituts für Deutsche Sprache 12). Berlin/New York: de Gruyter, pp. 55-70
- Storjohann, P. (2013). Korpuslinguistische und lexikografische Ansätze zur Beschreibung deutscher Paronyme. In H. Scheuringer, D. Sava (eds.) *Im Dienste des Wortes. Lexikologische und lexikografische Streifzüge*. Festschrift für Ioan Lăzărescu. (Reihe: Forschung zur deutschen Sprache in Mittel-, Ost- und Südosteuropa Vol. 3). Passau: Stutz, pp. 401-418.
- Vachková, M., Belica, C. (2009). Self-Organizing Lexical Feature Maps. Semiotic Interpretation and Possible Application in Lexicography. In *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis*, 13(2), pp. 223-260.
- Wiktionary. Das freie Wörterbuch*. Accessed at: <http://de.wiktionary.org> [27/3/2014].
- Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography. Tübingen: Niemeyer.
- Taylor, Ch. (2013). Searching for Similarities Using Corpus-assisted Discourse Studies. In *Corpora*, 8, pp. 81-113.
- Taylor, J. (2003). Near synonyms as coextensive categories: 'Tall' and 'high' revisited. In *Languages Science*, 25, pp. 263-284.



# Pragmatic Meaning in Lexicographical Description: Semantic Prosody on the Go

Mojca Šorli  
Trojina - Institute for Applied Slovene Studies  
mojca.sorli@trojina.si

## Abstract

The present paper focuses on ways in which the pragmatic (functional) meaning, known in corpus linguistics as semantic prosody, is treated in monolingual and also bilingual dictionaries. We have analysed a selection of lexicographical descriptions, as they are represented in the *Slovene Lexical Database* (SLD), comparing them to a number of English and Slovene lexical sources, and demonstrated how corpus-derived pragmatic meaning can become an integral part of dictionary definitions. This is particularly important for the treatment of phraseology and idiomatics, where greater involvement of pragmatics is demonstrated. A tentative typology was compiled for the purpose of this analysis in order to categorise lexical units according to their inner semantic-pragmatic relations, with a view to examining the implications for the monolingual dictionary treatment of individual items, as well as any potential strategies that could be applied, on the basis of the posited categories, to their translation. We have also pointed out the treatment of individual lexical units in the selected bilingual dictionaries.

**Keywords:** lexicographical description; lexical database; monolingual/bilingual dictionary; pragmatics; semantic prosody

## 1 Introduction

The present paper is based on some results obtained in the course of doctoral research into the ways in which the pragmatic (functional) meaning that arises from various contextual features, known in corpus linguistics as semantic prosody (Sinclair 1991, 1996; Louw 1993; Stubbs 1995, 2001; Partington 1998; Tognini-Bonelli 2001; Whitsitt 2005; Hunston 2007, etc.) can become an integral part of (monolingual and bilingual) dictionaries. We will attempt to demonstrate the value of the explicit description of pragmatic meaning, i.e., semantic prosody, as implemented in the Slovene Lexical Database (SLD),<sup>1</sup> while also presenting some conclusions based on the exploration of the possibilities of recording semantic prosody in a bilingual perspective. We posit various types of meaning that are codified in specific types of linguistic form or patterns of use. Of central concern to us is the meaning de-

---

1 Slovene Lexical Database (2008-2012): The project was co-financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia.

scribed as “peripheral” or “underspecified” (see Philip 2009). This can only be studied in context, as it is completely dependent on collocation and syntagmatic relations, and therefore cannot be attributed solely to a concrete word form. The basic pattern of language use is represented by collocation based on the distributional features of words, while a more abstract type of pattern is derived from “inter-collocational” generalisations, which include semantic prosody (Ellis et al. 2009: 89-90). In other words, we not only speak of lexicogrammatical patterns, but also of pragmatic patterns of language use. In the present paper, we adopt the view that the function of a dictionary should not be limited to presenting the “referential”, “denotative”, “cognitive”, “semantic”, “dictionary”, etc., meaning, but should contain a comprehensive description of inherent semantic features of words, as well as the pragmatic circumstances of their use. A number of successful (English) language learner’s dictionaries have been designed to take into account functional aspects of meaning. Although still lagging behind, bilingual dictionaries have also moved on from being mere “glossaries” expected to provide no more than “prototypical”, “systemic” or “cognitive” equivalents to not only corpus-based but “corpus-like” language resources, in which the user can explore words in real use. With space restrictions no longer in place, electronic lexicography now has the means and the opportunity to devote more attention to the textual and pragmatic dimensions of meaning, such as the complexities of semantic prosodies, which, as research shows (Hunston 2002; Zethsen 2006; Zhang 2009, etc.), contemporary monolingual lexical databases and dictionaries still fail to convey, typically implying them in examples of (typical) usage. Moreover, in bilingual dictionaries semantic prosody is typically ignored altogether.

## 2 Forms of Encoded Pragmatic Meaning: Semantic Prosody

Semantic prosody is an integral part of an (extended) unit of meaning, identifiable only by examining its repeated occurrences in a large amount of (corpus) data. For example, at first glance “situation” seems perfectly neutral, but examining a large number of contexts of situation shows that it typically occurs as the node of units of meaning in contexts conveying negative events, facts or features that evoke negative associations and carry negative semantic prosody. In the 112-million BNC reference corpus there are 19,569 hits (174,4 per million). Collocations are very dispersed, with some of them being seemingly neutral, e.g., *present, given, similar*, but the wider context reveals negative circumstances of meaning. The most frequent collocations are: *the situation is [complicated, worse, hopeless, conducive, analogous, desperate, unsatisfactory, unstable, confusing, confused, vacant, tense, grave, favourable, different]* and *[current, present, given, dangerous, similar, particular, intolerable, economic, stressful, difficult, ideal, financial, complex, deteriorating, worsening, etc.] situation*. In the first collocate set, we can identify 11, and in the second 5 out of 15 collocates that could be marked as “negative”, i.e., carrying negative implications and associations arising from the extended units of meaning. The concordance shows that prosody is neutral mainly in (semi)terminological contexts. Even though semantic preference and semantic prosody overlap to an extent in general

contexts, semantic prosody is usually about a particular “scenario” rather than merely a “preference” related to a semantic field. Amongst the verb collocates [*react, respond, adapt, adjust, correspond, apply, refer, relate, etc.*] stand out. These are again rather neutral at first glance, but are confirmed as predominantly negative upon the examination of the wider co-text and context:

- (1) *Clientelism is a strategy used by capitalists and workers to adapt to a **situation** where there is limited mobility.*  
 (2) *In parliament it was difficult to adjust to the new **situation**, whereby the party was supposed to abstain from all criticism of the government but had no say in its decisions.*

Of the verbs with “situation” as the prepositional object, the most typical are [*cope, deal, face, confront, compare, etc.*] *with the situation*; the first 6 collocates in the genitive relation, which indicate difficulty, are [*seriousness, gravity, urgency, reality, complexity, absurdity*] *of the situation*, while an even more explicit reference to concrete sociopolitical conditions is made by the first 6 telling collocates in the prepositional phrases *the situation in* [*Somalia, Yugoslavia, Gulf, Russia, Africa, Iraq*]. The research presented in the continuation is based on an examination of the possibilities of including pragmatic information in lexical-lexicographical descriptions from two perspectives: 1) as above, monolingually, identifying and recording pragmatic components in the context, and 2) exploring options for conveying pragmatic information, especially semantic prosody, in bilingual dictionaries. For this purpose, we have analysed a selection of meaning descriptions from the SLD and studied *a prima facie* translation equivalents of the relevant lexical units, drawn from the DANTE lexical database, the selected EFLs<sup>2</sup> and Collins English Dictionary (CED), which were then checked against a selection of bilingual (corpus-based) dictionaries.<sup>3</sup> The bilingual focus remains throughout on the Slovene-English rather than English-Slovene perspective. We did, however, consult two English-Slovene sources to check the degree of bidirectionality of the translations. A tentative typology was compiled in order to categorise lexical units according to their inner semantic-pragmatic relations, with a view to examining the implications for the monolingual dictionary treatment of individual items, as well as any potential strategies that could be applied, on the basis of the posited categories, to their translation. The schematic representation is not based on the structural relations between the components of the lexical units, but is driven by the semantic-pragmatic relations established by each category.

2 Monolingual learner’s dictionaries: COBUILD, MED, LDOCE and MWLD. See bibliography.

3 Bilingual Slovene-English dictionaries: Concise Slovenian-English Dictionary DZS (1<sup>st</sup> Ed.) (PSA), Slovenian-English Pocket Dictionary DZS (MSA), PONS Slovenian-English (& English-Slovenian) Dictionary, DZS-Oxford Comprehensive English-Slovenian Dictionary (VASS) and the automatically reversed VASS database with 120,000 entries, nicknamed OXZILLA DZS.

### 3 Pragmatic Analysis Based on the Different Situations of Meaning

#### 3.1 Meaning as an Inherent Markedness (at the Morphosyntactic Level): Connotation

Some words can be identified pragmatically as morphosyntactically non-neutral. In the case below, the SLD fails to convey the pragmatic meaning either with a label or in the description:

**Example 1:** *debeluh* – fatso; fatty

SLD: *debel* člověk (Eng.: a fat person)

(3) **Fatties** who lose a lot of weight talk about the need for mental adjustment. /.../<sup>4</sup>

The English sources label both lexical items, “fatty” and “fatso”, as informal or slang (DANTE and MWLD), also indicating their connotation with the label “insulting” or with the inclusion in the definition of “an insulting word for” (CED, LDOCE and MED).

DANTE: fatty: 1 n [**offens**] [**inf**] nickname or appellation or a fat person; fatso /

(4) He reviewed a gallery of the great **fatties** of all time, from Nero through Falstaff to Arbuckle.

(5) Go on **fatty**!

The appropriate semantic-pragmatic profile of this lexical unit can thus be created with the use of labels or, as in most sources, with a combination of the label and the definition. A key component of meaning is the speaker’s intent to insult, in the case of direct address, and, in the case of third person use, to establish a certain distance and/or to express disdain in relation to such a person. Below is a summary of monolingual and bilingual treatment of this lexical unit in the relevant dictionaries and lexical databases.

---

4 The examples taken from the SLD are translated into English as literally as possible to preserve the original meaning.

COBUILD	LDOCE	MED	MWLD	CED
<b>fatty</b>				
/	fatty [countable] informal. <b>an insulting word for</b> someone who is fat	noun [countable] informal. <b>an insulting word for</b> someone who is fat	[count] informal + <b>offensive</b> : a fat person	(informal) a fat person
<b>fatso</b>				
/	[countable] informal. <b>an insulting word for</b> someone who is fat	<b>an insulting word for</b> someone who is fat	[count] informal + <b>offensive</b> : a fat person	fatso: ( <b>slang</b> ) a fat person: used as an insulting or disparaging term of address
<b>bilingual</b>				
<b>PSA</b> <i>debeluh</i>	<b>MSA</b> <i>debeluh</i>	<b>OXZILLA</b> <i>debeluh</i>	<b>PONS</b> <i>debeluh</i>	<b>VASS</b> <i>fatty</i>
pejor. fatso, fatty	pejor. fatso, fatty	inf. Am. <b>fat-ass</b> ; inf. Br. <b>lard-ass</b> ; inf. pejor. <b>fatso, fatty</b> ; inf. Am. <b>blimp</b> ; [...]	fatty; fatso Am.	pog. žalj. baj(i), bajsja; <b>debeluh(ar)</b> , <i>debeluhinja</i>

Table 1: The treatment of *debeluh* in the selected monolingual and bilingual sources.

### 3.2 Meaning as a Matter of Attitude Towards a Pragmatic Situation

Another tentative, but lexicographically significant category has been posited that will not be treated in detail here due to a lack of space. In some situations of meaning, such as in *plezati* 'make one's way through/over/out of an uncomfortable position with effort', a set of circumstances has been identified that differs considerably in terms of prosodies from the main sense, thus calling for a separate (sub) sense:

**Example 2: plezati (čez/skozi kaj; iz česa)** – climb (over/through/from sth); clamber (over/across/into/out of)

SLD: če ČLOVEK pleza preko OVIRE, skozi ODPRTINO ali iz neudobnega POLOŽAJA, se skuša s pomočjo celega telesa premakniti v želeno smer, navadno s trudom ali težavo

(Eng.: if a HUMAN climbs over an OBSTACLE, through an OPENING, or from an uncomfortable POSITION s/he, using all limbs and her/his whole body, attempts to move in the desired direction, usually with great difficulty or with some effort)

(6) The studio personnel must sometimes **climb** over heaps of presents, nappies and various toys.

- (7) The track is in parts unsurpassable, we have to **climb** up on the dug up deviations.
- (8) At that point, the angry supporters at the west stand began to **climb** over the fence, while objects were being thrown to the area next to the play field where the referees were.
- (9) Thousands tried to come into the stadium **climbing** over the fences and closed gates.
- (10) Often they would have to **climb** through the windows, the side exits and run away from the girls through the kitchen.
- (11) The airbag had already emptied itself and released the pressure on my body. I slowly unfastened the safety belt and started **climbing** out of the vehicle – through the co-driver’s window!

In this case, with the subject typically human, the act of climbing is largely unplanned or undesired, bringing with it the semantic prosody of anguish or despair arising from the frustration at not being able to move faster or with greater ease. All of the examples of use above contain some element that indicates unfavourable circumstances surrounding the central event, accessible through the speaker’s attitude, which is not at all typical of climbing a ladder, a tree or a mountain; therefore, a separate sense is in place to capture adequately the identified pragmatic components.

COBUILD	LDOCE	MED	MWLD	CED
<b>climb</b>				
/	with difficulty [intransitive always + adverb/preposition] to move into, out of, or through something slowly and awkwardly	<i>[intransitive/transitive]</i> <i>to use your hands and feet</i> <i>to move up, over, down, or</i> <i>across something</i>	always followed by an adverb or preposition [no obj] : to move yourself in a way that usually involves going up or down	/
<b>bilingual</b>				
<b>PSA</b> <b>plezati</b>	MSA plezati	OXZILLA plezati	PONS plezati	VASS climb
(with difficulty) to clamber; to scramble	(with difficulty) to clamber; to scramble	clamber; scramble; climb	climb	splezati, (po) vzpeti se na; plezati po; vzpenjati se po

**Table 2: The treatment of *climb* in the selected monolingual and bilingual sources.**

### 3.3 Meaning as a Matter of Emphasis

The way meaning and use are in fact two distinct, yet inseparable, facets of language is illustrated by the analysis of the following example, which shows how meaning can arise from an emphasis on a

particular aspect of the (pragmatic) situation rather than from the word's inherent (semantic) features:

**Example 2: *bobnati (pri kom)*** – to drum with/in (=to be the drummer with a group)

The prototypical meaning defined by most (Slovene and English) sources as “to beat or play a drum, or a set of drums” (MWLD) or “to play a drum” (MED, LDOCE) (but not listed in COBUILD) is not the most frequent meaning of *bobnati*. On closer inspection and driven by a pragmatic function of meaning a subsense can be separated out, i.e., to play the drums in a particular band or group on a regular basis:

SLD\*: če ČLOVEK **bobna**, se posveča igranju na bobne kot stalni član glasbene zasedbe, navadno v določenem obdobju (Eng: if a HUMAN **drums** s/he is engaged in playing the drums as a regular member of a music group, usually for a period of time)

(12) The new drummer is Nenad Kostadinovski, who used to **drum** with groups such as Scuffy Dogs and Traffic Religion.

(13) Do you then have a drummer for your concerts? At the concerts we use Moreno Buttinar, who is Lara Baruca's drummer. We have also practiced with Janez, who **drums** with Miladojka Youneed.

(14) Meanwhile, Eva and Nataša sing and **drum** on Laibach's concert tour, while Darja has used the short period of lesser working intensity really well and has freshly fallen in love.

(15) Cecil Durkin was a druggie, a knife cutter and a prison rapist, but he also **drummed** in a few good jazz bands.

This use of the verb is distinct, as its colligational behaviour in particular indicates: it typically requires explicit temporal complementation (recently, at night, later, on tour, in the election time, sometimes, etc.) or the time of the action is implied in the context by the use of, typically, a past tense. It is commonly used with prepositional complementation (drum with, at, in, etc.) denoting individual people or groups with whom one drums; proper names, therefore, appear regularly in the co-text as part of listings and/or coordinate structures with “and”. Of course, the prototypical meaning “to play a drum” is still present, but it is now a secondary rather than the key component of the conveyed sense. The emphasis is on the fact of being engaged as member of a performing musical group, often in the context of other players who make up the group, i.e., on the role ensuing from the ability to play the drums. A colligational feature is that the verb cannot be pre- or postmodified, e.g., by an adverb. The semantic preference is for musical groups, players and settings, from which the association of an opportunity for success and fame emerges. In some cases semantic prosody, which seems to lie here first and foremost in “renewing the connection of this semantic information with the reality of language in use /.../,” (Philip 2009) relies more heavily on the colligates than the collocates alone, “if anything tending to favour the patterns and participants in verbal processes over lexical-semantic features per se” (ibid.).

COBUILD	LDOCE	MED	MWLD	CED
/	to play a drum	Music. to play a drum	to beat or play a drum or set of drums	to play (music) on or as if on a drum
<b>bilingual</b>				
<b>PSA</b> <b>bobnati</b>	MSA bobnati	OXZILLA bobnati	PONS bobnati	VASS to drum
(as a profession) to play the drums	(as a profession) to play the drums	to drum	to drum	bobnati

**Table 3: The treatment of *bobnati* in the selected monolingual and bilingual sources.**

### 3.4 Restricted Meaning in Semantically Analysable Units: Collocation

This is probably the most widespread, yet lexicographically somewhat neglected lexical category that lies at the very core of semantic prosody. Semantic analysability can be observed along a continuum stretching from collocations – commonly thought to be transparent, but so only in their restricted meanings, i.e., in only one of the possible meanings resulting from the various meanings of their components, as well as of their various combinations – to the most opaque idioms that lie at the other extreme (Philip 2009). The collocation below can be observed in its restricted meaning “dull weather”:

**Example 4: *kislo vreme*** – sour/grey/dull weather

SLD: *kislo VREME* je takrat, ko ni sonca ali dežuje (Eng: sour WEATHER is when it rains or the sky is overcast)

The only two collocates that stand out are *zagosti* and *pokvariti* (“to spoil”), paralleled in scarcity by collocational patterns. The semantic preference is for cultural and sports events, and, within a limited spectrum, for agricultural products, especially grapes which are expected to ripen and develop sugar in the sun. The semantic association of physical and mental discomfort caused by the weather conditions helps build up the semantic prosody based on imminent danger of poor turnouts at public events or people not going ahead with their outdoor plans, such as tourists cancelling their bookings: (16) A general characteristic is that the camping sites in Gorenjska are pretty full, and even the **sour weather** of recent days has not chased away the tourists.

(17) That is why the construction workers are working at a good pace, but, on the other hand, due to **sour weather** the owners of Bioterme are in no hurry to open the swimming pool.

(18) The expected **sour weather** will cause malaise or indisposition in many people.



(19) When the summer is sunny and September alike, the grapes will be sweet, but **sour weather** will give us grapes that will be hard to sell

An important fact about *kislo vreme* is that it is often used with its verbal collocates to express the opposite, i.e., to convey that unfavourable conditions did not, in fact, have the expected effect and did not put people off from coming and/or having a good time. Colligationally, therefore the use of negation is noticeable:

(20) The good spirits were not destroyed even by somewhat **sour weather**, which towards noon cleared up, so they set off for short or longer walks in the surrounding areas [...].

COBUILD dull	LDOCE grey/dull weather	MED dull	MWLD dull	CED dull
You say the weather is <b>dull</b> when it is very cloudy.	cloudy and not bright	if the weather is dull, there are a lot of clouds and it is rather dark	not sunny : having a lot of clouds ▪ <i>a dull winter sky</i>	(of weather) not bright or clear; cloudy
<b>bilingual</b>				
PSA kislo vreme	MSA kislo vreme	OXZILLA kislo vreme	PONS kislo vreme	VASS dull sky
bad, nasty, foul	bad, nasty, foul	/	/	oblačno nebo

Table 4: The treatment of *kislo vreme* in the selected monolingual and bilingual sources.

### 3.5 Meaning as Encyclopedic Knowledge: (Terminological) Compounds

There are some words and phrases that cannot be understood without knowledge of the real world or so-called encyclopedic knowledge, such as compounds (*kislo zelje* - sauerkraut) conveying different degrees of (semi)terminological meaning. Semantic prosody is the least prominent here, as the meaning has already been fixed by word-semantics and extra-linguistic knowledge. Due to this fact, and lack of space, we will not treat this category in more detail.

Meaning as (Pragmatic) Knowledge about Language Use: Idiomatic Expressions

In some expressions, where the salient (or metaphorical) meaning of either of words is insufficient for the reader to know their overall meaning, “word-semantics are redundant and yield entirely to the pragmatic reality of use, both textual and contextual, as the meaning of the phrase relies heavily on knowledge of semantic associations and semantic prosody” (Philip 2009):

**Example 5: *deklica za vse*** – dogsbody; girl/gal Friday

SLD: nekdo, ki kje opravlja najrazličnejša dela, od najnižjih do najzahtevnejših

(21) Having finished school, she, soon after World War II, got a job with the Slovenian Railways where she persisted for almost 34. She was a **general dogsbody**: worker, booking clerk and paymistress at the Head Office of the Slovenian Railways in Ljubljana.

(22) Jože Klemenčič is a **general dogsbody** in Slovenian langlauf: vodja, koordinator in pomočnik trenerja.

(23) He is now involved fixing computers, in the shop and on the terrain, but he calls himself a **dogsbody** as he does all sorts of jobs.

The concordance to *deklica za vse* shows that the expression is used both for men and women, and that, interestingly enough, it is not possible to identify a clearly negative attitude towards this enforced role. The emphasis is on the variety and unpredictability of the tasks that one is expected to perform, which in some contexts even conveys positive prosodies ensuing from the fact that such a person typically displays positive qualities such as dedication, resourcefulness and efficiency in the assigned tasks.

DANTE: dogsbody: n [inf] [non\_AmE] sb who has to do any unpleasant jobs that nobody else wants to do

(24) We pull over to the side of the road and 10 minutes later meet George Wolter who is to be our guide, host, translator, organiser and **general dogsbody** for the next 10 days. George is a freelance translator in English and Spanish and has been working for the Political Song Festival for the last 10 years.

(25) Then I suppose, when the festival had opened, I'd already gone to the site with the school, before I left, in the Easter of 1951 and then I started an apprenticeship with Vickers, but I had to wait some time before I could actually start, and I was given the job as a **dogsbody**, you know a fetcher and carrier.

(26) STRUCTURE N\_mod↔ He started off as '**kitchen dogsbody**' at 13 and went on to study at Westminster Catering College./.../

According to the English database, “dogsbody” can be premodified by an adjective, as in “general dogsbody”. In the English-Slovene PONS, we do indeed find “general dogsbody” with the translation *deklica za vse*, while for the entry *deklica za vse* it lists “jack-of-all-trades”, a unit less frequent and perhaps semantically more comparable with *mojster za vse*, which evokes associations of “to be a jack-of-all-trades (and master of none)” and thus conveys negative attitudes. The suggested translations are relatively consistent in all of the bilingual sources (dogsbody, Br. Eng./girl Friday, informal old-fashioned);

however, the interlingual differences as well as the monolingual differences (synonymy) should be explicitly highlighted. As LDOCE, MED and MWLD report, “girl Friday” is old-fashioned and seems to refer exclusively to women who do several different jobs in an office, or are, in fact, secretaries (e.g., MED, also MWLD). “Dogsbody”, on the other hand, is used for both sexes and has, in comparison with the Slovene *deklica za vse*, a much more evidently negative connotation (a person is forced to do the jobs that other people refuse to do). In addition, the functional (pragmatic) meaning of “dogsbody” is different to an extent, conveying feelings of bitterness due to being systematically exploited, which leads us to posit negative prosody.

COBUILD	LDOCE	MED	MWLD	CED
<b>dogsbody</b>				
<b>British, informal</b> A dogsbody is a person who has to do all the boring jobs that nobody else wants to do.	someone who has to do all the small boring jobs that no one else wants to do	<b>British informal</b> <i>someone who is forced to do all the jobs that no one else wants to do</i>	<b>informal + old-fashioned</b> : a woman who does many different jobs in an office	<b>informal</b> a person who carries out menial tasks for others; drudge
<b>girl Friday</b>				
	<b>British English old-fashioned</b> a girl or woman worker who does several different jobs in an office	<b>informal old-fashioned</b> <i>a female secretary</i>	<b>informal + old-fashioned</b> a female office assistant	a female employee who has a wide range of duties, usually including secretarial and clerical work
<b>bilingual</b>				
PSA deklica za vse	MSA deklica za vse	OXZILLA deklica za vse	PONS deklica za vse	VASS dogsbody
dogsbody BR; girl/gal Friday esp. AM	dogsbody BR; girl/gal Friday esp. AM	dogsbody; girl Friday; pejor. Am. ward-wheeler	jack-of-all-trades	deklica za vse; dekla, hlapec

**Table 5: The treatment of *deklica za vse* in the selected monolingual and bilingual sources.**

## 4 Discussion

In the SLD, an attempt was made to incorporate, to a maximum degree, corpus-derived pragmatic components, i.e., semantic prosody, into meaning descriptions by means of various definition strategies. One way of highlighting a particular (shade of) meaning is also to introduce a new (sub)sense. Thus the concordance for *plezati čez/skozi kaj, plezati iz česa* shows consistently unfavourable condi-

tions in which *plezati* occurs, expressing a negative attitude of the speaker towards the central act of “climbing”, which in the other listed senses appears to be neutral, if not positive. However, with the exception of LDOCE, the English monolingual sources do not identify this use, nor does the Dictionary of Standard Slovene, while both corpus-based Slovene-English dictionaries do, listing it as a new subsense introduced by the semantic indicator (*s težavo se premikati*) (to move with difficulty), which in most cases will yield the translation “clamber” rather than “climb”. In the selected subsense of *bobnati* treated in the SLD, the emphasis is on the fact of being engaged as member of a performing musical group, rather than on the actual physical act of playing a drum, which is indicated by a separate subsense. The bilingual sources do not recognise a separate sense, with the exception of PSA and MSA, which list the relevant subsense introduced by (*hot poklic*) (as a profession) and provide an alternative translation “to play the drums”. Collocational meaning, such as in *kislo vreme*, is not regarded as semantically transparent but restricted in semantically analysable units. The most extensive collocational range for “X weather” is provided by LDOCE and MED, with the most typical collocates being “dull” and “grey”. The semantic equivalent “sour weather” is practically non-existent, but there are some corpus occurrences, e.g., “the weather turned sour with thunderstorms and heavy rain,” confirming the translation, albeit hypothetically. The semantic prosody in the case of “dull/grey weather” cannot be paralleled to that of *kislo vreme* as described in 3.4, and there are not enough corpus examples of “sour weather” to establish its prosody, which could be interlingual. The type of meaning referred to as encyclopedic, such as in *kislo zelje* (*sauerkraut*), is very much dependent on real world knowledge, but much less sensitive to axiological aspects of meaning, which is generally the case in (semi)terminological lexical units. Determining interlingual equivalence is often rather straightforward and, in a monolingual perspective, is the least prone to language change. Finally, with reference to idiomatic units of meaning, which are normally treated as phraseological units, it can be said about *deklica za vse* that its functional equivalent “dogsbody” conveys the subject’s feelings of bitterness due to being systematically exploited, which leads us to posit negative prosody; “girl Friday”, on the other hand, can only refer to a woman and is, as COBUILD and MWLD demonstrate, old-fashioned. The choice of the translation – in our case “dogsbody”, “girl Friday” or “jack-of-all-trades” – is dependent on the circumstances of meaning, and on whether (or not) semantic prosody, which often plays a part in phraseology, is detected in the source language. The analysis of lexical units has shown that we are not dealing with full equivalents, but, at best, with comparable idiomatic units, where meaning requires greater involvement of pragmatic knowledge.

Based on a tentative typology of meaning we have discussed a number of strategies for tackling pragmatics monolingually: first we dealt with an example of connotation where in some sources an adequate label (pejorative/disparaging) was applied, while other sources combined labelling and defining (informal/slang + an insulting word for). The focus of the analysis, however, was on those instances that encode pragmatic meaning as part of their meaning, rather than that which is traditionally subsumed under connotation. In the SLD, some meanings that would traditionally be labelled as “humorous” or “showing disapproval” were, where possible, made part of the definition. Usually the first part

of the definition provides the general semantic-syntactic pattern and the second describes pragmatic circumstances, including the semantic prosody where appropriate. In other words, the first part of the definition is a straightforward explanation, while typically the second part tells us about the speaker's attitude to the meaning situation and, expressing the pragmatic function, clarifies why a particular lexical choice has been made by the speaker. We have also pointed out the treatment of individual lexical units in the selected bilingual dictionaries. Any indication of prosodic meaning is consistently implicit, even when the prosody in the two languages is different, or when only a particular (sub)sense of L1 carries prosodies and the translation differs accordingly. The identified ways of conveying pragmatics in the examined bilingual dictionaries are: functionally adequate translations, i.e., a new (sub)sense (with an appropriate semantic indicator), such as in *plezati*: (*s težavo se premikati*), examples of use and labels. The bilingual focus remains throughout on the Slovene-English rather than English-Slovene perspective. Considerable progress is noted in the treatment of language use in contemporary bilingual corpus-based dictionaries and databases, such as PSA and VASS.

## 5 Conclusion

Corpus-derived facts that concern axiological aspects of meaning, such as those related to “situation” and the other lexical units examined, can be explicitly described in monolingual lexical databases and dictionaries. Lexicography, by definition, is concerned with the investigation and recording of all aspects of lexical meaning. Semantic prosody can be viewed as a link between the lexical and the textual or discoursal levels. By including information not only on collocational but also lexical-textual co-selection, we are bound to improve the dictionary, this practical tool, equally serving language learners, translators and interpreters, as well as communicologists, copywriters, etc. It is well known (cf. Partington 1998: 72) that signals on semantic prosody are particularly important for second or foreign language learners, as they do not have the subconscious understanding of pragmatic meaning that, presumably, native speakers do. However, assuming that semantic prosody is part of a dictionary “sense” and, in many cases, the key to its actual identification, explicating prosodic meaning in native-speaker dictionaries cannot be considered insignificant. Establishing a difference between the lexis that requires labelling due to its connotational meaning that is morphosyntactically coded (*fat-ty, fatso, bunny*, etc.) and the lexis that displays other, contextual types of pragmatic meaning naturally leads to three basic solutions: a label, a definition or both. Collocational meaning, particularly semantic prosody, which is by nature delexical, functional, phraseological, textual and abstracted from various contextual features – which is why some authors have described the phenomenon as “collocational”, “discourse” or “pragmatic” prosody (e.g., Stubbs 1995, 2001), or “semantic harmony” (Lewandowska-Tomaszczyk 1996) – should be presented as part of the definition or in an additional gloss. The question remains as to how exactly semantic prosodies should be presented in bilingual dictionaries, and this is currently under investigation in the doctoral research.

## 6 References

### 6.1 Corpora, Dictionaries and Databases

*British National Corpus* (BNC). Accessed at: <https://the.sketchengine.co.uk/>.

*Collins COBUILD English Dictionary for Advanced Learners*, HarperCollins Publishers (COBUILD). Accessed at: <http://www.collinsdictionary.com/dictionary/english-cobuild-learners>.

*Collins English Dictionary* (CED). Accessed at: <http://www.collinsdictionary.com/dictionary/english>.

*DANTE - A Lexical Database for English*. Accessed at: <http://www.webdante.com/>.

*Dictionary of Standard Slovene* (SSKJ). Accessed at: <http://bos.zrc-sazu.si/sskj.html>.

*Longman Dictionary of Contemporary English Online* (LDOCE). Accessed at: <http://www.ldoceonline.com/>.

*Macmillan English Dictionary Online* (MED). Oxford: Macmillan Education. Accessed at: <http://www.macmillandictionary.com/>.

*Mali angleško-slovenski in slovensko-angleški slovar* (MSA) = English-Slovenian & Slovenian-English Pocket Dictionary. 1<sup>st</sup> edn. Zaranšek, P. ed. 2006. Ljubljana: DZS.

*Merriam Webster's Learner's Dictionary* (MWLD). Accessed at: <http://www.learnersdictionary.com/>.

*PONS Splošni slovensko-angleški slovar in angleško-slovenski slovar* (PONS)

Accessed at: <http://www.pons.si/pons-splosni-slovensko-angleski-slovar-p-518.php>.

*Priročni angleško-slovenski in slovensko-angleški slovar* (PSA) = Concise English-Slovenian & Slovenian-English dictionary. 1<sup>st</sup> edn. Drinovec Sever, N., Pogačnik, A., Žerak A. eds. 2010: Ljubljana: DZS.

*Slovene Lexical Database* (SLD). Accessed at: <http://www.slovenscina.eu/spletni-slovar>.

*Veliki angleško-slovenski slovar Oxford-DZS* (VASS) = Oxford-DZS Comprehensive English-Slovenian Dictionary. Krek, S. (ed.) [Vol. 1, 2005; Vol. 2, 2006]. Ljubljana: DZS.

All links accessed [05/04/2014].

### 6.2 Other Literature

Ellis, Nick. C., Frey, E. in Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody (1): Lexical Access. In *Exploring the lexis-grammar interface*, U. Romer, R. Schultze (eds.). Amsterdam/Philadelphia: John Benjamins.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: CUP.

Lewandowska-Tomaszczyk, B. (1996). Cross-linguistic and language-specific aspects of semantic prosody. *Language Sciences*, 18, pp. 153-178.

Louw, W. E. (1993). Irony in the Text or Insincerity in the Writer?: The Diagnostic Potential of Semantic Prosodies. In *Text and technology: in honour of John Sinclair*, M. Baker, G. Francis and E. Tognini Bonelli (eds). Amsterdam: John Benjamins, pp. 157-176.

Partington, A. S. (1998). *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. SCL 2. Amsterdam and Philadelphia: John Benjamins.

Philip, G. (2009). Why prosodies aren't always there: Insights into the idiom principle. Corpus Linguistics Conference. Liverpool. Great Britain. <http://ucrel.lancs.ac.uk/publications/cl2009/>. [Accessed 06/09/2012].

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. M. (1996). *The Search for Units of Meaning*. TEXTUS IX.

- Stubbs, M. (1995). Corpus evidence for norms of lexical collocation. In *Principle and Practice in Applied Linguistics*, Cook, G. and Seidlhofer, B. (eds.), Oxford: Oxford University Press, pp. 245-6.
- Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
- Šorli, M. (2013). Forms of encoded pragmatic meaning: semantic prosody: a lexicographic perspective. In *Lingue e linguaggi* [Online ed.], 2013, Vol. 10, pp. 95-111.
- Tognini Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John Benjamins.
- Whitsitt, S. (2005). A critique of the concept of semantic prosody. In *International Journal of Corpus Linguistics* 10, pp. 283-305.
- Zethsen, K. K. (2006). Semantic Prosody: Creating Awareness about a Versatile Tool. *Tidsskrift for Sprogforskning*, 4(1), pp. 275-294.
- Zhang, W. (2009). Semantic Prosody and ESL/EFL Vocabulary Pedagogy. *TESL Canada Journal/Revue TESL DU CANADA*. VOL. 26, NO 2, SPRING 2009





# **Bi-and Multilingual Lexicography**



# Linking a Dictionary to Other Open Data – Better Access to More Specific Information for the Users

Ulrich Apel  
Eberhard Karls University Tübingen  
ulrich.apel@uni-tuebingen.de

## Abstract

The project *WaDokuJT* (2013) has developed into the most comprehensive Japanese-German dictionary in regard to covered lemmata and translation equivalents. The dictionary is used by many users from German speaking countries and from Japan. The project was designed concentrating on German speaking users who want to read and translate Japanese texts. So, priority is given to a rather complete coverage of orthographic variations of Japanese headwords on the one side, and German translation equivalents on the other side. Definitions are given only when they are really necessary and only in German language; further, they are as short as possible.

This means that the dictionary project may have certain shortcomings for text production, for Japanese users or for users who need deeper encyclopedic explanations connected to a certain dictionary headword.

This paper presents an approach how to alleviate such deficiencies by providing hyperlinks or other references to data of other dictionaries and encyclopedia projects which may contain the information, users are looking for. For this aim, *WaDokuJT* data refers especially to open source projects, dictionaries, which aren't protected by copyright anymore, and collaborating projects, the data of which can be accessed easily at least in parts.

**Keywords:** References; Japanese; German; hyperlinks

## 1 *WaDokuJT* – Project overview

The Japanese-German dictionary project *WaDokuJT* (2013) started in 1998 as individual initiative at Osaka University (Apel, 2001). In the meantime it covers approximately 115,000 lemmata and around 275,000 database records. Records like derivations, compounds, examples of usage and example sentences are not main headwords.

Records contain the Japanese lemmata, their pronunciation written in Japanese syllable alphabet – *hiragana* – and translation equivalents. Orthographical variety in Japanese is represented in more than half a million different written forms of the entries. The *kana* transliteration is extended by a mark-up to calculate Rōmaji transcription following the new standard *DIN 32708 – Transliteration of*

*Japanese* (NA 009 Normenausschuss Bibliotheks- und Dokumentationswesen, 2013). The number of translation equivalents is more than half a million.

## 2 Weakness investigation of the project and possible improvements

The project was designed mainly for text reception of German users. Sometimes additional information like domain, definitions, explanations etc. are given in German, too. Some information for text production is added, like the Japanese pitch accent for correct pronunciation or conjugation types for Japanese verbs. A certain concession for Japanese users is the addition of a mark-up for the gender of German headwords within the German translation equivalents.

The dictionary's limitations seem to be not too serious, as the dictionary is used by Japanese users and also for text production.<sup>1</sup> Nevertheless, improvements are possible. Further information could be added directly to the dictionary with a specific mark-up and would be displayed only for certain user profiles or use situations. Unfortunately this would complicate the data management not to mention the lack of resources to provide such detailed information. Another way to make further information available without too much trouble for the users and lexicographers is to include links and references to other projects or dictionaries with additional information.

Giving the example of this dictionary project, the paper presents how such links can be added by hand or with automated suggestions of link candidates.

## 3 Linking to data in public domain or with open source license

### 3.1 Großes Japanisch-Deutsches Wörterbuch from 1937

The most comprehensive printed Japanese-German dictionary was edited by Kinji Kimura and was published first in 1937. It is still reprinted since then (*Großes Japanisch-Deutsches Wörterbuch*, 1952). The

---

1 Actually, the dictionary is often used as German-Japanese dictionary, since the German translation side can be easily searched, too. This of course is not recommended because users may get overwhelmed by possible Japanese translations, some of which are outdated or are used only in rather specific situations. Since there is quite a number of rather good German-Japanese dictionaries, there would be enough alternatives, although most of these dictionaries concentrate on the needs of text reception by Japanese users (e.g. *Shogakukan Großes Deutsch-Japanisches Wörterbuch*, 1998, or *Wörterbuch der deutschen und japanischen Gegenwartssprache. Deutsch-Japanisch*, 1989).

author deceased in 1948, and the copyright expired in 1998 – fifty years later according to Japanese copyright laws.

This dictionary seems to be written with the purpose to enable Japanese diplomats and writers to explain Japanese intentions and ways of thinking to Westerners and especially to speakers of German. Different meanings of polysemous words are defined in Japanese and the corresponding translation equivalents are given in German. The German equivalents are completed by articles to show gender, by verb conjugation, by noun and adjective declension or by adjective comparison etc.

This information is still rather valuable, especially for Japanese users and for German text production, even though the Kimura dictionary is outdated in certain aspects. For example, it doesn't reflect orthographic reforms in Japan and Germany or it contains many Japanese names for cities in Manchuria, which was *de facto* a Japanese colony at the time of the compilation. Nowadays, these names are of very limited use.

For German users the Kimura dictionary is also lacking for example the pronunciation of Japanese subentries, the conjugation of Japanese verbs and further German explanations.

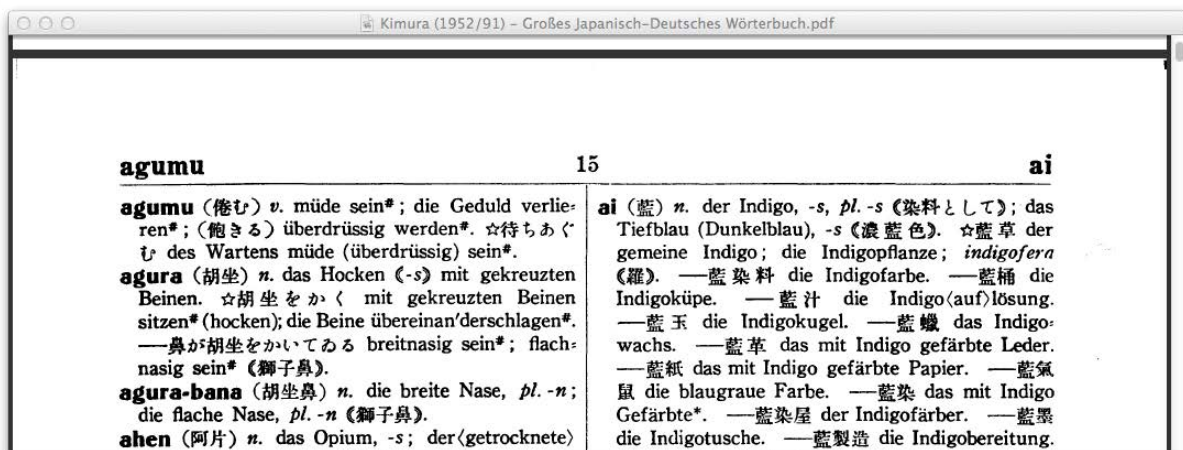
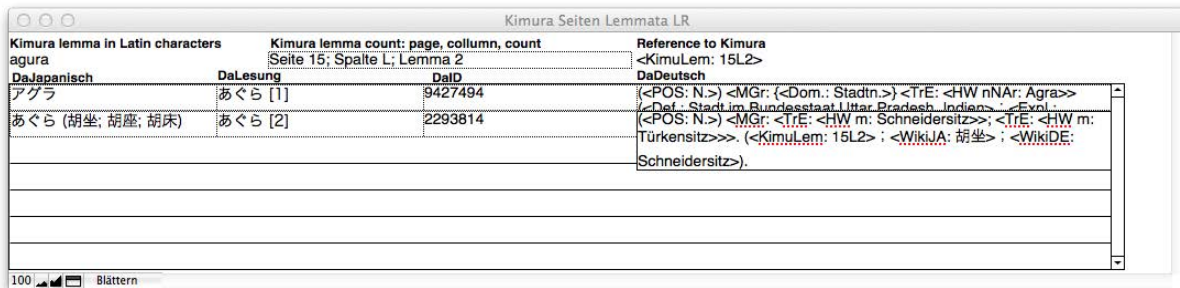


Figure 1: Screenshot of a sample page of the *Großes Japanisch-Deutsches Wörterbuch* (1952) in a digitalized version as PDF with the example entry of *agura*.

The *WaDokuJT* project is now adding gradually tags that refer from one *WaDokuJT* entry to the corresponding page, column and entry count of the Kimura dictionary. This process involves quite a lot of manual operations, but it can be supported by electronic means. A scan of the Kimura dictionary had enough quality to run an optical character recognition for Latin characters.<sup>2</sup> Since both dictionaries contain headwords in Rōmaji transcription, one can set up a database relation using the pronunciation transcription as a common key.

2 Japanese character recognition doesn't work well with pre-war orthography, since modern OCR technique concentrates on the most probable characters and prefers contemporary short forms. Further, Japanese OCR has problems with e.g. German umlauts.

Unfortunately, the Japanese language has a lot of homonyms, especially as a result of the intensive borrowing of loan words from Chinese which were adapted to Japanese pronunciation while losing some distinguishing features in the process. For example, the Kimura dictionary gives 38 entries with the pronunciation of *ko* or *kō*.



**Figure 2: Screenshot of the entry *agura* in a database layout relating to two *WaDokuJT* entries with the same pronunciation.**

A relational database being able to deal with homophony – or to be more precise with identical transcriptions of the pronunciation in Latin characters – is shown in Figure 2. The entry with the pronunciation *agura* relates potentially to two entries of the *WaDokuJT* dictionary. The screenshot shows them displayed in a portal of a database layout using a desktop application. A human editor can then enter the reference to the corresponding Kimura entry to the *WaDokuJT* entry via script and a defined hotkey.

A similar approach is used by a students’ project at the Institute of Asian Studies – Department of Japanese Studies of Tübingen university in an online version. The project makes serious and steady progress, and, since the first few thousand entries are already covered, it is only a matter of time until the whole dictionary is adapted in this way.

The next step is to give the users access from the *WaDokuJT* webpage to the Kimura information. We will add a hyperlink to the interface of the online dictionary which will open a scan of the page with the corresponding entry from the Kimura dictionary. This process will be very similar to the one that is explained in Paragraph 4.2, where a hyperlink from the online *WaDokuJT* dictionary opens a page in *Google Books*.

Linking the *WaDokuJT* dictionary with the Kimura dictionary will hopefully also lead to a new correction iteration of the *WaDokuJT* data. As for instance, missing translation equivalents or example sentences from the Kimura dictionary can be added more easily.

### 3.2 German and Japanese *Wikipedia*

The Japanese *Wikipedia* (2013) is the largest edition of *Wikipedia* in a non-European language. Although the worth of *Wikipedia* as a primary source or reliable references is disputed in academia, it is an easy accessible and important resource. In most cases, it gives a better overview on a certain topic

than would be possible in any ordinary bilingual dictionary. As the *WaDokuJT* dictionary offers mainly translation equivalents and only short hints on meaning or very short definitions, more extensive explanations might be of advantage to many users. *Wikipedia* often refers to corresponding entries in other languages, and Japanese-German pairs can be found, too.

In the source data of the *WaDokuJT* project, the *Wikipedia* article's title is used as unique reference with a mark-up for *Wikipedia* and its language version. As far as possible, *WaDokuJT* refers to both, the Japanese and the German *Wikipedia*. Unfortunately, in many cases a Japanese *Wikipedia* entry has no German equivalent, and often entries, that are marked as corresponding, differ relatively far in meaning and can definitely not be considered as translation equivalents.

The process of finding correspondent *Wikipedia* articles can be automated to a certain extend, although orthographic variety of Japanese makes this process more difficult. Sometimes, several trials with different orthographical forms are necessary. Names of plants and animals for example are written in *katakana* characters within the Japanese *Wikipedia* – as is customary with scientists in the domain of biology – while many traditional dictionaries use *kanji* writings. The *WaDokuJT* project tries to give the most frequent Japanese writing first. Doing so, it is possible to generate automatically links for example to the *Wikipedia* with high probability to work correctly.



**Figure 3: Screenshot of the database interface for working on links to the Japanese and German *Wikipedia*.**

The screenshot in Figure 3 shows a window from the database application that is used to add links in the *WaDokuJT* dictionary to *Wikipedia* from existing entries. The example entry has four Japanese writings and a generated link to the Japanese *Wikipedia*. A "Web Viewer" field displays the content of this link and enables the editor to check easily the correct correspondence.

A similar link to the German *Wikipedia* can be generated using the first German translation equivalent. If these links between *WaDokuJT* and different *Wikipedia* versions work correctly in spite of possible polysemy or homography, a certain mark-up shorthand as reference to the Japanese or the German *Wikipedia* can be added via script and a hotkey.

The advantage of such a reference for users is, besides getting the translation equivalents which are to be expected from a bilingual dictionary, an easy access to up-to-date encyclopaedic information as well and in many cases even directly both in Japanese and in German.

## 4 Cooperations with partially open source projects

Currently, the *WaDokuJT* project is cooperating with two domain specific dictionaries, which are the topic of this paragraph. Further cooperation with specialists in other fields is discussed, also. The domains are for example: disaster prevention, mechanical engineering, economy, medicine and life sciences, traditional craft professions and martial arts. Thanks to the great variety and most diverse possibilities, one can be rather optimistic about the results.

The main characteristic of the presented approach is that every project can keep its individuality and its own strengths. The projects shouldn't compete for users or financing, but complement each other in very productive way.

### 4.1 *Sōgō bukkyō daijiten* – a Buddhist encyclopedia

One cooperating dictionary project is the translation of a Buddhist encyclopaedia (*Sōgō bukkyō daijiten*, 1987; Aoyama et al. 2006; Aoyama et al. 2013). Since a number of years, the articles are translated and extended by information that is relevant for German speaking users. Recently, these data were converted into an online version stored on the same server of Tübingen University as the *WaDokuJT* dictionary.

The entries of the Japanese version of the lexicon are in the Japanese *a-i-u-e-o* order. The online version identifies them by page and count of the entry on the very page. *WaDokuJT* uses this information as unique ID to refer to the Japanese version of the Lexicon as well as to the German translation of the online dictionary. The data of the encyclopaedia will be easier accessible and *WaDokuJT* can concentrate on the translation equivalents and leave buddhological explanations to that lexicon.

### 4.2 Japanese-German archaeological dictionary and its new Japanese-English-German version

Another collaborating project is an archaeological dictionary project, of which a Japanese-German version already exists as a print-on-demand book: *Kleines Wörterbuch zur Japanischen Archäologie* (Steinhaus, 2010). This project addresses scientific archaeologists and tries to give the latest and often normative translation equivalents, reflecting the current state of the art in the field.



The *WaDokuJT* project has obtained an electronic version of the book from the author and now points from one *WaDokuJT* entry to the corresponding page of the archaeological dictionary. With this information, one can also generate a hyperlink which opens the dictionary page in *Google Books*.

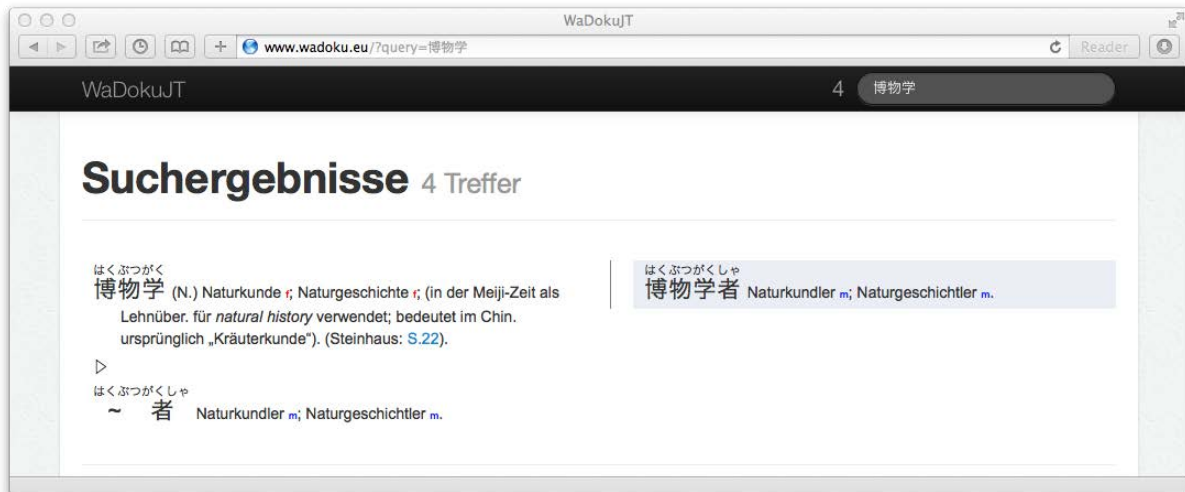


Figure 4: Web interface of *WaDokuJT*, displaying an entry with hyperlink to Steinhaus *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch* (2010) at *Google Books*.

Figure 4 is a screenshot from the online version of the *WaDokuJT* project *wadoku.eu* which shows an entry with reference to Steinhaus *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch* (2010). The page information is also a hyperlink which by clicking it opens the mentioned page of the book in its online version at *Google Books*.

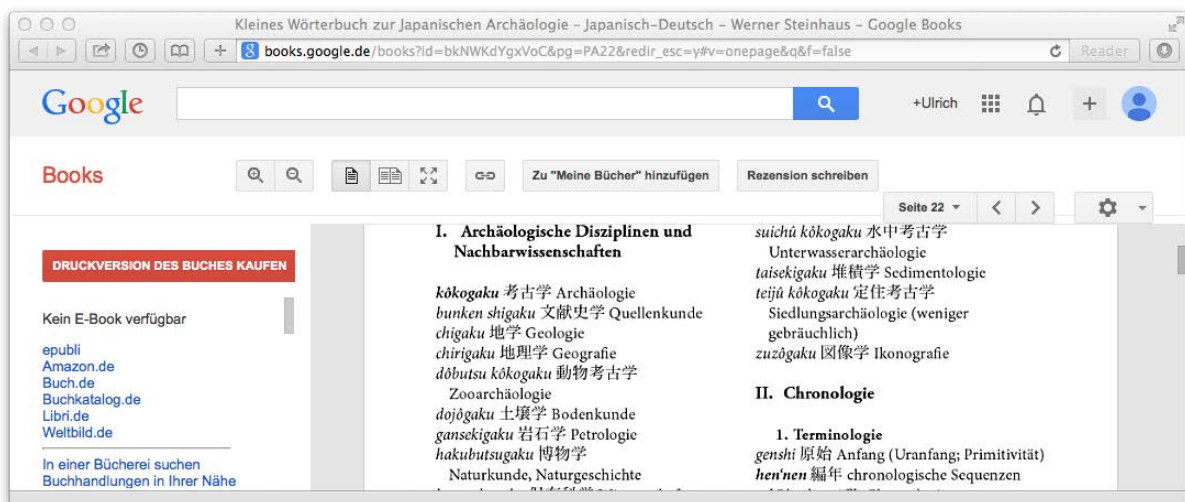


Figure 5: Target page at *Google Books* with the entry *hakubutsugaku* “Naturkunde, Naturgeschichte”.

Figure 5 shows a screenshot of the corresponding page of Steinhaus (2010) which is opened at *Google Books*. Here, the translation equivalents are more streamlined for the needs of archaeologists, who

aren't interested in e. g. linguistic explanations about when the word came into usage in Japan and what it meant in its Chinese version.

One main feature of the original archaeological dictionary is the arrangement of entries around certain topics. This means that the archaeological dictionary and the *WaDokuJT* project are no competitors. Further, *WaDokuJT* has no claims of being normative for certain scientific fields but gives possible translations. Through the reference to the archaeological dictionary, users get easy access to the normative translations, too.

The development of the archaeological dictionary goes on in a rather fast pace. In collaboration with the University of East Anglia, Norwich, Great Britain, it is extended to a Japanese-English-German version. In addition, it is planned, that the new version will be hosted at Tübingen as well.

## 5 Projects linked to *WaDokuJT*

Linking dictionaries to other projects doesn't need to be a one-way-street. Other projects link to the Japanese-German data. For example, *JMdict* (2013), a multilingual lexical database with Japanese as the pivot language uses *WaDokuJT* data for German translations.

Further, a new Korean-German dictionary *Handok.eu* (2013) refers to *WaDokuJT* entries via their ID. Korean and Japanese use a lot of common loan words from Chinese. The project founded by Benjamin Rusch, a former student of the Department of Japanese Studies, may not have the same amount of translation equivalents in the beginning and users may be pleased to get more choice from the *WaDokuJT* project.

## 6 Managing project derivatives

Linking data plays an important role in managing two forks of the *WaDokuJT* project, too. An online interface of the project developed a life of its own, as format changes have rendered it incompatible with the original data. Users can add new entries or suggest corrections of existing entries online. But also on the original data side, new entries and corrections have been added.

Both sides use unique IDs - seven figure IDs for common entries and eight figure numbers for new entries for the online version at *wadoku.de*. Via these IDs entries and changes can be at least monitored, what makes common improvements possible and has most benefit for the users. Data import from *wadoku.de* also includes another edition cycle and should lead to better overall quality of the project.

## 7 Conclusion

In our opinion, one dictionary alone cannot satisfy all possible users' needs. An attempt to do so may turn the data rather hard to manage and very difficult to use. For additional information, we suggest the approach to link one dictionary to other dictionaries and sources. Our examples come from the daily lexicographical praxis and are added one by one, entry by entry. An automatic system that suggests links, will be helpful for the lexicographer in many cases, but we mistrust a full automatic systems at the moment.

The ultimate aim is to provide users with more information that is better tailored to their specific needs without corrupting usability and manageability of the data.

## 8 References

- Aoyama, T., Paul, G., Schmidt-Glitzner, H., Schmithausen, L. and Wittern, C. (ed.) (2006). *Das Große Lexikon des Buddhismus – Erste Lieferung: A–Bai*. Munich: Iudicium.
- Aoyama, T., Paul, G., Rotermund, H. O., Schmithausen, L., Steineck, R. C. and Wittern, C. (ed.) (2013). *Das Große Lexikon des Buddhismus – Zweite Lieferung: Bait–D*. Munich: Iudicium.
- Apel, U. (2001): Ein elektronisches japanisch-deutsches Wörterbuch auf Datenbankbasis – Über das Finden von Wörterbucheinträgen im Computer-Zeitalter. In Gössmann, H. and Mrugalla, A. (eds.). *11. Deutschsprachiger Japanologentag in Trier 1999*. Bd. II. Hamburg: Lit. 627–644.
- Google Books. <http://books.google.com/> [10/11/2013].
- Großes Japanisch-Deutsches Wörterbuch* (1952). [First edition 1937, Kimura, K. (ed.)]. Tokyo: Hakuyūsha.
- Handok.eu. <http://handok.eu/> [founded by Benjamin Rusch; 10/11/2013]
- JMdict. EDRDG. <http://www.csse.monash.edu.au/~jwb/jmdict.html> [10/11/2013].
- NA 009 Normenausschuss Bibliotheks- und Dokumentationswesen (NABD) [draft for DIN 32708, Information und Dokumentation – Umschrift des Japanischen]: <http://www.nabd.din.de/projekte/DIN+32708/de/147167717.html> [10/11/2013]
- Shogakukan Großes Deutsch-Japanisches Wörterbuch* (1998): Kunimatsu K. (ed.) [revised edition]. Tokyo: Shōgakusan.
- Steinhaus, W. (2010): *Kleines Wörterbuch zur Japanischen Archäologie – Japanisch-Deutsch*. Berlin: Epubli. [Also: on line]. <http://books.google.de/books?id=bkNWKdYgxVoC> [10/11/2013].
- Sōgō bukkyō daijiten* (1987): Sōgō Bukkyō Daijiten Henshū Iinkai (ed.). Tokyo: Hōzōkan.
- WaDokuJT: <http://wadoku.eu/> or <http://wadoku.de/> [10/11/2013].
- Wikipedia – Die freie Enzyklopädie: <http://de.wikipedia.org/wiki/Wikipedia:Hauptseite> [10/11/2013]
- Wikipedia [Japanese version]: <http://ja.wikipedia.org/wiki/メインページ> [10/11/2013]
- Wörterbuch der deutschen und japanischen Gegenwartssprache. Deutsch-Japanisch* (1989). Schinzingler, R., Yamamoto, A. and Nanbara, M. (ed.): Tokyo: Sanshusha.



# Bilingual Word Sketches: the *translate* Button

Vít Baisa, Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý  
Lexical Computing Ltd, UK; Faculty of Informatics, Masaryk University, Czech Republic  
{xbaisa,jak,xkovar3,pary}@fi.muni.cz, adam@lexmasterclass.com

## Abstract

We present bilingual word sketches: automatic, corpus based summaries of the grammatical and collocational behaviour of a word in one language and its translation equivalent in another. We explore, with examples, various ways that this can be done, using parallel corpora, comparable corpora and bilingual dictionaries. We present the formalism for specifying equivalences between gramrels in the two languages. We show how bilingual word sketches can be useful for dictionary-making and we present additional functionality to make them more useful. We state the language pairs for which bilingual word sketches are currently available, and our plans for adding more pairs.

**Keywords:** word sketch; corpus lexicography; lexical computing

## 1 Introduction

Word sketches are one-page, automatic corpus-based accounts of a word's grammatical and collocational behaviour (Kilgarriff et al 2004). Since their introduction in 1998 they have come to be widely used in lexicography, often serving as the first port of call for a lexicographer analysing a word. Until recently, they have been monolingual. Bilingual lexicographers would like to see the word and its grammar and collocations, matched up with its translation and its grammar and collocations. In this paper we discuss how this might be done, and present the solution we have adopted and now make available within the Sketch Engine.

Two open questions are:

- how should the source word's translation be identified?
- how should the grammar and collocations be matched up?

We first describe three responses, based on different answers to the first question: *bics*, *bips* and *bims*. We then describe how we amalgamate all three to give a single, easy-to-use ***translate*** function, with a report as illustrated in Figure 1.



Figure 1: Bilingual word sketch for fire/Feuer. The user can click on alternative translations to see alternative sketches.

In all the approaches discussed, we start from the word sketch for one language (hereafter L1) and augment it with information from the other language (L2). Note that, here, L1 and L2 are neither ‘source’ and ‘target’ as understood by translators, nor ‘mother tongue’ and ‘language being learnt’ as in the language learning literature. They simply reflect the fact that the user (and algorithm) starts from one language and adds information from another. At some point we may develop symmetrical, direction-independent bilingual word sketches but we have not done so yet.

## 2 BIPs

Bips are bilingual word sketches based on parallel corpora. A dictionary is not needed because the connections between the languages can be inferred, by looking to see which <L1, L2> pairs of words are frequently found in aligned chunks <L1, L2> chunks (where the chunks are usually sentences). We first count occurrences in aligned chunks for all <L1, L2> word pairs, and then use the Dice coefficient to identify candidate translations.

Where a lemmatiser is available for the language, the corpus is first lemmatised and the dictionary-induction process is applied to lemmas rather than word forms.

This provides a bilingual dictionary, with each lemma in each language having a list of candidate translations, with confidence scores. The default setting is that the ten top candidates for each lemma are retained.

Similar methods are used in GIZA++ (Och and Ney, 2000) and other tools for statistical machine translation, though usually applying to word forms rather than lemmas, and with different statistics and objectives.

Once the bilingual dictionary is in place, the algorithm for creating the bip sketch is as follows:

- (1) the user inputs an L1 headword.
- (2) take the set of L1 collocates<sup>1</sup> from the L1 word sketch and translate them using the bilingual dictionary.
- (3) take the top translation candidate for the L1 headword: call it the L2 headword
- (4) take the set of collocates from the word sketch of the L2 headword
- (5) perform an intersection between the translations-of-collocates from step 2 and the collocates-of-translation from step 4
- (6) for each item in the intersection,
  - is there at least one pair of aligned chunks in the parallel corpus where
    - the L1 collocation occurs in the L1 chunk, and
    - the L2 collocation occurs in the L2 chunk?
  - if yes
    - present as a translation candidate for the L1 collocation,
    - illustrate with the aligned chunk in the two languages
- (7) For L1 collocates with no translation candidates in the intersection, or where there were items in the intersection but there were no instances of corresponding collocations in aligned chunks, present the L1 collocate monolingually, without any candidate translations.

A bilingual sketch produced using this method, also showing the aligned chunks with both L1 and L2 collocates, is shown in Figure 2.

declaration (noun) EUROPAL7, en freq = 5532	
déclaration (noun) EUROPAL7, fr freq = 11505	
use another candidate translation: <a href="#">déclarations</a> <a href="#">écrites</a> <a href="#">écrite</a> <a href="#">Déclarations</a>	
object_of	
write	<a href="#">118</a> The American president 's current negotiator , Robert Zoellick , wrote the 1990 transatlantic <b>declaration</b> . Subject : Remembrance of the Holocaust Following up the Swedish Prime Minister Göran Persson 's initiative in arranging the well-received Stockholm International Forum on the Holocaust in January 2000 , Parliament adopted a written <b>declaration</b> on 7 July 2000 on the remembrance of the Holocaust .
écrite	<a href="#">14</a> Objet : Souvenir de l' Holocauste Suite à l' initiative prise par le premier ministre suédois Göran Persson d' organiser en janvier 2000 l' important Forum International de Stockholm sur l' Holocauste , le Parlement européen a , le 7 juillet 2000 , adopté une <b>déclaration</b> écrite sur le souvenir de l' Holocauste .
sign	<a href="#">139</a> Neither signed the <b>declarations</b> of renunciation required by the constitutional office , because they have distanced themselves from the house of Habsburg ; instead they merely entered the following statement on their passport applications : " In order to avoid misunderstandings , let it be noted in this connection that I , the applicant , have not and do not question the republican form of government of Austria nor have I ever considered making any manner of claim to sovereignty " .
signer	<a href="#">31</a> Therefore , Madame President , in addition to calling upon you to make authoritative representations to the Slovenian Government , I also invite all the Members of Parliament to sign the <b>declaration</b> calling for the institution of a day dedicated to the Shoah , to the holocaust , to be observed throughout the European Union . C' est la raison pour laquelle , Madame la Présidente , je vous demande d' intervenir auprès du gouvernement slovène et invite tous les députés à signer la <b>déclaration</b> qui demande l' institution d' une journée consacrée à la Shoah , à l' Holocauste , au nom de toute l' Union européenne .
signé	<a href="#">26</a> It appears that a vast majority in the Greek parliament has signed a <b>declaration</b> which states that it is very regrettable that Mr Milosevic has been taken to the so-called Court of Justice in The Hague . Il apparaît que la très grande majorité du parlement grec a signé une <b>déclaration</b> dans laquelle il déplore vivement que M. Milosevic ait été transféré vers ce que l' on appelle la Cour de Justice de La Haye .
issue	<a href="#">99</a> The European Court of Auditors also issued a <b>declaration</b> of assurance for the budget of the European Development Fund , although it did find cause for complaint .

Figure 2: Bip word sketch for *declaration/déclaration*.

1 Our terminology here is that a *collocation* comprises a *headword* and a *collocate* (in a specific grammatical relation).



The ‘intersection’ method follows Grefenstette (1999): to find translations for compositional collocations like English *work group* (into, e. g., French) he looked up *work* and *group* in an English-French dictionary, where he found three translations for *work*, five for *group*. That gives  $3 \times 5 = 15$  possible combinations. He then checked to see which was commonest in a French corpus, and presented that as the candidate translation. This core method is explored in much comparable corpus work (see Sharoff et al 2013 for the state of the art).

Linguee<sup>2</sup> provides users with some similar functionality, with aligned bilingual concordance data together with the dictionary translations of the search and some matched pairs of collocations.

## 2.1 In Praise of EUROPARL

A central constraint on methods using parallel corpora is the quality, genre, size and availability of parallel data for each language pair. For the 22 official EU languages<sup>3</sup> and corresponding 253 language pairs, we are fortunate: the EUROPARL corpora are large, contain professional quality translations, and are of a text type - European parliamentary speeches - which, while far from perfect for general-language lexicography, is far more general than the language of, for example, software manuals or patient information leaflets, two other domains where parallel data is available in bulk. Moreover the EUROPARL corpora have been prepared and made ready for language technology use (Koehn 2005). We are currently only exploring parallel-corpus methods for EU language pairs, for this reason.

## 3 BICs

Bics are bilingual word sketches based on comparable corpora. They require a bilingual dictionary, as well as two comparable corpora, as input. Our first attempts at bilingual word sketches took a comparable-corpus approach, with dictionaries from publishers (Kilgarriff et al 2011). Our conclusion was that this left us too dependent on dictionaries from publishers, which were highly variable in availability and licence terms, not to mention format, size, quality and lexicographic approach. The approach was unviable for extending to multiple language pairs.

Now that we can build bilingual dictionaries for all EU-language pairs, these dictionaries can be used as free-standing resources for building bic sketches. The algorithm again uses the intersection method, and is as presented above for bips, except that step 6 is not available. Wherever an L2 headword’s collocate is a translation of an L1 collocate, it is presented as a candidate translation, as shown in Figure 3.

---

2 <http://www.linguee.com>

3 Excluding Irish: irish has a distinct status to the other 22 languages and there is far less data available.





Figure 3: Bic word sketch for *declaration/déclaration*.

## 4 BIMs

Bims are bilingual word sketches based on **manual** selection of headwords. In this approach the user chooses the two words, usually translation equivalents from two different languages, whose word sketches they want to compare, and the corpora to be used. The two word sketches are spliced together. This takes forward something that bilingual lexicographers have been doing since word sketches were first developed: opening two browser windows side by side, with one word sketch in each. A bilingual sketch for English *brown* and Portuguese *marrom* is shown in Figure 4.



Figure 4: Bim word sketch for *marrom/brown*.

## 4.1 Alignment of grammatical relations

The lexicographer would like to see collocations and their translation equivalents aligned. This is possible to some extent, and is attempted in bip and bic sketches, but is difficult and error-prone. Also the lexicographer would often like more control, and finds it straightforward to match, eg, brown *leather* and *couro marrom* where they are in columns next to each other. So in bim sketches we set ourselves the more limited ambition of matching up columns, so that collocates for corresponding grammatical relations are shown next to each other. (In word sketches, ‘grammatical relations’ or ‘gramrels’ are the relations such as *object*, *object\_of*, *modifier* that are specified in the sketch grammar and are showed at the heads of the columns in a word sketch.)

Where the gramrels have the same names for two languages, this is trivial: we simply show same-name gramrels next to each other. However sketch grammars for different languages are usually prepared independently of each other; the grammar of different languages is different, underlying part-of-speech taggers may use different conceptualisations and word classes; and gramrels will often be given names in the language that the word sketch is for: the French equivalent of *object\_of* is called *objet\_de*. So matching gramrels with identical names is not the standard case. We need a mapping between the gramrels of the two languages.

A mapping for each set of names to some ‘master’ set is preferable to a different mapping for each pair. In our current setting, we use the English names as the master. Both 1:1 and m:n mappings are possible.

The mappings are defined in the sketch grammar using a newly introduced processing directive \*UNIMAP. The following definition from a French sketch grammar

```
*DUAL
=objet/objet_de
*UNIMAP object/object_of
...
```

says that *objet* should be joined with *object* and *objet\_de* should be joined with *object\_of* (or the gramrels paired with English *object* and *object\_of* in other languages). The algorithm for finding a target language (TL) gramrel to display next to a source language gramrel X is:

- if there is one or more TL gramrel with a UNIMAP value matching the UNIMAP value of X, select that one/them
- else if there is a TL gramrel of the same name, select that one
- else, nothing is aligned with X.

Left-over, unaligned TL gramrels are shown after SL and aligned gramrels.

## 4.2 An inline form for finding missing items

For a user, some matching pairs are immediately evident from the bim sketch (*brown leather, couro marrom*) but others are not. We do not find anything equivalent to *brown rice* on the Portuguese side of Figure 5. The user then wants to know “how do you say brown rice in Portuguese?” To meet this need, a new function has been added: the user may click on *rice* to reveal a text-input box where they can input the missing target-language equivalent (here, *arroz*). A new bim word sketch appears, as illustrated in the same Figure 5. This feature helps the user find the missing translation: here, *arroz integral* for *brown rice*.

The screenshot shows two panels. The left panel is for the word 'marrom' (brown) in Portuguese, with a frequency of 14585 (4.5 per million). Below it, the word 'brown' is shown in English with a frequency of 374425 (28.9 per million). A table lists collocations for 'brown' in Portuguese: coloração (340, 7.49), anã (71, 7.4), corrimento (58, 6.92), linhaça (48, 6.9), borra (42, 6.55), and aranha (100, 6.35). A table for 'brown' in English lists: modifies (223,516, 0.3), rice (13,109, 9.28), trout (3,425, 8.26), sugar (11,161, 8.21), and bear (2,754, 7.11). A red arrow points from the 'rice' entry to a text input box labeled 'Lemma: arroz' and a 'Query' button. The right panel shows the word 'rice' (noun) in English with a frequency of 268089 (20.7 per million). Below it, the word 'arroz' is shown in Portuguese with a frequency of 86646 (26.7 per million). A table lists collocations for 'arroz' in Portuguese: parboilizado, cozido, carolino, integral (highlighted in a red box), arbóreo, doce, and polido. A table for 'arroz' in English lists: modifier (100,314, 0.1), brown (13,071, 9.53), fried (3,480, 9.31), basmati (1,592, 8.98), sticky (1,651, 8.06), glutinous (739, 7.88), pasta (1,590, 7.44), and jasmine (647, 7.42).

Figure 5: “Finding the missing translation” functionality.

## 5 Observations

Bics, bips and bims were developed in 2013.<sup>4</sup> In the course of presenting and beginning to use them, we made several observations:

- The bilingual dictionaries created from EUROPARL were of good quality. Most of the time, the top candidate translation was valid.
- For most collocations on the L1 word sketch, we did not find any strong candidates for L2 translation equivalents. The *declaration* examples above were selected because, there, the algorithm did find a translation candidate for many of the L1 collocations. More often, there were very few translations offered. This affected both bics and bips.
- Bims often worked well, but the access method did not: users were required to select, first, the L2; then, the L2 corpus from which the L2 word sketch should be drawn, and then, the L2 lemma. Unless they made all the right choices, they were going to be disappointed.

<sup>4</sup> They were presented at the e-Lxicography conference in Tallinn in October 2013.

## 6 The *translate* button

Most users just want to select the target language, and do not want to think about ‘which corpus’. The system developers are the people who know which corpus has the best word sketches for a language, so they should make that choice.

Many users would also rather not have to think of, and input, the L2 lemma. For all EU language pairs, the EUROPARL dictionary offers candidate translations, so here again, more can be done by the system leaving less work for the user. The user can be offered a BIM sketch in which the L2 word is the top candidate from the EUROPARL bilingual dictionary (with second and other candidates also offered.)

With these choices made, we can add a ‘translate’ button to the word sketch. The ‘translate’ button gives the user a choice of languages. The languages that the user can choose between are all of those where there is:

- a bilingual dictionary between L1 and the language
- high-quality word sketches for the language
- as at April 2014, there are twenty such languages
- a UNIMAP mapping

Then, when the user has selected the L2, they see the bilingual sketch directly, as in Figures 1 and 6:

amore <small>(noun)</small>		itTenTen10 freq = <a href="#">634049</a> (206.1 per million)	
Liebe <small>(noun)</small>		deTenTen10 freq = <a href="#">294973</a> (103.7 per million)	
<b>n_modifier</b>	<b>3,656</b> 0.6	<b>modifier</b>	<b>281,129</b> 0.5
sponsale	<a href="#">129</a> 9.84	wahr	<a href="#">2,537</a> 6.76
lesbico	<a href="#">116</a> 9.39	bedingungslos	<a href="#">583</a> 5.93
oblativo	<a href="#">41</a> 8.46	ewig	<a href="#">870</a> 5.86
karmico	<a href="#">34</a> 8.0	dein	<a href="#">3,017</a> 5.71
leopardiano	<a href="#">20</a> 7.15	unerfüllt	<a href="#">465</a> 5.7
lesbo	<a href="#">16</a> 6.7	unglücklich	<a href="#">539</a> 5.69
lesbici	<a href="#">10</a> 6.29	innig	<a href="#">468</a> 5.65
fato	<a href="#">31</a> 6.21	verboten	<a href="#">490</a> 5.58
anoressico	<a href="#">9</a> 5.87	romantisch	<a href="#">552</a> 5.45
cannibale	<a href="#">13</a> 5.85	heimlich	<a href="#">476</a> 5.4
conta	<a href="#">23</a> 5.7	groß	<a href="#">12,073</a> 5.38
cinefilo	<a href="#">9</a> 5.18	leidenschaftlich	<a href="#">423</a> 5.35
court	<a href="#">12</a> 5.05	viel	<a href="#">6,631</a> 5.33
<b>preN_V</b>	<b>120,341</b> 2.0	<b>VerbY+Subst...</b>	<b>24,736</b> 0.5
giurare	<a href="#">430</a> 6.71	gestehen	<a href="#">750</a> 7.54
dichiarare	<a href="#">1,080</a> 6.7	erwidern	<a href="#">139</a> 6.1
donare	<a href="#">573</a> 6.66	besingen	<a href="#">48</a> 5.38
chiamare	<a href="#">1,754</a> 6.62	entdecken	<a href="#">836</a> 5.03
Un	<a href="#">1,160</a> 6.49	beteuern	<a href="#">43</a> 4.87
cantare	<a href="#">601</a> 6.49	schwören	<a href="#">59</a> 4.45
nascere	<a href="#">1,388</a> 6.29	schenken	<a href="#">194</a> 4.27
provare	<a href="#">901</a> 6.27	entgegenbringen	<a href="#">24</a> 4.22
vivere	<a href="#">1,643</a> 6.21	wünschen	<a href="#">634</a> 4.22
confessare	<a href="#">311</a> 6.18	symbolisieren	<a href="#">45</a> 4.14
ricambiare	<a href="#">277</a> 6.13	empfinden	<a href="#">200</a> 4.12
sbocciare	<a href="#">260</a> 6.08	predigen	<a href="#">38</a> 4.09
trasmettere	<a href="#">584</a> 6.05	grüßen	<a href="#">85</a> 4.08

Figure 6: Bilingual word sketch for *amore/Liebe*.

The functionality is currently available for all combinations of English, French, German, Italian and Spanish.

## 7 Current and future work

- **Aligning collocates within BIMs**

A feature of bics and bips that bims were lacking, was the alignment of translation-equivalent collocations. Although, often, no alignments were found, where an alignment was found, it was usually valid and helpful. So, we can further enrich bim-style sketches, by re-ordering the collocates in the L2 word sketch table so that matched L1, L2 pairs are next to each other.

This is currently in progress.

- **More languages**

For monolingual or bilingual word sketches to work well, there are a number of prerequisites; first a very large corpus, then processing tools including a tokeniser, lemmatiser, part-of-speech tagger and sketch grammar. All components are currently in place for all major world languages, all EU languages, and a number of others; languages where we intend to get all components working well in the near future include Icelandic, Malay, Bahasa Indonesia and Burmese.<sup>5</sup>

- **More bilingual dictionaries and language pairs**

We are looking into parallel resources for other language pairs including English and the non-EU major world languages, in particular Arabic-English, Chinese-English, Japanese-English, Russian-English.

- **Evaluating EUROPARL-based bilingual dictionaries**

For Czech-English we are currently assessing our induced dictionary by comparison with a publisher's dictionary. We shall also compare the Dice algorithm with the Giza++ algorithm for dictionary induction. We shall then extend the evaluation to other language pairs.

## 8 References

- Grefenstette, G. (1999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. *Proceedings of Aslib Conference on Translating and the Computer 21*. London.
- Kilgarriff, A., Avinesh, P. V. S., & Pomikálek, J. (2011). Comparable Corpora BootCaT. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex*, 122-128.
- Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proc. Euralex*. Lorient, France.
- Koehn P. (2005). EuroParl: A Parallel Corpus for Statistical Machine Translation. *Proc. Machine Translation Summit*.
- Och, F. & Ney, H. (2000). Improved Statistical Alignment Models. *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447, Hong Kong, China.
- Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P., editors (2013). *Building and Using Comparable Corpora*. Springer.

---

5 In most cases the work has been a collaboration with linguists of the language in question.





# Creating a Bilingual Italian-English Dictionary of Collocations

Barbara Berti, Laura Pinnavaia  
Università degli Studi dell'Insubria, Università Statale di Milano  
barbara.berti@uninsubria.it, laura.pinnavaia@unimi.it

## Abstract

Collocations are among the most challenging issues that learners of a foreign language have to face. The main obstacle for Italian learners of English is encountered in active tasks, and especially, when translating or writing essays in English. The encoding of collocations in fact often results in the selection of wrong terms. There is no doubt that the production and translation of Italian collocations would be facilitated were students to have at their disposal a bilingual Italian-English Dictionary of Collocations. The aim of this work is to endorse the importance of the existence of such a tool and to show the beginnings of its compilation.

**Keywords:** collocations; bilingual lexicography; corpus linguistics.

It is not so much the words of English nor the grammar of English that makes English difficult ... The vague and undefined obstacle to progress ... consists for the most part in the existence of so many odd comings-together-of-words. (Cowie 1999: 52-53)

## 1 Introduction

At the intersection between syntax and semantics, collocations are among the most challenging issues that learners of a foreign language have to face. Not only are they massively present in languages (Hoey 2005), but they are also (or at least appear to be) arbitrary and highly specific for a given language. This poses problems in second language acquisition. In fact, at some stage of the learning process, learners will inevitably have to deal with the way in which words combine with each other and the restrictions that they are subject to. Moreover, their first language will be likely to interfere to some extent with the production of natural-sounding word combinations in the second language and will cause them to produce incorrect *verbum pro verbo* translations (e.g. *make a photo\** as a calque of *fare una foto*).

The main obstacle for learners of English is, in fact, encountered in production tasks, and, especially, when translating from Italian, or writing essays in English. In reception tasks, the meaning of English collocations can often be derived quite easily, especially if the context in which they occur is suf-

ficiently clear. For example, while in isolation the meaning of the collocation *to explode a myth* might be opaque to a learner of English, in an appropriate context, such as *scientists wanted to explode the myth of a monster living in the lake*, it becomes apparent. However, to produce the collocation from scratch would undoubtedly be a problem for the learner. It is highly likely that he/she would not select the correct terms, especially if, as in the case of Italian, the collocate of the noun *myth* is semantically unrelated to the verb *explode* (*sfatare un mito*). It thus becomes apparent that learners need to have at their disposal a resource that will help them to overcome such problems.

Nowadays, English monolingual lexicography does tackle the issue of collocations quite extensively, and especially dictionaries intended for learners of English; yet, there is evidence that Italian learners of English feel more at ease using bilingual tools rather than monolingual ones (Baxter 1980; Bensoussan, Sim and Weis 1984; Atkins 1985; Atkins and Knowles 1990; MacFarquhar and Richards 1983; Piotrowski 1989; Rundell 1999; Scholfield 1999). Unfortunately, however, Italian-English bilingual lexicographical resources poorly document this phenomenon from both a quantitative and a qualitative point of view (Berti 2010, 2012). That is why, in addition to implementing bilingual dictionaries with more and more carefully selected word combinations, a bilingual Italian-English dictionary of collocations would be very useful for Italian learners of English.

Successive to a theoretically-oriented study regarding the creation of a corpus-based Italian-English dictionary of collocations (Berti & Pinnavaia 2012), this work comes as a practical presentation of what selecting collocates entails, followed by two sample entries. The dictionary has as its principal target Italian learners of English, as well as professional translators. It will be unidirectional, presenting Italian collocations with their equivalents in the English language. The aim is not just to provide viable translations of Italian collocations, but also to offer users the most suitable English equivalents based upon frequency of use and situational appropriateness. This can only be achieved if the equivalents are drawn from linguistic corpora of English and carefully evaluated by the compilers, whose linguistic expertise must be bilingual.

## 2 Methodology

Nowadays, the use of corpora is of fundamental importance in lexicography (Atkins 1994), offering a more “objective” perspective on language. Over the years, in fact, it has become progressively clearer that resources based on the lexicographers’ *Sprachgefühl* are subjected to a number of limitations and that they cannot be regarded as fully representative of the current state of a language. Diastratic and diatopic variations, together with a speaker’s idiolect, exert an influence on the type of material that is included or excluded in and from a lexicographic tool, and even the work of a composite team of experts cannot guarantee an objective treatment of the language, or the selection of appropriate material.



In particular, as regards collocations, research shows that acceptability is extremely subjective and that the individuals' knowledge of word combinations in their language might vary to a great extent (Berti, *forthcoming*; Nesselhauf 2003). Acceptability depends, firstly, on semantics and, secondly, and most importantly, on usage. Indeed, it is usage that ultimately provides us with the intuition we need in order to discern between acceptable and unacceptable combinations in our mother tongue. Yet, given that the linguistic experience of each individual is absolutely unique and that the factors that have an influence on it are numerous, there will be a large grey area of disagreement in the judgement of word combinations. For this reason, it is nowadays essential to query corpora for the selection of the lexical material to be used for lexicographical purposes.

Since they are the parts of speech that occur more often in collocational patterns and are the most needed by students, it is nouns and verbs that we decided to select. This choice also seemed reasonable in virtue of the manner in which our thoughts are shaped. A speaker, unaware of the adjective + noun combination *broad daylight*, will more naturally think of the noun *daylight* before its qualifier. Similarly, the retrieval of any verb and adverb combination will normally imply the consultation of the verb before that of the adverb. This seems to be the rationale behind the organization of most dictionaries of collocations, among others, Rundell (2010) and the Oxford Dictionary of Collocations (see ODC 2002). The idea is thus to create an onomasiological dictionary that relies on the well-established distinction between base and collocate.

For the present study we took the Italian noun *odore* (a smell) and verb *pagare* (to pay) as our working examples retrieving the Italian collocations from the *Dizionario delle collocazioni* (Tiberii 2012). The decision to collect our sources from an already existing dictionary of collocations as opposed to extracting collocations from an Italian corpus was taken on recognizing the efficiency of this tool and after having ascertained the overall paucity of general Italian corpora compared to English.

We proceeded to find the collocates of the English equivalents of *odore* and *pagare*, *smell* and *pay* in the British National Corpus (BNC). Each Italian collocate was then matched with its English semantic equivalent. The explanation of the procedure for and the problems encountered with the pairing off of items ensues.

## 2.1 The Noun *Odore*

In Tiberii (2012), the following collocates for the noun *odore* were found:

**odore** *nm* accattivante, acre, acuto, aspro, buono, caratteristico, delicate, disgustoso, distintivo, fetido, forte, fresco, gradevole, inconfondibile, inebriante, intense, invitante, leggero, nauseabondo, nauseante, opprimente, particolare, penetrante pungente, ripugnante, rivoltante, sgradevole, sottile, stomachevole, strano, tenue, terribile, vago.

Considering that the Italian collocates for *odore* are all adjectives, we searched for adjective-noun collocational patterns in the BNC, obtaining the following one hundred collocates for the noun *smell*, listed in order of decreasing frequency:

- sweet
- strong
- pungent
- strange
- musty
- rich
- faint
- sour
- bad
- unpleasant
- acrid
- stale
- warm
- delicious
- fresh
- heavy
- horrible
- lovely
- different
- familiar
- good
- nice
- able
- clean
- damp
- distinctive
- sickly
- dank
- funny
- other
- overpowering
- wonderful
- awful
- hot
- new
- old
- pleasant
- aware
- earthy
- lingering
- peculiar
- thick
- close
- distinct
- metallic
- oily
- foul
- full
- masculine
- musky
- open
- particular
- sharp
- spicy
- chemical
- cold
- cloying
- comfortable
- dirty
- fishy
- fragrant
- free
- great
- green
- milky
- nasty
- putrid
- real
- slight
- smoky
- soft
- tangy
- usual
- very
- whole
- curious
- dangerous
- comforting
- clinical
- characteristic
- burning
- blue
- better
- bitter-sweet
- acute
- dry
- evil
- fine
- homely
- keen
- little
- noxious
- powerful
- rancid
- rotten
- savoury
- sick
- stronger
- supposed
- sure

The next step was to match each Italian collocate from Tiberii (2012) with an equivalent English one from the above list. For example, we paired off *odore accattivante* with *captivating smell*. On finding Italian collocations that correspond to two or more English ones, as in the case of *odore gradevole* which is equivalent to both *sweet* and *pleasant smell*, none were discarded and all were kept. This was done even when one collocation was found to be more frequent than another (i.e. *sweet smell*), out of scientific rigour, on the one hand, and in order to make up for the scant number of collocations provided in bilingual dictionaries, which were naturally also consulted, on the other. When, moreover, it was noticed that some of the collocates in Italian are synonymous, these were grouped together; for instance, *acre* and *aspro*, can both be rendered as *acid*, *sour* or *sharp* in English, just as *leggero*, *tenue*, *sottile*, *vago* can be translated as *slight* and *faint*. When the close examination of the list of English collocations revealed the absence of a number of Italian collocations in the source text, it was deemed opportune to introduce them, especially if the English equivalents are frequently occurring collocations. For example, it seemed important to include the collocations *cattivo odore* and *odore stucchevole*, missing from Tiberii (2012), aware of the fact that the collocations *unpleasant/musty smell* and *heavy/cloying smell* are commonly used in English. Lastly, when equivalents for Italian collocates were not found in the BNC, we consulted WebCorp, which owing to its larger size, often managed to provide suitable equivalent adjectives, such as *captivating*, *delicate*, *intoxicating* and *penetrating smell* for *accattivante*, *delicato*, *inebriante* and *penetrante*.

## 2.2 The Verb *Pagare*

Under the lemma *pagare*, Tiberii (2012) lists the following adverbial collocates:

**pagare** v. abbondantemente, anticipatamente, caro, comodamente, completamente, generosamente, immeditamente, integralmente, interamente, obbligatoriamente, pesantemente, profumatamente, prontamente, provvisoriamente, puntualmente, regolarmente, tempestivamente, volontariamente.

While the collocates we retrieved in the BNC for *pay* amount to forty-three, only eight have been reported here. This is because, unlike the list of collocates for *smell*, the list for *pay* included many that were not pertinent. It was necessary, in fact, to shortlist the relevant ones by examining the concordances. During this process, we also decided to exclude collocates with a single occurrence, following Sinclair's tenet that a single occurrence does not represent a "settled pattern in the language" (2003: 15).

- promptly
- dearly
- hereby
- handsomely
- proportionately
- compulsorily

- upfront
- punctually

As earlier, the Italian and English items were matched. For example, *pagare obbligatoriamente* was matched with *pay compulsorily*. Similarly, synonymous collocations were grouped together, as in *pagare abbondantemente/generosamente/profumatamente*, which were all associated to *pay handsomely*. However, the pairing off of collocates proved to be more problematic than it was for *odore/smell*. Firstly, because there are very few correspondences between the Italian and English collocates; secondly, because it is not always easy to understand the real meaning of certain Italian collocations on account of the absence of contextual data in Tiberii (2012). For example, *pagare anticipatamente/comodamente/integralmente/interamente/puntualmente/regolarmente/volontariamente* do not have a corresponding collocate in the BNC. For this reason, in order to provide an equivalent, not only did we have to consult various bilingual dictionaries along with WebCorp, but we also had to rely on our own knowledge of the language. At times this resulted in equivalent collocations (e.g. *pagare comodamente* = *pay easily*), at other times in lexical sequences that do not fit the verb + adverb pattern (e.g. *pagare anticipatamente* = *to pay in advance*), at other times in single words (e.g. *pagare anticipatamente* = *to prepay*). The results we obtained can be observed in the sample provided in the following section. Not having examples of usage to refer to made the translation of the items *pagare completamente* and *pagare provvisoriamente* particularly problematic: while the former sounds like a free combination, the latter is clearly semantically ambiguous. The way in which the collocates of *pagare/pay* have been rendered in English can be observed in the sample provided in the following section.

### 3 The Sample Entries

Here follow the sample entries for the noun *odore* and verb *pagare*.

#### **odore - smell n.**

##### ADJ. + NOUN

*accattivante* – captivating  
*acre/aspro* – acrid/sour/sharp  
*acuto/pungente* – pungent  
*buono* – good/nice/delicious  
*caratteristico* – characteristic  
*cattivo/sgradevole* – unpleasant/musty  
*delicato* – delicate  
*disgustoso/ripugnante/rivoltante/terribile* – bad/ horrible/nasty/awful  
*distintivo* – distinctive  
*fetido* – foul/putrid  
*forte* – strong/overpowering  
*fresco* – fresh  
*gradevole* – sweet/pleasant

inconfondibile – distinct  
inebriante – intoxicating  
intenso – powerful/keen  
invitante – warm  
leggero/tenue/sottile/vago – slight/faint  
nauseabondo/nauseante/stomachevole – sickly  
opprimente – overpowering  
particolare – curious/peculiar/particular  
penetrante – penetrating  
persistente – lingering  
strano – strange  
stucchevole – heavy/cloying

**pagare - pay v.**

ADV. + VERB

Abbondantemente/generosamente/profumatamente – handsomely  
anticipatamente – in advance/prepay (v.)  
caro/pesantemente – dearly  
comodamente – easily  
immediatamente/prontamente/tempestivamente – promptly  
integralmente/interamente – in full  
obbligatoriamente – compulsorily  
puntualmente – punctually  
regolarmente – regularly  
volontariamente – voluntarily

As the examples show, the entries provide a list of the Italian collocates, documented for the noun *odore* and for the verb *pagare*, along with their respective adjectival and adverbial equivalents for the English noun *smell* and verb *pay*. The Italian collocates appear listed in alphabetical order, and sometimes include more than one exemplar when they have similar meanings (e.g. *odore disgustoso/ripugnante/rivoltante/terribile*; *pagare immediatamente/prontamente/tempestivamente*). Next to them appear the equivalent English collocates, retrieved from the data in the BNC and WebCorp. As can be seen, the correspondences can be of various types. It is possible to have collocations in Italian and English that have the same number of collocates (e.g. *strano odore = a strange smell*; *cattivo/sgradevole odore = unpleasant/musty smell*; *pagare puntualmente = to pay punctually*); sometimes the Italian language has more (*nauseante, nauseabondo, stomachevole odore = sickly smell*; *pagare immediatamente/prontamente/tempestivamente = to pay promptly*); at other times, fewer (e.g. *odore buono = good/nice/delicious smell*; *pagare anticipatamente = to pay in advance/to prepay*). In all cases, it is our intention that dictionary-users have all the most frequent and most appropriate equivalent English collocations for the Italian ones.

## 4 Conclusions

While this work is still in progress and has still not fully solved all the problems involved in retrieving and selecting equivalent English collocates for Italian ones, the objective of compiling a bilingual dictionary of Italian and English collocations is, in our view, invaluable for the quantitative and

qualitative advantages it can offer next to general bilingual and monolingual dictionaries. Compared to a normal Italian-English bilingual dictionary, such as *Il Ragazzini2011*, a bilingual collocations dictionary can, first of all, provide much more information. While certainly being an efficient tool, *Il Ragazzini2011*, like any other bilingual dictionary, has to include all aspects of the Italian language. It cannot, evidently, devote as much space to collocations. In fact, it only provides the following English collocations for *smell* under the noun *odore*: *a good smell, a pleasant smell, a nice smell, a bad smell, a nasty smell* and an *offensive smell*. Despite offering a wide range of noun + verb collocations under the lemma *pagare*, with regards to the lexical pattern adverb + *pagare*, the dictionary only includes *pagare salato* and *pagare profumatamente*, both translated as *to pay dearly*. Clearly, bilingual dictionaries need to make a selection among the various collocations to include, whereas a dictionary of collocations can be much more exhaustive.

Of course, it could be argued that learners or translators, needing to find English collocations, can directly consult the monolingual tools already available. There is no doubt that monolingual English dictionaries of collocations are excellently compiled: once the user is aware of the English base, then he/she can consult such dictionaries to find collocates. It is also true, however, that this task in a monolingual dictionary takes longer and requires greater linguistic skills than it would do in a bilingual collocations dictionary. As seen above, the Italian-English bilingual collocations dictionary could more carefully account for meaning nuances in English, based on the Italian collocates. In finding the English collocations, *an acrid smell, a sour smell, a sharp smell, and a pungent smell*, an Italian learner or translator might be tempted to think that they are all equivalents. We have seen, however, that while the first three can be considered synonyms, the last, *a pungent smell*, means something a little different as its Italian equivalents endorse. The use of a bilingual collocations dictionary should, thus secondly, save the dictionary-user time in trying to individuate what collocation to choose, resulting in a more qualitative and effortless task. The combination of a fine lexicographic and linguistic analysis of Italian and English words will hopefully result in a bilingual reference work that will be quantitatively and qualitatively valid.

## 5 References

- Atkins, B. T. S. (1985). Monolingual and bilingual learners' dictionaries: A comparison. In R. Ilson (ed.), *Dictionaries, Lexicography and Language Learning*. Oxford: Pergamon Press and British Council, pp. 15-24.
- Atkins, B.T.S. (1994) A corpus-based dictionary. In *Oxford-Hachette English- French Dictionary* (Introductory section). Oxford: Oxford University Press: xix – xxxii.
- Atkins, B.T., Knowles, F.E. (1990). Interim report on the EURALEX/AILA research project into dictionary use. In I. Magay, J. Zigany (eds.), *Budalex 88 Proceedings*. Budapest, Akademiai Kiado, pp. 381-392.
- Baxter, J. (1980). The dictionary and vocabulary behavior: a single word or a handful? In *TESOL Quarterly* XIV, 3, pp. 325-336.

- Bensoussan, M., Sim, D. & Weiss, R. (1984). The effect of dictionary usage on EFL test performance compared with student and teacher attitudes and expectations. In *Reading in a Foreign Language*, 2, pp. 262-276.
- Berti, B. (2010). The treatment of lexical collocations of six adjectives related to feelings in a sample of bilingual dictionaries English-Italian. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*, Fryske Akademy, Leeuwarden. [contribution on CD-rom]
- Berti, B. (2012). Le collocazioni lessicali nel dizionario bilingue. Uno studio su tre dizionari italiano-inglese. In G. Gobber *Usi dei vocabolari nell'apprendimento delle lingue*, Milano: Vita e Pensiero, pp. 5-25.
- Berti, B. (forthcoming). 'This sounds odd to me'. Issues of collocational acceptability in a bilingual English-Italian dictionary". In *Proceedings of the conference Norma e uso nella lessicografia bilingue, 18-20 October 2012*, Ragusa.
- Berti, B., Pinnavaia, L. (2012). Towards a Corpus-driven Italian-English Dictionary of Collocations. In R. Facchinetti (ed.) *English Dictionaries as Cultural Mines*, Newcastle-Upon-Tyne: Cambridge Scholars Publishing, pp. 201-222.
- Cowie, A. (1999). *English Dictionaries for Foreign Learners: A History*. Oxford: Oxford University Press.
- Facchinetti (ed.) *English Dictionaries as Cultural Mines*, Cambridge Scholars Publishing, Newcastle-Upon-Tyne: 201-222.
- Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- MacFarquhar, P. D., Jack C. R. (1983). On dictionaries and definitions. *RELC Journal*, 14, pp. 111-124.
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. In *Applied Linguistics*. 24(2): pp. 223-242.
- Oxford Collocations Dictionary for Students of English* (2012). Oxford: Oxford University Press.
- Piotrowski, T. (1989). Monolingual and bilingual dictionaries: Fundamental differences. In Makhan L. Tickoo (ed.), *Learner's Dictionaries: State of the Art*. Singapore: Seamo Relc, pp. 72-83.
- Il Ragazzini 2011. Dizionario inglese-italiano, italiano-inglese*, Bologna: Zanichelli.
- Rundell, M. (1999). Dictionary use in production. In *International Journal of Lexicography*, 12, (1): pp. 35-54.
- Rundell, M. (2010). *Macmillan Collocations Dictionary for Learners of English*, Oxford: Macmillan Publishers Limited.
- Scholfield, P. J. (1999). Dictionary use in reception. In *International Journal of Lexicography* 12, (1): pp. 13-35.
- Sinclair, J. (2003). *Reading concordances. An introduction*. Harlow: Pearson Education Limited.
- Tiberii, P. (2012). *Dizionario delle Collocazioni. La Combinazione delle Parole in Italiano*, Bologna: Zanichelli.

### Websites

- BNC: *British National Corpus*, accessible at <http://www.natcorp.ox.ac.uk>.
- WebCorp, accessible at <http://www.webcorp.org.uk/live/>.





# La valencia del adjetivo en diccionarios bilingües alemán-español-alemán

Andreu Castell, Natàlia Català, María Bargalló  
Universitat Rovira i Virgili - Tarragona (España)  
andreu.castell.@urv.cat, natalia.catala@urv.cat, maria@bargallo@urv.cat

## Resumen

El verbo constituye el elemento fundamental en la teoría de valencias, por lo que es normal que se le haya prestado siempre una especial atención. Menos atención se le ha dedicado hasta ahora a la valencia del nombre y mucho menos aún a la del adjetivo. El estudio que se presenta forma parte de la fase inicial de un proyecto que se está desarrollando con el fin de establecer las bases teóricas y metodológicas para la elaboración de un diccionario sintáctico del adjetivo en alemán y en español. Se analiza la información sintáctica que ofrecen diversos diccionarios bilingües alemán-español-alemán sobre la complementación de ocho adjetivos que, en principio, poseen los mismos complementos argumentales en ambas lenguas. Se estudia, en concreto, qué información ofrecen respecto de una serie de criterios que desde el punto de vista de la gramática de valencias se deberían considerar fundamentales, así como hasta qué punto la información ofrecida, independientemente de los presupuestos de la teoría de valencias, resulta realmente útil al usuario. También se analiza el grado de coherencia interna que presentan los diccionarios estudiados en las diversas entradas.

**Palabras clave:** lexicografía; adjetivos; valencia del adjetivo

## 1 Introducción

La valencia del adjetivo se trata de forma más o menos detallada en casi todas las gramáticas de la lengua alemana aparecidas en la propia Alemania a partir de los años setenta del siglo pasado: desde Helbig/Buscha (1972: 283-285) y Duden (1973: 229-230) hasta Duden (2009: 361, 927-933), pasando por Eisenberg (1986: 93-95) y Engel (1988: 592-600 y 2004: 353-361). Este último llega a adjudicarle al elemento que nos ocupa ni más ni menos que 16 complementos argumentales distintos. Es más, para la lengua alemana incluso existe, aunque sea el único hasta ahora, un diccionario de valencias del adjetivo (Sommerfeldt/Schreiber 1974).

Respecto del español, a pesar del artículo publicado por Bosque (1983), la atención que se ha prestado a la valencia del adjetivo ha sido, como constatan Garriga/Bargalló (2000: 602), mucho menor: “En muchas ocasiones las gramáticas se limitan a comentar esta cuestión y ofrecen unos pocos ejemplos que no cubren todas las posibilidades, entre ellas, especialmente, la que aquí nos ocupa: el régimen del adjetivo”. No es hasta la aparición de las dos últimas grandes gramáticas de la lengua española (Bosque/

Demonte 1999; Real Academia Española 2009) que ha variado el panorama. Que un adjetivo posea complementos ya se insinúa en gramáticas anteriores: “La cualidad que expresa un adjetivo con respecto al sustantivo a que se refiere puede limitarse o concretarse por medio de un sustantivo precedido de preposición; p.e. *dócil a la advertencia* [...]” (Real Academia Española 1973: 414). Incluso se encuentran, aunque muy excepcionalmente, referencias al respecto en libros de texto para la enseñanza secundaria: “Al igual que el sustantivo, el adjetivo puede recibir complementación mediante un sintagma preposicional. Este sintagma se denomina complemento del adjetivo” (Llácer et al. 1996: 135). Sin embargo, que un adjetivo pueda poseer un complemento indirecto tan solo llega a afirmarse o a insinuarse en las dos gramáticas más recientes: “Concluimos, pues, que tanto sustantivos como adjetivos pueden aparecer con un complemento indirecto” (Bosque/Demonte 1999: 1563); [...] se plantea inmediatamente la cuestión de si los adjetivos y los sustantivos pueden tener complemento indirecto” (Real Academia Española 2009: 2659).

Se constata, en general, como señala Castell (2009: 183 y ss.), una recepción mucho mayor de la teoría de valencias en las gramáticas de la lengua alemana, por lo que podría suponerse que ello debería reflejarse también en los correspondientes diccionarios monolingües. Sin embargo, dejando de lado alguna excepción en el caso del alemán (como Langenscheidt 2003), lo cierto es que dicha suposición no resulta del todo correcta. En un trabajo todavía inédito, realizado como parte del proyecto en el que se enmarca el presente estudio, se constata que por ejemplo el diccionario de Moliner (2008) y el *Diccionario Salamanca* para el español y el *Deutsches Universalwörterbuch* (Duden 2008) y el diccionario *Deutsch als Fremdsprache online* de Pons para el alemán presentan carencias muy similares en cuanto a la información que ofrecen sobre la complementación del adjetivo.

En el presente estudio se analiza qué información sintáctica ofrecen respecto del adjetivo diversos diccionarios bilingües alemán-español-alemán. En principio cabe suponer que los usuarios de dichos diccionarios son fundamentalmente hablantes de la lengua de partida que, aún poseyendo un nivel muy alto de la misma, desconocen del todo o en gran medida cómo funciona la lengua de destino. Por lo tanto, este tipo de diccionarios debería ofrecer una información aún más detallada y específica que la que ofrecen los diccionarios monolingües.

Se analizan a este respecto diccionarios bilingües, tanto en forma impresa como electrónica o accesibles online: *Diccionario de las lenguas española y alemana* (Slaby/Grossmann 1993 y Slaby/Grossmann/Illig 1991) (SLG), *Hueber Wörterbuch. Diccionario Alemán-Español/Español-Alemán* (2007) (HUE), *Langenscheidt e-Taschenwörterbuch Spanisch-Deutsch-Spanisch* (Langenscheidt 2002) (LNG), *Pons Diccionario español-alemán-español online* (PON), *Larousse Diccionario español-alemán-español online* (LAR), *LEO Spanisch-Deutsch-Spanisch online* (LEO), *Reverso Diccionario Collins español-alemán-español online* (REV), *Beolingus online* (BEO) y *DIX online* (DIX).

Los diccionarios citados se han analizado en cuanto a la información sintáctica que ofrecen respecto de ocho adjetivos que, en principio, poseen los mismos complementos argumentales en ambas lenguas:

1) Adjetivos con un complemento argumental preposicional que puede ser realizado mediante una oración: *fähig/capaz*, con el significado de “estar dispuesto o atreverse a algo” e *interessiert/interesado*, con el significado de “mostrar interés por alguna cosa o alguien”:

- (1) Er ist zu *allem* *fähig*.
- (2) Es *capaz de todo*.
- (3) Ich bin *an dem Projekt* *interessiert*.
- (4) Estoy *interesado en el proyecto*.

El complemento argumental preposicional de *fähig/capaz* puede ser realizado también por una oración de infinitivo, el de *interessiert/interesado* admite su realización tanto mediante una oración de infinitivo como mediante una oración subordinada finita con *dass/que*:

- (5) Er war nicht *fähig, seine Meinung zu verteidigen*.
- (6) No fue *capaz de defender su opinión*.
- (7) Ich bin *daran interessiert, an dem Projekt teilzunehmen*.
- (8) Estoy *interesado en participar en el proyecto*.
- (9) Ich bin sehr *daran interessiert, dass Sie das Projekt befürworten*.
- (10) Estoy muy *interesado en que recomiende usted el proyecto*.

2) Adjetivos con un complemento argumental preposicional que no puede ser realizado mediante una oración: *reich/rico*, con el significado de “tener mucho de lo que se expresa”:

- (11) Das Land ist *reich an Energiequellen*.
- (12) El país es *rico en recursos energéticos*.

3) Adjetivos con un complemento argumental direccional o local que no puede ser realizado mediante una oración: *gebürtig/natural*, con el significado de “nacido en un lugar” y *wohnhaft/residente*, con el significado de “habitar en un lugar”:

- (13) Meine Nachbarin ist *aus Ungarn gebürtig*.
- (14) Mi vecina es *natural de Hungría*.
- (15) Beide waren *in Madrid wohnhaft*.
- (16) Ambos eran *residentes en Madrid*.

4) Adjetivos con un complemento argumental dativo y otro preposicional, con los que el último puede ser realizado mediante una oración: *dankbar/agradecido*, con el significado de “dar muestras de gratitud por algún favor recibido”:

- (17) Wir sind *Ihnen für Ihre Unterstützung dankbar*.
- (18) Le estamos *agradecidos por su apoyo*.

En alemán, el complemento argumental prepositivo puede ser realizado mediante una oración subordinada finita con *dass*, mientras que en español resulta más habitual una oración de infinitivo:

- (19) Wir sind Ihnen *dankbar dafür, dass Sie uns unterstützt haben*.
- (20) Le estamos muy *agradecidos por habernos apoyado*.

No está claro que el complemento prepositivo pueda ser realizado en español mediante una oración subordinada finita con *por que*. En el *Corpus de referencia del español actual* (Real Academia Española) se

encuentran numerosos ejemplos de realización mediante oraciones subordinadas de infinitivo, pero también se encuentran ejemplos como los siguientes:

(21) Estuve sonriendo, asombrado y *agradecido porque fuera tan fácil* [...].

(22) Le estaba muy *agradecido porque le hacía las cuentas* [...].

En ellos se tendería a analizar la oración subordinada como un complemento adjunto de causa. Sin embargo, debe tenerse en cuenta que ambas expresiones equivalen a:

(23) Estuve sonriendo, asombrado y *agradecido por el hecho de que fuera tan fácil*.

(24) Le estaba muy *agradecido por el hecho de que hiciera las cuentas*.

Aquí nos hallamos, sin duda alguna, ante un complemento argumental prepositivo, por lo que, en principio, no deberían descartarse como agramaticales expresiones como las siguientes (cfr. Brucart/Gallego (2009: 166-167) y Pavón (2012: §2.1)):

(25) ?Estuve sonriendo, asombrado y *agradecido por que fuera tan fácil*.

(26) ?Le estaba muy *agradecido por que le hiciera las cuentas*.

5) Adjetivos con un complemento argumental dativo que no puede ser realizado mediante una oración: *treu/fiel*, con el significado de “no faltar a sus ideas o a su compromiso con su pareja” y *untreu/in-fiel* como antónimos de los anteriores:

(27) Er ist *seinen Idealen* immer *treu* gewesen.

(28) Siempre ha sido *fiel a sus ideales*.

(29) Ich bin *ihr* immer *treu* gewesen.

(30) Siempre *le* he sido *fiel*.

## 2 ¿Qué información deberían ofrecer los diccionarios?

En los diez diccionarios analizados se ha estudiado lo siguiente:

1) En el caso de los complementos argumentales prepositivos, direccionales y locales se observa si se nombra de forma explícita la preposición correspondiente y, en cuanto al alemán, el caso que ésta rige, de modo que no haya que deducirlo exclusivamente de los ejemplos aportados. Se considera adecuada e imprescindible una indicación del tipo *interessiert an + dat = interesado en*, puesto que *an* puede regir más de un caso. Sin embargo también se consideraría adecuado que únicamente se ofreciera información sobre el caso que rige la preposición cuando ésta puede regir más de un caso, como en el ejemplo anterior, siempre que este proceder se diese de forma sistemática. Sería por tanto justificable que en el caso de *fähig zu* (*capaz de*) o *dankbar für* (*agradecido por*) no se indicara el caso, puesto que las preposiciones *zu* y *für* siempre rigen, respectivamente, dativo y acusativo.

2) En el caso de los complementos argumentales realizables también mediante una oración, se observa si se hace referencia explícita al tipo de subordinada, por ejemplo mediante indicaciones como *NS-dass* o *NSInf* para el alemán u *OSque* u *OSinf* para el español. También se considera aportada esta información si por ejemplo en el caso de *interessiert/interesado* se ofrecen paráfrasis del tipo *daran interessiert sein, etwas zu tun/estar interesado en hacer algo* y *daran interessiert sein, dass jemand etwas tut oder dass etwas stattfindet/estar interesado que alguien haga algo o que suceda alguna cosa*. En estrecha relación con ello, los diccionarios también deberían ofrecer información sobre si la oración principal alemana debe o no contener el elemento catafórico formado mediante *da(r)*+preposición. Esta información es importante, puesto que hay adjetivos que lo exigen (31) y otros en los que es facultativo (33):

(31) Wir sind sehr *daran* interessiert, an dem Projekt teilzunehmen.

(32) Estamos muy interesados en participar en el proyecto.

(33) Er war nicht (*dazu*) fähig, seine Mitarbeiter zu überzeugen.

(34) No fue capaz de convencer a sus colaboradores.

3) En el caso del complemento argumental dativo, se observa si se ofrece alguna paráfrasis que evidencie inequívocamente que se trata de este tipo de complemento. En el caso del alemán se consideraría adecuada una indicación del tipo *jemandem dankbar*, puesto que la forma declinada de *jemand* señala claramente que se trata de un dativo. En cambio, debería considerarse insuficiente, por lo que respecta al español, la indicación equivalente *agradecido a alguien*, puesto que el sintagma preposicional no garantiza que el usuario lo interprete inequívocamente como un complemento indirecto. Una indicación del tipo *agradecido a alguien/le* o *estar(le) agradecido a alguien* sería sin duda más adecuada.

4) En todos los casos se analiza si se indica de alguna manera el hecho de que el complemento argumental en cuestión es obligatorio o facultativo. Al usuario le puede interesar, por ejemplo, si los adjetivos *dankbar/agradecido* pueden aparecer con solo uno de los dos complementos argumentales que rigen o incluso sin ninguno de ellos, sin que por ello cambie el significado del enunciado:

(35) Ich bin Ihnen sehr dankbar für Ihre Hilfe.

(36) Le estoy muy agradecido por su ayuda.

(37) Ich bin Ihnen sehr dankbar.

(38) Le estoy muy agradecido.

(39) Ich bin sehr dankbar für Ihre Hilfe.

(40) Estoy muy agradecido por su ayuda.

(41) ?Ich bin sehr dankbar.

(42) ?Estoy muy agradecido.

O si los adjetivos *interessiert/interesado* pueden aparecer sin su complemento argumental prepositivo:

(43) ?Wir sind sehr interessiert.

(44) ?Estamos muy interesados.

Se considera información adecuada al respecto que se enuncie de forma explícita o que se indique, ya sea en la paráfrasis introductoria o en los ejemplos aportados, poniendo p. e. los complementos argu-

mentales facultativos entre paréntesis (previa explicación en las instrucciones de uso del diccionario). Así, en cuanto a los adjetivos *dankbar/agradecido*, si se considera que ambos complementos argumentales son realmente facultativos, podría indicarse mediante una paráfrasis como (*jemandem*) (*für etwas*) *dankbar/agradecido* (*a alguien/le*) (*por algo*) o poniendo entre paréntesis los complementos argumentales en los ejemplos (35) y (36). Del mismo modo también debería señalarse de alguna manera si el adjetivo debe aparecer como mínimo con uno de ellos, si no fuese posible la elisión de ambos a la vez.

5) También se analiza en todos los casos si se ofrece información sobre restricciones semánticas que afectan al complemento argumental en cuestión. Se considera información explícita al respecto que, independientemente de los ejemplos aportados, pueda deducirse de forma inequívoca si el complemento puede referirse a personas o a cosas o a ambas indistintamente. Se trataría de ofrecer p. e. una paráfrasis como la mencionada en el punto anterior en cuanto a los adjetivos *dankbar/agradecido* u otra del tipo *an etwas/jemandem interessiert /interesado en algo/alguien* en el caso de los adjetivos *interessiert/interesado*.

6) Finalmente se analiza en todos los casos si los diccionarios ofrecen ejemplos unívocos de todo lo anterior. Como tales se consideran únicamente oraciones enteras como *Wir sind an diesem Projekt interessiert/Estamos interesados en este proyecto* o *Wir sind daran interessiert, dass alle an dem Projekt teilnehmen/Estamos interesados en que todos participen en el proyecto* y, en ningún caso, posibles paráfrasis como *an etwas interessiert sein/estar interesado en algo*.

### 3 Los adjetivos y sus complementos en los diccionarios analizados

1) De los nueve diccionarios analizados, solo uno (BEO) prescinde por completo de informar sobre las preposiciones que pueden o deben intervenir en los complementos argumentales según lo definido en el punto (2) del apartado anterior. Eso no significa, sin embargo, que los demás sí lo hagan. En realidad, ninguno de ellos lo hace en cada una de las 12 ocasiones en que deberían informar al respecto (*fähig/capaz*, *interessiert/interesado*, *reich/rico*, *wohnhaft/residente*, *gebürtig/natural* y *dankbar/agradecido*). Si ordenamos los diccionarios de menor a mayor frecuencia, obtenemos el siguiente resultado: DIX (1), REV (2), LEO (6), LAR (6), SLG (6), PON (7), HUE (7) y LNG (9). Por otra parte, en los seis últimos se observa que la información no se ofrece de forma sistemática en las dos partes del diccionario. Valgan los casos de LNG y SLG como ejemplo: En el primero se ofrece la información pertinente de forma bidireccional respecto de *fähig/capaz*, *interessiert/interesado* y *reich/rico* tanto en la parte alemana como en la española; en cambio, respecto de *wohnhaft/residente*, *gebürtig/natural* y *dankbar/agradecido* únicamente se ofrece en la parte alemana. En el caso de SLG solo se ofrece información bidireccional en ambas partes con respecto a *gebürtig/natural*, mientras que la información correspondiente respecto de *fähig/capaz*,

*reich/rico* y *dankbar/agradecido* sólo se ofrece en la parte alemana y la de *wohnhaft/residente* e *interessiert/interesado* únicamente en la parte española.

En cuanto al caso que rige la preposición alemana, exceptuando LEO, que cuando ofrece información sobre la preposición lo hace indicando siempre el caso que rige, incluso cuando ésta solo puede regir un único caso, los demás diccionarios optan mayormente por ofrecer dicha información solo cuando la preposición puede regir dos casos diferentes. Así, en HUE encontramos junto a la indicación “zu etw. fähig sein ser capaz de algo”, en la que no se señala el caso, la indicación “reich an etw. dat rico en algo”, mediante la que se evidencia que la preposición *an* rige dativo. Mientras que algunos diccionarios proceden de forma sistemática al respecto, en otros, como en PON, el hecho de informar sobre el caso parece producirse de forma aleatoria: Los complementos argumentales prepositivos de *interessiert* y *reich* se introducen mediante la preposición *an*, que en principio podría regir tanto acusativo como dativo. Sin embargo sólo se indica el caso concreto en las entradas *interessiert* (“an etw dat/jdm interessiert sein estar interesado en [oder por] algo/alguien”) e *interesado* (“estar interesado en algo an etw dat interessiert sein”), mientras que en *reich* solo se dice: “reich an”. El adjetivo *gebürtig* posee un complemento direccional introducido por *aus*, preposición que siempre rige dativo. Mientras que en la entrada española correspondiente natural se indica “natural de gebürtig aus + dat”, en la entrada del adjetivo alemán solo se lee “gebürtig aus natural de”.

2) La información que ofrecen los diccionarios sobre las distintas posibilidades de realización del complemento argumental solo puede calificarse de muy deficiente. Son tres las parejas de adjetivos que también pueden ser realizados mediante oraciones subordinadas, ya sean finitas o de infinitivo: *fähig/capaz*, *interessiert/interesado* y *dankbar/agradecido*. La información debería ser aportada, pues, en un total de seis casos en todos los diccionarios, por lo tanto en un total de 54 ocasiones. Sin embargo solo se encuentra indicada o al menos insinuada la posibilidad de realización oracional en dos ocasiones y lo cierto es que en solo una de ellas aparece formulada de la manera que se ha descrito más arriba como adecuada. Se trata de la información aportada sobre *fähig* en HUE: “fähig sein, etwas zu tun ser capaz de hacer algo”, paráfrasis que aparece junto a la que haría referencia a la realización nominal o pronominal “zu etw. fähig sein ser capaz de algo”. La segunda se encuentra en PON, donde respecto de *interessiert* se ofrece la siguiente información, que en realidad no es la típica paráfrasis, pero que tampoco llega a la categoría de ejemplo: “ich bin nicht daran interessiert, dass ... no me interesa que... +subj”. Resulta curioso que en la correspondencia en español se opte por la expresión verbal “no me interesa que”, cuando a fin de hacer más comprensible la construcción alemana se habría poder optado perfectamente, al menos complementariamente, por la expresión adjetival paralela “no estoy interesado en que”. Lo cierto es que al consultar en el mismo diccionario la entrada *interesado* tampoco se encuentra referencia alguna a una posible realización oracional.

Ninguno de los diccionarios analizados incide de algún modo en la obligatoriedad o facultatividad del elemento catafórico que debe o puede aparecer en el caso de realización oracional en el caso del ale-



mán. De ahí que el usuario de PON no pueda deducir de ninguna manera si el adverbio pronominal *darán* en la indicación que se acaba de mencionar es obligatorio o no.

3) Son tres las parejas de adjetivos que poseen un complemento indirecto: *dankbar/agradecido*, *treu/fiel* y *untreu/infiel*. En SLG, LEO y DIX no se encuentra referencia alguna al respecto. En los demás, la información se ofrece de forma dispersa y poco sistemática. Ofrecen información parcial HUE (para *agradecido* y *fiel*), REV (para *treu*) y BEO (para *treu/fiel*). Se trata de indicaciones como “jemandem treu sein ser fiel a alguien” (BEO), en las que solo resulta evidente que se trata de un complemento indirecto en alemán. Solo en LNG (para *dankbar* y *untreu*) y en PON y LAR (para *treu* y *untreu*) se encuentran formulaciones que evidencian que también en español se trata de un complemento indirecto: “estar(le) a alg agradecido por a/c” (LNG) o “serle infiel a alguien” (LAR). La falta de coherencia interna de los diccionarios también resulta evidente aquí, al comparar la última información en LAR con la que el mismo diccionario ofrece respecto de *dankbar*: “jm (für etw) dankbar sein estar agradecido (f agradecida) a alguien (por algo)” (LAR). Nótese que aquí no se incluye, como en el caso de *infiel*, el pronombre *le*.

4) Una posible indicación en cuanto a la obligatoriedad o facultatividad del complemento argumental, tal y como se define como adecuada en el apartado 2, solo se encuentra en HUE y LAR, concretamente respecto de *dankbar/agradecido*. Así, en la paráfrasis definitoria de *dankbar* (no en la de *agradecido*) en LAR se dice: “jm (für etw) dankbar sein estar agradecido (f agradecida) a alguien (por algo)”. En HUE se encuentra el mismo tipo de indicación, pero en este caso en la entrada de *agradecido* (no en la de *dankbar*): “estar agradecido con alguien (por algo) jdm (für etw.) dankbar sein”. Podría deducirse de ello que los dos diccionarios consideran, al colocarlo entre paréntesis, que el complemento argumental prepositivo de *dankbar/agradecido* es facultativo en ambas lenguas. Sin embargo, no existe en los citados diccionarios ninguna instrucción de uso que explique el significado de estos paréntesis. Si se consultan en ellos otros adjetivos con los que el complemento argumental prepositivo debería considerarse facultativo, como por ejemplo *stolz/orgullosa* o *verliebt/enamorado*, no se halla ninguna indicación similar. Y lo mismo sucede si se consulta en ellos la descripción de verbos como por ejemplo *sich ärgern* (*enfadarse*). En ninguno de los dos diccionarios aparece el complemento argumental prepositivo entre paréntesis con el fin de indicar que es facultativo: “sich über etw./jdn ärgern enfadarse por algo/con alguien” (HUE); “sich über etw./jn ärgern enfadarse por algo/por culpa de alguien” (LAR). Podría deducirse de ello que los paréntesis arriba mencionados son debidos a la mera casualidad.

5) En todos los diccionarios, excepto en REV y DIX, se encuentra alguna información sobre restricciones semánticas que afectan a los complementos argumentales de los adjetivos *fähig/capaz*, *interessiert/interesado*, *reich/rico*, *dankbar/agradecido*, *treu/fiel* y *untreu/infiel*. Valga como ejemplo lo que se dice en LNG respecto de *dankbar*: “jemandem für etwas dankbar sein estar(le) a alg agradecido por a/c”. Pero también aquí se observa que los diccionarios no proceden de forma sistemática, como lo demuestra la simple comparación cuantitativa. Del total de 12 ocasiones en que podría ofrecerse este tipo de infor-



mación, ninguno de los diccionarios lo hace en todas ellas. Si se ordenan de menor a mayor frecuencia, el resultado es el siguiente: SLG (1), BEO (1), LEO (2), LNG (2), PON (4), LAR (7) y HUE (8). Incluso en el caso de este último resulta difícil de entender por qué en cuanto a *fähig/capaz*, *interessiert/interesado* y *reich/rico* se ofrece dicha información de manera bidireccional tanto en la parte alemana como en la española, mientras que en el caso de *dankbar/agradecido* y *treu/fiel* solo se hace en la parte española, prescindiendo además por completo de dicha información por lo que respecta a *untreu/infiel*. Lo cierto es que, incomprensiblemente, el adjetivo alemán *untreu* ni tan solo aparece en el diccionario.

6) Contrariamente a lo que sería deseable en diccionarios bilingües, la mayoría de diccionarios analizados prescinden totalmente (REV) o al menos en gran medida de ofrecer ejemplos de uso de los adjetivos estudiados. Entre los últimos se encuentran SLG y LAR, que solo ofrecen un único ejemplo de los 16 posibles, y LNG y DIX, que solo los ofrecen en dos ocasiones. Únicamente HUE y PON se acercan a lo que sería deseable, al ofrecerlos en ocho ocasiones. Se trata de ejemplos como: “Er ist zu allem fähig”, “Das Essen ist reich an Kohlehydraten”, “Ich wäre Ihnen für eine schnelle Antwort sehr dankbar” e “Ihr Mann war (ihr) untreu” (HUE) o “estos muchachos son capaces de todo diese Burschen sind zu allem fähig”, “er ist aus Valencia gebürtig es natural de Valencia” e “ich weiß, dass mein Mann mir treu ist sé que mi marido me es fiel” (PON). Cabe señalar que HUE solo ofrece ejemplos en alemán, proceder que podría considerarse justificado por el hecho de ir dirigido a usuarios que aprenden alemán y que por lo tanto ya saben español, pero que en última instancia puede resultar contraproducente, si se tiene en cuenta que dichos usuarios consultan el diccionario precisamente por desconocer dicho idioma, lo que puede comportar que no acaben de entender los ejemplos ofrecidos. LEO y BEO, que no se caracterizan precisamente por aportar una información detallada sobre la complementación, ofrecen un sinfín de ejemplos reales, que no pueden considerarse realmente útiles al ofrecerse de forma totalmente asistemática.

## 4 Conclusión

Todos los diccionarios analizados ofrecen una información claramente deficiente respecto a la complementación del adjetivo. Entre las deficiencias cabe destacar sobre todo que no ofrecen siempre información detallada sobre la preposición que introduce el complemento y que la información sobre la posible realización oracional de los complementos y sobre su obligatoriedad es prácticamente inexistente. En todos ellos se observa además que son poco sistemáticos, puesto que en determinados casos ofrecen cierta información al respecto, mientras que en otros prescinden totalmente de ello, sin que dicho proceder pueda atribuirse a unos parámetros determinados y lógicos. Por otra parte se constata que algunos se limitan a actuar como simples traductores de palabras sueltas, sin ofrecer al usuario indicaciones respecto a su uso. Del mismo modo se constata que la falta de ejemplos es una característica común a la gran mayoría de ellos. En los dos diccionarios en que se ofrece una gran variedad de

ejemplos reales (LEO y BEO), esta característica no puede considerarse positiva, puesto que al usuario le puede resultar muy difícil encontrar lo que busca.

## 5 Referencias

### 5.1 Diccionarios analizados

- Beolingu online*. Acceso: <http://dict.tu-chemnitz.de> [10/03/2014].
- DIX online*. Acceso: <http://dix.osola.com/index.es.php> [10/03/2014].
- Hueber Wörterbuch. Diccionario Alemán-Español / Español-Alemán (2007). Ismaning: Hueber.
- Langenscheidt (2002). *e-Taschenwörterbuch Spanisch-Deutsch-Spanisch* (CD Version 4.0). Berlin/München: Langenscheidt.
- Larousse Diccionario español-alemán-español online*. Acceso: <http://www.larousse.com/es/diccionarios/allemand-espanol> [10/03/2014].
- LEO Spanisch-Deutsch-Spanisch online*. Acceso: [http://dict.leo.org/esde/index\\_de.html](http://dict.leo.org/esde/index_de.html) [10/03/2014].
- Pons Diccionario español-alemán-español online. Acceso: <http://es.pons.eu/> [10/03/2014].
- Reverso Diccionario Collins español-alemán-español online*. Acceso: <http://diccionario.reverso.net/aleman-espanol> [10/03/2014].
- Slaby, R., Grossmann, R. (1993) Diccionario de las lenguas española y alemana I. Español-Alemán. Barcelona: Herder.
- Slaby, R., Grossmann, R., Illig, C. (1991) Diccionario de las lenguas española y alemana II. Alemán-Español. Barcelona: Herder.

### 5.2 Otras

- Bosque, I. (1983). El complemento del adjetivo. En *Lingüística Española Actual* V, pp. 1-14.
- Bosque, I., Demonte, V. (eds.) (1999). *Gramática descriptiva de la lengua española*. Barcelona: Ariel.
- Brucart, J.M., Gallego, Á.J. (2009). "L'estudi formal de la subordinació i l'estatus de les subordinades adverbials". En: *Llengua & Literatura*, 20, pp. 139-191.
- Castell, A. La recepción de la teoría de valencias en España. Reflexiones en torno a la adaptación al español de los términos 'Ergänzungen' y 'Angaben'. En *Revista de Filología Alemana* 17, pp. 183-204.
- Diccionario Salamanca de la lengua española*. Acceso: <http://fenix/cnice.mec.es/diccionario/> [10/03/2014].
- Duden (1973). *Grammatik der deutschen Gegenwartssprache*. Mannheim: Bibliographisches Institut.
- Duden (2008). *Deutsches Universalwörterbuch*. (CD Version 5.0). Leipzig: Bibliographisches Institut & F.A. Brockhaus AG.
- Duden (2009). *Die Grammatik*. Mannheim/Wien/Zürich: Dudenverlag.
- Eisenberg, P. (1986). *Grundriss der deutschen Grammatik*. Stuttgart: J.B. Metzlersche Verlagsbuchhandlung.
- Engel, U. (1988). *Deutsche Grammatik*. Heidelberg: Julius Groos Verlag.
- Engel, U. (2004). *Deutsche Grammatik – Neubearbeitung*. München: Iudicium.
- Garriga, C., Bargalló, M. (2000). Gramáticas y diccionarios: el régimen del adjetivo. En *Euralex 2000 Proceedings*, pp. 601-609.
- Helbig, G., Buscha, J. (1972). *Deutsche Grammatik*. Leipzig: VEB Verlag Enzyklopädie.
- Langenscheidt (2003). *e-Großwörterbuch Deutsch als Fremdsprache* (CD Version 4.0). Berlin/München: Langenscheidt.

- Llácer, I. et al. (1996). *Lengua Española. Nivel COU*. Fuente del Jarro-Paterna: Ecir Editorial.
- Moliner, María (2008). *Diccionario de uso del español* (CD versión 3.0). Madrid: Gredos.
- Pavón, M.V. (2012). *Estructuras sintácticas en la subordinación adverbial*. Madrid: Arco Libros.
- Pons. *Deutsch als Fremdsprache online*. Acceso: <http://www.pons.de/produkte> [10/03/2014].
- Real Academia Española (1973). *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa Calpe.
- Real Academia Española: Banco de datos (CREA) (en línea). Corpus de referencia del español actual. Acceso: <http://www.rae.es> [10/03/2014.]
- Real Academia Española – Asociación de Academias de la Lengua Española (2009). *Nueva gramática de la lengua española*. Madrid: Espasa Libros.
- Sommerfeldt, K-E., Schreiber, H. (1974). *Wörterbuch zur Valenz und Distribution deutscher Adjektive*. Leipzig: VEB Bibliographisches Institut.

### **Reconocimientos**

La investigación que subyace a este trabajo ha sido parcialmente financiada por el Proyecto FFI2012-32705.



# Esame storico dei “realia” nei dizionari bilingui italiano/ungheresi

Zsuzsanna Fábán

Università “Eötvös Loránd” di Budapest, Istituto di Romanistica

fabian.zsuzsanna@btk.elte.hu

## Abstract

Lo scopo del presente lavoro (di carattere lessico-contrastivo ma anche storico-lessicografico) è quello di analizzare l'evoluzione del trattamento dei “realia” (parole culturo-specifiche, *lexical/referential gaps*) in otto dizionari bilingui italiano-ungheresi e in sei ungherese-italiani. Per l'esame sono stati scelti 16 esponenti italiani e altrettanti ungheresi appartenenti al lessico della gastronomia (i lemmi sono presentati nelle due Appendici). Le singole equivalenze registrate nei dizionari sono state assegnate alle varie categorie delle possibili strategie traduttive quali l'uso di una parola di origine straniera (forestierismi, prestiti o calchi), quello di un equivalente (sintagmatico) esplicativo (distinto ev. anche graficamente, p.es. col corsivo o con la collocazione tra parentesi, da un “vero” equivalente), o di un equivalente generalizzante (iperonimo), l'adattamento o la sostituzione con un equivalente approssimativo/parziale di L2, ecc. Nella parte conclusiva dell'analisi sono state indagate le problematiche seguenti: Quali sono le strategie traduttive più frequenti nei dizionari bilingui tra italiano e ungherese? E' possibile individuare, nei singoli autori dei dizionari analizzati, una preferenza verso certe strategie traduttive? Nel caso di eventuali errori si possono registrare “copiature” e/o “correzioni”? Sussiste una connessione tra le strategie dominanti nei dizionari bilingui e la direzione delle lingue/tipo del vocabolario (attivo o passivo)?

**Keywords:** bilingual dictionaries; lexical gaps; Italian & Hungarian

## 1 I “realia” nella traduzione e come problema lessicografico

A partire da Vlahov & Florin,<sup>1</sup> il concetto dei “realia” è stato studiato da vari punti di vista e numerose sono le definizioni e le categorizzazioni (Rey 1986; Lendvai 1986; Tellingner 2003; Klaudy 2005 ecc.), spesso elaborate anche su basi empiriche (analisi contrastive di testi concreti tradotti). In questo saggio il termine *realia* sarà inteso sia nel senso di ‘referente caratteristico in un dato ambiente culturale’, sia come ‘segno linguistico indicante un referente di questo tipo’ (v. p.es. *Realienlexeme* in Kujamäki 2004: 920; *realialexéma* in Heltai 2013: 33; e ancora Klaudy 2005: 60). Nella letteratura internazionale il fenomeno è spesso designato, oltre che dal termine riassuntivo già evocato, anche da espressioni au-

1 Vlahov, S. & Florin, S. (1970). *Neperovodimoe v perevode. Realii*. In Sovetskij pisatel', Mosca, pp. 432-456.

tonome per i due diversi concetti quali *referential gap* (Svensén 2009: 271), *cultural gap* risp. *lexical gap* (v. p es. Svensén 2009: 271-275).

Le difficoltà che emergono a proposito dei realia si manifestano prima di tutto nel processo di mediazione tra due culture, e quindi, tipicamente, nelle traduzioni. È perciò evidente che una descrizione dettagliata e una ricca sistematizzazione, oggi generalmente condivisa dagli studiosi, sono state elaborate nell'ambito della traduttologia (Heltai 2001, House 2004, Kujamäki 2004, Klaudy 2005, 2013). Nello stesso tempo sono stati studiati anche i risvolti lessicografici dei problemi di equivalenza che le lacune lessicali pongono agli autori dei dizionari bilingui (p.es. Tomaszczyk 1984; Rey 1986; Marello 1989: 52-54; Rey 1991; Uzonyi 2006; Éber & Fata 2009-2010; Fata 2010). Alcuni studiosi analizzano anche i tipi delle commutazioni adoperate sia nella traduzione che nella prassi lessicografica (Szabó 2008; Fábíán 2013).

In questo saggio le categorie dell'esotizzazione (ingl. *foreignization*) e della localizzazione (o adattamento, ingl. *domestication*) saranno rispettivamente intese come "strategia traduttiva consistente nell'introdurre o conservare elementi culturali appartenenti a una cultura diversa da quella ricevente" (Osimo 2003: 201), e come "strategia traduttiva in base alla quale i traduttori eliminano gli elementi culturali appartenenti a una cultura diversa da quella ricevente [...] sostituendovi elementi della cultura ricevente" (Osimo 2003: 210; Varga-Mujzer 2009). Esaminando invece le diverse classificazioni dei possibili cambiamenti traduttivi (descritti nei dettagli da Osimo 2003; Klaudy 2005: 43-171, e ancora in Heltai 2001: 12; Uzonyi 2006: 121; Varga-Mujzer 2009: 6; Éber & Fata 2010: 41 sulle orme di Tellingier 2003; ecc.) possiamo ribadire che nei dizionari sono adoperati, più frequentemente, i seguenti metodi: uso di una parola straniera, conio e uso di un prestito, conio e uso di un calco, uso di un equivalente (sintagmatico) esplicativo, uso di un equivalente generalizzante (iperonimo), adattamento o sostituzione con un equivalente approssimativo o parziale di L2. Szabó (2008: 62) considera i procedimenti elencati "surrogati di equivalenza".

## 2 Descrizione del corpus scelto per l'analisi e della metodologia

Il presente lavoro si pone il duplice scopo di indagare i tipici trattamenti dei realia nei dizionari bilingui italiano / ungheresi e di osservare, nello stesso tempo, se e come i cambiamenti traduttivi siano mutati nell'arco dell'ultimo secolo, periodo in cui i vocabolari presi in esame sono stati pubblicati. Si tratta di otto dizionari italiano-ungheresi e di sei ungherese-italiani, a partire dal 1912. (Per la storia dei vocabolari italiano / ungheresi v. Fábíán 2011. Le sigle che contraddistinguono i dizionari nel presente lavoro sono elencate nella parte bibliografica.)

I vocaboli scelti per il corpus (16 parole-esponenti sia nell'una che nell'altra direzione) appartengono al lessico della gastronomia: un campo in cui la presenza dei realia culturo-specifici è prevedibile. Si tratta dei seguenti vocaboli italiani: *barbera*, *cannelloni*, *chianti*, *gorgonzola*, *grissino*, *lasagna*, *marsala*,

*mascarpone, mortadella, pandoro, panettone, panforte, pastasciutta, pizza, raviolo, tortellino*. Le parole ungheresi (seguite da una breve spiegazione del significato tra virgolette) sono invece le seguenti: *beigli* ('dolce tipico di Natale, a forma di rotolo, con ripieno di noci o di semi di papavero macinati'), *fröccs* ('una specie di spritz preparato con vino e acqua di seltz'), *gulyás* ('gulasch: zuppa a base di carne di manzo e verdure'), *halászlé* ('zuppa di pesce alla paprica'), *kadarika* ('tipo di vino rosso, da tavola'), *kovászos uborka* ('cetrioli interi preparati attraverso fermentazione'), *lángos* ('leggera pasta lievitata, fritta in abbondante olio, dalla forma rotonda o quadrata'), *paprikás csirke* ('pezzi di pollo in umido, alla paprica'), *pogácsa* ('pasticcino/focaccina rotondo/a, salato/a, ev. al formaggio o di altri gusti'), *rétes* ('strudel'), *szaloncukor* ('cioccolatino riempito di creme di vari gusti, tipico di Natale perché - incartato - adorna l'albero'), *tejberizs* ('riso al latte'), *tejföl* ('panna acida'), *tokaj(i)* ('vino di Tokaj'), *töltött káposzta* ('crauti/cavoli farciti'), *túrós csusza* ('pasta cotta a forma di rettangoli cc. 3 x 3 cm, con sopra ricotta e panna acida').

Quanto alla presenza dei lemmi selezionati per l'analisi possiamo affermare che nella direzione italiano-ungherese sono tutti compresi in KÓ, H e H & J. Il fatto che in U mancano invece 12, e in KK1 11, delle parole scelte per il nostro corpus,<sup>2</sup> può essere spiegato dal diverso formato dei dizionari: i primi tre sono i "grandi" dizionari, gli ultimi due invece di mole molto minore. Nella direzione ungherese-italiana tutti i vocaboli scelti sono presenti in G & S; nei "grandi" mancano invece due lemmi in KKJ e solo uno in KKJ & J.<sup>3</sup> Interessante è la situazione relativa ai due dizionari bidirezionali per turisti: mentre in F & V si trovano 13-13 lemmi nelle due direzioni, in H & I abbiamo 12 lemmi nella direzione italiano-ungherese ma solo 6 in quella ungherese-italiana,<sup>4</sup> fatto che attesta la non completa bidirezionalità di questa opera.

Le equivalenze indicate nei singoli vocabolari sono espone in dettaglio nelle due Appendici finali.

In questa analisi gli esponenti appaiono in corsivo (e possono risultare in corsivo anche le equivalenze, se presenti in questo carattere nei dizionari); le equivalenze sono precedute dal segno dell'uguaglianza (=). Le traduzioni delle parti in ungherese sono state collocate tra due apici ('...').

### 3 Risultati dell'analisi

#### 3.1 I cambiamenti traduttivi più frequenti nei dizionari bilingui

La scelta tra i diversi cambiamenti traduttivi operata dal lessicografo è connessa prima di tutto con la mole del dizionario: in uno più ridotto essa sarà vincolata ad una gestione più parsimoniosa dello

2 In U sono presenti solo gorgonzola, lasagna, mortadella, tortellino; in KK1 invece lasagna, marsala, mortadella, panforte, pizza.

3 In KKJ *töltött káposzta* e *túrós csusza* e in KKJ & J *túrós csusza*.

4 Sono presenti in H & I *fröccs*, *gulyás*, *halászlé*, *rétes*, *tejberizs*, *tejföl*.

spazio (si tenderà, per esempio, ad evitare l'uso delle circoscrizioni esplicative, a favore degli adattamenti o dei prestiti, delle soluzioni generalizzanti con un iperonimo, ecc.).

Non tutti i cambiamenti traduttivi possono essere adottati anche nel lavoro lessicografico. Non è qui possibile, per esempio, ricorrere al metodo del tralascio perché a questo equivarrebbe, nel dizionario, la cancellazione del lemma stesso. D'altronde, è possibile solo nel dizionario la combinazione di più cambiamenti traduttivi: è molto frequente, per esempio, la combinazione dell'uso di un prestito o di un adattamento localizzante con una spiegazione esplicativa.

Nei seguenti paragrafi saranno elencati i cambiamenti traduttivi più frequentemente adoperati nei dizionari, illustrati da esempi tratti dai dizionari in esame.

### 3.1.1 Uso di una parola di origine straniera

Ci sono più possibilità per l'uso di una parola di origine straniera anche come equivalenza in L2, ma viste le peculiarità delle due lingue e per il fatto che il lessicografo basa il suo lavoro più su materiali scritti che sul parlato, nella direzione italiano-ungherese sarebbe difficile, se non impossibile annoverare le equivalenze di questo tipo nelle classiche sottocategorie delle parole di origine straniera (forestierismo: senza adattamento a L2, prestito: adattato formalmente a L2, calco: traduzione). Gli esponenti italiani che appaiono anche come equivalenti in L2 sono *chianti* e *mortadella* (in 5 dizionari), *gorgonzola* (in 4), *pizza* (in 3), *tortellini* (in 2), *lasagna*, *mascarpone*, *panettone* (in 1). È H & J il dizionario in cui questo tipo è maggiormente presente (infatti, vi appaiono tutte le equivalenze indagate), ed è sempre questo dizionario ad usare per la prima volta *pizza* e *tortellini* anche in funzione di equivalenti ungheresi. Le equivalenze nel caso di *marsala* rientrano invece in una categoria mista: l'aggettivo derivato (con *-i*) dal nome della città siciliana appare sia nella forma originaria (= *marsalai* KK1, Ki) che in forma trascritta (= *marsalai* H, H & J), ed accompagna in tutti i casi il nome *bor* 'vino' (iperonimo). Potrebbe essere considerato un "falso calco" anche il caso di *pastasciutta* = *száraztészta* 'pasta non ancora cotta' (H, H & J). Nella direzione ungherese-italiano sono i due nomi di vini (*tokaji* in 4 e *hadarka* in 2 dizionari) e l'internazionalismo = *gulasch* per *gulyás* (in 4 dizionari) a rientrare in questa categoria delle equivalenze.

### 3.1.2 Uso di un equivalente sintagmatico esplicativo

In mancanza di un equivalente lessicale in L2 il lessicografo può scegliere di offrire una spiegazione sintagmatica che descriva i tratti essenziali del denotatum in L1. Questo metodo è adoperato molto spesso dai lessicografi, come attestato anche dagli innumerevoli casi riscontrati nel nostro corpus. Gli

---

5 Il caso di *pizza*, ormai radicato nell'ungherese, presenta i seguenti problemi. Quanto alla pronuncia, convivono oggi [pitstsa] (come in it. alzo) e [pizza] (come in it. viso), ma appare anche [pidzda] (come in it. gazza), di cui fortemente concorrenziali le prime due. Se nel tempo dovesse vincere la seconda, la scrittura potrebbe rimanere inalterata (ma nascerebbe "un falso amico", nella pronuncia, tra it. e ungh.). Se dovesse invece vincere la prima variante di pronuncia (quella "all'italiana"), per poterla annoverare tra i classici prestiti bisognerebbe trascriverla secondo le regole dell'ungh. in *picca* (difficilmente ipotizzabile, vista anche la ormai globale diffusione dell'originaria forma italiana).



equivalenti esplicativi sono spesso differenti tra loro. *Panettone* è, secondo Kó, = lisztből, cukorból, száfrányból készített és sörrel kelesztett sütemény ('dolce preparato con farina, zucchero, zafferano e lievitato con la birra'), per H è = mazsolás, cukrozott gyümölcsös milánói kalács" ('focaccia milanese con uva sultanina e canditi') e per H & I = *kuglófhoz hasonló, milánói eredetű karácsonyi sütemény*" ('dolce natalizio di origine milanese simile al Gugelhupf<sup>6</sup>'); *paprikás csirke* è = pollo in umido colla paprica" in G & S, = pollo con intingolo di paprica, spezzatino con paprika" in KKJ, e ancora = pollo alla paprika, spezzatino di pollo in KKJ & J e = pollo all'arrabbiata, pollo all'ungherese in F & V.

Le esplicazioni appaiono spesso anche come aggiunte agli adattamenti localizzanti (*panettone* = kuglóf (*karácsonyi édesség* 'dolce natalizio') (F & V)), agli iperonimi generalizzanti (*raviolo* = étel (vagdalthús- és túrósgombócból) 'pietanza (fatta da gnocchi di carne macinata o di ricotta') (Ki), e alle parole di origine straniera in L2 (*panettone* = panettone [karácsonyi milánói kuglóf 'Gugelhupf natalizio milanese'] (H & J).

### 3.1.3 Uso di un equivalente generalizzante (iperonimo)

Questo metodo appare molto frequentemente nei dizionari bilingui, prima di tutto in quelli di piccola mole. Un iperonimo può apparire come unico equivalente (spesso p. es. in Király: *gorgonzola* = sajt fajta 'tipo di formaggio', *panforte* = édestésztafajta 'tipo di dolce'). Siccome un unico equivalente iperonimico dice spesso troppo poco, come indicazione orientativa i lessicografi possono aggiungervi un attributo (*pizza* = zsíros sütemény 'un dolce che contiene grasso' Kó). Si cerca di indicare l'iperonimo più vicino, ma non mancano i casi in cui questo risulta troppo generalizzante (forse il lessicografo non conosce il denotatum?) perché situato, nella gerarchia tassonomica, diversi livelli sopra e l'equivalente ha dunque bisogno di un'aggiunta esplicativa (Ki: *raviolo* = étel (vagdalthús- és túrósgombócból) 'pietanza (fatta da gnocchi di carne macinata o di ricotta'). Nel caso dei nomi di vini (ad eccezione di *tokaji*) viene quasi sempre aggiunto l'iperonimo *vino* risp. *bor*.

Un iperonimo può servire anche come completamento di altri tipi di equivalenza, come nel caso di forestierismi non ancora diffusi in L2 (*gorgonzola* = gorgonzóla (sajt) 'formaggio' Kó, U; *mortadella* = mortadella (felvágott) 'affettato' H, H & J, F & V) o nel caso di eventuali fraintendimenti *tortellino* = húsos tásk<sup>8</sup> (étel) 'fardellino di pasta farcito di carne (pietanza)' Ki).

### 3.1.4 Adattamento o sostituzione con un equivalente approssimativo o parziale di L2

Nel caso dell'adattamento il lessicografo cerca di trovare l'equivalente che più somigli all'originale, che riesca a rievocarlo, ad avvicinarsi ad esso. Nel nostro corpus appartengono a questo sottotipo i casi seguenti: *cannelloni* = palacsinta 'crêpe' (H, F & V), *grissino* = sütemény 'dolce' (Kó), = ropi 'stick salati' (H & J), *mortadella* = disznósajt 'sopressata' (Kó, KK1, Ki), = májas hurka 'salsiccia al fegato' (U), *panettone*

6 Siccome i dizionari ungherese-italiani traducono ungh. *kuglóf* proprio con *panettone*, conviene usare qui il corrispondente tedesco.

7 Sic!

8 Il sintagma *húsos tásk* potrebbe infatti significare, per la ricca polisemia della parola *táska*, 'una borsa con dentro la carne' o 'destinata al trasporto, alla custodia della carne' ecc.

= sütemény ‘dolce’ (Kő), = kalács ‘pane dolce, lievitato, fatto a trecce’ (H), *pastasciutta* = makaróni ‘maccheroni’ (Kő, Ki), *pizza* = lángos ‘leggera pasta lievitata frita in abbondante olio’ (KK1, Ki, H, H & J), *lángos* = crostino (G & S), = pizza (KK2, KKJ), *rétes* = millefoglie (KKJ, KKJ & J).

In base ai casi riscontrati va ribadito che il metodo dell’adattamento può essere considerato “pericoloso” nei dizionari perché indirizza l’utente verso false direzioni. L’adattamento è quindi da evitare nei dizionari bilingui.

### 3.2 Preferenze dei singoli lessicografi quanto ai cambiamenti traduttivi

Nel primo “grande” vocabolario italiano-ungherese l’autore Kőrösi predilige ampie spiegazioni esplicative. Il metodo armonizza con lo stile generale del suo dizionario, ritenuto da Rezső Honti, anch’egli italianista e contemporaneo di Kőrösi, un’opera “di stile vivace, loquace” che vale la pena non solo di usare come vocabolario ma di leggere per puro diletto.<sup>9</sup> (Trattandosi di nomi di pietanze, è lecito dire che in molti casi le definizioni lessicografiche sfociano in vere e proprie ricette.<sup>10</sup>) In questo dizionario in più della metà dei lemmi esaminati viene usato il metodo esplicativo, e anche gli adattamenti e i forestierismi sono spesso accompagnati da aggiunte chiarificatrici (*pastasciutta* = (leves helyett tálalt) makaróni ‘maccheroni (che si mangiano come prima portata’)).

Nei dizionari di Urbanek, Gelletich & Sirola e di Koltay-Kastner 1 e 2, forse anche a causa della mole minore, domina invece il metodo dell’adattamento (*tortellino* U = táska, *pizza* KK1 = lángos, *gulyás* G & S = guazzetto).

Quanto a Király, colpisce una marcata preferenza per una soluzione generalizzante che viene usata all’incirca nella metà delle equivalenze (*gorgonzola* = sajtfafta ‘tipo di formaggio’, *panettone* = édestésztaféle ‘tipo di pasta dolce’): nei suoi lemmi troviamo infatti spesso gli elementi *-féle*, *-fafta* ‘un tipo di’ che seguono l’iperonimo.

Herczeg adoperava tutti i metodi possibili per le equivalenze, spesso utilizzandoli anche in combinazione. In H & J l’approccio diventa teoricamente più fondato e sistematico, nel senso che le spiegazioni sintagmatiche saranno distinte dalle vere e proprie equivalenze lessicali: le spiegazioni sintagmatiche metalinguistiche sono inserite tra i segni < >, le aggiunte agli adattamenti solo approssimativi sono invece collocate tra parentesi quadre: *chianti* = <toszkán vörösbor> chianti ‘vino rosso della Toscana’, *cannelloni* = [hússal, sajttal stb. töltött] tésztaarolád ‘rotolo di pasta [riempito di carne, formaggio ecc.]’.

Nei “grandi” dizionari tra ungherese e italiano di Koltay-Kastner (KKJ e KKJ & J) ritornano i metodi considerati sopra a proposito dei dizionari di Herczeg: *fröccs* = <vino e acqua gasata>. Tale affinità metodologica può essere spiegata con il semplice fatto che i dizionari di Herczeg e di Koltay-Kastner sono stati redatti e pubblicati all’incirca negli stessi anni e dalla stessa casa editrice.

9 Honti, R. (1912). Recensione a *Olasz-magyar szótár* di Kőrösi. In *Magyar Nyelvőr* X (3), p. 471.

10 P.es. *raviolo* = vajjal, túróval, fűszerszámmal és vagdalt hússal töltött toklóhúsból készült étel (‘pietanza a base di carne di agnello, con burro, ricotta e spezie’)

Infine, nei due dizionari destinati ai turisti dominano le spiegazioni esplicative in corsivo. F & V è il primo dizionario tra italiano e ungherese in cui queste sequenze metalinguistiche sono segnalate in corsivo, per distinguerle nettamente dalle equivalenze vere e proprie: *panforte* = (*sienai édesség*) ‘dolce tipico di Siena’, *fröccs* = (*due decilitri di vino con un decilitro di acqua di seltz*). Il metodo del corsivo (ma senza parentesi) sarà adoperato anche nella direzione italiano-ungherese in H & I: *panforte* = *gyümölcskenyérhez hasonló sienai karácsonyi sütemény* ‘dolce di Natale simile al panfrutto tipico di Siena’.

Sarebbe fuorviante dedurre da questa breve panoramica sulle strategie traduttive adottate nei dizionari bilingui che esse siano rimaste immutate nell’arco dei cento anni trascorsi: pare che fossimo partiti, quanto ai realia, dal metodo esplicativo (usato con virtuosismo dal Kőrösi) e che sia proprio questo anche oggi quello adoperato di preferenza dai lessicografi (v. i due dizionari per turisti). Bisogna tuttavia considerare l’ultima fase del processo descritto come risultato dell’evoluzione della teoria lessicografica, in quanto le equivalenze esplicative appaiono nei dizionari di oggi come sequenze metalinguistiche (e quindi in corsivo e/o tra parentesi ecc.) e non come equivalenze sintagmatiche di tipo circoscrittivo.

### 3.3 Connessioni tra il tipo del referente e la strategia traduttiva

Considerato il numero relativamente esiguo degli esempi presi in esame, possiamo solo ribadire che i nomi dei vini risalenti a toponimi, quali i deonomastici chianti, risp. tokay, tokaj ecc. sono adoperati, in ambedue le direzioni, come forestierismi in L2.

### 3.4 Errori di equivalenza: “copiature” e “correzioni”

Il carattere diacronico dell’analisi non solo rende palesi le equivalenze poco fortunate o addirittura sbagliate, che sono state eventualmente riprese dai posteri, ma mette in luce anche le innovazioni o le correzioni degli errori. Le prime caratterizzano tipicamente i prodotti precoci della direzione italiano-ungherese e sono connesse all’utilizzo, da parte del lessicografo, della strategia dell’adattamento localizzante (spesso accompagnato da aggiunte esplicative).

Di seguito sono elencati alcuni altri casi che mostrano anche le fasi della correzione delle equivalenze.

Per *grissini Kó* adopera l’equivalente generalizzante = *sütemény* ‘dolce’, e anche se vi aggiunge l’attributo *ropogós* ‘croccante’, si tratta di una soluzione sbagliata; Ki introduce un altro equivalente fuorviante: = *rúdkenyér* ‘pane a forma allungata, filoncino, baguette’, che viene solo mutato nell’inesistente e sempre sbagliato = *kenyerrúd* ‘ibid.’ in H (e ripreso ancora in F & V nell’equivalente sintagmatico esplicativo); in H & J e H & I, invece, gli equivalenti esplicativi risultano finalmente corretti. (È da notare come *grissini* stia, nel frattempo, diventando prestito nell’ungherese.)

La parola *lasagna* viene interpretata da Kó come ‘pasta per minestra,<sup>11</sup> in forma di larghe tagliatelle’, e – forse sulla sua scia – anche i posteriori (U, KK1, Ki, H) la fanno equivalere a = metélt ‘tagliatelle, fettuccine’. Senza ulteriori passaggi intermediari l’errore viene poi corretto in H & J (siamo nell’anno 2000), dove appare per la prima volta l’equivalente *lasagna* come prestito dall’italiano. Più cauta è invece la posizione adottata in F & V e H & I, dove si fa ricorso a descrizioni esplicative metalinguistiche.

La *pizza* è un’altra volta = sütemény ‘dolce’ per Kó (con l’aggiunta *zsíros* ‘grasso’); per KK1 e Ki si tratta invece di = lángos ‘una specie di focaccia’ (che nelle cucine ungheresi viene fritta in abbondante olio e ha solo una parca guarnizione); H riprende la parola = lángos aggiungendo anche = lepény ‘schacciata’; è con H & J che la *pizza* diventa = pizza, quindi la parola arriva in Ungheria (anno 2000); nell’uso della sola parola italiana come equivalente in F & V e H & I si vede, infine, che *pizza* è ormai saldamente radicato nell’ungherese come prestito e non ha quindi bisogno di adattamenti o di spiegazioni.

Nella direzione contraria è la parola *gulyás* a mostrare un consolidamento dello stesso tipo: dopo i tentativi di equivalenza con = guazzetto (G & S), = spezzatino (KK2, KKJ e con l’aggiunta alla ungherese ancora in KKJ & J) è la volta di = gulasch, che può essere ritenuto oggi un prestito ormai radicato nell’italiano.<sup>12</sup>

Nel caso di *pogácsa* ‘piccola focaccia rotonda e salata’ (F & V) sono fuorvianti le soluzioni con = focaccia (G & S, KK2, KKJ, senza le precisazioni necessarie) e con = schiacciata (KKJ, KKJ & J), anche se l’equivalenza in quest’ultimo dizionario è relativizzata, giustamente, dall’abbreviazione kb ‘all’incirca’.

Anche nel caso di *rétes* ‘strudel’ è fuorviante l’uso (da parte di KKJ e KKJ & J) come equivalente di = millefoglie (tipo di pasta che viene usata per dolci di diverso tipo).

Gli esempi riportati in questo paragrafo e altri casi nel corpus attestano che un’analisi diacronica nell’ambito dei dizionari bilingui può rivelare anche le fasi del processo in cui una parola straniera entra nel lessico di un’altra lingua. I dati raccolti rivelano che nell’ungherese di oggi possono essere considerati prestiti dall’italiano *pizza*, *chianti* e (come calco-traduzione) *marszalai bor*, e sono probabili candidati *mortadella* e *tortellino*,<sup>13</sup> nell’altra direzione sono prestiti (gli ormai storici) *gulasch* e *tokaj(i)*.<sup>14</sup>

11 Per Kórosi *minestra* si riferisce ai ‘vari tipi di pasta (come maccheroni ecc.) cotti in acqua, pietanza consumata come prima portata nei pranzi degli Italiani’.

12 La parola entra nell’italiano dal ted. (in cui è di origine ungh.) nel 1892 (*Vocabolario Zingarelli della lingua italiana* 2008).

13 Benché non ancora registrato dai dizionari analizzati, negli ultimi anni si nota una grande diffusione in Ungheria anche di *panettone*.

14 Ungh. *gulya* ‘mandria di bovini’ → *gulyás* ‘mandriano’ (con sottinteso *hús* ‘carne’), quindi: ‘la carne del mandriano’; nell’italiano attraverso il tedesco, 1892 (*Zingarelli* 2008). Ungh. *tokaj* ‘località vinicola sull’alto Tibisco’ → *tokaji, tokaj* ‘vino biondo oro molto pregiato ivi prodotto’, 1709 (*Zingarelli* 2008).

### 3.5 Strategie traduttive e vocabolari attivi / passivi

In base ai dati dell'analisi può essere ribadito che il rapporto tra le strategie traduttive adoperate e la direzione delle lingue nei vocabolari risulta contraddittorio. Da una parte, il largo uso dei prestiti e dei forestierismi di origine italiana quali equivalenti nell'ungherese è, in base alle affermazioni della letteratura specifica, segno dell'esotizzazione, la quale si pone però in contrasto con lo scopo del dizionario passivo italiano-ungherese che dovrebbe essere orientato verso la localizzazione (e viceversa, v. Fábíán 2013: 70). Dall'altra parte, la presenza e gli effetti di una cultura più prestigiosa (quale, nel nostro caso, l'italiana) nei confronti di un'altra (l'ungherese) sono sempre di tipo esotizzante: ciò è attestato dal grande numero di vocaboli di origine italiana nell'ungherese che, di conseguenza, sono presenti anche nei dizionari italiano-ungheresi (v. p.es. House 2004, Klaudy 2013: 88).

## 4 Conclusione

L'analisi diacronica sui realia nei dizionari italiano / ungheresi ha non solo presentato l'evoluzione delle strategie traduttive adoperate dai lessicografi ma anche le loro preferenze soggettive, ed ha inoltre contribuito alla descrizione delle fasi dell'addomesticamento dei rispettivi prestiti in L2. Per poter arrivare a conclusioni ancor più precise al riguardo, converrà estendere l'esame ad una cerchia tematicamente più vasta dei realia (p.es. espressioni del linguaggio amministrativo, nomi di parentela ecc.) e raffinare ulteriormente la metodologia dell'esame.

## 5 Bibliografia

### 5.1 I Vocabolari analizzati (in ordine cronologico)

- KÓ = Kőrösi S. (1912). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Budapest: Lampel.
- G & S = Gelletich, V. & Sirola, F. (1914). *Magyar-olasz szótár* [Dizionario ungherese-italiano]. Fiume: Mohovich.
- U = Urbanek, S. (1915). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Fiume: Mohovich.
- KK1 = Koltay-Kastner, J. & Virányi, E. & Szabó, M. (1940<sup>2</sup>). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Pécs: Danubia.
- KK2 = Koltay-Kastner, J. & Virányi, E. & Szabó, M. (1943<sup>2</sup>). *Magyar-olasz szótár* [Dizionario ungherese-italiano]. Pécs: Danubia.
- Ki = Király, R. (1944). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Budapest: Szent István Társulat.
- H = Herczeg, Gy. (1978<sup>3</sup>). *Olasz-magyar szótár* I-II. [Vocabolario italiano-ungherese]. Budapest: Akadémiai Kiadó.
- KKJ = Koltay-Kastner, J. (1981). *Magyar-olasz szótár* I-II. [Vocabolario ungherese-italiano]. Budapest: Akadémiai Kiadó.

- H & J = Herczeg, Gy. & Juhász, Zs. (2000). *Olasz-magyar szótár* [Vocabolario italiano-ungherese]. Budapest: Akadémiai Kiadó.
- KKJ & J = Koltay-Kastner, J. & Juhász, Zs. (2000). *Magyar-olasz szótár* [Dizionario ungherese-italiano]. Budapest: Akadémiai Kiadó.
- F & V = Fábián, Zs. & Vásárhelyi, J. (2001). *Olasz-magyar és magyar-olasz útiszótár* [Vocabolario italiano-ungherese ed ungherese-italiano per turisti. Ed. riveduta ed aggiornata]. Budapest: Akadémiai Kiadó.
- H & I = Hessky, E. & Iker, B. (2011). *Olasz-magyar útiszótár* [Vocabolario italiano-ungherese ed ungherese-italiano per turisti]. Szeged: Grimm Kiadó.

## 5.2 Saggi

- Éber, B. & Fata, I. (2010). Reáliák és fordítói stratégiák vizsgálata német középfokú iskolatípusok és végzettségek magyarra fordítása kapcsán [Analisi dei realia nelle strategie traduttive dei concetti relativi ai tipi di scuola e ai nomi dei diplomi tedeschi]. In J. Dróth (ed.) *Szaknyelv és szakfordítás. Tanulmányok a szakfordítás és a fordítóképzés aktuális témáiról 2009–2010*. Gödöllő: Szent István Egyetem, pp. 40-56. Accesso: [http://tti.gtk.szie.hu/datadir/content/file/idegen\\_nyelvi\\_k%C3%A9pz%C3%A9s\\_%C3%A1llom%C3%A1nyai/kiadvanyaink/szakford\\_book\\_2010\\_v%C3%A9gleges.pdf](http://tti.gtk.szie.hu/datadir/content/file/idegen_nyelvi_k%C3%A9pz%C3%A9s_%C3%A1llom%C3%A1nyai/kiadvanyaink/szakford_book_2010_v%C3%A9gleges.pdf) [10/03/2014]
- Fábián, Zs. (2011). La lessicografia ungherese / italiana. In Zs. Fábián (ed.) *Hungarian Lexicography I. Bilingual Dictionaries*. Budapest: Akadémiai Kiadó, pp. 93-108. [Serie Lexikográfiai füzetek 5.]
- Fábián Zs. (2013). Reáliák a kétnyelvű szótárban régen és ma: megfigyelések és következtetések [I realia nei vocabolari bilingui ieri e oggi: osservazioni e conclusioni]. In V. Bárdosi (ed.) *Reáliák a lexikológiától a frazeológiáig*. Budapest: Tinta Könyvkiadó, pp. 61-72.
- Fata, I. (2010). *Német nyelvű reáliák megjelenítése kétnyelvű szótárakban – esettanulmány*. [La presentazione dei realia tedeschi in dizionari bilingui]. Short paper al congresso della MANYE a Debrecen, 26-28 agosto 2010. Accesso: [http://www.manyexx.unideb.hu/sites/default/files/manyexx\\_absztraktok.pdf](http://www.manyexx.unideb.hu/sites/default/files/manyexx_absztraktok.pdf) [10/03/2014]
- Heltai, P. (2001). Lexikai átváltási műveletek irodalmi és szakfordításban [Strategie traduttive lessicali nelle traduzioni letterarie e speciali] In *Fordítástudomány X* (1), pp. 5-17.
- Heltai, P. (2013). Kultúraspecifikus kifejezések és reáliák [Espressioni culturo-specifiche e realia]. In *Fordítástudomány, XV* (1), pp. 32-53.
- House, J. (2004). Culture-specific elements in translation. In H. Kittel et al. (eds.) *Übersetzung / Translation / Traduction*. Berlin: De Gruyter, pp. 494-504. [Serie HSK 26.1.]
- Klaudy, K. (2005). *Bevezetés a fordítás gyakorlatába* [Introduzione alla prassi della traduzione]. Budapest, Scolastica.
- Klaudy, K. (2013). Nyelvi és kulturális aszimmetria a reáliák fordításában [Assimmetria linguistica e culturale nella traduzione dei realia]. In V. Bárdosi (ed.) *Reáliák a lexikológiától a frazeológiáig*. Budapest: Tinta Könyvkiadó, pp. 85-92.
- Kujamäki, P. (2004). Übersetzung von Realienbezeichnungen in literarischen Texten. In H. Kittel et al. (eds.) *Übersetzung / Translation / Traduction*. Berlin: De Gruyter, pp. 920-925. [Serie HSK 26.1.]
- Lendvai, E. (1986). *A „lefordíthatatlan elem” megfeleltetési lehetőségei* [Possibilità di equivalenza degli “elementi intraducibili”] Tesi per il titolo di “candidato” dell’Accademia delle Scienze Ungherese. Università di Pécs, HU.
- Marello, C. (1989). *Dizionari bilingui*. Bologna, Zanichelli.
- Mujzer-Varga, K. (2009). *Honosítás és idegenítés Örkény Egyperces novelláinak fordításaiban* [Localizzazione ed esotizzazione nelle traduzioni di “Racconti di un minuto” dello scrittore Örkény]. Tesi per il titolo di PhD. Università “Eötvös” di Budapest, HU. Accesso: <http://doktori.btk.elte.hu/lingv/mujzervargakrisztina/tezis.pdf> [10/03/2014]



- Osimo, B. (2003). *Il manuale del traduttore*. Milano: Hoepli.
- Rey, A. (1986). Les écarts culturels dans les dictionnaires bilingues. *Lexicographica* 2: 33-42.
- Rey, A. (1991). Divergences culturelles et dictionnaire bilingue. In F. J. Hausmann, F. J. et al. (eds.) *Wörterbücher. Ein internationales Handbuch zur Lexikographie*. Berlin: De Gruyter, pp. 2865-2870. [Serie HSK 5.3.]
- Svensén, B. (2009). *A Handbook of Lexicography*. Cambridge: Cambridge University Press.
- Szabó, H. (2008). A szótári ekvivalencia és a fordítási ekvivalencia viszonya. [Sul rapporto dell'equivalenza lessicografica e dell'equivalenza traduttiva] In *Fordítástudomány X* (1), pp. 61-70.
- Tellingier, D. (2003). A reáliák fordítása a fordító kulturális kompetenciája szemszögéből. [La traduzione dei "realia" dal punto di vista della competenza culturale del traduttore]. In *Fordítástudomány V* (2), pp. 58-70.
- Tomaszczyk, J. (1984). The culture-bound element in bilingual dictionaries. In R. Hartmann (ed.) *LEXeter '83 Proceedings*. Tübingen: Niemeyer, pp. 289-297. [Lexicographica. Series Maior.]
- Uzonyi, P. (2006). A forrásnyelvi és célnyelvi adatok ekvivalenciájának nehézségeiről [Sulle difficoltà delle equivalenze tra i dati di partenza in L1 e quelli di arrivo in L2]. In T. Magay (ed.) *Szótárak és használói*. Budapest: Akadémiai Kiadó, pp. 117-126. [Serie Lexikográfiai füzetek 2.]

## Appendice 1

- barbera**: Kő: piemonti vörös bor-faj, U: -, KK1: -, Ki: borfajta, H: Asti környéki vörös bor, H & J: *bor*<sup>15</sup> <piemonti vörösbör>, F & V: -, H & I: -
- cannelloni**: Kő: a makaróninál vastagabb csöves tészta, U: -, KK1: -, Ki: -, H: (hússal, sajttal stb. töltött) tésztarolád, palacsinta, H & J: *konyh*<sup>16</sup> [hússal, sajttal stb. töltött] tésztarolád], F & V: húsos töltött palacsinta, H & I: *ujjnyi hosszú tésztacsövek, melyeket töltve készítene* el
- chianti**: Kő: 2. Chianti-bor, U: -, KK1: -, Ki: toszkánai vörösborfajta, H: toszkán vörösbör, chianti, H & J: <toszkán vörösbör> chianti, F & V: chianti (*toszkán vörösbör*), H & I: (*az azonos nevű toszkán vidékről származó vörösbör*) chianti
- gorgonzola**: Kő: gorgonzola (sajt), U: gorgonzola (sajt), KK1: -, Ki: sajt fajta, H: gorgonzola sajt, H & J: *konyh* gorgonzola sajt, F & V: (*sajtféle*), H & I: *Lombardiából származó kéhpenezes sajt*
- grissino**: Kő: hosszúkás ropogós sütemény Piemontban, U: -, KK1: -, Ki: ropogós kenyérrúd, H: hosszú, vékony rúdkenyér, H & J: *konyh* ropi <hosszú, vékony ropogtatnivaló>, F & V: (*hosszú és vékony rúdkenyér*), H & I: *kenyértésztából készült, ropogós pálcika*
- lasagna**: Kő: széles szalagokra vagdalt levestészta v. más minesztrához való tészta, U: szalagmetélt, KK1: metélt, Ki: szélesmetélt, H: széles metélttészta, H & J: lasagna; *lasagne al forno* csöben sült lasagne v. rakott tészta, F & V: (*lerakott tésztaféleség*), H & I: *lapos, széles, téglalap alakú tésztalapok, melyekből rakott ételeket készítene*
- marsala**: Kő: 2. marszáli bor, U: -, KK1: marsalai bor, Ki: marsalai bor, H: (magas szesztartalmú) marszalai fehér bor, H & J: [magas szesztartalmú] marszalai fehér likörbor, F & V: -, H & I: -
- mascarpone**: Kő: -, U: -, KK1: -, Ki: -, H: *észol*<sup>17</sup> (Camemberthez hasonló) lombardiai puha sajt, H & J: mascarpone [lombardiai sajt], F & V: -, H & I: *tejszínes, fehér krémsajt*
- mortadella**: Kő: disznósajt, U: májas hurka; mortadella, KK1: disznósajt, mortadella, Ki: disznósajtféle, H: mortadella (felvágott), H & J: *konyh* mortadella [felvágott], F & V: mortadella (*felvágottféle*), H & I: -
- pandoro**: Kő: -, U: -, KK1: -, Ki: édestésztafajta, H: veronai kuglóf, H & J: *konyh* <karácsonyi veronai kuglóf>, F & V: kuglóf (*karácsonyi édesség*), H & I: *kuglófhoz hasonló, veronai eredetű karácsonyi sütemény*
- panettone**: Kő: lisztből, cukorból, sáfrányból készített és sörrel kelesztett sütemény, U: -, KK1: -, Ki: édestésztaféle, H: mazsolás, cukrozott gyümölcsös milánói kalács, H & J: *konyh* panettone [karácsonyi

15 *bor* 'vino'

16 *konyh* 'termine gastronomico'

17 *észol* 'settentrionale'

milánói kuglóf], F & V: kuglóf (*harácsonyi édesség*), H & I: *kuglófhoz hasonló, milánói eredetű harácsonyi sütemény*

**panforte:** Kő: mandulából, kakaóból stb. készített borsos sütemény, U: -, KK1: sienai mézeskalács, Ki: édestésztafajta, H: lapos, kerek, sienai sütemény, H & J: *konyh* <lapos, kerek sienai sütemény>, F & V: (*sienai édesség*), H & I: *gyümölcskenyérhez hasonló sienai harácsonyi sütemény*

**pastasciutta:** Kő:<sup>18</sup> (leves helyett tálalt) makaróni, U: -, KK1: -, Ki: makaróni, H: száraz tészta, főtt tészta, H & J: száraztészta [élelmiszer], főtt tészta [fogás], F & V: főtt tészta, H & I: (száraz)tészta

**pizza** Kő: 2. zsíros sütemény, U: -, KK1: lángos, Ki: lángos, H: lángos, lepény, H & J: 1. pizza; ~ *alla napoletana* [...] 2. lángos; lepény, F & V: pizza, H & I: pizza

**raviolo** Kő: vajjal, túróval, fűszerszámmal és vagdalt hússal töltött toklóhúsból készült étel, U: -, KK1: -, Ki: étel (vagdalthús- és túrósgombócból), H: húsos táska; derelye, H & J: *konyh* húsos táska, derelye, F & V: húsos derelye/táska, H & I: *hússal, zöldséggel, ricottával töltött tésztatatyu*

**tortellino** Kő: tyúkhússal töltött, levesbe való táska, U: táska (levesbe), KK1: -<sup>19</sup>, Ki: húsos táska (étel), H: darált hússal töltött táska, H & J: *konyh* tortellini [sonkával, hússal töltött tészta], F & V: tortellini (töltött tésztagyűrű), H & I: *hússal, sajttal töltött gyűrű alakú téasztatatyu*

## Appendice 2

**beigli:** G & S: -, KK2: -, KKJ: -, KKJ & J: *konyh* rotolo dolce ai semi di papavero o ai noci [sic!], F & V: -, H & I: -

**fröccs:** G & S: -, KK2: -, KKJ: due decilitri di vino con un decilitro di acqua di seltz, KKJ & J: <vino e acqua gasata>, F & V: (*due decilitri di vino con un decilitro di acqua di seltz*), H & I: spritz

**gulyás:** G & S: guazzetto, KK2: spezzatino, KKJ: gulasch, spezzatino alla ungherese, KKJ & J: gulasch, spezzatino alla ungherese, F & V: gulasch, H & I: gulasch

**halászlé:** G & S: salsa/zuppa di pesce, KK2: zuppa di pesce, KKJ: minestra/brodetto di pesce (all'ungherese), KKJ & J: *konyh* zuppa di pesce (all'ungherese), F & V: zuppa di pesce, minestra di pesce, H & I: zuppa di pesce

**kadarka:** G & S: -, KK2: -, KKJ: kadarka (uva, vino leggero [sic!] rosso ungherese), KKJ & J: kadarka <uva/vino rosso leggero ungherese>, F & V: -, H & I: -

**kovácsos uborka:** G & S: cetriolo sotto aceto, KK2: -, KKJ: cetriolo lievitato, KKJ & J: cetriolo in salamoia a base di lievito, F&V: -, H & I: -

**lángos:** G & S: crostino, KK2: pizza, KKJ: pizza, *toszk*: schiacciata, *róm*: piada, KKJ & J: *toszk*: schiacciata <pasta lievitata fritta in abbondante olio>, F & V: (*una specie di schiacciata*), H & I: -

**paprikás csirke:** G & S: pollo in umido colla paprica, KK2: -, KKJ: pollo con intingolo di paprica, spezzatino con paprika, KKJ & J: pollo alla paprika, spezzatino di pollo, F & V: pollo all'arrabiata, pollo all'ungherese, H & I: -

**pogácsa:** G & S: focaccia, KK2: focaccia, KKJ: schiacciata, focaccia, KKJ & J: *konyh kb*<sup>20</sup> schiacciata, focaccia, F & V: (piccola focaccia rotonda e salata), H & I: -

**rétes:** G & S: pasta avvolta con ripieno, KK2: strudel, KKJ: *kb* millefoglie, KKJ & J: *konyh* strudel, millefoglie, F & V: strudel, H & I: strudel<sup>21</sup>

**szaloncukor:** G & S: -, KK2: -, KKJ: confetto/fondente incartato, KKJ & J: <confetto/fondente incartato per ornare l'albero di Natale>, F & V: -, H & I: -

**tejberizs:** G & S: riso in latte, KK2: riso col latte, KKJ: riso col latte, KKJ & J: riso al latte, F & V: (*riso cotto in latte dolce*), H & I: riso al latte

18 Nel lemma di *asciutto*.

19 Come lemma abbiamo: *tortello* húsos táska.

20 *kb* 'all'incirca'

21 Nello stesso tempo nei lemmi con collocazioni appaiono anche altre soluzioni: G & S: *háposztás rétes* pasticcio di cavolo, *almás rétes* dolce di mele, KKJ: *túrós rétes kb* pasta sfoglia di ricotta, KKJ & J: *túrós rétes* pasta sfoglia di ricotta, strudel alla ricotta.



**tejföl/tejfel:** G & S: panna, crema, fior di latte, KK2: fiore di latte, KKJ: fiore di latte, crema, KKJ & J: fiore di latte, crema/panna acida, F & V: panna acida, H & I: panna acida

**tokaji:** G & S: (bor) (vino) di Tokaj, KK2: -, KKJ: tokay, tocai, KKJ & J: tokaj, tocai, F & V: tocai, H & I: -

**töltött háposzta:** G & S: -, KK2: -, KKJ: -, KKJ & J: cavoli/crauti ripieni, F & V: cavoli ripieni, H & I: -

**túrós csusza:** G & S: -, KK2: -, KKJ: -, KKJ & J: -, F & V: (*tagliatelle con ricotta e panna acida*), H & I: -



# Quello che i dizionari possono fare: l'esempio dei Dizionari di Tedesco (Giacoma/Kolb – Zanichelli/Klett)

Luisa Giacoma  
Università di Torino  
luisa.giacoma@unito.it

## Abstract

Partendo dall'affermazione di Hausmann (1993) che dichiara la non apprendibilità del lessico e che vede nei dizionari l'unica possibilità di salvezza per l'apprendente, si avverte l'esigenza di indagare su cosa possa fare la lessicografia per far fronte a questa gravosa responsabilità. Da una parte bisogna educare il lettore ad un uso proficuo del dizionario, mentre dall'altra è necessario innovare i dizionari, affinché possano aiutare l'apprendente anche nell'uso attivo della lingua. Per potersi esprimere adeguatamente in L2 egli deve infatti saper padroneggiare i legami semantici, sintattici e pragmatici esistenti tra le parole ed essere consapevole dei notevoli vincoli che questi pongono alla libertà combinatoria del parlante (Lo Cascio 1997). Se teoricamente il parlante può abbinare liberamente le parole, in realtà solo poche combinazioni sono possibili, poiché esse sottostanno a limitazioni sintattiche, semantiche, enciclopediche e combinatorie dettate dall'uso e dal contesto. Un'analisi dei principali problemi rimasti irrisolti nella lessicografia tradizionale, orientata per lo più verso l'uso passivo della lingua, ha portato allo sviluppo di un nuovo modello lessicografico alla base dei dizionari di Tedesco a cura di Luisa Giacoma e Susanne Kolb e pubblicati dalla casa editrice italiana Zanichelli e tedesca Klett, che costituiscono una risposta concreta alle nuove esigenze degli apprendenti.

**Keywords:** Dizionari; Tedesco; Italiano

## 1 Introduzione

Quando l'aereo della compagnia di bandiera italiana, poco dopo il decollo, abbandonò l'assetto della salita vertiginosa per assumere quello orizzontale della crociera, mi venne servito un tè che mi fece trasalire. Sul bicchierino di carta bianca campeggiavano a caratteri cubitali due scritte verdi e rosse a completamento dei colori nazionali, una delle quali recitava: IL NOME DELLA BEVANDA PIÙ DIFFUSA DEL MONDO PUÒ ESSERE SCRITTO IN ITALIANO IN DUE MODI TÈ O THÈ. Rilessì nuovamente, ma l'errore continuava a rimanere lì in bella evidenza a fine frase. Ora, ammettiamo pure che qualcuno possa non sapere che in italiano la più nota bevanda inglese si scrive "tè" oppure "the", e che quindi la "e" possa scegliere se andare a braccetto con l'"h" oppure portare un accento calcato sulle ventitrè come un cappello, ma se si decide di scrivere una frase come quella, dove l'oggetto è proprio la corretta grafia di una parola, non bisognerebbe come minimo controllarla sul dizionario? In un caso

come questo, ovviamente, nemmeno il migliore dei dizionari può fare qualcosa, ma in moltissimi altri casi può fare invece la differenza.

## 2 Il ruolo del dizionario nella didattica delle lingue

Se Hausmann (1993: 471) afferma che non si può imparare il lessico di una lingua perché è sterminato, totalmente idiomatico e caotico e vede nei dizionari l'unica via di salvezza, diventa allora cruciale disporre di strumenti ben costruiti, che forniscano per ogni lemma non solo le informazioni morfologiche, semantiche, sintattiche e i tradimenti usualmente reperibili nei dizionari, ma anche una sorta di mappa delle possibilità combinatorie del lemma, con tutte le informazioni sull'uso solitamente appannaggio dei madrelingua. Il recente spostamento del focus lessicografico dalla descrizione della lingua in sé alle esigenze del lettore ha avuto come conseguenza evidenti vantaggi per l'apprendimento e il passaggio dell'interesse dall'uso prevalentemente passivo a quello attivo.

L'intero corpo di una lingua, pur entro limiti stabiliti, si trova solo in un dizionario, che va quindi sempre più concepito "come strumento centrale per la conoscenza e la descrizione sistematica della lingua nel suo pieno uso, superando la sua vecchia immagine di strumento per conoscere il 'significato' (o poco più) delle parole" (Sabatini 2008: 112).

Nonostante vi siano ancora posizioni antitetiche sull'apprendibilità del lessico, nella didattica delle lingue, dopo un periodo di ingiustificato ostracismo, è ormai indiscusso il ruolo centrale del dizionario bilingue come strumento per l'apprendimento. Resta però ancora molto da migliorare sul fronte delle possibilità del lettore di utilizzarlo efficacemente. Da una parte viene fatto troppo poco per insegnare a usare bene i dizionari, mentre dall'altra il contributo della metalessicografia non si è ancora sufficientemente diffuso nella pratica lessicografica e quindi sono ancora pochi i dizionari centrati sui bisogni effettivi dell'apprendente.

Come è già stato evidenziato in ricerche sul tema (Giacoma 2011, 45-46), la conoscenza "spontanea" dell'uso del dizionario è uno dei tanti pregiudizi molto diffusi su questi strumenti: varrebbe invece la pena di dedicare, possibilmente fin dall'inizio del corso, qualche ora per avviare gli studenti ad un utilizzo efficace del dizionario. Poiché non si può vedere quello che non si conosce, come affermò il fondatore della casa editrice per guide da viaggio Dumont, si potrebbe sacrificare qualche ora di spiegazione riservata ad argomenti grammaticali minori e sfruttare questo tempo a vantaggio di un uso migliore dei dizionari, dato l'impatto che essi hanno sull'apprendimento delle lingue straniere.

### 3 Dalla parte del lettore: alcuni problemi irrisolti nella lessicografia bilingue tradizionale

I maggiori problemi della lessicografia bilingue tradizionale del secolo scorso, ma ancora presenti nelle successive edizioni di dizionari di Tedesco-Italiano concepiti in quel periodo, sono da attribuirsi principalmente alla mancata ricezione dei risultati ottenuti negli studi di linguistica e di lessicografia contrastiva, risultando così inadeguata a soddisfare le sempre crescenti necessità dei lettori, soprattutto di quelli di livello più avanzato, che si trovano a usare la lingua in modo molto più (inter)attivo rispetto a quanto avvenisse in passato, quando i dizionari venivano consultati quasi esclusivamente per la decodificazione dalla L2 verso la propria. I limiti più evidenti dei dizionari tradizionali sono (Giacoma 2011):

- l'insufficiente o inadeguata registrazione delle collocazioni,<sup>1</sup> che non compaiono come tali, ma come esempi alle singole accezioni o nel blocco fraseologico al fondo della voce, se non addirittura in luogo delle spiegazioni
- l'inadeguata differenziazione tra i diversi equivalenti della L2, che vengono generalmente elencati uno dopo l'altro, senza alcun elemento che indichi quale sia da preferire all'interno di un certo contesto, come se fossero sinonimi intercambiabili in tutti i contesti, anche se gli apprendenti tendono a fare molti errori proprio nel momento in cui costruiscono frasi o parti di esse combinando tra loro elementi che non possono essere abbinati
- la mancanza di indicazioni sistematiche e ben evidenziate sull'intorno sintattico del lemma, nonostante la trasposizione in L2 delle strutture della propria lingua sia all'origine di molti errori
- l'inadeguata registrazione dei fraseologismi
- le scarse informazioni sulla morfologia delle parole tedesche, fonte di frequenti dubbi per l'apprendente italofono.

Naturalmente ci sarebbero anche altri problemi da discutere, ma si rimanda ad altro luogo per ragioni di spazio.

### 4 L'evoluzione della voce lessicografica nei dizionari di Tedesco-Italiano

L'avvento nella pratica lessicografica di corpora come *COSMAS II* dell'Institut für Deutsche Sprache di Mannheim, consultabile gratuitamente in rete, ha notevolmente migliorato le basi materiali sulle quali possono essere costruiti oggi i dizionari, permettendo l'accesso a dati statisticamente significa-

---

1 Qui le collocazioni vengono intese in senso ampio, adatto a scopi lessicografici. Si confronti a questo riguardo la definizione data nel dizionario per apprendimenti del tedesco *LGWDaF* = Dieter Götz et al., *Langenscheidts Großwörterbuch Deutsch als Fremdsprache*, Langenscheidt, Berlin 1993, S. XX „[...] collegamenti tipici di più parole che costituiscono un'unità sintattica [...] poiché mostrano [...] 'partner' con il quale il lemma occorre frequentemente,„ [trad. LG].

tivi sull'uso reale della lingua. Anche le acquisizioni della metalessicografia sono potenzialmente di grande aiuto, ma il dialogo tra linguisti e lessicografi è una conquista relativamente recente. Nied Curcio (2006, 61), nella sua analisi dei dizionari di Tedesco, lamenta che i dizionari di tipo tradizionale come il Sansoni e il DIT “non si basano sugli studi della linguistica contrastiva, ma si limitano ad essere un insieme di informazioni utili solo per la consultazione”. Già nel 1989 Marellò criticava la scarsa ricaduta sui dizionari bilingui di Tedesco-Italiano del vivace dibattito metalessicografico sulle valenze che aveva avuto luogo in Germania.

## 5 Possibili soluzioni e loro realizzazione

I dizionari di Tedesco a cura di Luisa Giacoma e Susanne Kolb, vale a dire l'intera serie di dizionari scritti dalle autrici nell'arco degli ultimi vent'anni e pubblicati dalla casa editrice italiana Zanichelli e da quella tedesca Klett (d'ora in avanti GK), nati sulla base di approfondite riflessioni sui limiti della lessicografia bilingue di Tedesco-Italiano degli anni '90, hanno cercato di risolvere tali problemi con l'introduzione di alcuni strumenti innovativi, offrendo risposte concrete alle necessità degli apprendenti. A tal fine è risultato particolarmente utile poter contare sulla teoria della valenza come riferimento teorico e valutare le applicazioni pratiche in campo lessicografico, che proprio in quegli anni venivano realizzate in Germania.

Chi ascolta, scrive, legge, parla o traduce una lingua si relaziona nella quasi totalità dei casi con testi, mentre nel dizionario trova parole singole. A questo proposito Sabatini (2008: 112) afferma che “di questa globalità, accolta in un unico contenitore anche se scomposta nella lemmaticità alfabetica, c'è bisogno per cogliere il funzionamento della lingua, incrociando le diverse prospettive (morfologica, sintattica, semantica, testuale [...])”.

Per riuscire a superare il gap tra il testo e le parole è necessario che la voce lessicografica dia il maggior numero di informazioni sull'intorno lessicale e sintattico della parola, su quali parole cioè si trovano frequentemente in compagnia del lemma (i *collocatori*) e su come lemma e *collocatori* si combinano tra loro (con o senza preposizione, ecc.). Il modello valenziale di descrizione del comportamento sintattico della singola parola ha mostrato enormi potenzialità nella sua applicazione alla voce lessicografica, con importanti risvolti per l'apprendente. Inoltre esso è particolarmente utile quando si mettono a confronto lingue diverse perché evidenzia le divergenze e/o convergenze, che sono spesso fonte di errore per gli apprendenti, soprattutto nei momenti di codificazione in L2 (Curcio 1999).

A questo proposito Fischer e Mollica (2012) osservano che l'approccio valenziale, così come quello costruzionistico, vede i lessemi e le costruzioni in stretta relazione tra loro, ma ciò che distingue il primo approccio dal secondo è la prospettiva “bottom-up”, che parte cioè dal lessema per arrivare alla costruzione. Si può quindi trarre la conclusione che proprio questo sia ciò che serve alla lessicografia, che parte dai singoli lemmi per descrivere costruzioni, frasi, testi e la lingua in genere. In Germania si hanno i primi dizionari valenziali già alla fine degli anni '60 con il *Wörterbuch zur Valenz und Distributi-*

on *deutscher Verben* di Helbig/Schenkel del 1969 (Nied Curcio 2012, 175). Non stupisce pertanto che le prime applicazioni lessicografiche della teoria della valenza filtrino in Italia proprio grazie ai contributi di germanisti italiani (Bianco, Curcio, Soffritti) o romanisti attivi in Germania (Blumenthal, Rovere) sotto forma di dizionari valenziali su base contrastiva, come quello dei verbi tedeschi e dei loro equivalenti italiani (Bianco 1996), oppure quello dei verbi italiani con traduenti tedeschi (Blumenthal/Rovere 1998), o ancora della lingua (tedesca e italiana) parlata (Curcio 1999), degli aggettivi (Soffritti, DIVA 2005) e dei circa 3000 lemmi del lessico di base presenti in ELDIT. Ma una nuova generazione di dizionari generali si stava affacciando nel panorama della lessicografia bilingue di Tedesco-Italiano. I dizionari GK nascono in un periodo di grande innovazione lessicografica, come si evince dall'uscita in Germania nel 1993 del primo dizionario monolingue per apprendenti *Langenscheidt Großwörterbuch Deutsch als Fremdsprache* (LGWDaF) e del dizionario bilingue *Pons Großwörterbuch Französisch Deutsch* (PONS) nel 1996. Entrambi applicano la teoria della valenza per un pubblico di apprendenti stranieri e hanno costituito un ottimo modello di riferimento per i dizionari GK. Essi sono i primi nei quali la teoria della valenza trova spazio per la descrizione di *tutta* la lingua italiana e tedesca. Da questo punto di vista essi possono essere considerati, in quanto dizionari valenziali, *dizionari di studio*, vale a dire strumenti articolati e sistematici di riflessione sulle proprietà sintattiche e semantiche della lingua. Seguendo il modello del LGWDaF e del PONS, essi introducono una sorta di *sintassi della parola* basata sulle teorie grossiane (1967, 1975) che, al compito svolto anche dai dizionari tradizionali di descrivere con puntigliosa precisione ogni singola parola, aggiungesse quello non meno importante di dare informazioni esplicite su come le parole possono o debbono essere combinate tra loro. Sono proprio queste informazioni sintagmatiche ad essere spesso carenti nei dizionari di tipo tradizionale. Il metodo scelto per realizzare questa *sintassi della parola* è stata l'introduzione sistematica di due nuove categorie: i *collocatori* e le *formule di struttura*, entrambe *desiderata* della lessicografia teorica (Fontenelle 1997).

## 5.1 I collocatori

Uno degli aspetti meno soddisfacenti dei dizionari di tipo tradizionale è la scarsa registrazione di *collocatori* e soprattutto il modo confuso col quale questa avviene. A questo riguardo Marellò/Rovere (1999: 198) sottolineavano, due anni prima dell'uscita del primo dizionario GK, la necessità di indicare le collocazioni come tali e di registrarne un numero maggiore nei dizionari. Per il lettore i principali vantaggi della registrazione dei collocatori da un punto di vista lessicografico sono: la differenziazione dei traduenti, la registrazione di un maggior numero di contesti tipici, una migliore strutturazione della voce, nonché una maggiore possibilità di controllo. Nei dizionari GK si è tenuto conto fin dall'inizio del progetto della possibilità di consultazione elettronica ed è pertanto stato dedicato un campo di ricerca ai soli collocatori. Nel GK (2014<sup>3</sup>) è sufficiente digitare una parola, ad esempio *cavallo*, all'interno di tale campo nella maschera della ricerca avanzata per ottenere la lista delle parole che ad essa più frequentemente si accompagnano. Sulla sinistra della maschera di ricerca vi è la colonna a

scorrimento con l'elenco dei collocatori numerati in ordine progressivo. Il numero totale (193) è in basso. Facendo scorrere l'elenco si possono leggere tutti i 193 collocatori della parola *cavallo*. Sulla destra è invece aperta la voce del collocatore evidenziato nella lista. La parola per la quale è stata fatta la ricerca appare all'interno di un riquadro evidenziato in rosso. Dizionari così concepiti permettono di rispondere a domande che rimarrebbero quasi sempre senza risposta con la consultazione, anche elettronica, di un dizionario tradizionale (Fontenelle 1997). Se vogliamo ad esempio sapere che verso fa il cavallo in un dizionario senza collocatori possiamo solo sperare di trovare un esempio al lemma *cavallo* o al lemma *verso*, ma non avremmo altri aiuti per arrivare a *nitrito*. Anche la ricerca elettronica di *cavallo* non ha dato esiti né nel DIT, né nel Sansoni.

In questo caso particolare osserviamo che anche nel GK manca *nitrito* nella lista dei collocatori di *cavallo*, ma si può facilmente risalire ad esso grazie alla presenza del verbo *nitrire*. Il potenziale lessicografico dei collocatori è però molto più evidente se desideriamo sapere, ad esempio, quali azioni può compiere un cavallo (*ansimare, cadere, calciare, caracollare, disarcionare, galoppare, imbizzarrire, impennare, inalberare, nitrire, rinculare, rompere, sbuffare, scartare, sgroppare, stramazzone, trottare, trotterellare, volteggiare ...*), o quando il cavallo è l'oggetto dell'azione (*addestrare, addomesticare, allenare, ammaestrare, azzoppare, bardare, cavalcare, domare, dopare, dressare, ferrare, frenare, guidare, imbrigliare, incavezzare, legare, montare, sbalzare, scendere, scommettere, sfiancare, spronare, strigliare ...*). Possiamo inoltre elencare quali caratteristiche può avere un cavallo semplicemente selezionando tra i collocatori gli aggettivi (*addestrato, baio, balzano, belga, berbero, bizzarro, bizzoso, domabile, drogato, favorito, fuoriclasse, furioso, irrequieto, matto, ombroso, perdente, pezzato, piazzato, pomellato, purosangue, ricalcitante ...*) o trovare i termini corrispondenti alle singole parti del cavallo (*callo, cantone, coda, criniera, fianco, garretto, groppa, manto, nodello, retrotreno, schiena, spalla, unghia, ventre, zampa ...*) o al suo frame (*ambio, bardatura, briglia, paraocchi, sperone, ...*).

## 5.2 Le formule di struttura

Le *formule di struttura* (o semplicemente *strutture*) vengono riprese dai due dizionari di lingua tedesca serviti da modello (LGWDaF e PONS) e modificate per adattarle alle esigenze di un pubblico di apprendenti italiani. Anche nei dizionari GK esse riducono a formula le *reggenze* della grammatica tradizionale dando informazioni esplicite su come i collocatori e il lemma si combinano tra loro, ma diversamente dal LGWDaF e PONS, rinunciano ad esempio all'uso della terza persona singolare nei verbi. Tale scelta è motivata dal fatto che l'apprendente italiano ha dimestichezza con la forma all'infinito *scrivere qc a qu* e non a quella flessa *qu scrive qc a qu*. Quest'ultima formula ha però l'indubbio vantaggio di dare informazioni anche sul soggetto. Come ovviare, allora, a tale perdita di informazione? Si è optato per dare il soggetto come collocatore e differenziarlo dagli altri attraverso l'uso del corsivo. L'uso del grassetto corsivo mette in evidenza nei dizionari GK le *formule di struttura* permettendo al lettore di rendersi immediatamente conto del contesto sintagmatico del lemma (alla voce *cadere*: ***cadere da qc, cadere in qc, cadere su qu/qc, cadere a qc***, ecc.), differenziando spesso anche un'accezione dall'altra.



In riferimento alla necessità dell'apprendente di avere informazioni sui possibili collegamenti sintattici del lemma per poter produrre frasi grammaticalmente corrette, Schafroth (2004: 119) sottolinea come questo non costituisca generalmente un problema se il dizionario registra sistematicamente *formule di struttura* in modo esplicito. L'inserimento delle strutture all'interno della voce sono un'ulteriore aiuto per orientare il lettore e obbligano il lessicografo ad un lavoro minuzioso di strutturazione della glossa lessicografica. Esse hanno anche il vantaggio di essere facilmente controllabili perché date in modo esplicito, anziché implicitamente, come negli esempi.

Come si può evincere da quanto sopra riportato, la presenza delle *formule di struttura* non è solo un modo diverso di organizzare le informazioni, ma porta in primo piano la sintassi della parola enucleandola dal contingente e facendone brillare la quintessenza. Inoltre, a differenza di quanto possibile fare con gli esempi, le *formule di struttura* segnalano anche la facoltatività o meno delle *Ergänzungen* sia nella lingua di partenza che in quella di arrivo, con evidenti vantaggi per il lettore soprattutto quando una *Ergänzung* è obbligatoria in una lingua e facoltativa in un'altra.

Nied Curcio (2012), a proposito dei vantaggi dell'introduzione delle strutture nella voce lessicografica, afferma che, fornendo indicazioni esplicite, esse permettono un utilizzo più efficace del dizionario, diminuendo così in misura considerevole le possibilità di errore da parte di chi lo utilizza.

### 5.3 Corredo di maggiori informazioni sulle espressioni idiomatiche

Per quanto riguarda invece la registrazione delle *espressioni idiomatiche* Dobrovolskij (2009) afferma che la soluzione fraseografica tradizionale di giustapporre nei dizionari bilingui ad una *espressione idiomatica* della lingua di partenza una *espressione idiomatica* della lingua di arrivo è insufficiente se non addirittura fuorviante, per via delle differenze semantiche, sintattiche e pragmatiche che spesso si rilevano all'interno di queste coppie di espressioni. Ad esempio a ***etw im Auge haben*** nel DIT viene semplicemente fatta seguire la traduzione 'avere qcs in testa', senza nessuna ulteriore informazione.

Ai fini di permettere al lettore di comprendere e usare correttamente le *espressioni idiomatiche* è necessario corredare queste ultime di informazioni riguardanti la semantica, la sintassi e la pragmatica delle stesse. Le regole per l'uso delle *espressioni idiomatiche*, e le restrizioni alle quali esse sottostanno, devono costituire un commento all'equivalente, soprattutto in considerazione del fatto che le equivalenze sono quasi sempre parziali.

Ogni *espressione idiomatica* andrebbe quindi descritta nelle sue peculiarità semantiche, pragmatiche e combinatorie (tra le quali quelle sintattiche) per permetterne un uso corretto all'interno di un contesto.

Un contributo in tale direzione è quello fornito alla lessicografia bilingue italo-tedesca dal metodo utilizzato nel Giacoma/Kolb, sia dal punto di vista dell'ordinamento che delle informazioni sintattiche, semantiche e pragmatiche che accompagnano le *espressioni idiomatiche*. Si confronti ora, ad esempio, la registrazione dell'*espressione idiomatica etwas im Auge haben* nel DIT e nel Giacoma/Kolb.

(1) DIT

***etw im Auge haben*** ‘avere qcs in testa’

(2) Giacoma/Kolb

***etw im Auge haben*** {INSEKT, STAUBKORN}, avere qc nell’occhio;

(*etw im Sinn haben*) {SEINEN EIGENEN VORTEIL}, guardare solo a qc;

{EIN BESTIMMTES MODELL}, aver pensato a qc di preciso

Si può osservare come, in quest’ultimo caso, oltre all’informazione standard generalmente registrata nei dizionari, e cioè la traduzione, ve ne siano numerose altre che vale la pena di commentare. Per cominciare, il significato proprio viene aggiunto in prima posizione e corredato dei collocatori {INSEKT, STAUBKORN} e della traduzione ‘avere qc nell’occhio’. Successivamente il significato idiomatico viene introdotto da una spiegazione (*etw im Sinn haben*). Per concludere i due equivalenti suggeriti dal dizionario ‘guardare solo a qc’ e ‘aver pensato a qc di preciso’ vengono accompagnati dal loro contesto in forma di collocatori {SEINEN EIGENEN VORTEIL} e {EIN BESTIMMTES MODELL}. Da questo semplice esempio si può dedurre quanto rischiosa sia l’equivalenza ***etw im Auge haben*** = ‘avere qcs in testa’ se non viene accompagnata da altro. Il lettore può infatti essere tratto in inganno in più modi, sia essendo indotto a pensare che il traduttore proposto ‘avere qcs in testa’ possa valere anche per il significato proprio, sia non impedendogli di selezionare, sul piano idiomatico, un equivalente inadeguato al contesto. L’equivalente ‘avere qcs in testa’ non si adatterebbe infatti alla maggior parte dei contesti citati in (2).

## 5.4 Le tabelle di flessione

Le indicazioni morfologiche che tradizionalmente accompagnano il lemma sono solo una minima parte delle informazioni necessarie all’apprendente quando deve esprimersi in L2.

A partire dalla seconda edizione, grazie alla collaborazione con la società Canoo di Basilea, il dizionario GK dispone nella sua versione elettronica della tabella di flessione di tutti i lemmi tedeschi, comprendente la coniugazione dei tempi semplici per i verbi regolari e irregolari, il nominativo, genitivo, dativo e accusativo singolare e plurale per i sostantivi e la flessione completa degli aggettivi, anche al grado comparativo e superlativo. Naturalmente questo è stato possibile grazie ai supporti elettronici, che permettono di memorizzare un’enorme quantità di dati.

## 6 Orizzonti vicini: uno sguardo verso un futuro realizzabile

Il mestiere del lessicografo è certamente difficile ed estremamente faticoso, ma possiede l'indubbio fascino di poter migliorare continuamente negli anni la propria opera, apportando cambiamenti, a volte anche rilevanti, da un'edizione all'altra.

Oltre alle soluzioni realizzate finora e sopra descritte, ve ne è una ulteriore piccola parte nella terza edizione del Dizionario di Tedesco. Sono state aggiunte infatti 600 note sui falsi amici, note d'uso e di civiltà come aiuto concreto per comprendere persone di altre culture, comunicare con loro, recepire e produrre testi in lingua straniera, partecipare attivamente alla vita sociale e culturale di altri paesi. La versione cartacea è così stata sfruttata al massimo delle sue potenzialità e non ha permesso purtroppo ulteriori aggiunte, ma nella versione elettronica è stato possibile utilizzare lo spazio disponibile e le caratteristiche del supporto per arricchire il dizionario con la pronuncia sonora di tutte le parole del tedesco.

Cosa sarebbe utile realizzare ancora in una quarta, quinta, ecc. edizione? Poiché i dizionari sono testi complessi e ricchi di abbreviazioni, un modo per renderli di più facile lettura sarebbe la possibilità di leggere al passaggio del mouse la forma estesa delle abbreviazioni. Potrebbe essere infatti di aiuto, per uno studente italiano, leggere "lessico giuridico" passando sull'etichetta *jur*, che di per sé potrebbe essere non del tutto trasparente. Nella mia esperienza di docente mi è capitato spesso di constatare come proprio le etichette tendano ad essere saltate dal lettore, soprattutto se inesperto. Esse sono invece di grande importanza perché orientano nella lettura, evidenziando la struttura della voce e permettono una lettura verticale della glossa per arrivare rapidamente al punto che interessa, senza essere costretti a leggerla tutta.

Altro punto che, a mio avviso, dovrebbe essere migliorato data la sua importanza all'interno della lingua, è il trattamento della fraseologia. Oltre a quanto già fatto finora nei dizionari GK, sarebbe necessario innanzitutto evidenziare le espressioni idiomatiche come categoria a sé stante e chiaramente individuabile. Sarebbe inoltre auspicabile aggiungere informazioni ed esempi attivabili con un clic. Ecco sull'esempio dell'espressione idiomatica tedesca *etw im Auge haben* un confronto tra la situazione attuale e quella che potrebbe essere realizzata in future edizioni:

### Giacoma/Kolb 2014<sup>3</sup>

***etw im Auge haben*** {INSEKT, STAUBKORN}, avere qc nell'occhio; (*etw im Sinn haben*) {SEINEN EIGENEN VORTEIL}, guardare solo a qc; {EIN BESTIMMTES MODELL}, aver pensato a qc di preciso

### Giacoma/Kolb (future edizioni)

***etw im Auge haben*** {INSEKT, STAUBKORN}, avere qc nell'occhio;  
***jd/etw im Auge haben*** (*jd/etw im Sinn haben*)

{SEINEN EIGENEN VORTEIL}, guardare solo a qu/qc;

{EIN BESTIMMTES MODELL}, aver pensato a qu/qc di preciso

> Fenati/Rovere/Schemann (2009, 2011<sup>2</sup>)

jn./etw. (immer/...) **im Auge haben** 1. tenere d'occhio qu, + non perdere d'occhio qu, stare con gli occhi addosso a qu 2. avere sott'occhio qc

3 avere ben presente qc

> 1. Die Kleine muß Du beständig im Auge haben, sonst läuft sie auf die Straße und es gibt einen Unfall.

2. Seien Sie unbesorgt, ich habe das Karrussell immer im Auge. Da passiert nichts, was ich nicht sofort bemerke.

3. Wenn Du den Kerngedanken der Sache nicht immer im Auge hast, verlierst du dich in Einzelfragen.

#### **Link 1**

*Institut für Deutsche Sprache (2010): COSMAS II. Corpus Search, Management and Analysis System.* <http://www.ids-mannheim.de/cosmas2/projekt/referenz/korpora.html>

#### **Link 2**

<http://it.bab.la/dizionario/tedesco-italiano/im-Auge-haben>

In questo caso specifico sarebbe da migliorare la differenziazione tra il significato proprio e quello idiomatico anche attraverso l'evidenziazione delle strutture diverse (come una semplice ricerca in *COSMAS II* dimostra, il significato idiomatico prevede infatti anche un complemento oggetto animato e quindi questo andrebbe aggiunto). Si potrebbe inserire poi subito dopo un campo apribile con un clic nel quale trovare materiale proveniente dai dizionari fraseologici bilingui come ad esempio quello di Fenati/Rovere/Schemann. In questo primo campo potrebbero esserci solo i traduttori, mentre in un successivo, apribile se fosse necessario proseguire la ricerca, potrebbero trovarsi anche i relativi esempi d'uso. Al fondo potrebbero essere aggiunti link a banche dati, come ad esempio *COSMAS II* o il sito <http://it.bab.la/dizionario/tedesco-italiano> nelle quali il lettore possa trovare ulteriori informazioni.

## **7 Conclusioni**

I dizionari bilingui sono centrali nella didattica delle lingue ma, per essere veramente utili, devono essere pensati per le esigenze specifiche dell'apprendente. Il dialogo da poco iniziato tra linguisti e lessici-

cografi si è rivelato un'enorme spinta per quell'innovazione lessicografica difficile ma necessaria che non poteva essere ulteriormente procrastinata. Accogliendo l'invito rivolto dai linguisti ai lessicografi, Luisa Giacoma e Susanne Kolb hanno introdotto nella lessicografia bilingue italo-tedesca la registrazione sistematica dei collocatori in spazi ben identificati anche graficamente (carattere maiuscolotto tra parentesi graffe nella versione cartacea con l'aggiunta del colore rosso in quella elettronica). Hanno reso altresì possibile, nella versione elettronica disponibile sia su cd-rom che on line, la consultazione per collocatori attraverso un campo di ricerca ad essi dedicato.

Ai collocatori sono state affiancate anche oltre 40.000 formule di struttura che descrivono in modo esplicito come il lemma e i collocatori possono combinarsi tra loro. In questo modo è stato possibile superare i confini della lessicografia tradizionale costruendo dizionari di nuova generazione che forniscono informazioni esplicite e sistematiche sul contesto nel quale il lemma compare. A proposito del metodo sviluppato nel dizionario Giacoma/Kolb Gobber (2003, 448) afferma: "[...] thanks mainly to the methodology adopted, this could well be considered one of the best bilingual dictionaries of any language pair anywhere. The editors are to be congratulated on such a work, which will be relevant and useful for years to come".

Nelle versioni elettroniche più recenti sono state anche aggiunte le tabelle di flessione per tutta la lingua tedesca, compiendo un altro importante passo verso il lettore, che frequentemente ha dubbi proprio sulla morfologia del tedesco.

Naturalmente le rapide evoluzioni non solo tecnologiche del futuro permetteranno di migliorare continuamente i dizionari, non senza però la quotidiana collaborazione tra lessicografi, linguisti e case editrici per poter essere sempre di più dalla parte del lettore.

## 8 Bibliografia

- Bianco, M. T. (1996). *Valenzlexikon Deutsch – Italienisch*, Dizionario della valenza verbale. Heidelberg: Julius Groos Verlag.
- Blumenthal, P., Rovere, G. (1998). *Wörterbuch der italienischen Verben. Konstruktionen, Bedeutungen, Übersetzungen*. Stuttgart: Klett.
- Curcio, M. L. (1999). *Kontrastives Valenzwörterbuch der gesprochenen Sprache Italienisch-Deutsch. Grundlagen und Auswertung*. Mannheim: Institut für Deutsche Sprache.
- DIT = (2008<sup>4</sup>). *Dizionario Tedesco – Italiano, Italiano – Tedesco*. Torino: Paravia. Berlin [u.a.]: Langenscheidt.
- DIVA = Soffritti, M., Heinrich, W. (2005). *Dizionario valenziale degli aggettivi/Deutsch – Italienisches Valenzwörterbuch der Adjektive*. (Applicazione software).
- Dobrovól'skij, D. (2009): Zur lexikografischen Repräsentation der Phraseme (mit Schwerpunkt auf zweisprachigen Wörterbüchern). In C. Mellado Blanco (Hg.), *Theorie und Praxis der idiomatischen Wörterbücher*. Tübingen: Niemeyer, 149-168.
- ELDIT. (1999ss.) *Elektronisches Lern(er)wörterbuch Deutsch – Italienisch/ Dizionario elettronico per apprendenti italiano – tedesco*. Bolzano: Europäische Akademie ([www.eurac.edu/eldit](http://www.eurac.edu/eldit)).
- Fischer, K., Mollica, F. (2012). *Valenz, Konstruktion und Deutsch als Fremdsprache*. Frankfurt a.M. [u.a.]: Peter Lang.
- Fontenelle, T. (1997). *Turning a bilingual dictionary into a lexical-semantic database*. Tübingen: Niemeyer.

- Giacoma, L. (2011). Übersetzungsfehler und Gebrauch von zweisprachigen Wörterbüchern Deutsch – Italienisch: ein Erfahrungsbericht. In S. Bosco, M. Costa, L. Eichinger edd. *Deutsch/Italienisch: Sprachvergleiche*. Heidelberg: Winter Verlag, pp. 45-65.
- Giacoma, L., Kolb, S. (2006). L'utilità dell'introduzione sistematica delle collocazioni nella voce lessicografica bilingue. L'esempio del *Dizionario di Tedesco* (GIACOMA/KOLB, Zanichelli/Klett, 2001). In E. Corino, C. Marello, C. Onesti edd. *Atti del XII Congresso Internazionale di Lessicografia – Torino, 6-9 settembre 2006*. Alessandria: Edizioni dell'Orso, pp. 967-978.
- GK = Giacoma, L., Kolb, S. edd. (2014<sup>3</sup>, 2009<sup>2</sup>, 2001). *Il Nuovo dizionario di Tedesco*. Bologna: Zanichelli/Stuttgart: Klett.
- Giacoma, L., Kolb, S. edd. (2010). PONS Wörterbuch Studienausgabe Italienisch-Deutsch. Deutsch-Italienisch. Stuttgart: Klett.
- Giacoma, L., Kolb, S. edd. (2011). *Il Tedesco smart*. Bologna: Zanichelli/Stuttgart: Klett.
- Gobber, G. (2003). Review – *Dizionario Tedesco Italiano – Italiano Tedesco/Wörterbuch Deutsch Italienisch – Italienisch Deutsch*, hrsg. v./a cura di Luisa Giacoma e Susanne Kolb. Bologna: Zanichelli/Stuttgart: Pons Klett. 2001. In *International Journal of Lexicography* 16/4, pp. 445-448.
- Gross, M. (1967). *Analyse formelle comparée des complétives en français et en anglais*, Thèse de troisième cycle. Université de la Sorbonne. Paris.
- Gross, M. (1975). *Méthodes en syntaxe*. Paris: Hermann.
- Hausmann, F.J. (1993). Ist der deutsche Wortschatz lernbar? In *Informationen Deutsch als Fremdsprache* 20, 471-485.
- LGWDaF = Götz, D. et al. edd. (1993). *Langenscheidt Großwörterbuch Deutsch als Fremdsprache*. Berlin [u.a.]: Langenscheidt.
- Lo Cascio V. (1997). Semantica lessicale e i criteri di collocazione nei dizionari bilingui a stampa e elettronici. In T. De Mauro, V. Lo Cascio (a cura di), *Lessico e grammatica. Teorie linguistiche e applicazioni lessicografiche. Atti del convegno interannuale della Società di Linguistica Italiana*. Roma: Bulzoni.
- Marello, C. (1989). *Dizionari bilingui*. Bologna: Zanichelli.
- Marello, C., Rovere, G. (1999). Mikrostrukturen in zweisprachigen Wörterbüchern Deutsch – Italienisch/Italienisch – Deutsch. In H. E. Wiegand, *Germanistische Linguistik* 143-144 (Studien zur zweisprachigen Lexikographie mit Deutsch IV), pp. 177-206.
- Nied Curcio, M. (2006). La lessicografia tedesco-italiana: storia e tendenze. In: F. San Vicente, *Lessicografia bilingue e traduzione: metodi, strumenti, approcci attuali*. Monza: Polimetrica International Scientific Publisher, pp. 57-70.
- Nied Curcio, M. (2012). Die Valenz in der zweisprachigen Lexikographie Italienisch-Deutsch. Wohin führt der Weg? In *Studi germanici* 1/12, pp. 175-191.
- PONS = (1996). *Pons Großwörterbuch Französisch Deutsch*. Stuttgart: Klett.
- Sabatini, F. (2008). La grammatica in un dizionario. In: S. Vanvolsem, L. Lepschy, *Nell'officina del dizionario. Atti del Convegno Internazionale organizzato dall'Istituto Italiano di Cultura. Lussemburgo 10 giugno 2006*. Stuttgart: Ibidem Verlag, pp. 111-112.
- Sansoni = (2006<sup>6</sup>). *Dizionario Sansoni Tedesco-Italiano, Italiano-Tedesco*, Milano: RCS Libri.
- Schafroth, E. (2004). Anmerkungen zur lexikographischen Dimension der Lernersprachen Italienisch und Deutsch. In: *daf-werkstatt* 3, pp. 109-124.

# Bilingual Dictionary Drafting. The Example of German-Basque, a Medium-density Language Pair

David Lindemann<sup>1</sup>, Iker Manterola<sup>2</sup>, Rogelio Nazar<sup>3</sup>, Iñaki San Vicente<sup>2</sup>, Xabier Saralegi<sup>2</sup>

<sup>1</sup>UPV-EHU University of the Basque Country, <sup>2</sup>Elhuyar Foundation,

<sup>3</sup>Pontificia Universidad Católica de Valparaíso

david.lindemann@ehu.es, i.manterola@elhuyar.com, rogelio.nazar@ucv.cl,

i.sanvicente@elhuyar.com, x.saralegi@elhuyar.com

## Abstract

This paper presents a set of Bilingual Dictionary Drafting (BDD) methods including manual extraction from existing lexical databases and corpus based NLP tools, as well as their evaluation on the example of German-Basque as language pair. Our aim is twofold: to give support to a German-Basque bilingual dictionary project by providing draft Bilingual Glossaries and to provide lexicographers with insight into how useful BDD methods are. Results show that the analysed methods can greatly assist on bilingual dictionary writing, in the context of medium-density language pairs.

**Keywords:** bilingual dictionary drafting; comparable corpora; Natural Language Processing; open lexical resources; parallel corpora

## 1 Introduction

For a bilingual dictionary project that starts from scratch, from no or little previous lexicographical work and no or little bilingual glossaries (BG) existing on their language pair, a lexicographer lacks a useful set of guidelines for Bilingual Dictionary Drafting (BDD) strategies. A Dictionary Draft, i.e., lexicographical data obtained by automatic or semi-automatic methods, is useful in the lexicographical process as it may ease the editing of macro- and microstructural lexicographical data and save human resources.

In this article, we present a set of BDD methods and their evaluation on the example of German-Basque as language pair: Direct extraction of bilingual glossaries from existent lexicographical databases and Corpora based Natural Language Processing extraction methods. The evaluation of the obtained glossaries is done (1) quantitatively for the covering of German lemmata against a corpus based frequency lemma list adapted from DeReWo-40.000 (IDS 2009) as gold standard, (2) quantitatively for the amount of Basque Translation Equivalent (TE) obtained, and (3) qualitatively for the adequateness of the TE pairings, against manually edited German-Basque dictionary entries from EuDeLex, the lem-

malist of which is adapted from DeReWo-40.000, as gold standard (3406 German lemmata starting with A), and for the adequateness of the TE's part of speech (POS)<sup>1</sup>.

Our aim is twofold: to give support to the bilingual dictionary project EuDeLex by providing draft BG and to provide useful information related to BDD methods for lexicographers working on medium-density language pairs.

#### A Word about Density

Density, understood as “the availability of digitally stored material” in a language (Varga et al. 2005) is a factor not to be neglected in corpus-based lexicography. In most cases, the number of speakers of a language and its size on the web serve as approximation indicators for density, and the availability of electronic language resources is also a factor to be considered. Following Varga et al. (*op. cit.*), we group languages according to density as follows:

- (1) High-density languages: languages with a hundred million speakers or more (about 12)
- (2) Low-density languages: small languages with less than half a million speakers (more than 5000)
- (3) Medium-density languages: languages that lie between these two extremes (about 500)

Basque is one of the latter ones; Table 1 carries a comparison of density approximation indicators for German, one of the high-density languages, and Basque.

In the bilingual context, it is the density of the smaller language of the pair that determines by which methods Dictionary Draft data can be gathered and to what extent those methods lead to useful results.

Approaches for obtaining BG that rely on statistical Natural Language Processing methods (part 2.2) and that provide reliable results in higher density language pairs, in our case may lead to a much more limited success, and we shall ask whether the reasons for a more limited success of NLP methods are the quantitative and qualitative limitations of parallel and comparable corpora available for our language pair, or whether a lack of performance is explained also by the employed NLP tools themselves, which do lead to good results for the (high density) languages they were designed for. In the case of Basque, as it is not official in EU, we can not recur to parallel corpora based on EU legal documents, as we would in the case of other medium-density European languages (cf. Steinberger et al. 2006). Independently from this fact, German-Basque parallel corpora compiled from movie subtitles and software localization files may reach considerable sizes in a near future, as promised in the OPUS Corpus project (cf. Tiedemann 2012).

---

1 EuDeLex is currently being developed at UPV-EHU (cf. Lindemann 2014). The manual editing of German letter A (around 10% of the planned lemmalist based on DeReWo) has been finished. The intersecting set of EuDeLex and DeReWo (German Letter A) covers more than 90% of both. EuDeLex is available at <http://www.ehu.es/eudelex/>.



	German	Basque
Speakers	98 million	0,8 million
Biggest Corpus (token counts)	5,4 billion	0,12 billion
Wikipedia Pages	4,5 million	0,37 million
Web contents	5.7%	< 0.1%
ELRA Products	444	6

**Table 1: Some density approximation indicators for German and Basque.**

Approaches that rely on lexicographical databases maintained by human lexicographers (part 2.1) also presumably suffer from a density-bias: Wikimedia content is crowd-edited by the collaborating communities of volunteers, and those of a high-density language like German largely outnumber volunteers in Basque<sup>2</sup>. On the other hand, lexical databases maintained in an academic context by human lexicographers like WordNet, may be on a par in terms of quantity and quality, disregarding density across languages.

## 2 Bilingual Dictionary Drafting Methods

### 2.1 Extraction of BG from existent lexicographical databases

#### 2.1.1 Wikimedia

For the page titles that match to our gold-standard lemmalist, we extract the whole Wikipedia / Wiktionary page content if existent. Redirect pages are also taken into account. From the page content, we extract the Interlanguage-Link for Basque (Wikipedia) and the Basque translation links (Wiktionary).

#### 2.1.2 WordNet

In this experiment, we align German WordNet lexical units (GermaNet 8.0, see Hamp & Feldweg 1997) with Basque WordNet (EusWN 3.0, see Pociello 2007) lexical units using Princeton WordNet (PWN 3.0, see Fellbaum 1998) as pivot. GermaNet synsets (to which  $n$  lexical units belong) are referred to PWN synsets in the GermaNet Interlingual Index records (ILI). On the other hand, the EusWN datasets carry links to PWN. By parsing the WordNet data into the same XML file, we get a structure like the one shown in Fig. 1. From each of those aligned datasets, German-Basque glossary entries are extracted by pairing all German lexical units to all Basque lexical units present in the synset.

2 For instance, German *wiktionary* counts with 78634 user accounts and 199 active members over the last month, while the Basque *wiktionary* only has 1982 accounts with 11 active members (statistics from 15.09.2013).

```

<EngSynset Synset="eng-30-00042757-n">
  <PWN30 EngLexUnit="departure_1"></PWN30>
  <PWN30 EngLexUnit="going_1"></PWN30>
  <PWN30 EngLexUnit="going_away_1"></PWN30>
  <PWN30 EngLexUnit="leaving_1"></PWN30>
  <GermaNet80 GerLexUnit="Abfahrt"></GermaNet80>
  <GermaNet80 GerLexUnit="Weggang"></GermaNet80>
  <GermaNet80 GerLexUnit="Aufbruch"></GermaNet80>
  <GermaNet80 GerLexUnit="Distanzierung"></GermaNet80>
  <EusWN30 EusLexUnit="irteera_6"></EusWN30>
  <EusWN30 EusLexUnit="joanaldi_1"></EusWN30>
  <EusWN30 EusLexUnit="joate_1"></EusWN30>
</EngSynset>

```

Fig. 1: Aligned data of 3 WordNets.

## 2.2 NLP Methods

In this work, three different tools were used in order to extract lexical correspondences. Each tool depends on different NLP methods and resources. On the one hand, we applied a tool called *Pibolex*, which relies on pivoting over existing bilingual dictionaries and combines their structure and comparable corpora based methods for selecting correct translations of source words. On the other hand, we made use of two tools for bilingual lexicon extraction from parallel corpora: Giza++ and Bifid. The following sections describe those tools and the experiments we conducted with them.

### 2.2.1 Pibolex: Pivot techniques + comparable corpora word-alignment

Pivot-based bilingual dictionary building is based on merging two bilingual dictionaries which share a common language (e.g. LA-LB, LB-LC) in order to create a dictionary for a new language pair (e.g. LA-LC). In our case, we merged the German-English *Beolingus*<sup>3</sup> dictionary (Lde-en) with the English-Basque *Elhuyar*<sup>4</sup> dictionary (Len-eu), obtaining Lde-eu. However, this process may include wrong translations due to the polysemy of words. A pivot word can lead to wrong translations corresponding to senses not represented by the source word. These senses can be either completely different or related but with a narrower or wider meaning. For pruning the wrong translations, in this work we apply the Pibolex tool (Saralegi, Manterola & San Vicente 2011) which uses two different methods adequate for medium-density language pairs because they depend on resources that can be easily obtained:

- (a) Inverse Consultation (IC1) (Tanaka & Umemura 1994): this algorithm uses the structure of the source dictionaries to measure the similarity of the meanings between a source word and its translation candidates. The IC1 method counts the number of pivot words in language B between a source word in LA and its TE candidate in LC. The more pivot words found, the stronger is the evidence for the candidate to be correct.
- (b) A pruning method based on cross-lingual distributional similarity (DS) computed from a bilingual comparable corpus. Different authors (e.g. Fung 1995; Rapp 1999) have proposed to extract bilingual

3 <http://dict.tu-chemnitz.de>

4 <http://hiztegiak.elhuyar.org>

equivalents from monolingual or comparable corpora because, despite offering lower accuracy than those extracted from parallel corpora, they can be an alternative for medium and low density language pairs where parallel corpora are scarce. The underlying idea is to identify as TEs those words which show similar distributions or contexts across two corpora of different languages, assuming that this similarity is proportional to the semantic distance. The method we apply here is described in detail in Saralegi, San Vicente & Gurrutxaga (2008). Following the “bag-of-words” paradigm, a word  $w$  is represented by a vector composed of weighted collections of words. Those words are extracted from the contexts where the word  $w$  appears in the corpus. The context words are weighted with regard to  $w$  according to the Log-likelihood ratio measure. Once we have vector representations of the words in both languages, the algorithm computes for each source word in LA the cosine similarity between its context vector and the context vectors of all TE candidates in LC. However, we can not directly compare vectors in different languages. In order to overcome this problem, we translate vectors of words in LA to LC by means of the noisy bilingual dictionary Lde-eu, which is the only bilingual dictionary available at this stage of the process.

The IC1 algorithm suffers from low recall, which makes it rather inadequate for the task at hand. But the combination of it with the DS based method may be a way to tackle this problem. DS results vary depending on the corpora used for computing the cross-lingual similarities. The more comparable the corpora, the better. With that in mind, experiments were conducted over two different comparable corpora:

- (4) News comparable corpus: the first experiment was conducted using a comparable corpus composed of news articles extracted from Die Zeit<sup>5</sup> newspaper in German (29M tokens) and from Berria<sup>6</sup> newspaper in Basque (36M tokens). No effort was done to match news topics or publications dates.
- (5) Wikipedia comparable corpus: Wikipedia has been extensively exploited with NLP methods a comparable corpus (eg, Tomás et al. 2008; Paramita et al. 2012). In this case we constructed a corpus by gathering all articles that have both Basque and German versions, connected through wiki interlanguage links. The corpus has 61.484 articles per language and 91M tokens (72.5M tokens in German and 18M tokens in Basque). Although this corpus is highly comparable with respect to the topics (each article has its counterpart), it is important to note that the amount of tokens per language is unbalanced. This can lead to a decrease in the comparability degree of the corpus, because the German part holds much more information.

---

5 <http://www.diezeit.de>

6 <http://berria.info>

Table 2 shows the dictionaries used in the process and their statistics.

	#entries	#pairs
Lde-en (A) - Beolinguus	146,451	171,775
Len-eu (B) - Elhuyar	17,672	43,201
Lde-eu (A+B, no pruning)	12,939	48,097
Lde-eu (IC1)	4,305	7,211
Lde-eu (IC1+DS wiki)	7,878	18,641
Lde-eu(IC1+DS news)	7,821	20,014

**Table 2: Pivot dictionary process.**

## 2.2.2 Parallel Corpus word-alignment

There is a large tradition of parallel corpus processing in computational linguistics, starting with the work of Gale & Church (1991), Brown, Lai & Mercer (1991), McEnery & Oakes (1995) and others (see Véronis 2000 for an overview). Different methods and tools have been proposed to align parallel texts and extract lexical correspondences from them. In this investigation, we used two word alignment tools based on these methods: Giza++ (Och & Ney 2000) and Bifid (Nazar 2012).

The fact that Basque is a medium density language unlike German represents an added difficulty for any attempt in the line of Resnik (1999), who proposed to download parallel corpora by mining the web for translated pages. In our experiments we used two parallel corpora of different sizes: a German-Basque Literary Corpus created in the context of recent research (Sanz Villar 2013; Zubillaga 2013) and another built by aligning Basque and German translations of the Bible<sup>7</sup>. The first one was compiled using the content of 81 digital or digitized and OCR-ed literary German originals and their official direct translations into Basque (146,457 segment pairs). In the case of the second, after removing the books not included in both Bible versions, a parallel corpus of 30,440 segment pairs was obtained using the verse as segment unit. This is an easily built resource for medium-density language pairs because the Bible is available for a wide variety of languages (Lardilleux, Gosme & Lepage 2010; Resnik, Olsen & Diab 1999) and is, therefore, an adequate baseline parallel resource for evaluating other extraction methods over this kind of language pairs.

The aforementioned extractions tools (Giza++ and Bifid) were used in order to obtain translations pairs. The resulting figures are shown in table 3.

---

7 Basque (Elizen Arteko Biblia 1994) and German (1984 revision of the Luther Bible).

	# of seg.	# of EU tokens	# of DE tokens	# of candidates GIZA++ $p(b g) > 0.1$	# of candidates BIFID
Literary Corpus	146,457	1,948,504	2,203,307	266,678	4,838
Bible Corpus	30,440	639,581	810,671	49,443	2,926

**Table 3: DE-EU parallel corpora.**

For word alignment with Giza++, the default sequence of models were used (IBM model 1, HMM-based model, IBM model 3 and IBM model 4). The German corpus was lemmatised and POS tagged by using TreeTagger (Schmid 1995) and the Basque one using Eustagger (Ezeiza et al. 1998). Then, each word of the source corpus was substituted by a chain including the corresponding lemma and POS category. Punctuation marks and words with POS regarded as possible source of noise in the alignment process were removed from both corpora. Specifically, words excluded from the German part were those with POS tags such as APPR (preposition), APPRART (preposition with article), ART (article), KOKOM (particle of comparison), KOUS (subordinating conjunction), PRELS (relative pronoun), VAFIN (auxiliary verb, finite form) and VAINF (auxiliary verb, infinitive). From the Basque corpus, in turn, the excluded units were those with the ADL tag (auxiliary verb)<sup>8</sup>.

Giza++ returns two files of word alignments (DE-EU and EU-DE) including a probability for each word alignment. For the draft BG, BibleGiza and LitGiza word alignments with a probability  $p(b|g)$  greater than 0.1 were selected from both Bible and Literary corpora.

In order to reduce noise, the BG obtained by Giza++ was then submitted to a filtering process using a stoplist consisting of the 150 most frequent Basque words<sup>9</sup>. Two versions of these BG have been evaluated: (1) Giza++ BG after stoplist filtering, and (2) after stoplist and a filtering that only allows BG entries with the POS-tags mapping to each other in one of the following ways (see table 4):

TreeTagger flag	Eustagger flag
NN (noun)	IZE (noun)
VV (verb)	ADI (verb)
ADV (adverb)	ADB (adverb)
AD (adjective)	ADB (adverb) <sup>1</sup>
AD (adjective)	ADJ (adjective)

**Table 4: mapping of POS-tags TreeTagger (German) and Eustagger.**

The other alignment tool, Bifid, is part of a larger project comprising the analysis of language pairs where no prior knowledge is available, which means that all forms of external resources are excluded from the processing. This tool incorporates modules for the integral process of analysing a set of do-

<sup>8</sup> VAFIN and VAINF would be German equivalents of Basque ADL. The other removed German POS would be a morpheme in Basque words in almost all cases.

<sup>9</sup> The Basque stoplist has been obtained from Basque ETC Corpus data (UPV-EHU, Sarasola et al. 2013).

cuments in unknown languages with the only assumption that such set consists of a parallel corpus in two languages. In its original version, this tool separates the set of documents in the two languages, aligns each document with its most probable translation and then proceeds to align the segments inside the documents (assuming that the newline character is the segment separator). Finally, from this segment alignment, it extracts an initial bilingual vocabulary which is then used for a realignment of the corpus at the segment level. The process is iterated in this way  $n$  times to improve the quality of the alignment at all levels.

In the case of this paper, however, we only used the bilingual lexicon extraction module because our parallel corpora were already aligned at the sentence level. The corpora were also lemmatized with the above mentioned tools, but no mapping exploiting the POS tags was used because this information is not used by the algorithm<sup>10</sup>. The bilingual vocabulary extraction module of Bifid uses a combination of strategies that include co-occurrence statistics as well as length and orthographic similarity metrics. As in its original version this extracted vocabulary was intended to be used for realignment, the program is very conservative in its lexical alignment in order to avoid the reproduction of errors in subsequent steps. As a consequence, it favors precision over recall, with fewer aligned pairs having a higher probability of being correct. Further experimentation will determine the right thresholds for the best compromise between noise and silence, meaning larger sets of aligned pairs with the maximum possible purity.

## 3 Evaluation

### 3.1 Comparison of German Lemma lists

Germanet offers the best recall on DeReWo lemmata. Wikipedia and Wiktionary on their own offer a similar recall; the intersections of both of these with DeReWo also reaches a very high level (see Table 5 below):

---

<sup>10</sup> As the motivation behind Bifid is to be a language independent alignment tool, it does not use any kind of language-specific resources such as lemmatization or POS-tagging.

	Derewo		GermaNet		Wikipedia		Wiktionary	
$\cap$ Derewo			32,199	33,73%	19,461	2,22%	22,028	7.01%
$\cap$ GermaNet	32,199	80,50%			47,588	5,43%	34,309	10.93%
$\cap$ Wikipedia	19,461	48,65%	47,588	49,86%			46,968	14.96%
$\cap$ Wiktionary	22,028	55,07%	34,309	35,94%	46,968	5,36%		
$\cap$ WikiORWikt	29,164	72,91%	59,995	62.86%				
Lemma total	39,998		95,449		876,309		314,016	

**Table 5: DeReWo and existing lexicographical databases: German lemma counts (A-Z) and intersecting sets.**

The best recall is offered by LitGiza, with a notable difference regarding the rest of drafts, even BibleGiza. Pibolex recall is second best but far from LitGiza. BibleBifid and LitBifid offer a very low recall (see table 6).

	Derewo	Bible Giza Stop	Bible Giza StopPos	LitGiza Stop	LitGiza StopPos	Bible Bifid	Lit Bifid	Pibolex Wiki	Pibolex News
Derewo		4,639 (34.97%)	3,500 (38.69%)	15,775 (23.30%)	12,846 (24.93%)	1,007 (40.84%)	2,995 (67.27%)	5,812 (77.34%)	5,753 (77.14%)
BibleGiza Stop	4,639 (11.60%)		9,047 (100.00%)	5,001 (7.39%)	3,879 (7.53%)	2,372 (96.19%)	1,122 (25.20%)	1,868 (24.86%)	1,851 (24.82%)
BibleGiza StopPos	3,500 (8.75%)	9,047 (68.19%)		3,763 (5.56%)	3,248 (6.30%)	1,699 (68.90%)	957 (21.50%)	1,518 (20.20%)	1,504 (20.17%)
LitGiza Stop	15,775 (39.44%)	5,001 (37.70%)	3,763 (41.59%)		51,533 (100.00%)	1,125 (45.62%)	4,389 (98.58%)	4,571 (60.83%)	4,549 (60.99%)
LitGiza StopPos	12,846 (32.12%)	3,879 (29.24%)	3,248 (35.90%)	51,533 (76.12%)		935 (37.92%)	3,864 (86.79%)	3,960 (52.69%)	3,955 (53.03%)
Bible Bifid	1,007 (2.52%)	2,372 (17.88%)	1,699 (18.78%)	1,125 (1.66%)	935 (1.81%)		526 (11.81%)	544 (7.24%)	546 (7.32%)
Lit Bifid	2,995 (7.49%)	1,122 (8.46%)	957 (10.58%)	4,389 (6.48%)	3,864 (7.50%)	526 (21.33%)		1,404 (18.68%)	1,398 (18.74%)
Pibolex Wiki	5,812 (14.53%)	1,868 (14.08%)	1,518 (16.78%)	4,571 (6.75%)	3,960 (7.68%)	544 (22.06%)	1,404 (31.54%)		7,309 (98.00%)
Pibolex News	5,753 (14.38%)	1,851 (13.95%)	1,504 (16.62%)	4,549 (6.72%)	3,955 (7.67%)	546 (22.14%)	1,398 (31.40%)	7,309 (97.26%)	
Lemma total	39,998	13,267	9,047	67,699	51,533	2,466	4,452	7,515	7,458

**Table 6: DeReWo and BG German entries intersections (A-Z, NLP methods).**

### 3.2 Quantitative and Qualitative Evaluation of BG

Table 7 shows results of the qualitative evaluation carried out manually by a human lexicographer. EuDeLex is set as gold standard for comparison, regarding lemmalist and evaluation of TE appropriateness, in terms of (a) a full matching as suitable Basque TE for one of the word senses of the German BG headword (OK), (b) a semantic mismatch (FALSE), (c) a semantic (fuzzy) matching without being the TE a valuable equivalent to cite in a dictionary entry (NEAR), or (d) as PART, when a BG entry is a correct TE for a lemma as part of a Multi Word Expression in the other language. A second variable, part of speech (POS) is evaluated as matching (OK) or mismatching (FALSE).

The BG obtained from aligned WordNet synsets offers a relatively high recall on GS lemmata and, in absolute figures, the largest proportion of correct TEs. Wikipedia and Wiktionary extraction is less effective in terms of recall, but more effective with regard to TE adequateness.

Among the NLP methods, Giza and Pibolex BG offer the largest number of TE evaluated as correct, far ahead of Bifid, which on the other hand returned very little false TE and POS among the results. Giza BG, and, to a lower extent, Pibolex BG, suffer from a high percentage of inadequate, noisy TE.

	LitGiza Stop	LitGiza StopPos	BibleGiza Stop	BibleGiza StopPos	LitBifid	BibleBifid	Pibolex News	Pibolex Wiki	Wikipedia	Wiktionary	WordNet	ANY DRAFT
DE Letter A Lemma with EU TE	5,361	3,826	1,349	818	265	230	682	730	5,457	396	1,076	
Intersection with EuDeLex GS	1,276	696	434	323	177	49	559	567	265	147	774	2,928
<b>Recall on 3406 GS lemmata</b>	<b>35,41%</b>	<b>19,31%</b>	<b>12,04%</b>	<b>8,96%</b>	<b>4,91%</b>	<b>1,36%</b>	<b>15,51%</b>	<b>15,73%</b>	<b>7,35%</b>	<b>4,08%</b>	<b>21,48%</b>	<b>81,24%</b>
Ø TE per Lemma	2,27	1,73	2,35	1,46	1,07	1,22	2,64	2,82	1,00	1,33	2,57	2,28
Evaluated Lemma	1,276	696	434	323	177	49	559	567	265	147	774	4,248
Evaluated TE	2,901	1,206	1,019	470	189	60	1,476	1,601	265	195	1,988	9,694
• OK	746	537	256	215	178	43	939	1,007	236	189	1,654	5,248
• FALSE	1,923	519	728	230	3	15	414	513	6	3	188	3,793
• PART	186	122	11	6	6	1	14	7	4		17	246
• NEAR	46	28	24	19	2	1	109	74	19	3	129	407
Evaluated POS	2,901	1,206	1,019	470	189	60	1,476	1,601	265	195	1,988	9,694
• OK	1,071	965	448	422	186	52	1,321	1,286	245	195	1,983	6,787
• FALSE	1,830	241	571	48	3	8	310	315	21		5	3,063
Lemma with 1+ TE OK	609	439	213	189	168	39	443	454	236	147	700	2,081
Lemma with all TE OK	315	294	111	149	168	35	316	322	236	142	601	1,123
Lemma with something usable	444	232	129	62	7	6	165	164	23	5	143	1,056
Lemma with all TE FALSE	517	170	194	112	2	8	78	82	6		30	371
<b>TE OK</b>	<b>25,72%</b>	<b>44,53%</b>	<b>25,12%</b>	<b>45,74%</b>	<b>94,18%</b>	<b>71,67%</b>	<b>63,62%</b>	<b>62,90%</b>	<b>89,06%</b>	<b>96,92%</b>	<b>83,20%</b>	<b>54,14%</b>
<b>POS OK</b>	<b>36,92%</b>	<b>80,02%</b>	<b>43,96%</b>	<b>89,79%</b>	<b>98,41%</b>	<b>86,67%</b>	<b>89,50%</b>	<b>80,32%</b>	<b>92,45%</b>	<b>100,00%</b>	<b>99,75%</b>	<b>70,01%</b>
<b>Lemma with 1+ TE OK</b>	<b>47,73%</b>	<b>63,07%</b>	<b>49,08%</b>	<b>58,51%</b>	<b>94,92%</b>	<b>79,59%</b>	<b>79,25%</b>	<b>80,07%</b>	<b>89,06%</b>	<b>100,00%</b>	<b>90,44%</b>	<b>48,99%</b>
Lemma with all TE OK	24,69%	42,24%	25,58%	46,13%	94,92%	71,43%	56,53%	56,79%	89,06%	96,60%	77,65%	26,44%
Lemma with something usable	34,80%	33,33%	29,72%	19,20%	3,95%	12,24%	29,52%	28,92%	8,68%	3,40%	18,48%	24,86%
Lemma with all TE FALSE	40,52%	24,43%	44,70%	34,67%	1,13%	16,33%	13,95%	14,46%	2,26%	0,00%	3,88%	8,73%

Table 7: Manual Qualitative Evaluation.

## 4 Conclusions

### 4.1 Discussion



Combining all BDD methods presented here, we obtain a BG that covers more than 80% of a dictionary lemmalist based on DeReWo, and provides one or more correct TE for about a half of those.

For BDD purposes, correct TE must be separated from not suitable (noisy) BG entries; TE adequateness has to be the key criterion for lexicographical needs, before the amount of gathered data. In the ongoing editing process of EuDeLex, draft data will be divided in three groups, (1) methods with no or very little results evaluated as FALSE; the data obtained by those may be included in dictionary entries and published, without manual post-editing, (2) methods with high precision results (low degree of noise); the BG obtained by these methods could be pasted into the bilingual lexicographical database for manual post-editing, and (3) methods with a larger proportion of noisy results; the BG obtained by those will have to be post-processed in order to reduce false TEs; the POS-mapping approach presented here for Giza is a first step in that direction, enriching a Basque stoplist for results proposed by Giza from the list of Basque TEs that repeatedly have been evaluated as FALSE will be the next.

We propose to group the methods presented in this paper according to the criteria mentioned above as follows:

- (1) Wiktionary, Wikipedia, LitBifid
- (2) BibleBifid, WordNet
- (3) LitGizaStop, BibleGizaStop, LitGizaStopPos, BibleGizaStopPos, Pibolex News, Pibolex Wiki

As we found out in this investigation, more than two thirds of the DeReWo list, on which a lemmalist for EuDeLex that covers the whole alphabet will base on, are linked to a dataset in Wiktionary and/or Wikipedia. The high rates in our qualitative evaluation reached by these methods encourage us to make use of them, and thanks to their open licence, it is possible. While other draft data needs human post-editing before inclusion in published bilingual dictionary entries, relevant data from those sources may be directly included in a dictionary search result webpage<sup>11</sup>. The recall these sources offer for German-Basque TE is still limited; it will mainly depend on the growth and activity of the Basque editor communities to increase it, which is, supposedly, a matter of time. Measurements like those proposed in this investigation may serve to monitor that process.

The Basque WordNet EusWN has been actively developed by human lexicographer teams at UPV-EHU, and it is today the largest and trustworthiest Basque lexical resource available with open data sources. The approach to align its synsets with GermaNet synsets using Princeton WordNet as pivot has been the one which delivered the largest proportion of correct TE for more German lemmata.

Pibolex overall results are similar to those obtained for other language pairs, confirming the tool performs robustly across languages. With respect to the corpora used in the experiments, the news corpus achieves slightly better results. This means that the Wiki corpus, although more comparable in terms of topics, suffers from the difference in amount of text between languages. The results obtained by both word alignment tools from two parallel corpora of different size show, as was to be expected,

---

11 Data found in these sources relevant to a bilingual dictionary is not only a TE, but also encyclopedic, phonetic, morphological, syntactic and pragmatic information or audiovisual material about a lemma.

that recall rates relate to corpus size, and the same is true for result precision. A further development of German-Basque parallel corpora is strongly desired.

In spite of German-Basque being a medium-density language pair with limited bilingual lexical and corpus resources, the amount of adequate BG entries gathered during the presented experiments is considerably high, and it will help saving human resources in dictionary writing. There is no need to say, however, that human lexicographers are still the key factor for a German-Basque dictionary writing that would meet acceptable quality standards.

## 4.2 Future Work

A future line of work would be to create higher comparability degree corpora, taking care of maintaining balance in terms of size, topics and genres across languages, without decreasing the overall size of the corpora. Further research about BDD would also include rendering optimization of the applied corpus based methods for the language pair German-Basque (enhancements of corpus tagging, word alignment stoplists and parameter tuning), a sophistication of these (e.g., by making use of syntactic information), as well as reproducing these experiment sets for other language pairs, which would allow for a comparison of results. Other goals not achieved at the present stage are the inclusion of multiword expressions in the experiments and measurements of polysemy covered by draft BG. We are now centering our efforts in developing a new method for the exploitation of Wikipedia as a comparable corpus using the frequency distribution of lexical units in the articles. We are representing the relative frequency of words in the articles as curves, and then comparing the curves in a purely geometrical fashion using Euclidean distance. We assume that German and Basque words with similar frequency curves will be equivalent, however we still need to find the way to make up for the already mentioned asymmetry in the amount of text in the corresponding articles in both languages.

## 5 References

- Brown, P.F., Lai, J.C. & Mercer, R.L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 91, Stroudsburg, PA.: Association for Computational Linguistics, pp. 169–176.
- Ezeiza, N., Alegria, I., Arriola, J.M., Urizar, R. & Aduriz, I. (1998). Combining stochastic and rule-based methods for disambiguation in agglutinative languages. In *Proceedings of the 17th international conference on Computational linguistics*. Association for Computational Linguistics, pp. 380–384.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.
- Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the 3rd Workshop on Very Large Corpora*. Boston, MA., pp. 173–183.
- Gale, W.A. & Church, K.W. (1991). Identifying Word Correspondences in Parallel Texts. In *Proceedings of the ACL Workshop on Speech and Natural Language*. ACL 91, Stroudsburg, PA: Association for Computational Linguistics, pp. 152–157.

- Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid: Association for Computational Linguistics, pp. 9–15.
- Lardilleux, A., Gosme, J. & Lepage, Y. (2010). Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*. LREC 2010, Malta, pp. 252–256.
- Lindemann, D. (2014). Zweisprachige Lexikographie des Sprachenpaares Deutsch-Baskisch. In Domínguez Vázquez, M.J., Mollica, F. & Nied, M. (eds.) *Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik*, Lexicographica Series Maior. De Gruyter.
- McEnery, A.M. & Oakes, M.P. (1995). Sentence and word alignment in the CRATER project: methods and assessment. In *Proceedings of the EACL-SIGDAT Workshop: from texts to tags, Issues in Multilingual Language Analysis (ACL)*. Dublin, pp. 77–86.
- Nazar, R. (2012). Bifid: un alineador de corpus paralelo a nivel de documento, oración y vocabulario. In *Linguamática*, 4, 45–56.
- Och, F.J. & Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 00, Hongkong: Association for Computational Linguistics, pp. 440–447.
- Paramita, M.L., Clough, P., Aker, A. & Gaizauskas, R.J. (2012). Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles. In *Proceedings of the Eighth conference on International Language Resources and Evaluation*. LREC 2012, Istanbul, pp. 790–797.
- Pociello, E., Agirre, E. & Aldezabal, I. (2011). Methodology and construction of the Basque WordNet. In *Language Resources and Evaluation*, 45, 121–142.
- Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 99, College Park, MD.: Association for Computational Linguistics, pp. 519–526.
- Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL 99, College Park, MD.: Association for Computational Linguistics, pp. 527–534.
- Resnik, P., Olsen, M.B. & Diab, M. (1999). The Bible as a parallel corpus: Annotating the “Book of 2000 Tongues.” In *Computers and the Humanities*, 33, 129–153.
- Sanz Villar, Z. (2013). Hacia la creación de un corpus digitalizado, paralelo, trilingüe (alemán-español-euskera). In Sinner, C. & Van Raemdonck, D. (eds.) *Fraseología contrastiva del alemán y el español. Traducción y lexicografía, Études linguistiques Linguistische Studien*. München: Peniope, pp. 43–58.
- Saralegi, X., San Vicente, I. & Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *Proceedings of the 1st workshop on Building and using Comparable Corpora (BUCC)*. LREC 2008, Marrakech.
- Saralegi, X., Manterola, I. & San Vicente, I. (2011). Analyzing Methods for Improving Precision of Pivot Based Bilingual Dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP 2011, Edinburgh.
- Sarasola, I., Landa, J. & Salaburu, P. (2013). Egungo Testuen Corpora. UPV-EHU University of the Basque Country
- Schmid, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*. Dublin, pp. 47–50.
- Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D. & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. LREC 2006, Genoa.
- Tanaka, K. & Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 297–303.

- Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth conference on International Language Resources and Evaluation*. LREC 2012, Istanbul, pp. 2214–2218.
- Tomás, J., Bataller, J., Casacuberta, F. & Lloret, J. (2008). Mining wikipedia as a parallel and comparable corpus. In *Language Forum*, 34.
- Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. & Trón, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*. Borovets, pp. 590–596.
- Véronis, J. (2000). *Parallel Text Processing: Alignment and use of translation corpora*. Kluwer.
- Zubillaga, N. (2013). *Alemanetik euskaratutako haur- eta gazte-literatura: zuzeneko nahiz zeharkako itzulpenen azterketa corpus baten bidez*. PhD Thesis. Vitoria-Gasteiz: UPV-EHU.

### **Acknowledgements**

This study has been supported by the following projects: IT665-13, Zubiak (Saiotek-SA-2013/00308) and Ber2tek (Etor tek-IE12-333), funded by the Basque Government; and project EC FP7/SSH-2013-1 AThEME (613465), funded by the European Commission. Funding is gratefully acknowledged.

# Illustrative Examples and the Aspect of Culture: The Perspective of a Tshivenda Bilingual Dictionary

Munzhedzi James Mafela  
University of South Africa  
mafelmj@unisa.ac.za

## Abstract

Few Tshivenda dictionaries make use of illustrative examples to define lexical entries. Among those which use illustrative examples to define lexical entries is *Venda Dictionary: Tshivenda – English*, which makes use of simple sentences, idioms, riddles and proverbs. Through the use of illustrative examples lexicographers show the headword in use in typical contexts (Katzaros 2004). In many instances illustrative examples in *Venda Dictionary: Tshivenda -- English* reflect on the culture of the Vhavenḁa; information which is valuable to Tshivenda learners. Learners of a foreign language find it difficult to understand meanings of cultural words. Bartholomew (1995:4) writes that a cursory reading of the illustrative sentences not only helps the reader to understand the entry word, but it also gives him a feel for the kind of life led by the speakers. This paper seeks to highlight the importance of illustrative examples in familiarising dictionary users with the culture of the Vhavenḁa. Data from *Venda Dictionary: Tshivenda – English* will be used to facilitate the discussion of illustrative examples and the aspect of Tshivenda culture.

**Keywords:** Culture; illustrative example; phrase; proverb; idiom

## 1 Introduction

Bilingual dictionaries deal with the translation of entry words from the source language into the target language. In addition to the provision of words with equivalent meaning in the target language, a bilingual dictionary defines the entry word in order to make the meaning clearer to dictionary users. Sometimes such a definition of the entry word will be difficult to comprehend unless it is explained in context. This poses a challenge to dictionary users, especially those who are learners of the language, who may not find the meaning they are looking for. In order to avoid such misunderstandings, some lexicographers make use of illustrative examples to make the meaning of a word clearer. As Kavanagh (2000:101) writes: “We need to know about words and their meanings, but we also need to know about attitudes, manners, and social norms”. The functional inclusion of examples, illustrating actual everyday language usage, is of prime importance because it adds to the quality and user-friendliness of a dictionary (Prinsloo & Gouws 2000:139). Illustrative examples give dictionary users the opportunity to learn more about the way of life of the source language community, including its cultu-

ral aspects. Such knowledge of the community's culture provides a key to understanding the meaning of lexical entries. Illustrative examples may serve as vehicles that clarify the meaning of culturally specific words.

An examination of the cultural setting and the equivalence in bilingual dictionaries reveals that language does not exist in a vacuum but occurs in context. Hence, relationships between the various languages as well as within a language become clear. One of the important aspects is the cultural context in which a word is used. (Gangla-Birir 2005:40)

The inclusion of cultural aspects in the definition of a lexical entry in a dictionary would be of great benefit to language learners in South Africa, a country with many racial and ethnic groups. Kavanagh (2000:101) asserts: "In South Africa, a country where people are being positively encouraged to learn another language and to communicate across cultures, the need to develop cultural as well as linguistic awareness should be emphasised".

The purpose of this paper is to highlight the importance of illustrative examples in familiarising dictionary users with the culture of a community. As Whitcut (1995:255) argues: "... it is in the use of examples that our cultural presuppositions become most obvious". Examples from the *Venda Dictionary: Tshivenda - English* will be cited to illustrate how learners of Tshivenda can be helped to acquire information on cultural aspects of the lives of the Vhava.

## 2 Illustrative Examples

Illustrative examples can take the form of sentences or phrases. On the use of sentences as illustrative examples, Bartholomew (1995:3) cites Bernard and Salinas Pectraza (1989) who state: "The dictionary sentences provide a series of snapshots of the daily life of the Mezquital Otomi Indians. They don't constitute ethnography as such but the information they supply is consistent with a fuller ethnography written by an Otomi speaker." The message in this quotation applies not only to Mezquital Otomi Indians but to all communities. Good illustrative sentences help to define the entry word by using it correctly in a typical context (Bartholomew 1995:3). When dictionary users read illustrative examples they can imagine the kind of life led by the source language community. Learners get to know how the example works and how it collocates (Katzaros 2004). Language has a close connection to the culture that produced it.

A language is not solely composed of words interacting with one another; it has a close connection to the culture that produced it. Thus, learning a language is also getting to know the particular aspects of the culture behind it. Examples can play an essential part in showing culture by illustrating words that have a specific cultural dimension. (Katzaros 2004:491)

A bilingual dictionary with illustrative examples is user-friendly because such examples show the entry word in context, and distinguish one meaning from another, illustrating grammatical patterns and typical collocations, and indicating appropriate registers (Drysdale 1987 cited by Al-Ajmi 2008:16).

### 3 Illustrative Examples and the Aspect of Culture in Venda Dictionary: Tshivenda – English

Illustrative examples in the *Venda Dictionary: Tshivenda – English* take several different forms. Some of the examples are in the form of proverbs, whereas others are in the form of phrases, riddles and idioms. This discussion will focus on proverbs, idioms and phrases as they are used to highlight cultural information about a language community. Illustrative examples in the *Venda Dictionary: Tshivenda – English* are informative because they illustrate the use of the word and enhance the dictionary user's understanding of its semantic range and the culture of the Vhava (Mafela 2008:31). Kavanagh (2000:112) comments thus on the presentation of cultural items in the definition of a lexical item in Van Warmelo's dictionary: "In the *Tshivenda – English Dictionary* edited by Van Warmelo cultural snippets follow the translation". Through these illustrative examples, dictionary users learn more than the mere meaning of a lexical entry.

#### 3.1 Phrases

Van Warmelo (1989) has a tendency to use phrases to present the lexical entry in context. However, only those entry words whose meanings are figurative or cultural are explained in context. One such word is the lexical entry *-bva*. This entry is partly defined as follows:

- (1) *-bva* 1 go out, come out, emerge, issue, come from. -- *malofha* bleed. -- *ngomu* escape. -- *phaḍi* get scabies (which come out of the skin). *Nḍila i* -- *nga vhukati ha tsimu* path goes right through the garden. –  
*ḥhangu* go to consult a diviner.

The lexical entry *-bva* can be defined in three senses, i.e. going out or coming from, persevere, and pass. However, the English equivalents of *-bva* in the sense of going out or coming from are: go out, come out, emerge, issue, and come from. A number of illustrative examples in this regard have been provided to show its use in context: *-bva malofha* (bleed), *-bva ngomu* (escape), *-bva phaḍi* (scabies coming out of the skin). Only the last of these, *-bva ḥhangu* (go to consult a diviner) reveals an aspect of culture. The phrase *-bva ḥhangu* has been used as an illustrative example to distinguish its meaning from any other. According to the Tshivenda culture, if a member of a family dies, his/her relatives go out to consult a diviner to find out the cause of the death. This is a popular belief among the Vhava, as it is in some other African communities. These groups believe that people do not die of natural causes; they believe that death is brought about by another person or by the ancestors. This action is taken not only when a death takes place, but also when the family encounters misfortunes. A diviner is an important person in the lives of the Vhava in particular, and in African communities in general. Through this

illustrative example, learners of Tshivenda come to understand some of the cultural beliefs and religious activities of the Vhavana.

The use of a phrase as an illustrative example to reveal the culture of the Vhavana is also realised in the definition of the lexical entry *dzekiso*; which is defined this as follows:

- (2) *dzekiso* 5 bride-price for a wife, (or, usually, one animal towards it), given by one's father or other close relative, see *-dzekisa*. *Musadzi wa --* great wife, whose son will be heir (Van Warmelo 1989:35)

There is only one English equivalent for the lexical entry *dzekiso*, and that is bride-price of a wife. However, this is not an ordinary bride-price; it is the price paid for a wife who is highly regarded in the family. Polygamy is not a foreign practice among the Vhavana and it is perfectly acceptable to the community for a man to marry more than one wife. However, a woman whose bride-price is paid by the father of the husband or a close relative in the family is known as *Musadzi wa dzekiso* (the great wife). This woman is the one who gives birth to the heir in the family. In the case of the royal house, *Musadzi wa dzekiso* (the great wife) gives birth to a future king/chief. Normally, *dzekiso* (bride-price) will be paid in the form of animals. This knowledge, beliefs, and practices of a particular society are reflected in a language (Kavanagh 2000:103).

The definition of the entry word *-kuvha* shows, too, how supplementary information regarding the meaning of a word can be provided through the use of illustrative examples:

- (3) *-kuvha* 1 wash, as clothes. 2 (Kar. Idem) stop bearing (of trees at end of season), cease to be in season (of fruit). *Mapapawa a vho -- zwino pawpaw* are coming to an end. 3. finish payment, make up the balance outstanding, pay in full. -- *misho* pay all the presents due to bride's parents. *Ndi do -- dzothe* I shall take all. *Engedza thanu uri dzi kuvhe zwothe* pay a further five to make up the total

*Muthu o felaho kule na haya u a kuvhiwa*. If a man dies far from home and cannot be buried there, a sheep is killed and its head buried near his ancestors' graves by a doctor and the family. This is a substitute, and they say: "Mudzimu o vhuya hayani" (Van Warmelo 1989: 122).

Four meanings of the lexical entry *-kuvha* have been identified in the definition above. Meanings 1, 2 and 3 are clearly spelt out. However, the illustrative example in 3 is accompanied by information on a cultural aspect. The example *-kuvha misho* and the word *misho*, in particular, means to *pay all the presents due to a bride's parents*. Through this illustrative example, dictionary users learn that a bride-



groom is expected to produce all the presents before he is allowed to take his bride. If not, the bride will not be given permission to join the in-laws. *Kuvha* in this instance means pay.

Another interesting aspect of culture is revealed in illustrative example 4 (*Muthu o felaho kule na haya u a kuvhiwa* - If a man dies far from home and cannot be buried there, a sheep is killed and its head buried near his ancestors' graves by a doctor and the family). This illustrative example is not numbered and the meaning of the entry *-kuvhiwa* is not easily identified. Its use in this context reveals a cultural aspect of the lives of the Vhavanḁa. The Vhavanḁa believe that when a man dies far from home, he should be buried with his ancestors. This is not possible if the corpse is not brought back home. Instead, family members and the doctor will kill a sheep as a substitute. If this does not happen, something bad will befall the family. This is the reason a sheep is killed and buried near his ancestors' graves. In this regard, the Vhavanḁa say "Mudzimu o vhuya hayani", which means "God has come back home". Through this illustrative example, dictionary users are shown an aspect of Vhavanḁa culture they may not have realised if the example had not been provided. *Kuvhiwa* in this regard can be explained as: a sheep is killed and its head buried near the graves of the deceased's ancestors. *Kuvhiwa* in means replaced. The vocabulary of a language can provide some evidence of what is considered culturally important (Kavanagh 2000:103).

### 3.2 Idioms

Van Warmelo uses not only simple phrases to illustrate the meaning of lexical entries. He also makes use of idioms. In defining idioms Guma writes:

Idioms are characteristic indigenous expressions, whose meanings cannot be ordinarily deduced or inferred from a knowledge of individual words that make them up. They are native to a language, and have the stylistic effect of giving it a typical native ring that is characteristic of its mode of expression (Guma 1977:66).

Many idioms in Tshivenḁa depict the way the Vhavanḁa live, including their cultural beliefs. Even if their meanings cannot be ordinarily deduced from the knowledge of individual words that make them up, a number of idiomatic expressions have been used to supplement the information on the meaning of a lexical entry in a definition. Some idiomatic expressions used by Van Warmelo are *u bva dzimamuḁi* (go out of the village) and *u tevhele maḁamu* (collect fees due to a chief for Vhusha rites). The lexical entry *dzimamuḁi* is defined as follows:

- (4) *dzima-muḁi* (or *dzimu-muḁi*, *dzimuḁi*) in *-bva* -- go outside the village (sc. to stay there during an illness, to get fresh air, to be undisturbed, or for other reasons); *vha mu bvisa* -- they are taking the patient outside the village, away to relatives (Van Warmelo 1989:40)

*Dzimamuḍi* is a compound noun meaning “going outside the village”. One may go outside the village for various reasons, such as to stay away during an illness, to get fresh air, or to be undisturbed. However, the Vhavenḍa believe that when one is ill, one is more likely to get better if one goes outside the village. The illustrative example that serves as a vehicle to teach the traditional way of doing things is *vha mu bvisa muḍi* (they take him/her outside the village). According to the Vhavenḍa, when a person is seriously ill, he/she will not be healed if kept in the village. This is attributed to many factors, among them witches that might live in the village. The Vhavenḍa believe that a person is bewitched by someone closest to him/her, i.e. a relative. When he/she leaves the village, relatives will not have the opportunity to exacerbate the illness. The Vhavenḍa even have a proverb to support this belief: *Mutsinda ndi khwine shaka ndi bulayo* (A stranger is better than a relative; a kinship may be the death of one). This means that one will be safer outside the village than within it. The Vhavenḍa trust strangers more than they do kinsmen.

Another example of an idiomatic expression used as an illustrative example to supplement the information in the definition of a lexical entry is *U tevhela maḍamu* (collect fees due to a chief for Vhusha rites). The lexical entry *ḍamu* is defined as follows:

- (5) *ḍamu* 5 1 udder, female breast, hence e.g. *thungo ya ḍamuni* the mother’s side (of family, of one’s relatives). 2 node or hand of bananas, i.e. those growing from the same level on the stem of the bunch. 3 -tevhela maḍamu (follow the breasts) collect fees due to a chief for Vhusha rites, from girls who have gone through these rites elsewhere. 4 -vha na -- be in calf, in foal (ruminants, equines) (Van Warmelo, 1989:20).

The English equivalents of *ḍamu* are udder and female breast. However, the word can be used in different contexts to mean different things. Our concern in this discussion is with the context in which it has been used in meaning 3. -tevhela maḍamu, loosely translated as follow the breasts. This reveals a cultural aspect of the life of the Vhavenḍa. When Vhavenḍa children grow up they undergo various rites to prepare them for the challenges of adult life; in this case it is the *Vhusha* rite. *Vhusha* is a puberty rite for girls. In order to attend this rite, girls must pay a fee. *U tevhela maḍamu* means to collect fees due to the chief for the *Vhusha* rite from those girls who have undergone these rites elsewhere. Girls who attend these rites in other villages are made to pay a fee to their chief, which is referred to as *maḍamu*. The most important point here is that learners of Tshivhenḍa learn about this cultural activity through the definition of the lexical entry *ḍamu*. Besides the meanings given, a learner receives this added information.

### 3.3 Proverbs

Proverbs express a community's commonly held ideas and beliefs. They are concerned with those things that the people know in their daily lives, and not with things that fall outside the scope of their experience (Guma 1977:66). However, proverbs are found in various communities of the world. Comparisons of proverbs found in various parts of the world show that the same kernel of wisdom may be gleaned under different cultural conditions and languages (McHenry 1992:749). As indicated above, many proverbs have been used as illustrative examples in *Venda Dictionary: Tshivenda – English*. Some of these are discussed below.

(6) -vhanga 1 wrongfully lay claim to what is another's inherited right or entitlement, as vhukoma chieftainship, headship of family, ifa inheritance; dispute, challenge or fight over such right. 2 make oblique cut with axe, as to facilitate straightening a pole; make oblique cuts towards one another to remove wedge & get at inside of wood (Mbani i vhangwa mitandani the mbani bee is got out of wood with oblique cuts); hammer and flatten to make thinner, as wire. 3 cause, bring about, aggravate, as sickness or accident. Goloi yo mbangela khombo the cart did me a mischief (as by overturning). Zwi mbangela mushumo this makes work for me (that I don't want)

Prov: Ha sa vhangwa, a vhu lalami "Unless a chief has had to fight for his position, he does not keep it long" (Van Warmelo 1989:451)

The lexical entry *-vhanga* has three meanings. All these meanings have been well defined; they do not confuse the dictionary user. The lexicographer has gone a step further by explaining meaning 1 by providing an illustrative example in the form of a proverb. The proverb comments on chieftainship, that is, if his position is not fought for, the chief will not keep it long. It is a well-known fact that among the Vhavanḁa a chief is born into the position. However, according to Vhavanḁa culture, the heir should fight for this position. This is why, before the installation of a chief/king, there will be conflicts where different groups will fight for their favourite to be installed as chief/king of the people. Such conflicts make the winner a stronger person. According to Van Warmelo (1989:151), "No chieftainship that succeeds, has not survived the test of force at its inception. This quarrel over the succession forces the leading spirits to show their colours, and the victor can then eliminate his enemies. His victory convinces the passive majority that he is indeed the right man. He has "got something", and his ancestors clearly support him." At the end of the quarrels the rightful person is the victor, and his reign lasts a long time.

Another illustrative example which takes the form of a proverb is *Wa kokodza luranga, mafhuri a a tevhe-la* (You pull at the calabash creepers and the calabashes follow). This illustrative example is used to add meaning to the lexical entry *luranga* which is defined as follows:

- (7) *luranga* 11 (pl. *thanga* 10) cultivated cucurbit, plant of any var. of pumpkin, calabash, melon  
Prov. *Wa kokodza luranga, mafhuri a a tevhela* “You pull at the calabash creepers and the calabashes follow,” i.e. to get at the important facts of a case one discusses irrelevant details (Van Warmelo 1989:151)

This proverb makes a comment on the importance of details that may seem irrelevant but that can be valuable in supporting the important facts. Irrelevant details are not supposed to be ignored when one is discussing important facts. This proverb is usually used by the *Vhavenḁa* to comment on the actions of individuals, mainly men. In *Tshivenḁa* culture when a man gets married to a woman who has children by another man, he is expected to take all her children as his. These children are not to be ignored or left behind with the in-laws; they are to be treated as his own and they in turn should address him as “father”. He is not called “stepfather”, nor are the children called “stepsons/daughters” because the concepts of “stepfather” and “stepson/daughter” do not exist among the *Vhavenḁa*. This is such an important aspect of the culture that the children change their surname/ last name and adopt that of their new father.

Among the *Vhavenḁa* women are traditionally regarded as minors. This is supported by some of the proverbs in use in their daily life. In the definition of the lexical entry *tsadzi*, a proverb *Khuhu tsadzi a i imbi (ambi) mutsho, i imbaho (ambaho) ndi ya nduna* sums it all up. The entry word is defined as follows:

- (8) *tsadzi* 9,10 of -*sadzi* female; of female nature, origin. *O dzula thaka -- (= ya tshisadzini)* he inherited from the female side, i.e. from his mother.

Prov: *Khuhu tsadzi a i imbi (ambi) mutsho, i imbaho (ambaho) ndi ya nduna* “A hen does not announce the dawn, it is the cock that crows” i.e. a woman is a minor and may not discuss matters in public proceedings (Van Warmelo 1989: 385)

As women are considered minors, they are not usually allowed to participate at the *khoro* where public discussions are held and cases heard. Only men participate in such gatherings and they decide and hear cases on behalf of women. Women, on the other hand, accept this cultural aspect and abide by all decisions made by men.

## 4 Conclusion

The above exposition has shown that the *Venda Dictionary: Tshivenḁa - English* has a wealth of illustrative examples which have been used to convey cultural information to learners of the language. Cultural words exist in context. The knowledge, beliefs and practices of a society are revealed in the defi-

nitions of lexical entries. As a result, when looking for meanings of lexical entries, users of this dictionary also learn about the beliefs and practices of the Vhavenḁa. Cultural information merits inclusion in dictionaries and this can be achieved through the use of illustrations (Gangla-Birir, 2005:39). Prinsloo and Gouws (2000:139) have this to say: “Illustrative examples play a vital role in dictionaries, and the dictionary conceptualisation plan of any new lexicographic project should make provision for a systematic presentation of this data type in the data distribution structure”. Bilingual dictionaries, as tools for learning foreign language and culture should include illustrative examples for the purposes of enhancing users’ understanding of cultural words.

## 5 References

- Al-Ajmi, H. (2008). The Effectiveness of Dictionary Examples in Decoding: The Case of Kuwaiti Learners of English. *Lexikos* 18. pp.15 – 26.
- Bartholomew, D. (1995). Otomi Culture from Dictionary Illustrative Sentences. In Kachru, B.B. & Kahane, H. Tubingen: Max Niemeyer Verlag. pp.3 – 7.
- Gangla-Birir, L. (2005). The Use of Pictorial Illustrations in Africa Language Dictionaries. *Lexikos* 15. pp.38 – 51.
- Guma, S.M. (1977). *The Form, Content and Technique of Traditional Literature in Sesotho*. Pretoria: Van Schaik.
- Katzaros, V. (2004). The Different Functions of Illustrative Examples in Learners’ Bilingual Dictionaries. In Williams, G. and Vessier, S. (eds.), *Proceedings of the Eleventh EURALEX International Congress, July 6 – 10, 2004. Volume 11*. Lorient: Université de Bretagne-Sud. pp.487 – 494.
- Kavanagh, K. (2000). Words in a Cultural Context. *Lexikos* 10. pp.99 – 118.
- McHenry, R. (ed). (1992). *The New Encyclopaedia Britannica. Volume 9*. Chicago: Encyclopaedia Britannica, Inc.
- Mafela, M.J. (2008). Proverbs as Illustrative Examples in a Tshivendḁa Bilingual Dictionary: A Reflection of Meaning and Culture. *South African Journal of African Languages, Volume 28, No. 1*. pp.30 – 35
- Prinsloo, D.J. & Gouws, R.H. (2000). The Use of Examples in Polyfunctional Dictionaries. *Lexikos* 10. pp.138 – 156.
- Van Warmelo, N.J. (1989). *Venda Dictionary: Tshivendḁa – English*. Pretoria: J.L. Van Schaik.
- Whitcut, J. (1995). Taking it For Granted: Some Cultural Preconceptions in English Dictionaries. In Kachru, B.B. & Kahane, H. *Cultures, Ideologies, and the Dictionary: Studies in Honor of Ladislav Zgusta*. Tubingen: Max Niemeyer Verlag. pp.253 – 257.



# Corpus, Parallélisme et Lexicographie Bilingue

Adriana Zavaglia, Gisele Galafacci  
Université de São Paulo  
zavaglia@usp.br, gisele.galafacci@usp.br

## Résumé

Dans le cadre de la lexicographie bilingue brésilienne, surtout sur le statut de la relation entre l'entrée et ses traductions, il n'est pas rare que les dictionnaires proposent à l'utilisateur une liste de traductions possibles pour une entrée sans contextualisation ni rubrique d'usages. L'utilisateur de ses ouvrages se pose fréquemment des questions sur l'usage le plus approprié, sur la frontière entre les différents sens des mots proposés et leur champ notionnel. Les traductions répertoriées et décontextualisées n'aident donc pas à résoudre les ambiguïtés concernées. Pour essayer de collaborer là-dessus, nous proposons une approche différentielle (portugais - français) de la lexicographie bilingue à l'aide de la linguistique de corpus. Composé d'originaux et de traductions, le corpus parallèle utilisé dans ce travail est à la base de la contextualisation authentique de la polysémie des entrées, des définitions et des exemples bilingues. Compte tenu de ces aspects, ce travail fournit des commentaires sur les méthodologies des dictionnaires bilingues consultés en considérant le mot grammatical « bem » du portugais brésilien et présente la méthodologie que nous appelons « parallèle » et les conséquences pour l'utilisateur concernant l'entrée en question, en particulier dans les contextes d'apprentissage du français langue étrangère et de la traduction.

**Mots-clés:** traduction; corpus parallèle; lexicographie bilingue

## 1 Introduction

Dans le cadre de l'élaboration de dictionnaires et glossaires bilingues, il n'est pas rare que les dictionnaires disponibles - monolingues, bilingues ou multilingues - soient la principale source du lexicographe, en plus de ses propres connaissances linguistiques, méta-linguistiques et extra-linguistiques. Cependant, cette méthodologie de travail, encore courante aujourd'hui au Brésil pour les langues portugais-français, objet de notre intérêt, présente des résultats insuffisants. Comme l'a déjà dit Baldinger (1973: 63) dans le cadre de la lexicographie historique, et ceci peut être adapté à notre objectif, « les dictionnaires se copient [et se traduisent] mal... ». Par conséquent, les références aux méthodologies adoptées pour la constitution de la macro- et la micro-structure des œuvres lexicographiques bilingues ne sont pas claires, pour ne pas dire nébuleuses, dans les paratextes de ces œuvres (introductions, avant-propos, préface, etc.) ; les critères de sélection et d'enregistrement n'étant donc pas mis en évidence.

De ce fait résulte une persistance des problèmes observés dans la micro-structure de ces ouvrages, surtout dans le cas où l'utilisateur se voit proposer une liste de traductions possibles pour une entrée sans contextualisation ni rubrique d'usages. Face à cet éventail d'options pour traduire l'entrée, l'utilisateur est dès lors confronté à un dilemme: quel est l'usage le plus approprié? Comment délimiter la frontière entre les différents sens des mots? Dans quel champ notionnel ces traductions possibles s'insèrent-elles?

En appliquant les concepts de correspondance et d'équivalence (Delisle 2003) à la micro-structure des articles bilingues, il peut être déduit que les traductions répertoriées et décontextualisées correspondraient à une entrée dans un environnement statique. Mais comment rendre compte de la polysémie mouvante des mots dans les dictionnaires bilingues étant donné l'indétermination du langage? Sans oublier le facteur statique de leur description par le lexicographe et en même temps le facteur dynamique de l'énonciation du traducteur ou de l'utilisateur des dictionnaires / glossaires bilingues, comment en choisir un seul mot correspondant, en l'introduisant dans la traduction, dans un environnement dynamique d'équivalence où la question énonciative de l'aspect intersubjectif des interprétations se pose? Et que dire encore des mots grammaticaux?

Pour essayer de collaborer là-dessus, notamment sur le statut de la relation entre l'entrée et leurs traductions dans ce contexte, nous avons proposé une approche différentielle de la lexicographie bilingue (portugais - français) à l'aide de la linguistique de corpus (Zavaglia 2004, 2006, 2008, 2009, 2010, 2012) pour traiter un problème difficile pour la lexicographie en général et pour la lexicographie bilingue plus spécifiquement: la description des mots grammaticaux. Composé d'originaux et de traductions, le corpus s'est avéré être un bon allié. Pour les dictionnaires semi-bilingues, le corpus parallèle offre aux lexicographes une contextualisation authentique de la polysémie des entrées, c'est-à-dire une contextualisation du point de vue documentaire des sources. Considérant également la grande facilité actuelle pour la construction de dictionnaires ou lexiques bilingues électroniques, des définitions et des exemples bilingues peuvent être inclus facilement dans la micro-structure des articles, ce qui diminue considérablement les ambiguïtés et ainsi, augmente le degré de fiabilité.

Les avantages d'une telle approche ne sont pas nouveaux. Comme disait déjà Langlois (1996, Chapitre I: «Les corpus et les bitextes»)<sup>1</sup> :

La valeur des corpus en linguistique n'est pas à prouver. En effet, les langues étant particulièrement complexes à décrire, les chercheurs ne peuvent se fier à l'introspection seule pour les étudier. C'est pourquoi l'étude des corpus présente de nombreux avantages, que Jan Svartvik (1992: 8) et Geoffrey Leech (1992: 106) résumant bien.

Et l'auteur ajoute à la fois qu'« il est clair, cependant, que les corpus ne livrent pas que des renseignements intéressants ; il faut que le lexicographe analyse ces données empiriques pour en tirer le meilleur parti possible. » (Langlois 1996, Chapitre I: «Les corpus et les bitextes»). De toute façon, les diction-

---

1 Les pages de la version électronique de la thèse de Lucie Langlois ne sont pas numérotées, c'est pourquoi nous n'indiquons ici que le chapitre du travail d'où la citation a été retirée.



naires bilingues portugais-français brésiliens ne sont pas encore fabriqués à partir de corpus, ce qui fait l'intérêt de notre approche.

Dans ce cadre, la traduction d'une entrée donnée n'est ni considérée comme un correspondant, ni comme un équivalent, mais plutôt comme un parallélisme. Ceci est le résultat d'une méthodologie différente de celle adoptée dans la construction des articles de dictionnaires bilingues français-portugais brésilien disponibles. Compte tenu de ces aspects, ce travail (1) fournit des commentaires sur les méthodologies des dictionnaires bilingues consultés et leurs conséquences pour l'utilisateur en considérant le mot grammatical *bem* du portugais brésilien et (2) présente la méthodologie que nous appelons «parallèle» et les conséquences pour l'utilisateur concernant l'entrée en question, en particulier dans les contextes d'apprentissage du français langue étrangère et de la traduction.

## 2 Observations Générales sur les Dictionnaires Bilingues Consultés

Pour la réalisation de ce travail, nous avons sélectionné les ouvrages disponibles aux étudiants de langue française ou de traduction au Brésil, ceux qui sont utilisés comme outil d'apprentissage dans les cours de langue ou de traduction. Pour l'analyse, nous avons travaillé sur les contenus informationnels de l'entrée *bem* présentés dans les dictionnaires sélectionnés (voir tableau 1).

D'emblée, il est possible de repérer la grande variation de l'extension de contenu informationnel présenté par les différents ouvrages. En plus, ceux-ci ne fournissent pas d'informations précises aux usagers quant à leurs critères de sélection et d'enregistrement de données mis en œuvre lors de la production de ces outils d'apprentissage. En général, il n'existe aucune indication sur leurs pages indiquant la source des options offertes (textes authentiques et leurs traductions, dictionnaires monolingues ou bilingues?). Les lexicographes n'assument pas non plus la responsabilité des traductions de leurs dictionnaires. Nous pouvons dire que ces œuvres conservent une posture traditionnelle, plus impressionniste que scientifique, en ce qui concerne les options de traduction apparaissant dans la microstructure des articles: il s'agit d'apparier un mot avec un autre de façon directe. Cette perspective, très simpliste et aussi illusoire, que nous ne pouvons répertorier ici par manque d'espace, est un euphémisme pour l'équivalence, car il n'y a pas, dans la relation entre deux unités linguistiques, de correspondance directe au niveau des langues.

Cependant, après avoir observé les micro-structures de *bem*, nous pouvons avancer qu'aucun dictionnaire n'a présenté des citations pour exemplifier l'usage des correspondants proposés. De ce fait, il y a un grand désaccord entre les objectifs de conception de ces outils et les critères utilisés dans leur production. Dès lors, si la conception de dictionnaire bilingue est considérée comme outil d'apprentissage, ce désaccord est inadéquat aux buts pédagogiques intrinsèques. Selon Tarp:

Un dictionnaire d'apprentissage est un dictionnaire dont le vrai but est de répondre aux besoins d'informations lexicographiques qui sont importantes aux étudiants dans une série de situations extra-lexicographiques pendant leur processus d'apprentissage d'une langue étrangère. (Tarp 2006: 300, notre traduction)

Si les dictionnaires sélectionnés sont utilisés au Brésil dans des contextes d'apprentissage (de langue, de traduction), il est en effet moins productif de proposer une liste de correspondants sans une contextualisation de leur usage aux étudiants que de leur proposer un seul correspondant. Ceci s'explique par le fait que le manque d'exemples les empêche de choisir l'option la plus adéquate pour la construction du sens envisagé, surtout dans les contextes de production oraux et écrits. Même si ces dictionnaires n'ont pas été conçus avec ce but précis, ils sont utilisés dans ce contexte, et c'est à partir de ce contexte, apprentissage/usager, que nous avons repéré les problèmes déjà mentionnés.

Ainsi, en considérant chacun des dictionnaires consultés, avec quelques exceptions concernant Gálvez (2008) et Burtin-Vinholes (2003), qui présentent une quantité raisonnable d'exemples parmi les ouvrages analysés, nous pourrions dire qu'ils ne se fondent pas sur de vraies bases lexicographiques pour ce qui est de la conception de leurs micro-structures. Cependant, ceci ne remet pas en cause le fait que ces dictionnaires sont des sources importantes d'information pour les étudiants et les professionnels de la langue en général au Brésil. Mais, à présent, avec les progrès scientifiques et technologiques applicables à la lexicographie, les dictionnaires ont besoin de se construire sur une base théorique solide qui dirige sa constitution interne, avec une base de données consistante permettant l'extraction du corpus en tenant compte des valeurs de fréquence, d'usage, de complexité et de cohérence linguistique. Ceci est également le cas pour l'emploi d'outils informatiques robustes qui conduisent à une meilleure utilisation des données à analyser, à la fois en termes de temps et de qualité. C'est pour cette raison que nous nous concentrons sur ce problème méthodologique en essayant, à partir de certaines expériences, de présenter des solutions possibles.

<b>Avolio &amp; Faury 2002</b>	bem sm 1 bien. 2 bens pl biens. adv bien. ela dança muito bem/elle danse très bien. bem feito! bien fait! muito bem! bravo! nem bem nem mal ni bien ni mal. se bem que bien que. Veja nota em mieux.
<b>Burtin-Vinholes 2003</b>	BEM, s. m. Bien, bon, utile, avantageux. Bienfait, grâce; service; avantage; profit. POSSUIR ALGUNS -NS, posséder des biens. PAGAR O - COM O MAL, rendre le mal pour le bien. PESSOA DE -, honnête personne. FAZER O -, faire le bien, être charitable. FAZER -A ALGUÉM, rendre service à quelqu'un. QUERER O - DE ALGUÉM, désirer le bien à quelqu'un. QUERER - A SEUS PAIS, avoir de l'amour, de l'attachement pour ses parents. LEVAR A -, approuver. MEU -!, mon amour! POR -, volontiers.
<b>Florenzano n.d</b>	BEM, adv. Bien, beaucoup, fort, très. Convenablement, comme il faut. Certainement, assurément. ESTÁ -, c'est bien, c'est bon. AINDA -, à la bonne heure. É - FEITO, c'est bien fait. OS NEGÓCIOS VÃO -, les affaires vont bien. ELE ESTÁ MUITO -, il est dans l'aisance, il a de la fortune. MUITO -, très bien. NÃO VI MUITO -, je n'ai pas bien vu. NÃO A CONHEÇO -, je ne la connais pas beaucoup. NÃO SEI LÁ MUITO - O QUE ELE DISSE, je ne sais pas au juste ce qu'il a dit. - QUE, quoique, bien que.
<b>Gálvez 2008</b>	bem, s.m. e adv. bien.

<b>Marote 2004</b>	bem ◊ m bien m; praticar o bem faire le bien. ◊ adv 1. [ger] bien; dormiu bem? tu as bien dormi?; fez bem? tu as bien fait?; sente-se bem? tu te sens bien?; estar bem être bien; [de saúde] aller bien; queria uma bebida bem gelada je voudrais une boisson bien glacée; quero um quarto bem quente je veux une chambre bien chaude; é um quarto bem grande c'est une chambre bien grande; é um lugar bem bonito c'est un endroit bien beau; foi bem ali
<b>Rónai 1989</b>	bem. S.m. 1. Bien. bens. Pl. 2. Chose matérielle susceptible d'appropriation, propriété, possession, domaine. Bens de raiz. Biens-fonds..Fazer uma coisa por bem. Faire une chose pour le bien. Haver por bem. Daigner, vouloir bien. Levar a bem. Trouver bon, prendre en bonne part. Meu bem. Mon chéri, ma chérie. Pagar o bem com o mal. Rendre le mal pour le bien. Adv 3. Bien, comme il faut. Bem mais. Beaucoup plus. Ainda bem! À la bonne heure! Está bem! C'est bon! c'est bien! Falar bem de alguém. Dire du bien de quelqu'un. Pois bem! Eh bien! Querer bem a alguém. Aimer bien quelqu'un, avoir une grande estime pour quelqu'un. Se bem que Loc. conj. Bien que, quoique.
<b>Signer 1998</b>	bem adv. Bien ; - que loc. conj. quoique, bien que; ainda -! à la bonne heure!; adj. bien; s.m. bien; por - ou por mal bon gré mal gré; pl. bens, fortune f.; bens imóveis biens immeubles.
<b>Valdez 2000</b>	bem adv bien; bon; juste. está bem = c'est bien/c'est bon. falar bem = dire du bien. bem no meio = au beau milieu. bem falante = beau parleur. é bem do lado = c'est juste à côté. loc conj se bem que = encore que. nm bens de consumo = biens de consommation. bem imóvel = bien-fonds. bem de raiz = immeuble. bens trazidos pelo cônjuge, pelo sócio, etc. à sociedade = apport. pl avoires; patrimoine. Interj muito bem! = bravo!. Fam. tudo bem = ça colle/ça roule.

**Tableau 1: Entrée *bem* dans les dictionnaires bilingues.**

### 3 Corpus et Parallélisme

Les outils d'informatique dédiés à la manipulation de données linguistiques rendent possible le renouvellement de la méthode de production de dictionnaires. Ils permettent le stockage des corpus constitués d'exemples d'usage des unités lexicales qui peuvent être récupérés par les lexicographes avec le but de contextualiser les propositions de correspondance suggérés dans les entrées des œuvres lexicographiques. De plus, le travail sur les occurrences dans un corpus bien adapté au type de dictionnaire qui se développe peut aussi aider dans le choix des entrées qu'il en aura et dans la composition de leurs micro-structures.

Cette possibilité de manipulation et récupération de données linguistiques authentiques de manière rapide et fiable a contribué au développement d'un nouveau concept de dictionnaire bilingue, nommé dictionnaire bilingue contrastif. Selon Durão (2009: 18), un dictionnaire bilingue contrastif diffère d'un dictionnaire bilingue car il ne se limite pas à des propositions d'équivalence, mais fournit en plus des informations qui permettent à l'utilisateur de confronter les caractéristiques constitutives, les règles d'utilisation et les traits sémantiques des unités lexicales dans les deux langues. Cette approche permet d'exploiter les possibilités de transfert et offre des avertissements sur des possibles in-

terférences d'une langue à l'autre, car il est destiné aux personnes qui apprennent une langue étrangère avec la même langue maternelle ayant comme but de reconnaître les différences entre les deux langues.

Ainsi, nous proposons un travail lexicographique pour les langues portugais brésilien-français qui utilise des corpus parallèle, c'est-à-dire, comme nous l'avons déjà dit, une base de données composée de textes authentiques en portugais et leurs respectives traductions en français. Ce corpus, nommé CorPPoFranco – Corpus Parallèles du Portugais au Français de la Francophonie, est en train d'être développé pour ce projet de recherche. Avec le but d'être utilisé dans la production de dictionnaires bilingues, ce corpus aura l'adéquation nécessaire à la réalisation de cet objectif, car, selon Sardinha (2004: 29), la recherche avec l'appui de corpus a du sens seulement si les questions posées et les corpus sont adéquats aux objectifs du travail.

Cette méthodologie permet de traiter de façon plus efficace l'hétérogénéité qui entoure la relation complexe entre deux langues, ce qui permet tant l'observation de la distribution morphologique et structurale d'une lexie, que la visualisation de son profil, de ses collocations et de sa prosodie sémantique. L'élaboration de paradigmes adoptés dans le développement de la recherche sur la lexicographie bilingue dans cette perspective différentielle inclut des procédures spécifiques de la linguistique de corpus (comme par exemple la concordance et l'alignement), de la linguistique énonciative, de la lexicographie et des études de traduction. De ceci a émergé une nouvelle entité empirique (les traductions offertes dans les micro-structures) et théorique (la nature et la fonction des traductions). Cette nouvelle entité de la lexicographie bilingue demande un nom, une définition et une conceptualisation, que voici :

Le Parallélisme est la relation entre deux lexies, complexes ou non, décrite dans la micro-structure d'articles de dictionnaires bilingues de langue générale construits à partir de corpus parallèle (originaux et traductions publiées) ayant la fonction d'explicitier la variation sémantique du mot-vedette, avec ses usages spécifiques, en contexte bilingue authentique, avec ses traductions possibles.

## 4 Illustrations

Pour illustrer cette méthodologie parallèle, nous avons utilisé un échantillon du CorPPoFranco, dont nous avons sélectionné quelques occurrences de la lexie *bem* pour réfléchir sur les propositions de traductions offertes par les dictionnaires bilingues décrits sommairement avant et les parallélismes qui résultent de l'utilisation de cette méthodologie.

Nous n'allons prendre, pour manque d'espace, que quelques aspects sur la relation entre *bem* et *bien* remarquables sur cette traduction directe suggérée par tous les dictionnaires bilingues consultés: les lexicographes n'ont pas pris en compte dans leurs articles les informations sur les usages spécifiques, les

questions de syntaxe et de collocation concernant cette relation. Les exemples ci-dessous confirment leur relevance par rapport:

- à la morphologie :

(1) À mesa do café eu me enquadrava com meu robe branco, meu rosto limpo e bem esculpido, e um corpo simples

(1a) Assise à ma table, dans ma robe de chambre blanche, avec mon visage net et bien sculpté, mon corps tout simple

(2) Revi o rosto preto e quieto, revi a pele inteiramente opaca que mais parecia um de seus modos de se calar, as sobrancelhas extremamente bem desenhadas, revi os traços finos e delicados que mal eram divisados no negror apagado da pele.

(2a) Je revis le visage noir et placide, la peau parfaitement opaque qui semblait davantage une autre façon de se taire, les sourcils extrêmement bien dessinés, je revis les traits délicats et gracieux, mais à peine lisibles dans la noirceur éteinte de la peau.

Dans les exemples ci-dessus, avec *bem* adverbe traduit par *bien*, on peut voir, au-delà de sa fonction d'intensif, son positionnement, devant le participe en fonction d'adjectif, mais surtout son caractère invariable, car le parallélisme (1)/(1a) le montre avec un participe adjectif au singulier et le (2)/(2a), au pluriel.

- au positionnement :

(3) Mas era tão inegável sentir aquele nascimento de dentro da poeira - que eu não podia senão seguir aquilo que eu bem sabia que não era loucura, era, meu Deus, uma verdade pior, a horrível.

(3a) Mais c'était si incomparable de sentir cette naissance au sein de la poussière que je ne pouvais que suivre ce dont je savais bien que ce n'était pas de la folie; c'était, mon Dieu, une vérité pire, l'horrible vérité.

(4) Queria saber se era verdade que ele lhe arrancara a calcinha com os dentes, queria saber isto e aquilo, queria saber o que já sabia muito bem

(4a) Il voulait savoir s'il était vrai qu'il lui avait arraché sa petite culotte avec les dents, il voulait savoir ceci et cela, il voulait savoir ce qu'il savait déjà parfaitement

Dans les exemples ci-dessus, nous voyons une question d'usage en ce qui concerne les différentes positions de *bem* avec le verbe *saber*: devant le verbe, avec un usage absolu; après le verbe, dans une construction avec un adverbe qui l'intensifie. Dans les traductions, il est possible de remarquer tout de suite les différences en français: *bien* avec *savoir* après le verbe, avec la possibilité d'un autre parallélisme quand il est intensifié, comme dans (4a).

- à la polysémie :

(5) De meu próprio mal eu havia criado um bem futuro.

(5a) J'avais, avec mon propre mal, créé un bien à venir.

(6) A partilha dos bens foi realizada como previsto.

(6a) Le partage des biens a été réalisé comme prévu.

(7) No quarto de Ariela, Benjamim pagaria a Lorna para prosseguir falando “tu és meu bem, tu és meu bem, tu és meu bem”

(7a) Dans la chambre d’Ariela, Benjamin paierait Lorna pour qu’elle continue à dire « tu es mon ange, tu es mon ange, tu es mon ange »

Dans les parallélismes ci-dessus, *bem*, nom, est traduit par *bien* ; et pourtant, les contextes montrent clairement que, du point de vue du sens, la relation n’est pas biunivoque, mais polysémique, car dans (5)/(5a), le « bem », au-delà d’avoir son origine ancrée dans le « mal », ce qui exclut par elle-même toute relation simpliste d’antonymie, est un bien abstrait parmi d’autres qui ne peut pas être déterminé, à part son caractère futur ; dans (6)/(6a), *bens*, pluriel, couramment utilisé dans le vocabulaire juridique, se rapporte aux biens concrets, meubles ou immeubles. En plus, le parallélisme (7)/(7a) montre que *bem* concret peut faire référence à une personne, généralement quand il est précédé d’un possessif, avec une possibilité de traduction, *ange*, dans « mon ange ».

D’autres parallélismes ont été trouvés dans le corpus. Cependant, leurs usages spécifiques sont attachés à des structures particulières, lesquelles valident la construction du sens envisagé. Pour cela, la simple suggestion de ces mots dans un article de dictionnaire dans une liste de possibilités se révèle improductive pour l’usager qui a besoin d’informations plus détaillées sur chaque traduction proposée. Les parallélismes montrent que, en fonction de l’adverbe que *bem* intensifie, une traduction différente est convoquée (« bem devagar », « tout doucement » ; « bem mais », « beaucoup plus »), de même dans d’autres relations, *bem/bon* (« Bem, além de fixar... » / « Bon, en plus de fixer ... »), ou formations (lexies complexes: « bem-estar »/« bien-être » ; locutions : « se bem que »/« pourtant » ; binômes : « bem ou mal »/« bien ou mal »). Voici quelques exemples extraits de l’ouvrage *A paixão segundo G.H.*, de Lispector (1964/1998) et de sa respective traduction:

(8) Aquelas pessoas que, só elas, entenderão bem devagar que este livro nada tira de ninguém.

(8a) Ces personnes, et elles seules, comprennent tout doucement que ce livre n’enlève rien à personne.

(9) Desculpa eu te dar isto, eu bem queria ter visto coisa melhor.

(9a) Pardonne-moi pareil cadeau, je préférerais tellement avoir vu une chose plus agréable.

(10) É bem mais que uma elegância.

(10a) C’est beaucoup plus qu’un snobisme.

(11) Mas tendo aos poucos, por meio de dinheiro razoavelmente bem investido, enriquecido o suficiente, isso impediu-me de usar essa minha vocação: não pertencesse eu por dinheiro e por cultura à classe a que pertenço, e teria normalmente tido o emprego de arrumadeira numa grande casa de ricos, onde há muito o que arrumar.

(11a) Mais, pour avoir su placer adroitement mon argent, j’ai acquis peu à peu une certaine alliance, ce qui m’a empêchée de réaliser ma vocation profonde: si je n’appartenais pas, par la culture et l’argent, à la classe à laquelle j’appartiens, j’aurais dû normalement être femme de chambre chez des gens riches, dans une grande maison où il y a beaucoup à ranger.

- (12) Animeei-me com uma idéia: aquele guarda-roupa, depois de bem alimentado de água, de bem enfiado nas suas fibras, eu o enceraria para dar-lhe algum brilho, e também por dentro passaria cera pois o interior devia estar ainda mais esturrado.
- (12a) Une idée me réconforte: cette armoire, une fois complètement imbibée, gorgée d'eau dans toutes ses fibres, j'allais la cirer pour lui donner un peu de brillant et je passerais de la cire à l'intérieur aussi car l'intérieur devait être encore plus calciné.
- (13) De encontro ao rosto que eu pusera dentro da abertura, bem próximo de meus olhos, na meia escuridão, movera-se a barata grossa.
- (13a) Tout contre mon visage passé dans l'ouverture de la porte, tout près de mes yeux, dans la demi-obscurité, un énorme cafard avait bougé.
- (14) Até que - enfim conseguindo me ouvir, enfim conseguindo me comandar - ergui a mão bem alto como se meu corpo todo, junto com o golpe do braço, também fosse cair em peso sobre a porta do guarda-roupa.
- (14a) Jusqu'à ce que - parvenant enfin à m'entendre, parvenant enfin à me donner un ordre - je levai la main très haut comme si mon corps tout entier, entraîné par le mouvement du bras, allait tomber lui aussi de tout son poids sur la porte de l'armoire.
- (15) Bem, além de fixar as dunas com eucaliptos, eu tinha que não esquecer, se viesse a ser necessário, que o arroz prospera em solo salobre, cujo alto teor de sal ajuda a desbastar; disse eu também me lembrava das leituras de antes de dormir que eu, de propósito, procurava que fossem impessoais para me ajudarem a adormecer.
- (15a) Bon, en plus de fixer les dunnes avec des eucalyptus, il ne fallait pas que j'oublie, le cas échéant, que le riz pousse sur un terrain saumâtre dont la haute teneur en sel facilite le sarclage; j'avais aussi retenu cela de mes lectures d'avant de m'endormir que je choisissais, exprès, impersonnelles pour m'aider à trouver le sommeil.
- (16) O tédio profundo - como um grande amor - nos unia. E na manhã seguinte, de manhã bem cedo, o mundo se me dava.
- (16a) L'ennui profond - comme un grand amour - nous unissait. Et le lendemain, le matin très tôt, le monde s'offrait à moi.
- (17) Usarei, sim, o vestido azul novo, que me emagrece um pouco e me dá cores, telefonarei para Carlos, Josefina, Antônio, não me lembro bem em qual dos dois percebi que me queria ou ambos me queriam, comerei crevettes ao não importa o quê", e sei porque comerei crevettes, hoje de noite, hoje de noite vai ser a minha vida diária retomada, a de minha alegria comum, precisarei para o resto dos meus dias de minha leve vulgaridade doce e bem-humorada, preciso esquecer, como todo o mundo.
- (17a) Je mettrai, oui, ma nouvelle robe bleue qui m'amincit un peu et me donne des couleurs, je téléphonerai à Carlos, Joséfina, Antonio, je ne me souviens qu'il me désirait, ou si c'était tous les deux, je mangerai des crevettes « à la je ne sais quoi », et je sais pourquoi je mangerai des crevettes, ce soir, ce soir je vais reprendre ma vie quotidienne, celle de ma joie ordinaire ; j'aurai besoin

pour le restant de mes jours de ma légère vulgarité douce et bon enfant, j'ai besoin d'oublier comme tout le monde.

Cependant, comment inclure dans la micro-structure bilingue portugais-français de *bem* toutes les informations des observations et analyses faites à partir de cette méthodologie parallèle? Une fois que l'utilisateur cherche une solution immédiate dans un dictionnaire bilingue? Dans notre article, qui sera électronique et disponible en ligne, cette solution immédiate est présentée après la transcription phonétique et la classification morphologique, par ordre de fréquence dans le corpus. Par exemple:

bem [bê]

adv. bien, très, tout, beaucoup, complètement, tellement, adroitement.

n. bien, pl. bens, biens.

ap. bon.

Pour avoir plus de détails (de positionnement, d'usage ou de contexte), l'utilisateur peut cliquer sur l'une des traductions présentées. Par exemple, en cliquant sur « beaucoup », une nouvelle fenêtre s'ouvre avec les informations suivantes :

[DEFINIÇÃO] Intensificador comparativo [DEFINITION] intensif comparatif

[COLOCAÇÃO] bem mais [COLLOCATION] beaucoup plus

[SINÔNIMO] pt. muito mais [SYNONYME] fr. bien plus

[POSIÇÃO] bem + adverbe [POSITIONNEMENT] beaucoup + adverbe

[USO] neutro [USAGE] neutre

[CONTEXTO] É bem mais que uma elegância. [CONTEXTE TRADUIT] C'est beaucoup plus qu'un snobisme.

## 5 Conclusion

Implicitement, l'utilisateur aura une description bilingue de *bem* en raison du co-texte et du contexte dans lequel cette lexie apparaît. Cela lui permettra également la possibilité de saisir le caractère idiomatique qui tourne autour de ses parallélismes, à savoir les relations authentiquement bilingues entre *bem* et ses traductions. Le développement du concept de parallélisme, comme mentionné plus haut, a une origine naturelle lors de la mise en place de la méthodologie présentée, radicalement différente de celle observée dans les dictionnaires bilingues portugais-français utilisées au Brésil. De plus, le rapport entre l'entrée et les traductions proposées dans la micro-structure des articles, il faut le dire, n'est pas fixe. En effet, ce rapport se déplace et se modifie en fonction du co-texte et du contexte, comme deux lignes parallèles qui affluent vers l'infini sans envahir le domaine l'une de l'autre, conservant chacune ses propres caractéristiques, et met en avant les valeurs de la fréquence de l'utilisation de la lexie ou de sa complexité et, plus précisément, l'explicitation de sa polysémie, qui se construit et se réalise toujours dans un contexte véritablement bilingue.



## 6 Références

- Avolio, J.C., Faury, M.L. (2002). *Michaelis Minidicionário Francês-Português, Português-Francês*. São Paulo: Melhoramentos.
- Baldinger, K. (1973). Le DEAF en tant que dictionnaire diachronique. Problèmes théoriques et pratiques, *Meta*, 18, 61-85.
- Burtin-Vinholes, S. (2003). *Dicionário Francês-Português/Português-Francês*. São Paulo: Ed. Globo.
- Delisle, J. (2003). *La traduction raisonnée*. Ottawa: University of Ottawa Press.
- Durão, A.B.A.B. (ed.) (2009). *Por uma Lexicografia Bilíngüe Contrastiva*. Londrina: UEL.
- Florenzano, E. (n.d.) Dicionário Ediouro Francês-Português/Português-Francês.
- Gálvez, J.A. (ed.) (2008). Dicionário Larousse Francês-Português, Português-Francês: mini. São Paulo: Larousse do Brasil.
- Langlois, L. (1996). *Bitexte, bi-concordance et collocation*. Thèse de doctorat. Université d'Ottawa. En ligne <http://www.dico.uottawa.ca/theses/langlois> [13/08/2013].
- Lispector, C. (1998). *A paixão segundo G.H.* Rio de Janeiro: Rocco. (Original work published 1964).
- Lispector, C. (1978). *La passion selon G.H.* Trad. Claude Farny. Paris: Des Femmes.
- Marote, J.T.O. (2004). Minidicionário Francês-Português, Português-Francês. São Paulo: Ática.
- Rónai, P. (1989). *Dicionário Francês-Português/Português-Francês*. Rio de Janeiro: Nova Fronteira.
- Signer, R. (1998). Dicionário Brasileiro Francês-Português, Português-Francês. São Paulo: Oficina de Textos.
- Tarp, S. (2006). Lexicografia de Aprendizaje. In *Cadernos de Tradução*, n.18. Florianópolis: UFSC, pp. 295-317.
- Valdez, J.F. (2000). *Dicionário Francês-Português/Português-Francês*. Belo Horizonte: Garnier.
- Sardinha, T.B. (2004). *Linguística de Corpus*. Barueri: Manole.
- Zavaglia, A. (2004). Lingüística de cópús e lexicografia bilíngüe: o caso experimental de *como* e suas traduções para o francês. In *Crop*, v. 10. São Paulo: FFLCH-USP, pp. 211-224.
- Zavaglia, A. (2006). Lexicografia bilíngüe e corpora paralelos: procedimentos e critérios experimentais. In *Cadernos de Tradução*, v. XVIII. Santa Catarina: UFSC, pp. 19-39.
- Zavaglia, A. (2008) Apresentação das bases do Dicionário Relacional (português-francês) DIRE. In Isquierdo, A.N., Finatto, M.J.B. (eds.). *Ciências do Léxico: Lexicologia, Lexicografia e Terminologia*, v. 4. Campo Grande: Ed. UFMS, pp. 233-254.
- Zavaglia, A. (2009). Linhas gerais para a elaboração do Dicionário Relacional – DIRE (português-francês). In Rezende, L.M., Silva, B.C.D. & Barbosa, J.B. (eds.). *Léxico e Gramática: dos sentidos à construção da significação*. São Paulo: Cultura Acadêmica, pp. 185-201.
- Zavaglia, A. (2010). Sinonímia e lexicografia. In Isquierdo, A. N., Barros, L. A. (eds.). *As Ciências do Léxico: Lexicologia, Lexicografia e Terminologia*, v. V. Campo Grande: Editora da UFMS, pp. 189-199.
- Zavaglia, A. (2012). Por uma descrição esquemática do léxico: o caso de *mediante*. In Isquierdo, A.N., Seabra, M.C.TC. (Eds.). *As Ciências do Léxico: Lexicologia, Lexicografia, Terminologia*, v. VI. Campo Grande: Editora da UFMS, pp. 73-84.



**Lexicography for  
Specialised Languages, Technology  
and Terminography**



# EcoLexicon

Pamela Faber, Miriam Buendía Castro  
University of Granada  
pfaber@ugr.es, mbuendia@ugr.es

## Abstract

EcoLexicon (<http://ecolexicon.ugr.es>) is an environmental knowledge base which is based on the premises of Frame-based Terminology (FBT) (Faber 2009, 2011, 2012). EcoLexicon represents the conceptual structure of the specialized domain of the Environment in the form of a visual thesaurus in which environmental concepts are configured in semantic networks. The various terminological designations for a concept are offered in six languages: Spanish, English, German, French, Russian, and Greek. So far, EcoLexicon contains 3,527 concepts and 18,617 terms. It provides conceptual, linguistic, and administrative information for each entry, and we are now beginning to include phraseological information for terms as well (Buendía 2013; Buendía & Sánchez 2012, Sánchez & Buendía 2012). EcoLexicon is designed to meet the needs of different user types, such as a student of science wishing to acquire specialized knowledge about a certain concept, a translator seeking translation correspondences in a language, or a specialist interested in text production.

**Keywords:** knowledge base; specialized language; environment

## 1 Introduction

EcoLexicon (<http://ecolexicon.ugr.es>) is a visual online thesaurus of environmental science, which currently contains 3,527 concepts and 18,617 terms in English, Spanish, German, French, Russian and Modern Greek. It provides conceptual, linguistic, and administrative information for each entry. This information as well as the corpus of specialized texts is stored in a private database, which allows members of EcoLexicon to add, eliminate, and/or modify conceptual and terminological information. In EcoLexicon it is assumed that up to a certain level, its potential users are familiar with scientific language and its usage in English or Spanish at least, since these are the interface languages. Potential users should thus possess a good command of any of the six languages in the knowledge base, as well as a minimum of scientific knowledge (López, Buendía & García 2012: 62).

EcoLexicon is based on the premises of Frame-based Terminology (FBT) (Faber 2009, 2011, 2012), a cognitive approach to Terminology. The FBT approach to Terminology applies the notion of frame, defined as a schematization of experience (a knowledge structure), which is represented at the conceptual level and relates elements and entities associated with a particular culturally embedded scene, situation or event from human experience.

The public version of EcoLexicon is freely available online. The new version of EcoLexicon includes the following new features, which are an improvement over the previous version.

- It is compatible with all modern browsers and does not require Java.
- It has more interactive and configurable maps, which allow the users to do the following:
  - Change the scale of the map.
  - Select the relations to be represented.
  - Eliminate nodes in order to make other nodes more prominent.
  - Adjust the position of nodes.
  - Go backwards or forwards.
  - Establish a direct link to a concept or term.
  - Look up a concept or term on Google or Google Images.
- It has new representation modes:
  - Hierarchical tree structure
  - Map of the shortest path between concepts
- It allows user registration to do the following:
  - Personalize results.
  - Store user preferences between sessions.

## 2 The micro and macrostructure of entries in EcoLexicon

Figure 1 displays the entry for *alluvial fan* in EcoLexicon. As shown in Figure 1, when EcoLexicon is opened, three zones appear:

- (1) The top bar that allows access to different functionalities, such as the term/concept search or changing the language of the interface. It also permits to personalize the search by means of the contextual domains generally associated with a specialized knowledge field (e.g. Geology, Coastal Engineering, Environmental Law, etc.). This allows users to focus on the knowledge area or specific domain and to eliminate irrelevant information. It also allows users to create an account. This permits the storage of their preferences and options, independently of the computer used to access the database.
- (2) The central area that includes a dynamic network that displays the search concept/term and links it to all related items in terms of a closed inventory of conceptual relations. When users click on any of the terms or concepts in the map, the network rearranges itself. In this new map, the term/concept that was clicked on is at the center and is connected with all of the entities directly related to it. As shown (Figure 1), in the lower left corner of the map, there is a text box with captions that allow users to identify the three categories of conceptual relations in EcoLexicon: (i) hyponymic (generic-specific) relations; (ii) meronymic (part-whole) relations; (iii) non-hierarchical relations. The conceptual relations used in EcoLexicon include a set of 17 hierarchical (hyponymic



López 2009). The complete list of the resources for each concept is shown in this box. Users can easily identify the type of resource by means of the icons beside each of the resource listed. In order to access more information regarding the resource (title, description, source, etc.) users can place the cursor on the resource and a new window will open with all this information.

- Conceptual categories. Each concept in EcoLexicon is associated with one or more conceptual categories, which are shown as a list. If users click on one of these categories, this opens a window with a list of all the concepts included. Furthermore, this box includes a *Category hierarchy* icon, which, when clicked on, shows the concepts in a hierarchical format in which nodes can expand or retract. If one of the categories in the hierarchy is clicked on, a window appears with all the concepts associated with that category.

The screenshot displays the sidebar for the term 'alluvial fan' in EcoLexicon, organized into four main sections:

- Definition:** A dropdown menu labeled 'Definition' is open, showing the text: "alluvial fan: fan-shaped sediment deposit created where a stream flows out onto a gentle plain."
- Terms:** A dropdown menu labeled 'Terms' is open, listing translations in various languages:
  - alluvial fan (English)
  - abanico aluvial (Spanish)
  - alluvialer Schuttfächer (German)
  - Schwemmkegel (German)
  - Geröllfächer (German)
  - аллювиальный веер (Russian)
  - éventail alluvial (French)
  - αλλουβιακό ρητιδίο (Greek)
- Resources:** A 'Resource information' window is open, displaying:
  - Title: Alluvial fan
  - Description: Aerial view of Lost River Mountains, alluvial fan, and floodplain of Big Lost River, Butte and Custer counties, Idaho, U.S.A.
  - Image: A small thumbnail image of the landscape.
  - Source: <http://www.gly.uga.edu/railsback/Fieldimages/LostRiver/AlluvialFan2.jpeg>
 To the right of this window is a larger image showing an aerial view of a wide, flat alluvial fan extending from a mountain range.
- Conceptual Categories:** A dropdown menu labeled 'Conceptual categories' is open, showing:
  - C.2.2 Modified coastal area
  - C.1.1.1 Coast feature
 Below the list is a blue button labeled 'Categories hierarchy'.

Figure 2: Extract of the side bar associated for each search item in EcoLexicon.



**Phraseology**

**Term information**

Term:	hurricane
Language:	English
Term type:	main term
Context:	hurric3a.td
Part of speech:	common noun

[View concordances](#)

**Phraseological section**

**ACTION**

to_come_against_sth_with_sudden_force	batter blast hit strike
---------------------------------------	-------------------------

**CHANGE**

to_cause_to_change_for_the_worse	affect damage demolish destroy devastate injure ravage sweep away wreck
----------------------------------	---

[Phraseological entry](#)

**Details verbo**

Verb: damage

Usage examples

1. In 1985 another cyclone killed 10,000 people, destroyed 17,000 homes and damaged a further 122,000.
2. Hurricane Gilbert damaged more than 100,000 low-income homes in 1988, producing costs of \$558 million.
3. Over 7000 homes were damaged by the hurricane.
4. This hurricane damaged the coast from Texas to eastern Louisiana.

Note: The PATIENT is normally a construction, or area.

**Phraseology**

**Nuclear meaning**: CHANGE

**Meaning dimension**: to\_cause\_to\_change\_for\_the\_worse

**Dimension specification**: NATURAL DISASTER causes a PATIENT to change for the worse.

**Verbs**

affect damage demolish destroy  
devastate injure sweep away wreck  
ravage

Figure 3: Extract of the phraseology box associated with *hurricane* in EcoLexicon.

- Phraseology. If the central element in the map is a term, this box shows a list of verbs most commonly used with the term. Phraseological information can also be accessed via the term entry in certain concepts. Verb collocations are classified and described according to meaning. For this reason, they were primarily classified in terms of their lexical domain (i.e. the nuclear meaning), and subsequently in terms of the frame activated within each lexical domain (i.e. meaning dimension). Once the lexical domain and frames are stated, the verbs are specified. By clicking on the verbs, the user has access to the usage sentences for the verb in question, as well as a note section with information about meaning restrictions.

### 3 Conclusion

As shown, Frame-based Terminology provides a full account of the information necessary to describe a specialized knowledge unit in a terminological entry. The practical application of FBT, the environmental knowledge base EcoLexicon, provides conceptual, linguistic, and administrative information for each specialized knowledge unit. To enhance knowledge acquisition,

conceptual information in EcoLexicon is stored and represented in different ways. Specialized environmental knowledge is represented by means of conceptual networks codified in terms of conceptual propositions in the form of a triplet (concept relation concept). The conceptual relations used in EcoLexicon include a set of 17 hierarchical (hypernymic and meronymic) and non-hierarchical relations. Conceptual information is also shown in the form of natural language definitions in English and Spanish, which are based on the conceptual propositions established by the concept to be defined. Additionally, domain-specific knowledge is also presented in the form of images, documents and videos, which complement the previously entered conceptual information. Linguistic information is now being enhanced by the introduction of verb phraseological information regarding each term. In a near future, we hope to be able to provide a complete phraseological description of every term.

## 4 References

- Buendía Castro, M. (2013). *Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources*. PhD Thesis, University of Granada, Spain.
- Buendía Castro, M., & Sánchez Cárdenas, B. (2012). "Linguistic knowledge for specialized text production". In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul: ELRA, pp. 622-626.
- Faber, Pamela. (2009). *The Cognitive Shift in Terminology and Specialized Translation*. *MonTI. Monografías De Traducción e Interpretación* 1(1), pp. 107-134.
- Faber, Pamela. (2011). *The Dynamics of Specialized Knowledge Representation: Simulational Reconstruction or the Perception-action Interface*. *Terminology* 17(1), pp. 9-29.
- Faber, Pamela (ed.). (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin, Boston: Mouton de Gruyter.
- López Rodríguez, C.I., Buendía Castro, M. & García Aragón, A. (2012). *User Needs to the Test: Evaluating a Terminological Knowledge Base on the Environment by Trainee Translators*. *Jostrans. The Journal of Specialized Translation* (18), pp. 57-76. Accessed at: [http://www.jostrans.org/issue18/art\\_lopez.pdf](http://www.jostrans.org/issue18/art_lopez.pdf). [20/12/2012].
- Prieto Velasco, J.A. (2009). *Traducción e Imagen: La Información Visual En Textos Especializados*. Granada: Tragacanto.
- Prieto Velasco, J.A. & López Rodríguez, C.I. (2009). *Managing graphic information in terminological knowledge bases*, *Terminology* 15(2), pp. 179-213.
- Sánchez Cárdenas, B., & Buendía Castro, M. (2012). *Inclusion of Verbal Syntagmatic Patterns in Specialized Dictionaries: The Case of EcoLexicon*. In R. Vatvedt Fjeld, J. Matilde Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: EURALEX, pp. 554-562.

### **Acknowledgements**

This research was carried out within the framework of the project RECORD: Representación del Conocimiento en Redes Dinámicas [Knowledge Representation in Dynamic Networks, FFI2011-22397], funded by the Spanish Ministry of Science and Innovation.



# Experts and Terminologists: Exchanging Roles in the Elaboration of the Terminological Dictionary of the Brenner Base Tunnel (BBT)

Elena Chiocchetti, Natascia Ralli  
European Academy of Bolzano/Bozen (EURAC research)  
echiocchetti@eurac.edu, nralli@eurac.edu

## Abstract

This paper presents the Italian-German Terminological Dictionary of the Brenner Base Tunnel (BBT) produced and printed in 2011 by the European Academy of Bolzano. In particular, we describe the role played by domain experts and terminologists during the elaboration of the dictionary, which was different from the one that is usually assigned to them in terminological projects. This switching of roles had several consequences on the structure and content of the dictionary, which we will discuss. We also briefly illustrate the challenges faced and how we approached specific problems, e.g. the structure of the definitions, the managing of synonyms and variants, the varied nature of terms selected for the dictionary and their specific treatment with a view to the needs of the target users. The two main target groups consist of experts and semi-experts from various professions who are confronted daily with terminology pertaining to a large array of specialised domains (from environmental to engineering to legal terminology).

**Keywords:** Terminological dictionary; Domain experts' role; Terminology workflow

## 1 Introduction

The Brenner Pass is located between Italy and Austria in a favourable position, making it one of the main thoroughfares connecting Central and Northern Europe with the Italian peninsula and the Mediterranean area. Passenger and freight traffic on the motorway and railway crossing on this rather low Alpine pass (1372 m above sea level) is particularly intense all year round and in constant growth (cf. Maino et al. 2011: 10). The ensuing social, economic and environmental consequences call for an improvement of the railway line between Munich and Verona in order to favour a modal shift from road transportation to rail transportation for goods, in order to achieve a more efficient mobility system and relieve the population along the railway line from transit traffic (Maino et al. 2011: 8). The Brenner Base Tunnel (BBT) is the most important element of this improvement project and will be 55 km long, connecting Innsbruck in Austria with Fortezza in Italy. Its completion is currently scheduled for the year 2025.

For the two neighbouring countries of Austria and Italy to be able to cooperate efficiently in the realisation of the project, all relevant documents must be available in two languages, i.e. German and Italian. This calls for a large number of translations, with technical, legal, administrative and commercial content produced over a very long period. For smooth communication, all these texts need to use correct and coherent specialised terminology pertaining to a vast array of domains in both languages. Therefore, the scope of the BBT project, its duration, the number of different professionals involved, the need for overcoming language barriers as well as legal/administrative barriers, together with the complex and varied nature of the terminology needed, represent the background for the creation of a dedicated bilingual dictionary. This dictionary is the Italian-German dictionary *Dizionario Terminologico della Galleria di Base del Brennero/Terminologisches Wörterbuch zum Brenner Basistunnel*, which aims at supporting transnational communication and cooperation for all issues related to the BBT.

The dictionary was planned and created by domain experts and terminologists in close cooperation. However, unlike standard terminology projects in which terminologists perform most of the activities and are ideally supported by domain experts as revisers and consultants, the BBT terminological dictionary was compiled by experts, while the terminologists took over the role of consultants and quality controllers. The knowledge possessed by the domain experts on the special language and conceptual structure of the domains treated was channelled and directed into a precise terminological working method. This exchange of roles with respect to the most common workflow in terminology proved to be challenging but stimulating for both groups. In section 5 we discuss the main challenges faced.

## 2 The BBT dictionary

The BBT dictionary was promoted and financed by the Brenner Base Tunnel Societas Europae (BBT SE). The BBT Society commissioned the creation of a bilingual dictionary to a group of mainly monolingual Italian or German speaking experts in territorial planning, land management, engineering and mobility, who sought advice from a group of bilingual terminologists. The terminologists supported the dictionary authors in defining two main aspects before starting work, i.e. the target users of the dictionary and the methodology to be adopted during compilation, so as to avoid inconsistencies and the need for later adjustments.

The structure of the definitions and the treatment of variants/synonyms, for example, consistently follow a clearly onomasiological rather than semasiological approach. The onomasiological method is typical of terminology work, as it considers the key concepts of a specialised domain and the relations existing between them as central elements for the selection and definition of the terms to be included in the dictionary. The lemmas in the dictionary being concept-based, all synonyms or variants (e.g. a full form and its corresponding acronym) are listed and defined together, since they all designate the same concept. Unlike dictionaries that follow a semasiological approach, two synonyms are not

explained separately in distinct lemmas: the synonym only has a clear reference to the lemma of the main term, where all necessary information is given, including all equivalents in the other language. The authors and sponsor deemed it necessary to target the dictionary mainly at experts and semi-experts participating in all phases and aspects of the BBT project. The function of the BBT dictionary is therefore to support both groups of users in understanding and producing texts in the foreign language. Experts “will have no reception problems within [their] own field. [They] may have to acquire new knowledge, but [they are] not likely to find this in any lexicographical dictionary” (Bergenholtz & Kaufmann 1997: 102). Thus, they will predominantly use the dictionary for text production and retrieving foreign language terms. Semi-experts are potential dictionary users who come from other related subject fields that are relevant but not strictly specific to railway tunnel projects, e.g. engineers, surveyors, geologists, installers, etc. These professionals work in the public or private sector and daily interact with the world of railway construction and the BBT (cf. Bergenholtz & Kaufmann 1997: 101). They will need to find all synonyms and variants of a term as well as definitions in order to get a better picture of each concept.

The BBT dictionary contains about 2000 terms in German and Italian from diverse domains, ranging from railway construction to tunnel building, economy, energy, geology, telecommunications, transport/mobility, social issues and environmental terminology, thus encompassing – next to strictly technical terminology – also organisational, administrative and legal terminology (cf. Maino et al. 2011: 14).

### 3 Experts and terminologists: standard forms of cooperation

Keine Terminologie ohne Fachleute – keine Fachleute ohne Terminologie<sup>1</sup> (RaDT 2013: 9)

The quotation above expresses the essence of terminology. Terminologists and experts must work side by side and create synergies to achieve a reliable and high quality terminological product. In principle, terminologists retrieve, reference and record the terms that pertain to a specialised domain in one or more languages (RaDT 2004: 2; Chiocchetti et al. 2013: 41-42), while domain experts are normally consulted for explanations and information (RaDT 2013: 7-8). Often they also revise the final product of the terminologists.

Usually terminologists start by studying and delimiting the domain to be processed terminologically, which is often subdivided into smaller subdomains. To this purpose, terminologists acquire and skim relevant and up-to-date reference material, which may also be collected or selected as indicated by domain experts. The material is collected in all languages that are to be included in the terminologi-

---

1 “No terminology without domain experts – no domain experts without terminology” (translation by the authors).

cal product, making sure that it is original material written by expert native speakers, so that it reflects the actual language used by the community of experts of a given domain.

This material serves as a basis for understanding the key concepts of the domain under analysis, which are organised in concept systems to illustrate the relation between each concept within the specific domain or sub-domain. The material is also used to retrieve all the terms that designate the concepts of a specialised domain. “A term is a designation consisting of one or more words representing a general concept in a special language in a specific subject field” (ISO 2009:704: 7.2.1). There might be more than one designation for the same concept, i.e. there might be synonyms (ISO 2009:704: 7.2.4). Also term variants, e.g. abbreviated forms like clippings and acronyms, are common in specialised domains (cf. ISO 2009:704: B.2.4).

Based on this preparatory research and analysis, terminologists then compile (or update) fully fledged terminological entries with relevant information, i.e. with definitions, contexts of use, term variants/synonyms, usage notes etc. The most important element of a terminological entry is the definition, as it conveys the meaning of a concept within the specialised domain to which it belongs. All the other pieces of information contribute to explaining how the terms and variants that designate a concept are employed within a specialised domain.

Finally, with the help of all information gathered in the source language and the structure of the relevant concept-system, terminologists retrieve the equivalents in the target language. Target language terms are then processed terminologically in the same way as the terms in the source language are. If no equivalent exists, terminologists may propose new terms (translation proposals) to fill the terminological gaps (see section 5.5).

Written and human sources (i.e. domain experts) may be consulted by terminologists for information and explanations at any time during the entire process. Domain experts are preferably involved in the planning and realisation of terminology projects from the very beginning. Usually their role consists in supporting terminologists (RaDT 2013: 7-8). At the beginning, they can help to plan and organise terminology projects, especially by providing information on relevant reference material and selecting it. Their initial input is also important when delimiting the domains and subdomains to be processed terminologically.

As terminology work proceeds, domain experts may be asked to select the terms that were extracted from the written material by the terminologists, i.e. to choose which terms shall become lemmas of a dictionary or terminological entries in a database. Experts can also check concept systems and verify the correctness of the terms, synonyms and associated definitions. Domain experts usually provide competent advice in case of any doubts. Finally, being part of a scientific community and/or of a practical community, they represent the ideal channel for disseminating the results of the terminology work to their peers.

As we have seen, domain experts normally act as consultants and/or revisers. They are of paramount importance for the success of terminology work and represent a precious source of information, since they assess the quality of source documentation, explain the meaning of concepts belonging to their domain of expertise and/or check whether the terminology collection is correct and complete. However, what they usually do not do is compile terminological entries themselves.



## 4 Experts and terminologists: exchanging roles for the BBT dictionary

For the creation of the BBT dictionary domain experts and terminologists exchanged roles. This time the terminologists acted as consultants and revisers, while the experts took over most of the work concerning the compilation of the dictionary entries.

The advisory role of terminologists started with helping the experts select the most up-to-date and authoritative sources. As the experts are able to judge personally which pieces of information are correct and precise, they tend to disregard their provenance and to treat all types of reference material in the same way, from highly specialised technical manuals to commercial web pages. By applying strategies for source evaluation and selection learnt from the terminologists, the authors managed to produce a more homogeneous and complete dictionary. Current scientific and technical sources were consistently preferred to general language dictionaries or encyclopaedias. In any case, information was always double-checked and carefully evaluated.

Terminologists provided support also in the initial term selection phase. In fact, they strived to agree with the experts on a coherent set of related terms. For the dictionary this sometimes meant discarding a very specialised – albeit possibly useful – lemma and including a maybe less tricky lemma instead, in order to ensure that most concepts in the concept field of a specific subdomain were represented. This compromise in term selection likewise allowed reaching a higher level of homogeneity in the dictionary.

Great efforts were made by the terminologists in revising definitions. They convinced the experts that the typical structure of terminological definitions could represent a useful definition strategy, especially for semi-experts. Terminological definitions start with stating the superordinate concept and then list all the characteristics that distinguish the concept under analysis from its related concepts nearby (see section 5.6). In this way they provide essential information in a very compact form and help users to quickly understand the position of a concept within its specialised domain. Achieving a more systematic and coherent structure for all definitions also allowed the dictionary to be turned into a more consistent product.

Terminologists took over several other methodological aspects, e.g. by supervising the rigorous treatment of synonyms and abbreviated forms and their cross references to the main entry. They performed the terminological revision, i.e. they checked the consistency of terms and definitions. Most important, they checked whether the terms and definitions in Italian and German actually all referred to the same concept, thus making up for the lack of language competences of the domain experts. In some cases they actually retrieved the equivalents in the target language or advanced translation proposals. Finally, terminologists performed the linguistic revision in both Italian and German.

This swapping of roles was essentially born out of necessity, as the domain experts lacked the necessary competences and methodological basis for dictionary-making as well as (partly) lacking the linguistic competences in both working languages, while the terminologists could not take over their

standard role due to lack of time and financial means. However, despite the challenges faced, the co-operation in this “reversed form” proved stimulating and fruitful for both parts.

## **5 Challenges faced and compromises reached**

Before and during the compilation phase of the dictionary the domain experts were trained by the terminologists to follow the basic principles of terminology and of dictionary-making. In some areas it was particularly difficult to create a common knowledge base and achieve consensus on a methodology, so that both parts had to agree on compromises, as explained in detail in the following paragraphs.

### **5.1 Hierarchy of sources**

Many different types of sources can be used for terminology work (cf. ISO 10241-1:2011: 4.3.5.2). Their respective relevance will vary according to the aim, type and content of terminology work, the domain(s) treated, the languages considered, the end users, etc. (Chiocchetti et al. 2013: 18). Usually terminologists give preference to authoritative sources like legal documents, standards, documents generally recognised by the scientific community (e.g. textbooks) (ISO 10241-1:2011: 4.3.5.2). Often the reference material is classified hierarchically, with the most official and authoritative sources at the top (e.g. laws, standards, etc.) and the less authoritative ones at the bottom (e.g. private webpages, commercial material, etc.). Information retrieved in sources classified at a higher level of the hierarchy will be preferred to information found in documents filed at a lower level. Contrary to this terminological practice, the experts working on the BBT dictionary often used material from general encyclopaedias and from popular websites (e.g. Wikipedia), which are easily retrieved on the Internet but cannot always be considered reliable or specific enough. This is the reason why they are generally avoided by terminologists or at most used for information retrieval, but seldom quoted.

For the BBT dictionary, since the expertise of the authors allowed them to assess the correctness and quality of definitions found in “unconventional” reference material, many definitions were accepted and cited within the dictionary whenever no other source of information was available. This compromise allowed a faster compilation of parts of the dictionary without any notable loss of quality.

### **5.2 Subdivision into glossaries**

According to standard terminological practice, when treating large or very diverse sets of data (see section 2), work is subdivided into thematic glossaries to facilitate compilation and revision. This practice was new to the authors who nevertheless quickly grew accustomed to the method. In the final version of the dictionary all terms could still be listed in alphabetical order, thus ensuring imme-

diate retrievability of each term as well as of all variants or synonyms by the end-users (Maino et al. 2011: 14).

### 5.3 Term selection

Due to restrictions in space, it was not possible to compile complete concept systems for all glossaries and accommodate all relevant terms in the dictionary. Term selection was therefore guided by the relative relevance for the specific BBT project and not strictly by conceptual diagrams of each domain, as terminologists normally strive to do with the aim of treating all domains equally.

Catering to two different groups of users likewise required a series of compromises in the selection of terms and variants/synonyms to be considered. Experts tend to use highly specialised vocabulary (e.g. also consisting of acronyms, initialisms and formulas) that is monosemic and unambiguous to avoid problems in interpretation (cf. Sobrero 1993; Cortelazzo 1994). Semi-experts, however, call for what is defined as a “variationist approach” in terminology, i.e. an approach where all synonyms and variants used to designate a concept are considered, whether they be “full forms, such as simple, compound or complex terms [and] [...] all their variations” (Bertaccini & Lecci 2009), or abbreviated forms (see Fig. 1). To this aim, the dictionary builds a network of references from all synonyms and variants (e.g. acronyms) to the main entry containing the definition of the concept (see Fig. 2 and 3).

<hr/> <b>tunnel boring machine</b> <b>Sinonimo:</b> fresa (2), talpa <b>Sigla:</b> TBM <b>Definizione:</b> Macchina che permette lo scavo meccanizzato con fresa a tenuta idraulica ed il contemporaneo rivestimento con conci prefabbricati. <b>Fonte:</b> <a href="http://www.snamretegas.it:17.11.2004">http://www.snamretegas.it:17.11.2004</a> <hr/>	<hr/> <b>Tunnelbohrmaschine</b> <b>Synonym:</b> Fräse, Tunnelfräse <b>Abkürzung:</b> TBM <b>Definition:</b> Gerät zum Lösen von Gebirge im Zuge des Tunnel- oder Stollenbaus. Das Lösen erfolgt hierbei mechanisch und über den vollen Querschnitt des Ausbruchs. <b>Quelle:</b> <a href="http://www.bauwerk-verlag.de:17.11.2004">http://www.bauwerk-verlag.de:17.11.2004</a> <hr/>
---	--

Fig. 1: Entry tunnel boring machine/Tunnelbohrmaschine with synonyms (= sinonimo/Synonym) and initialisms (= sigla/Abkürzung).

### 5.4 Managing synonyms and variants

The consistent treatment of synonyms and also variants —with the respective references from the synonym/variant to the main defined terms — was taken over and managed by the terminologists. Experts tend to disregard the importance of terminological variation, because they have all the synonyms and variants in mind. But for a semi-expert dictionary user it might, for example, not be so easy to read the full form behind an acronym used in a text. By listing all designations that refer to the

same concept in separate lemmas as well as in alphabetical order in the BBT dictionary, with consistent cross-references to their respective full forms and/or main terms (see Fig.2 and 3), maximum retrievability of information for all end-users could be ensured.



Fig. 2: Entry *fresa (2)*, cross-reference to the main Italian term *tunnel boring machine*.

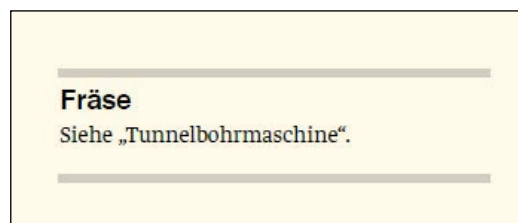


Fig. 3: Entry *Fräse*, cross-reference to the main German term *Tunnelbohrmaschine*.

## 5.5 Diverging treatment of legal/administrative terminology from purely technical terminology

The nature and origin of the terminology used within the BBT project is quite diverse (see section 2). Legal terminology, which is for example contained in building regulations, poses a particular challenge in such a context, since it is much more difficult to find conceptual equivalence in the legal/administrative than in the technical domain. Legal and administrative terminology is always strongly connected to a specific legal system with its own cognitive and conceptual structures, as well as its written or oral sources (cf. Gambaro & Sacco 1996: 9; Sandrini 1996: 138; Šarčević 1997: 232). For this reason, it might not always be possible to find an equivalent in the target language; the concept designated by the term might be specific to the source legal system and source language and be completely unknown in the target legal system and language. This situation creates a terminological gap. The presence of a terminological gap causes the need to look for strategies of translation other than equivalence (e.g. paraphrase, neologism, etc.). In the BBT dictionary, terminological gaps in the legal/

administrative terminology were faced by offering translation proposals that conveyed the meaning of the concept in the source legal system and source language into the target legal system and language (see Fig. 4). Due to the intrinsic connection of legal terminology with its legal system, it was not possible to borrow legal terms from other legal systems.

<p><b>piano comunale di classificazione acustica</b> <b>Sigla:</b> PCCA <b>Definizione:</b> Piano che definisce le norme finalizzate alla tutela dell'ambiente e della salute pubblica dall'inquinamento acustico prodotto dalle attività antropiche, disciplinandone l'esercizio al fine di contenere la rumorosità entro i limiti normativamente stabiliti. <b>Fonte:</b> <a href="http://www.arpat.toscana.it:03.02.2011">http://www.arpat.toscana.it:03.02.2011</a></p>	<p><b>Lärmklassifizierungsplan der Gemeinde</b> <b>Quelle:</b> EURAC (Übersetzungsvorschlag)</p>
---	--

**Fig. 4: Translation proposal in German for the Italian administrative term *piano comunale di classificazione acustica*.**

Another approach, however, was followed for gaps in the technical terminology. In this case, problems mainly concerned a different level of evolution of the tunnel building techniques in Italy and in Austria, with the ensuing absence of some terms that designate very specific concepts. As the technical terminology used in Switzerland in both Italian and German is very complete and up-to-date, several Swiss sources helped to fill the presumed gaps (see Fig. 5). For similar reasons, some sources of information from Federal Germany could be referred to for the German language terminology.

The terminologists had to explain and discuss this diverging treatment with the dictionary authors. Since none of them was a legal expert, the implication of legal comparison across national borders was not immediately clear and had to be motivated and explained.

<p><b>Bohrkopf</b> <b>Definition:</b> Vorderster Teil einer Vortriebsanlage, der mit Rollenmeißeln bestückt ist und der beim Abschlag den Fels löst. <b>Quelle:</b> <a href="http://www.islisbergtunnel.ch:29.04.2005">http://www.islisbergtunnel.ch:29.04.2005</a></p>
---

**Fig. 5: Entry *Bohrkopf* with definition from a Swiss website about tunnel building.**

## 5.6 Definitions

Experts have the tendency to provide longer encyclopaedic definitions rather than much more compact classical terminological definitions. They tend to give more information than the amount strictly necessary for a mere definition of the concept, tending to add explanations on how a defined concept is employed within their domain of expertise (see Fig. 6).

The traditional, most explicit and precise definition in terminology is the intensional definition, stating “the superordinate concept immediately above [the concept that is being defined], followed by the delimiting characteristic(s)” (ISO 704:2009: 6.2) that distinguish it from coordinate or from other related concepts. This type of definition allows full and systematic identification of a concept with respect to all others in the specific domain (Sager 1990: 42). Yet it is very brief and not always sufficient for laypersons or semi-experts to really understand the meaning and usage of the concept defined.

In this case the terminologists gave way to the desires by the dictionary authors for a more in depth explanation of some key concepts. In the dictionary, which has a very compact structure without any notes or comments, definitions were allowed a more flexible structure, sometimes leaving room for the inclusion of necessary additional information and clarifications.

<p><b>Schwermetall</b>  <b>Definition:</b> Metall mit einer Dichte von über 5 g/cm<sup>3</sup>, z.B. Zink, Kupfer, Zinn, Chrom, Cadmium, Blei, Quecksilber u. a. Eine Reihe von Schwermetallen sind in Spuren für biologische Vorgänge lebensnotwendig. Hohe Konzentrationen der meisten Schwermetalle sind dagegen äußerst giftig und können über den Abfall, Verbrennungsgase und Abwasser zu erheblichen Umweltproblemen führen, wenn sie über Boden und Pflanzen in die Nahrungskette gelangen.  <b>Quelle:</b> Köppen D., 1998. Definitionen für agrarökologisch relevante Sachverhalte. Fachbereich Agrarökologie, Rostock.</p>	<p><b>metallo pesante</b>  <b>Definizione:</b> Metallo con densità maggiore di 5 (zinco, cadmio, mercurio, stagno, piombo, cromo, manganese, ferro, cobalto, nichel e rame), nonché i relativi composti che, pur essendo naturali componenti della crosta terrestre, tramite numerose attività umane vengono mobilitati e concentrati a livelli pericolosi per la salute e l'ambiente.  <b>Fonte:</b> Gamba G., Martignetti G., 1995. Dizionario dell' ambiente. Formulazione di responsabilità. ISEDI, Torino.</p>
---	---

Fig. 6: Definition of *Schwermetall* with additional information.

## 6 Conclusions

This role-switching exercise proved very fruitful for both sides. The experts became familiar with the basic principles of terminology work and dictionary-making; particularly the evaluation of source material and the definition-writing skills were considered useful to them beyond the BBT dictionary itself. The terminologists learnt how to find pragmatic solutions to practical problems, as well as how



to reach compromises between standard terminological practice and the limitations of a printed reference work that is aimed primarily at experts and semi-experts, rather than at translators and other language professionals.

The result of this exchange of roles is a dictionary based on many more compromises than other terminological projects. Terminologists usually take over most of the work and the role of the experts is limited to sporadic interventions as advisers and proofreaders. As a consequence, terminologists often have the last word on dictionary structure and content, even though the opinion of the experts is always greatly considered and systematically taken into account. However, for the BBT dictionary, the advice of the terminologists during dictionary compilation was generally limited to more formal aspects, such as revising the structure and wording of the definitions, and to methodological issues, e.g. the consistent treatment of synonyms and short forms.

Great compromises were reached, for example, concerning definitions. As we have seen, in the BBT dictionary definitions still follow the classical terminological structure whenever possible, but their content goes beyond the mere identification of superordinate concept and delimiting characteristics. Definitions thus often include additional (technical) information that the authors considered necessary and useful for either peers or semi-experts. Another compromise was reached for the treatment of linguistic information that is normally given by terminologists in their work, i.e. grammatical information, example sentences, notes distinguishing the contexts of use of different synonyms and term variants, language register, collocations, etc. This type of information is basically absent from the dictionary, due to the fact that the target groups of the dictionary do not primarily include translators and language professionals. While translators might wish for specific linguistic information, this is often unnecessary for experts, so more space was devoted to treating a larger number of lemmas.

The BBT dictionary project has finally proven that, given the different backgrounds and approaches of terminologists and domain experts, it is advisable to provide detailed guidelines on how to handle specific aspects (e.g. the structure of definitions) in order to work along common guidelines and principles from the very beginning (Chiocchetti et al. 2013: 46). It has also shown that it is indeed possible to envisage different forms of cooperation between domain experts and language experts, still ensuring a high level of quality of the final product.

## 7 References

- Bergenholtz, H. & U. Kaufmann (1997). Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. In *Hermes, Journal of Linguistics*, 18/1997, pp. 91-125.
- Bertaccini, F. & C. Lecci (2009). Conoscenze e competenze nell'attività terminologica e terminografica. In "Terminologia, ricerca e formazione", *Publifarum*, 9. Accessed at: [http://www.publifarum.farum.it/ezone\\_articles.php?art\\_id=107](http://www.publifarum.farum.it/ezone_articles.php?art_id=107) [10/11/2013].

- Chiocchetti, E., Heinisch-Obermoser, B., Löckinger, G., Lušický, V., Ralli, N., Stanizzi, I. & T. Wissik (2013). *Guidelines for Collaborative Legal/Administrative Terminology Work*. Bolzano: EURAC. Accessed at: [http://www.eurac.edu/en/research/institutes/multilingualism/Documents/Guidelines\\_for\\_collaborative\\_legal\\_administrative\\_terminology\\_work.pdf](http://www.eurac.edu/en/research/institutes/multilingualism/Documents/Guidelines_for_collaborative_legal_administrative_terminology_work.pdf) [10/11/2013].
- Cortelazzo, M.A. (1994). *Lingue speciali. La dimensione verticale*. Unipress: Padova.
- Gambaro, A. & R. Sacco (1996). *Sistemi giuridici comparati*. Torino: UTET.
- ISO 704:2009. Terminology Work – Principles and Methods.
- ISO 10241-1:2011. Terminological Entries in Standards – Part 1: General Requirements and Examples of Presentation.
- Maino, F., Cavallaro, F. & M. Wagner (eds.) (2011). *Dizionario Terminologico della Galleria di Base del Brennero Italiano-Tedesco/Terminologisches Wörterbuch zum Brenner Basistunnel Deutsch-Italienisch*. Bolzano: EURAC/BBT.
- RaDT – Rat für Deutschsprachige Terminologie (2004). *Berufsprofil Terminologin/Terminologe*. RaDT: Bern. Accessed at: [http://www.iim.fh-koeln.de/radt/Dokumente/RaDT\\_Berufsprofil.pdf](http://www.iim.fh-koeln.de/radt/Dokumente/RaDT_Berufsprofil.pdf) [10/11/2013].
- RaDT – Rat für Deutschsprachige Terminologie (2013). *Terminologisches Basiswissen für Fachleute*. RaDT: Köln. Accessed at: [http://www.iim.fh-koeln.de/radt/Basiswissen-RaDT2013-16s\\_ebook.pdf](http://www.iim.fh-koeln.de/radt/Basiswissen-RaDT2013-16s_ebook.pdf) [10/11/2013].
- Sager, J. C. (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.
- Sandrini, P. (1996). *Terminologearbeit im Recht. Deskriptiver begriffsorientierter Ansatz vom Standpunkt des Übersetzers*. Vienna: TermNet (IITF-Series no. 8).
- Šarčević, S. (1997). *New Approach to Legal Translation*. The Hague: Kluwer Law International.
- Sobrero, A. (1993). *Lingue speciali*. In Sobrero, A. (ed.): *Introduzione all'italiano contemporaneo. La variazione e gli usi*, vol. 2. Bari: Laterza, pp. 237-277.



# Cloud Terminology Services Facilitate Specialised Lexicography Work

Tatiana Gornostay, Andrejs Vasiljevs

Tilde

tatiana.gornostay@tilde.lv, andrejs@tilde.lv

## Abstract

In this software demonstration paper we present an innovative cloud-based platform TaaS “Terminology as a Service” developed in an EU-funded project.<sup>1</sup> The TaaS platform provides language workers and language applications (human and machine users, accordingly) with the services to foster the creation, validation, harmonisation, sharing, and application of terminology resources. Under language workers we understand language professionals, for example, technical writers, editors and proof-readers, translators and localisers, terminologists and domain specialists, lexicographers and terminographers, and others. Under language applications (or machine users in other words), in the first place we consider computer-assisted translation (CAT) tools and machine translation (MT) systems (also, knowledge organisation systems in library and information science, search engines, and others). TaaS provides the following terminology services: terminology search in various sources, terminology identification in and extraction from user-uploaded documents, terminology visualisation in user-uploaded documents, translation equivalent lookup in and retrieval from various sources, terminology refinement and approval by users, terminology sharing with other users, collaborative working environment, and terminology reuse in other applications. Among other benefits for language workers, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based specialised terminology work. TaaS fills the gap of innovative environment to speed up the development of specialised dictionaries.

**Keywords:** terminology service; terminology work; specialised lexicography

## 1 Introduction

Lexicography, as the theory and practice of dictionary development, is one of the most labour-intensive human activities. The creation of a new dictionary from the scratch and its delivery to an end user requires many resources in terms of time, labour, and finance. The main drawback of a conventional *paper* dictionary is its static and out-of-date content. In specialised lexicography, it is even more critical since terminology is developing rapidly along with its subject field and science in general.

---

1 The TaaS Beta was officially launched on November 1, 2013 and is publicly available for open Beta testing at <https://demo.taas-project.eu>.

To overcome the shortcomings of conventional lexicography, an electronic punch-card machine was first used to create a prototype of a modern electronic dictionary by Roberto Busa in the 20th century. His first work was based on the automatic linguistic analysis of the works of Saint Thomas Aquinas (he lemmatised the texts). Roberto Busa compared the invention of an “electronic book” (instead of a printing book) to the introduction of a printing book by Gutenberg (instead of a manuscript). Since that time automated lexicography has been developing boomingly.

Nowadays, with the evolution of information technologies, the Web, and data (for example, open data, linked data, free language resources etc.), the task of automated specialised lexicographic work is put in the first place. Routine processes have been delegated to a computer. An electronic, or computer-based, specialised dictionary is easy to update and manage, and its main advantage is its flexible, dynamic, and extensible (for example, in terms of new languages) character. Moreover, the new era of information technologies offers the new ways of dictionary representation, for example, on a tablet, mobile, and other devices, and the usage patterns of a dictionary (including a specialised dictionary) are changing with the course of time.

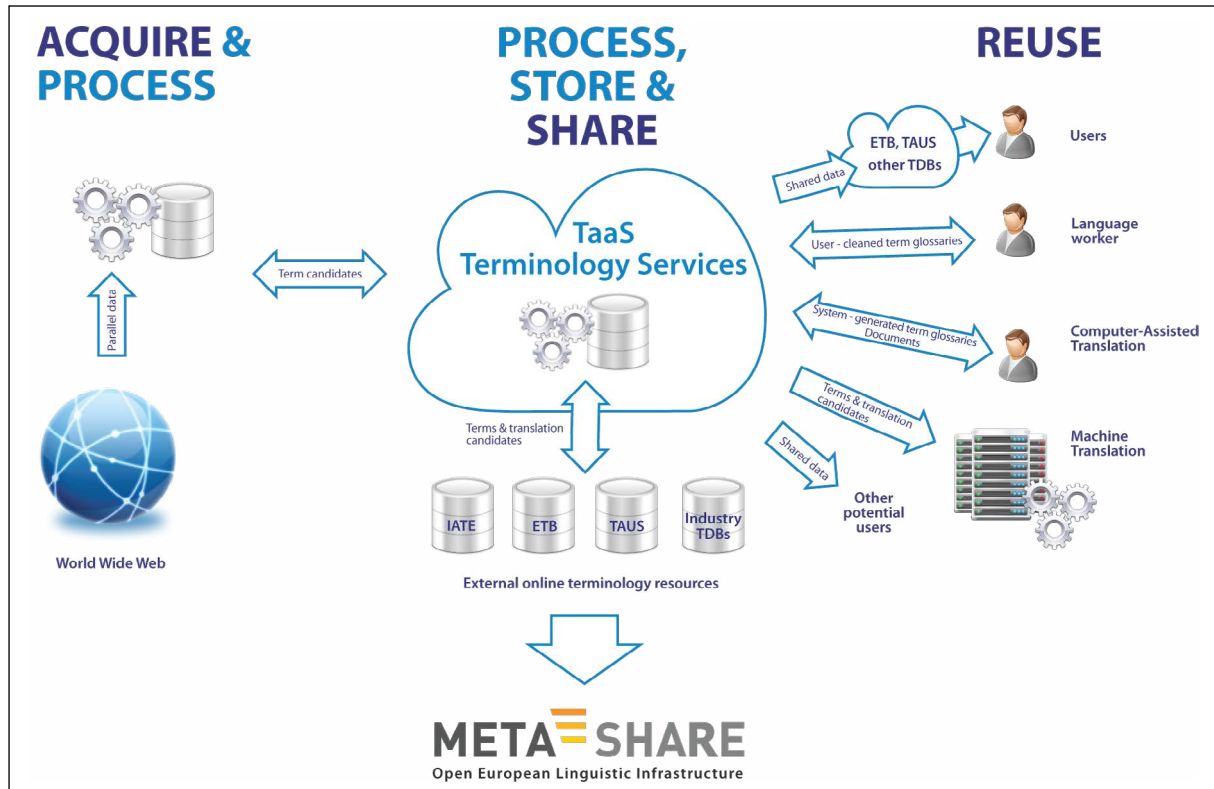
The integration of natural language processing tools within a lexicographer’s working environment have made it possible for him/her to linguistically and semantically analyse and tag data and then to extract required pieces of information from it. In specialised lexicography it is possible to identify and extract term candidates automatically for further processing (for example, refinement, approval, sharing and reuse). Thus a lexicographer can consider hundreds thousands of terms in a certain subject field in comparison with that time when only several thousands (usually no more than 2000) could be included in a conventional specialised paper dictionary. This opportunity is critical particularly in emerging domains.

## **2 TaaS: Terminology as a Service**

In this software demonstration paper we present an innovative platform TaaS “Terminology as a Service”. The platform provides language workers and language applications (human and machine users, accordingly) with the services to foster the creation, validation, harmonisation, sharing, and application of terminological data. Among others, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-oriented, collaborative, multilingual, interoperable, portable, and cloud-based work. TaaS fills the gap of innovative environment to speed up the development of specialised dictionaries.

TaaS is being developed within an industry-research collaborative project under the EU Seventh Framework Programme for Research and Technological Development. The main objective of TaaS is to address the need for instant access to the most recent terms and for direct user involvement in the creation, harmonisation, and sharing of terminological data. The Beta version of TaaS was officially launched on November 1, 2013 and is publicly available for open Beta testing.

The concept of innovative cloud terminology services for language workers and language applications is presented in Figure 1 below.



**Figure 1: TaaS innovative cloud terminology services for language workers and language applications.**

TaaS provides user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based terminology services to perform the following tasks:

- Search terminology in various sources;
- Identify term candidates in user-uploaded documents and extract them automatically applying linguistic and statistical processing;
- Visualise term candidates in user-uploaded documents;
- Look up translation equivalent candidates in various sources (for example, existing external terminology resources EuroTermBank<sup>2</sup>, IATE<sup>3</sup>, TAUS Data<sup>4</sup>, and possible other resources, as well as automatically extracted bilingual terminological data stored in the TaaS Shared Term Repository and used as an additional internal source for target translation lookup);
- Refine term candidates and their translation equivalent candidates;
- Approve refined terminology;

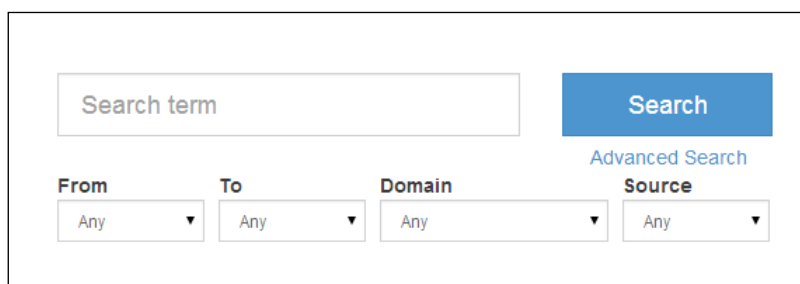
2 [www.eurotermbank.com](http://www.eurotermbank.com)

3 <http://iate.europa.eu>

4 [www.tausdata.org](http://www.tausdata.org)

- Share terminology with other users;
- Collaborate with colleagues in user-friendly working environment;
- Use terminology in other applications via TermBase eXchange ISO-standardised format (TBX), tab-separated value (TSV), and comma-separated value (CSV) export formats and via the TaaS Application Program Interface (API).<sup>5</sup>

To perform most of his/her work in TaaS, the user has to sign up for the services. However, for its non-signed users, TaaS provides the service for terminology search in two sources – TaaS database, which consists of TaaS users’ terminology collections made public by its users, and EuroTermBank, which is the largest European online term bank, providing access to more than 2 million standardised terms from more than 100 national terminology resources in 27 languages. For advanced search, the user has to select the source and target language, domain (a.k.a. subject field), and the source to be searched in (see Figure 2).



The image shows a search interface for TaaS. At the top, there is a text input field labeled "Search term" and a blue button labeled "Search". Below the "Search" button is a link labeled "Advanced Search". Underneath, there are four dropdown menus labeled "From", "To", "Domain", and "Source". Each dropdown menu has "Any" selected and a downward arrow.

**Figure 2: Search form in TaaS.**

For signed users the work in TaaS is organised in projects. A signed user gains access to full TaaS services. To start his/her work, the user has to create a new project indicating the source and target language and the domain the user works in (it is relevant to user document(s) domain). More than 10 input format for user documents are supported. The user might also want to specify optional properties, such as product, customer, project description, and the business unit (in case of a corporate user) (see Figure 3).

---

5 For TaaS API contact TaaS team via e-mail [langserv@tilde.com](mailto:langserv@tilde.com).

The screenshot shows a web form titled "Create Project". It contains the following elements:

- Name \***: A text input field.
- Description**: A larger text input field.
- Source Language \***: A dropdown menu.
- Target Language \***: A dropdown menu.
- Domain \***: A dropdown menu with a help icon.
- Product**: A text input field.
- Customer**: A text input field.
- Business Unit**: A text input field.
- Collection Status**: Two radio buttons: "Public collection (available for search and lookup by other TaaS users)" and "Private collection (available for search and lookup by project users)".
- Buttons**: A blue "Save" button and a "Back to projects" link.

**Figure 3: Creation of a new project in TaaS.**

TaaS also provides a default project with project properties already set for demonstration purposes. Finally, the user has to set the status of his/her project – private or public. If the status of a project is public, the user’s approved terminology will be available for search and lookup by other TaaS users; if the status of a project is private, the user’s approved terminology will be available only to project users.

The user can start his/her work with TaaS by using the default project or creating a new project. In both cases, the user is an administrator of his/her project.

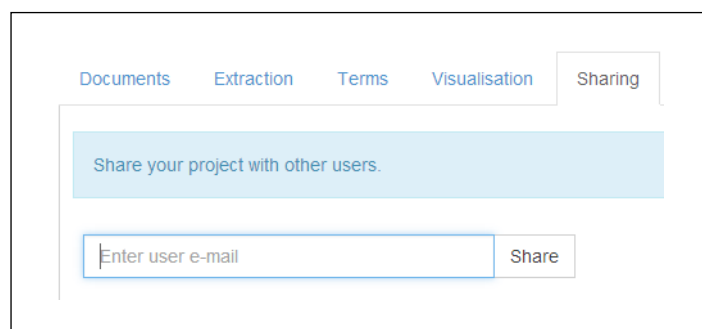
TaaS provides facilities for project sharing among users if they work in a team. This functionality that typically involves an interchange of non-confidential, non-competing, and non-differentiating terminology across various actors is highly rated by users. Recent surveys have shown that up to 60% of terminology resource users would share their resources with the community. The concept of sharing, unfortunately, is not present in the current management of major terminology databases and term banks. Instead of providing the opportunity for users to contribute their data, major term banks typically keep to the traditional one-way communication of their high-quality pre-selected terminological data.

To share his/her project with other users, the user has to add their e-mails and assign their roles. There are three available roles to a new user of the shared project: administrator, with full access rights; editor, with limited access to editing rights; and reader, with limited access to reading rights. One project can have more than one administrator; however, the owner of the project (the user, who has created the project) must consider assigning the administrator’s role to other users of his/her project as these users will get full access, including the right to delete the project and its terminology collection. The Administrator’s role is usually assigned to the project manager in the translation team, who adds documents to the project, and these are later processed by a terminologist, translator(s), editor(s), and other translation team members (see Figure 4).

The main usage scenario for the TaaS services is when the user uploads his/her document(s) under the created project, in order to then execute the terminology processing. TaaS supports user document upload in more than 10 formats including the most widely used MS Word, Excel, and Power Point formats as well as the Portable Document Format (PDF), the XML Localisation Interchange File

Format (XLIFF), and others. The open Beta version has certain limitations in terms of file and project size.

The terminology extraction service performs automatic extraction of monolingual term candidates from user- uploaded documents using generic or language specific terminology extraction techniques.



**Figure 4: Project sharing in TaaS.**

The user can customise the terminology extraction process. He/she can select one or more available (on the platform) terminology extraction tools for term candidate identification in user-uploaded documents. There are two term identification tools integrated into TaaS at the moment. These are the Tilde Wrapper System for CollTerm (TWSC)<sup>6</sup> that includes language specific patterns and morphological analysis and Kilgray Term Extractor that applies generic statistical approach to all supported languages. It is recommended to select the first tool; however, the statistical tools might also be of help in certain cases, for example, when linguistic processing produces insufficient results.

The platform provides the service for automatic retrieval of translation equivalents (for the extracted monolingual term candidates) in user-defined target language from different public and industry terminology databases.

The following terminology resources are available for translation equivalent lookup for term candidates identified in user-uploaded documents:

- TaaS public collections shared by other TaaS users;
- Terminology collections owned by the user;
- EuroTermBank;
- Inter-Active Terminology for Europe (IATE), an inter-institutional terminology database of the European Union<sup>7</sup>;
- TAUS Data that stores shared translation memories;

<sup>6</sup> See the ACCURAT Toolkit 3.0 at [www accurat-project.eu](http://www accurat-project.eu).

<sup>7</sup> <http://iate.europa.eu/>

TaaS database of raw bilingual terminological data automatically extracted from original and translated texts (a.k.a. comparable and parallel corpora) on the Web.

In the dynamic pace of technological developments and societal changes, new terms are coined every day by industry, translation and/or localisation agencies, collective and individual authors. Although these terms can be found in different online and offline publications, the inclusion of new terms in online public terminology databases and term banks takes months or even years, if it happens at all. As a result, terminology databases and term banks fail to provide users with extensive up-to-date multilingual terminology, especially for terms in under-resourced languages or specific domains that are poorly represented in online public terminology resources.

At the same time many new terms and their translations can be found on the Web – in multilingual websites, online documents, support pages, etc. TaaS provides four bilingual terminology extraction workflows for Web data: one workflow for terminology extraction from parallel data and three workflows from comparable data. The latter three are customised to collect terms from comparable news corpora, from multilingual Wikipedia, and from focused comparable corpora, respectively.

Web data are collected and then automatically processed. As a result, a list of bilingual raw term candidate pairs are extracted and fed into the TaaS terminology repository. During the execution of a terminology project at the translation candidate lookup step, these data are retrieved and proposed to the user for his/her validation. Thus the TaaS aligns the speed of terminology resource acquisition with the speed at which the content is created by mining new terms directly from the Web.

The data collection process is ongoing constantly feeding the TaaS repository with new terms. By April 2014, the TaaS database included more than 8 M bilingual term pairs extracted from the Web data. TaaS provides facilities for cleaning up raw terminological data extracted automatically that is noisy and needs validation by users. The process of validation can be regarded as a three-step procedure:

- monolingual validation (deletion of “unwanted” and/or unreliable term candidates, definition of termhood, term variant identification, deduplication, deletion of “incorrect” extraction, for example, a part of a longer noun group, synonym identification etc.);
- bilingual validation (bilingual checking of term candidates and their translation candidates, defining the right translation for the source term, deletion of irrelevant and/or incorrect translations, etc.);
- validation in context.

As soon as extraction finishes, the user can see extracted terms from his/her documents and their translation equivalents retrieved by TaaS. The user can hover over terms to get additional information, such as grammar, source, and context (see Figure 5).



Figure 5: Clean-up and validation of raw terminological data in TaaS.

The user can approve terms with a single click and add translations him-/herself, if the right translation from proposed translation candidates is not found.

An extracted term with its translation equivalent(s) forms a terminology entry. For advanced purposes, the user might want to edit a term entry in full entry view using the term entry editor and to add additional information about terms, for example, definitions, notes, grammatical information, and usage properties, such as term type, register, administrative status, temporal qualifier, geographical usage, and frequency. The history of editing is saved and is seen in the full entry view. The user might also want to see term candidates identified by TaaS in his/her documents, and the visualisation functionality is available for this purpose (see Figure 6).

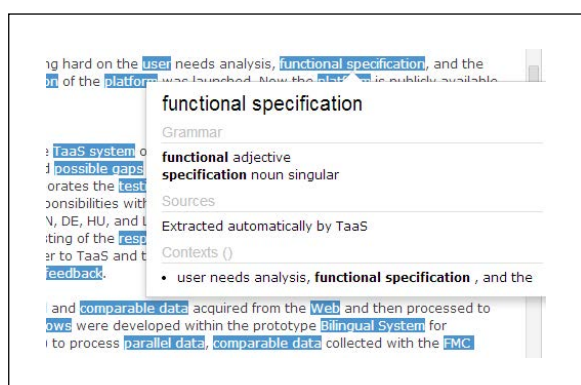


Figure 6: Visualisation of term candidates in the user's document in TaaS.

Validated terminological data can be exported and then reused in other working environments. During the analysis of user needs and requirements, we also proved our hypothesis that terminology, as a language resource, is central for the second large group of users – language applications (the first user group is represented by language workers). We have already performed first successful experiments on the integration of terminological data acquired within TaaS into the statistical MT system. At the time, the memoQ CAT tool<sup>8</sup> owned and developed by Kilgray, the TaaS project partner, is being integrated with TaaS via the TaaS API developed in the project and available for machine users.

8 See the description at <http://kilgray.com/products/memoq>.



TaaS demonstrates the efficacy of its terminology services within the following practical usage scenarios:

- For language workers: simplification of processing, storage, sharing, and application of task-specific multilingual terminology.
- For computer-assisted translation (CAT) tools: instant access to term candidates and translation equivalent candidates via the TaaS API.
- For statistical machine translation (SMT) systems: support for domain adaptation by a dynamic integration with TaaS-provided terminological data via the TaaS API.

At EURALEX the usage scenario for language workers with the emphasis on specialised lexicography work will be demonstrated online.<sup>9</sup>

### 3 Conclusion

In this software demonstration paper we have presented an innovative cloud-based platform TaaS “Terminology as a Service” developed in an EU-funded project. At the present time, TaaS is a unique dynamic cloud-based solution that provides a wide range of terminology services. We foresee the potential of the established platform for a wide range of user groups, both language workers and language applications. Among other benefits for language workers, TaaS serves the needs of specialised lexicography, or terminography, facilitating user-friendly, collaborative, multilingual, interoperable, portable, and cloud-based specialised terminology work. TaaS fills the gap of innovative environment to speed up the development of specialised dictionaries. This opportunity is critical particularly in emerging domains. At the EURALEX Congress the platform is demonstrated in real-time during the three days of the event.

#### **Acknowledgements**

Research within the TaaS project, leading to these results, has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 296312.

The TaaS platform is a result of fruitful collaborative work of the project partners – coordinator and lead developer Tilde (Latvia), research partners Cologne University of Applied Sciences (Germany) and University of Sheffield (UK), industry partners Kilgray (Hungary) and TAUS (Netherlands).

---

9 Live demonstration requires Internet access.



# Good Contexts for Translators—A First Account of the Cristal Project

Amélie Josselin-Leray\*, Cécile Fabre\*, Josette Rebeyrolle\*, Aurélie Picton\*\*, Emmanuel Planas\*\*\*

\*CLLE-ERSS, University of Toulouse & CNRS, France; \*\*FTI, University of Geneva, Switzerland;

\*\*\*LINA, University of Nantes, France

josselin@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr, rebeyrol@univ-tlse2.fr,

aurelie.picton@unige.ch, emmanuel.planas@univ-nantes.fr

## Abstract

This paper questions the notion of good contexts for translators and describes an experiment which tests the usefulness of two specific kinds of contexts in a translation task, namely (1) contexts that provide conceptual information about a term, and (2) contexts that provide linguistic information about the collocational profile of this term. In the experiment, trainee translators are asked to use several types of resources, including a set of pre-annotated contexts of various types, and to identify the contexts that they consider to be the most relevant for their task. We present the first results of this experiment, which confirm our general assumption about the usefulness of such rich contexts and indicate some differences regarding the use of contexts in the source and target language. This study takes place in the CRISTAL project whose aim is to retrieve from bilingual comparable corpora the contexts that are the most relevant for translation and to provide them to users through a CAT tool.

**Keywords:** CAT tools; corpus resources for translators; Knowledge-Rich Contexts

## 1 Introduction

Even though it is widely acknowledged as being essential to the translator, the very idea of context in translation is hard to define (Baker 2006: 321) and it also “lacks a definition that can be applied in the everyday work of a professional translator” as stated by Melby & Foster (2010: 1). Therefore, when one wants to provide translators with tools that better meet their needs—such as improved CAT tools—, one should in the first place wonder about what makes a context relevant for them. In other words, what is a ‘good context’ for translators? This is one of the questions the CRISTAL project<sup>1</sup> tries to give an answer to. The main aim of the CRISTAL project, an acronym that stands for “Knowledge-Rich Contexts for Terminological Translation” (“Contextes Riches en Connaissances pour la Traduction Termi-

---

1 CRISTAL is a three-year project (2012-2015) funded by the French National Agency for Research (ANR-12-CORD-0020). It involves four partners: a computing research team at the University of Nantes, France (LINA), a linguistics research team at the University of Toulouse, France (CLLE-ERSS), the Translation Technologies team from the Faculty of Interpreting and Translation at the University of Geneva in Switzerland, and a firm specializing in multilingual text management (Lingua et Machina).

nologique” in French), is to automatically retrieve from bilingual comparable corpora the contexts that are the most relevant for translation and to provide them to users through the CAT tool developed by Lingua & Machina, the Libellex Platform.

The first part of this paper reviews the translators’ needs regarding context. It seems necessary to first identify which type of information is essential for translators, to see how this information is recorded in the tools most commonly used by translators, i.e. dictionaries, term banks and corpora, and how satisfied translators are about the way that information is recorded. In order to refine the notion of “good contexts” for translators, in part 3 we investigate what a “good example” in lexicography and what a “Knowledge-Rich Context” in terminology are, and introduce the distinction between conceptually-rich and linguistically-rich concepts. Part 4 then focuses on one aspect of the methodology of the CRISTAL project: an experimentation involving trainee translators in order to refine our idea of a “good context for translators”. Finally, part 4 presents the very first results of the experiment.

## 2 Some Facts about the Needs of Translators Regarding Context

As stated by Rogers & Ahmad (1998: 195), “one of the translator’s prime needs is for context-sensitive information”. We may wonder what the notion of context-sensitive information encompasses and what sources of information translators can rely on—or not.

### 2.1 What do Translators need Contextual Information for and Where do they Find it?

#### 2.1.1 Context in Translation: a Preliminary Definition

As thoroughly explained by Melby & Foster (2006), specialists in many fields (e.g. philosophy, psychology, pragmatics, and functional linguistic) have discussed the notion of context, and various definitions have been written. The three facets of context as defined by Halliday (1999), i.e. *context of situation*, *context of culture* and *co-text* are all particularly relevant in translation. However, in this paper, we will only focus on what Halliday calls *co-text*. While both *context of situation* and *context of culture* are outside of language itself, *co-text* specifically pertains to language in use. It can broadly be defined as the surrounding discourse of an utterance. Therefore, our definition of “context” in this paper will be limited to *co-text*, and will rely on the definition provided by Fuchs<sup>2</sup>:

What is called *context* is the linguistic environment of an element (phonetic unit, word or group of words) within an utterances; i.e. the units that precede and follow it. Thus, in the utterance “Marie est jolie comme un cœur”, the element *comme* has as its immediate context “jolie...un cœur” and its

---

2 <http://www.universalis.fr/encyclopedie/concept/>. Consulted on April 2, 2014.

wider context “Marie est jolie...un cœur”. By extension, the word *context* is also applied to the utterance(s) which precede(s) and follow(s) a given utterance in discourse.<sup>3</sup>

### 2.1.2 What do Translators Need Context for?

Following Roberts & Bosse-Andrieu (2006: 203), let us remind here that the translation problems translators have to face can be considered as source text-related (for comprehension of the source-text) or target-text related (for transfer into the target text) and classified into three main categories: *encyclopedic*, *linguistic* or *textual*. *Encyclopedic* problems encompass “general subject-related problems as well as more specific problems dealing with proper nouns—that is, a lack of familiarity with the topic of the text or with specific places or people mentioned in the text”; *linguistic problems* are defined as “those attached to specific words and phrases—that is, problems related to the comprehension or translation of a given word or phrase”; finally, *textual* problems are those concerned with text types and the internal organization or reproduction of a given text type.

Bowker (2011, 2012) draws a list of those items of contextual information that can prove “useful” for the translator to solve his source-text and target-text-related problems. They can be summed up as follows: (i) information about usage; this of course includes collocations, in particular which general-language words collocate with terms (see also Roberts 1994: 56), (ii) information about the frequency of use of a particular word or term, (iii) information about lexical and conceptual relations (such as synonymy, meronymy, hyperonymy etc.) (see also Marshman, Gariépy & Harms 2012, Rogers & Ahmad 1998), (iv) pragmatic information about style, register and genre (see also Varantola 1998), (v) information about usages to avoid. It thus seems to us that the items that are not situation-linked fall into the following categories: *conceptual* information and *linguistic* information.

To solve those problems and to make decisions, translators need external help, which they typically get by consulting other human experts and conventional resources such as dictionaries and term banks (Rogers & Ahmad 1998: 198).

### 2.1.3 Where do Translators Find Contextual Information?

As mentioned by Varantola (2006: 216), “the translator’s problem-solving techniques have changed dramatically over the past decade or so”. In addition to the above-mentioned conventional resources (monolingual and bilingual dictionaries—which have undergone radical changes—; term banks), translators now also partly rely on the information provided by corpora.

- Dictionaries

It is mostly through examples that dictionaries provide contextual information. The empirical study on scientific and technical words (i.e. terms) in general bilingual and monolingual dictionaries carried out by Josselin-Leray (2005) has shown that up to 80.3% of users turned to dictionaries to find information about how to use the term in a sentence and that bilingual dictionaries always

---

3 We translated the quotation.

ranked higher in that respect. Among the respondents who chose that answer, it was the « language professionals » user group (which includes translators) that was mostly represented.

- Term Banks

Term banks typically provide contextual information through the “context” section of the terminological record. The importance given to contextual information in terminological resources by translators is confirmed by the findings of the survey by Duran-Muñoz (2010): examples were considered to be “essential data” by the respondents, and among “desirable data”, one found “a greater variety of examples” and “semantic information (semantic relations, frames)”.

- Corpora

Although corpus data is obviously intrinsically made of contextual information, it is a resource which seems still quite scarcely used by translators, as shown by the results of the survey by Duran-Muñoz (2010): only 5.09% of the participants quoted (parallel) corpora as being a terminological resource they “used *more*<sup>4</sup> when translating”. However, 41.8% of the respondents to the Mellange Survey<sup>5</sup>, which was carried out in 2005-2006 among trainee translators and professional translators, do claim they use corpora in their translation practice (the most frequent type being the corpora in the target language).

Although there seems to be a wide array of resources translators can turn to when they need contextual information, these resources do not necessarily meet the translators’ needs.

## 2.2 The Shortcomings of Existing Resources regarding Context

### 2.2.1 The Dissatisfaction of Translators regarding Contextual Information in Existing Resources: a Hard Fact

We found it relevant to first look at the findings of various empirical surveys on the use of conventional resources by translators.

- Dictionary Use

Before starting to compile the *Bilingual Canadian Dictionary*, Roberts (1994: 56) carried out a survey among its potential users in order to clearly identify their needs and reached the following conclusion: “Between one-third and one-half the members in each user group [of sophisticated second language users] appreciated, to varying degrees, the number of examples presented in their present most frequently used dictionaries. But between one quarter and one half of each group felt that improvement was needed in that respect”. The study by Josselin-Leray (2005) reached the same conclusion: although users were overall satisfied by the examples provided by their dictionaries (between 41.3% and 67.5% of users said they were satisfied), the level of satisfaction was lower for bilingual dictionaries, and lower among the “language professionals” user group.

---

4 We added the italics.

5 <http://mellange.eila.jussieu.fr/Mellange-Results-1.pdf>. Consulted on April 10, 2014.

- Term Banks

In the survey by Duran-Muñoz (2010), translators also had the opportunity to give their opinions regarding their needs; the second most repeated argument was “include more pragmatic information about usage and tricky translations”, and the fifth one was “provide examples taken from real texts”. The conclusion of her findings, found in Duran-Muñoz (2012: 82) is straightforward: “we can confirm that most of the terminological resources that are currently available (especially in electronic format) do not fulfill their requirements”. Why is that so?

### 2.2.2 Why are Translators Dissatisfied?

Varantola (e.g. 1994) has written at great length about the *context-free* vs. *context-bound* dilemma faced by translators: dictionaries and term banks only provide context-free examples, i.e. examples that are perceived as prototypical and frequent, while what the translators need to find the suitable equivalent(s) is typically context-bound. Moreover, the examples/contexts provided are not varied enough. This is especially true of term banks, as underlined by Bowker (2011: 214-215) who explains the information found on those records is rather limited and usually consists in definitions and terms presented out of context, or in only a single context. She pinpoints a paradoxical situation in which the advances of research on terminology (especially the work on Knowledge-Rich Contexts, which we will introduce in 3.2) have not been integrated into the tools translators most commonly use, i.e. term banks.

However, Varantola (2006: 217) says the context-free/context-bound dilemma should now be qualified since “context-free definitions of concepts within a particular domain [which] were for a long time the theoretical ideal in terminological theory [...] are now replaced by less rigid, contextually relevant definitions”. She ascribes it to the availability of large corpora, whose role is also now central in dictionary-compiling. Some dictionaries now even “provide access to more examples in the form of concordances from the corpus data that lie behind the dictionary”. Corpora are no panacea, though. One of arguments against corpora is that they are “tools of shallow intelligence” (Varantola, 2006: 223) when they are raw and non-tagged or POS-tagged, since the user “is left to handle the manipulation, dissection and interpretation of results”. In other words, compiling the corpus and analysing the corpus can be too tedious a task for translators who often work under tight constraints<sup>6</sup>.

---

6 Another point worth mentioning is that, in the survey carried out for the Mellange, even though 94.4% of the respondents said they used Google to research terminology, 10.2% found that Google was limited for finding information on language use because the “search results [did] not provide enough context to be useful”.

## 2.3 The Translator's Ideal Workstation?

In 1996, Atkins already suggested (p. 526) that the dictionary of the future should “give its users the opportunity to make their own decisions about equivalences” : the users “should be able to consult as many examples as they need of words used in their various senses, each in a variety of contexts with a variety of collocative partners”. More recently, Bowker (2011: 215) suggested: “it would be more helpful for translators to have access not simply to term records that provide a single ‘best’ term with a solitary context, but rather to information that would allow them to see all possible terms in a range of contexts and thus find the solution that works best in the target text at hand”. She insists on the fact that looking at a wide range of contexts should not be considered as a waste of time, and that this has been made easier thanks to corpus-analysis tools that present information in an easy-to-read format. She goes even further by suggesting (Bowker 2012: 391) that translators have access to the whole of the information that lexicographers usually rely on when devising a dictionary entry:

In order to arrive to that entry, lexicographers have gone through a number of intermediary steps, where they learn about the various characteristics of the words and concepts being described, such as their grammatical and collocational behaviours, the different relationships that hold between words and their underlying concepts, and the characteristics that are necessary and sufficient for distinguishing one concept in an intensional definition.

However relevant that objective might be, it seems rather ambitious and difficult to achieve in the very near future, all the more so as “lexicographers’ needs are very different from translators’ corpus needs” (Varantola 2006: 217). Narrowing down that objective to providing more corpus-based context data in a way that is more in keeping with the actual working conditions of translators seems more feasible, which is why the main aim of the CRISTAL project is to help design a CAT tool that provides translators with customized contexts automatically retrieved from comparable corpora.

In order to reach that goal—which can also sound ambitious, we first decided to refine the notion of ‘good contexts for translators’ by doing two things: (i) we first looked at the way lexicographers deal with examples in dictionaries and terminographers deal with “Knowledge-Rich Contexts” (part 3), (ii) we devised an experiment with trainee translators focusing on what we thought to be “good contexts” (part 4).

## 3 Dictionary Examples and Knowledge-Rich Contexts

The thoughts of lexicographers and terminologists on “good examples” or “Knowledge Rich Contexts” provide some valuable insight into what a good context for translators might be. After examining those two aspects, we give our own definition of Conceptually Rich Contexts and Linguistically Rich Contexts.



### 3.1 Good Dictionary Examples

Many studies have underlined the importance of the illustrative component in dictionaries as a means to provide typical contexts about a word's meaning and usage (Atkins & Rundell 2008). In monolingual as well as in bilingual dictionaries, examples are meant to help the dictionary user both in the production and the comprehension process. They have therefore diverse functions (Rey-Debove 2005, Roberts 1994, Siepmann 2005): they can provide syntagmatic information about word patterns and collocations, together with paradigmatic information about words that are semantically-related (synonyms, hyperonyms, etc.). They may also give pragmatic and stylistic indications about registers and specific uses, or be used as a more concrete and accessible complement to definitions, with an epilinguistic dimension.

Authentic examples that meet at least some of these requirements are very difficult to extract from corpora:

Finding good examples in a mass of corpus data is labour-intensive. For all sorts of reasons, a majority of corpus sentences will not be suitable as they stand, so the lexicographer must either search out the best ones or modify corpus sentences which are promising but in some way flawed (Rundell & Kilgariff 2011).

Kilgariff et al. (2008) have developed a method to automatically collect sentences that are good candidate dictionary examples, using two criteria: readability (judged from sentence length and average word length; it penalizes sentences with infrequent words, more than one or two non-a-z characters, or anaphora) and informativeness (judged from the density of collocates in the sentence).

### 3.2 Knowledge Rich Contexts (KRCs) in Terminology

Knowledge Rich Contexts play a very important role in identifying terms in specialized texts because they show conceptual relationships between terms. It is within that framework that Ingrid Meyer defined Knowledge Rich Contexts as “a context indicating at least one item of domain knowledge that could be useful for conceptual analysis.” (Meyer 2001: 281). These contexts are used in order to develop knowledge extraction tools for text-based terminology and ontology building (Condamines & Rebeyrolle 2001). Rich contexts for terminologists typically contain terms that are specific to the domain together with linguistic patterns that signal the conceptual relations between these terms as illustrated below in Meyer (2001):

- (1) Compost is an organic material deliberately assembled for fast decomposition.
- (2) Compost contains nutrients, nitrogen, potassium and phosphorus.

This type of information helps building networks of terms, generally focusing on hyperonyms (example 1) and meronyms (example 2).

### 3.3 KRCs for Translators: Conceptually vs. Linguistically Rich Contexts

Based on the type of information needed by translators as described in 2.1.2, and the type of information provided by dictionary examples and Knowledge-Rich Contexts as detailed in 3.1 and 3.2, we decided to extend the notion of KRC in our experiment, considering under this category two types of contexts: those that provide ‘conceptual’ information about a given term—called “Conceptually Rich Contexts” (CRCs) in our study, and those that provide ‘linguistic’ information about that term “Linguistically Rich Contexts” (LRC). In our experiment, the contexts which are neither conceptual nor linguistic are considered as *poor* (cf. Reimerink *et al.* 2010: 1934).

## 4 An Experiment Centered on “Good Contexts” within the CRISTAL Project

The main aims of the experiment were (i) to check that rich contexts extracted from corpora are useful to translators, (ii) to identify which types of rich contexts (CRCs or LRCs) are the most useful to them.

A pilot study was carried out in December 2013 at the University of Geneva (Switzerland). This allowed us to test the protocol and to make the necessary adjustments for the two experiments we conducted in March 2014: one at the Université Catholique de l’Ouest (Angers, France), and one at the University of Toulouse le Mirail (Toulouse, France). We will now describe the main aspects of the protocol designed for the experimentation.

### 4.1 Protocol

#### 4.1.1 Participants

For both the pilot study and the two experiments, participants were all trainee translators<sup>7</sup>. 7 students from the Faculty of Translation and Interpretation of the University of Geneva took part in the pilot study. There were 4 Masters’ students and 3 PhD students. As for the experiments in Angers and Toulouse, the participants (42 in total) were students in their final year of a Master’s Translation program.

#### 4.1.2 Translation Task

The participants were asked to translate a text from English into French (i.e. from L2 into L1 for most students). The text is around 150 words long, it is a popular-science text on volcanology entitled “Cinder Cones”<sup>8</sup>. It was chosen because it is well-structured (the two phases of the building of a cinder

---

7 This is the case in many empirical studies on translation: see for example Bowker 1998, Künzli 2001, and Varantola 1998.

8 It was taken from *What’s so hot about volcanoes?* by Wendell A. Duffield (2011), Mountain Press.

cone are described), because it contains a certain number of terms whose translation might be complex for a translator, even if they are not highly specialized (*basalt cinder cone, fountaining stage...*), and a number of syntactic patterns or collocations that are particularly tricky to translate (e.g. *bubble off*; transitive use of the verb *erupt*). Only one group out of the two was already a little familiar with the field of volcanology. The participants were allocated around 2 hours to translate the text, choose the relevant contexts and fill out an online questionnaire about the main translation difficulties and the use and usefulness of conventional resources and KRCs. Then several group interviews and a couple of one-to-one interviews were conducted.

### 4.1.3 Resources

Since we wanted the conditions of the experiment to be as close to a real-life context as possible for translators,<sup>9</sup> the participants had at their disposal the same type of resources as the ones they usually have when they translate in professional environment, i.e. various dictionaries and term banks. What made the experiment specific is that we added an extra resource, i.e. shortlisted contexts.

### 4.1.4 The Argos Interface

The participants used a customized interface, Argos, with four different windows (see figure 1): (i) one for the source-text, (ii) one for the target-text, (iii) one with several icons allowing access to term banks (*Termium, le Grand Dictionnaire Terminologique*), a specialized bilingual dictionary of earth sciences, a general bilingual dictionary, (iv) one with a list of shortlisted contexts.

---

9 This is what Ehrensberger-Dow & Massey (2008) call “ecological validity”.

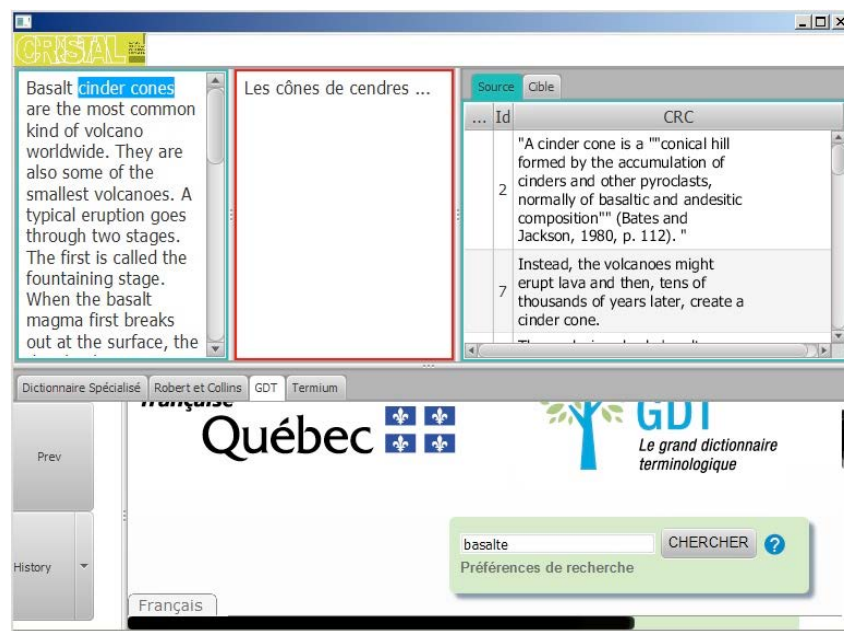


Figure 1: the Argos interface

#### 4.1.5 The Contexts

Participants were provided with contexts in the source language, and contexts in the target language, which were presented in random order. These contexts had been carefully chosen beforehand according to a classification devised by the team of linguists: details about the selection of contexts and the type of contexts provided will be discussed in the next section (4.2). During the translation process, the participants had to choose the contexts that had been most useful to them when translating by clicking on them. Once the translator had typed in a word (in the source language or the target language) in the search window for contexts, the target text window was blocked in order to ensure the translator chose at least one context, or explicitly chose that none was useful.

#### 4.1.6 Extra Data Compiled

In addition to the final translations themselves and the answers to the online questionnaires, the data compiled comprise the following:

- screen-recordings performed through specific software<sup>10</sup>
- logs: all keyboard activity, as well as change of windows shifts, was recorded
- audio recordings of the two types of interviews<sup>11</sup> during which the participants were asked to give more detail about the usefulness of some given contexts.

<sup>10</sup> BBFlashback Express.

<sup>11</sup> In the one-to-one interview, the participant viewed part of the screen recording (the extracts corresponding to his translation of two terms for which contexts were provided: *cinder* and *fountaining stage*) and was asked to verbalize what he was doing, following the methodology used by Ehrenbersger-Dow and Massey (2008).

## 4.2 The List of Contexts

The list of contexts provided to the participants was created with two principles in mind: we wanted to compile a large enough set of contexts, in order to limit the chance that the translator would look for contextual information on one particular word and get no results. At the same time, considering that the selection of contexts is a labor-intensive task, we wanted to limit ourselves to a reasonable number of words, and to keep only words for which our definition of Rich and Poor Contexts applies (see 3.3 above): in particular, the notion of Conceptually Rich Contexts (CRCs) is irrelevant for very familiar words that are not characteristic of the field of volcanology from which the text is drawn. This compromise was difficult to reach, so we took advantage of the pilot study to adjust and complement the list of terms that we had initially created.

### 4.2.1 Term Selection

For the pilot study in Geneva, we compiled a first set of contexts illustrating the use of:

- 7 words (noun, verbs and adjectives) in the source language, namely some lemmas from the text to be translated that we regarded as terminological units, or at least as words related to the field of volcanology (*basalt magma, blobs, cinder, cinder cone, fountaining, scoria, vesicles*)
- 11 words in the target language, selected among the possible equivalents of the corresponding source words.

We considered both simple and multiword units.

For example, contexts are provided for the word in bold type in the following sentence:

(3) As the **cinders** fall back to the Earth, they form layers that pile up into a cone-shaped hill.

One outcome of the pilot study was that the initial list of words proved to be very insufficient, especially in the target language: the logs compiled thanks to Argos (cf. 4.1.6) provided us with a much larger list corresponding to words that had actually been typed in the search window by the participants in order to get contexts. We thus decided to complement the first list with words that have been searched for by at least 2 participants. The final list for the experiments in Angers and Toulouse contains contexts for 22 English words and 41 French words.

In the same sentence as the one mentioned above, contexts were provided for four words (in bold type) instead of just one (i.e. 2 nouns, 1 verb, 1 adjective):

(4) As the **cinders** fall back to the Earth, they form **layers** that **pile up** into a **cone-shaped** hill.

### 4.2.2 Context Selection

The contexts were selected from several sources:

- we preferentially used a 800,000 word corpus of volcanology composed of specialized and popular science texts, which is used for the CRISTAL project as a whole,
- this source was complemented by a variety of web sites, giving priority where possible to texts dedicated to the presentation of volcanology to a wide audience.

We chose not to test the readability dimension of contexts (3.1.), so we only selected contexts that meet the criteria of readability (well-formed, not too long, with no anaphora elements, etc.). Contexts are one or two sentences long.

As explained before, our aim is to test whether the opposition between rich and poor contexts as defined in section 3 is relevant for the translation task. As a consequence, we annotated contexts according to this dimension. In figure 1, we give 3 examples illustrating (1) a linguistically rich context for the word *basalt* (with the presence of the term *basalt lava*), (2) a conceptually rich context providing a definition of the term, (3) a poor context. Note that linguistic and conceptual richness can combine in some contexts, which is not the case here. When possible, the conceptually rich contexts were classified into the following subcategories: definition, meronymy, hyponymy and co-hyponymy.

	Term	Rich or Poor	CRC	LRC	Type of CRC	Context
1	basalt	rich	no	yes	n/a	Shield volcanoes are made of thousands of thin basalt lava flows.
2	basalt	rich	yes	yes	def	Basalt is dark volcanic rock made up of small crystals and glass.
3	basalt	poor	no	no	n/a	When basalt enters water passively, it forms pillow basalt.

**Figure 5: Examples of contexts for the word *basalt*.**

We intended to balance the number of poor and rich contexts for each word. This proved impossible to achieve in many cases, since the great majority of contexts exhibit at least one relevant collocate. We collected 10 contexts per word, including no less than 2 poor contexts, totalling 222 contexts in English and 441 contexts in French.

### 4.3 First Results

We report here some preliminary observations about the results.

For the source language, 48% of the available contexts were chosen by at least one participant (108 contexts) versus 36% for the target language (152 contexts). This is an indication that the contexts are perceived as helpful, but the data are very dispersed, since about 40% of the selected contexts were chosen by only one participant in either language. Some terms are found several times in this list (fountain, cinder cone, basalt magma, and their French counterpart), showing specific translation problems.

To get a first picture of the results, we have chosen to focus on the 20 contexts that were selected most in either language. Each context was chosen between 4 and 14 times. The following examples show two of the most frequent ones.

(5) Hawaiian Eruptions are types of volcanoes and types of eruptions wherein basaltic lava is normally thrown up the air in jets. This process is called **fountaining**.

(6) During an eruption of gas-rich magma, small **blobs** of magma are ejected.

Example 5 is a conceptually rich context, more specifically a definition. Example 6, which contains several collocations (*blobs of magma, blobs ejected*), is a linguistically-rich context.

	Rich contexts	CRCs (definitions)	LRCs
<b>Source language</b>			
All available contexts	69%	25% (11%)	52%
20 most selected contexts	90%	65% (60%)	25%
<b>Target language</b>			
All available contexts	70%	29% (14%)	49.5%
20 most selected contexts	90%	55% (40%)	35%

**Table 2: Distribution of the contexts.**

Table 2 makes a comparison between this subset of contexts and all the contexts that were made available. First, this shows that the great majority of contexts that are considered as helpful by the participants are rich contexts (90%). Second, participants show a strong preference for conceptually-rich contexts, mainly definitions, as opposed to linguistically-rich contexts. Yet we can observe that if this overall pattern applies to both languages, there are some differences: the distribution between LRCs and CRCs is different when the users are exploring source (English) and target (French) contexts. They seem to give a higher priority to CRCs and definitions in the source language. This is consistent with the idea that the CRCs should provide help for the comprehension of terms and LRCs should be more useful when checking the usage of the words in the target language.

These are encouraging results: they confirm our assumption that rich contexts are seen as helpful by the participants, and they suggest differences in the way the translator uses these contexts in the source and target language. However, these first observations must be confirmed and complemented by the analysis of the entire set of data and the analysis of the replies to the questionnaires.

## 5 Conclusion

The notion of context is where lexicography, terminology and translation meet. Even though the specific needs of translators regarding their resources now seem quite clearly identified, addressing them still seems quite challenging. We hope the findings of the CRISTAL project will help tailor the tools according to the translator's profiles in one aspect, that of contextual information.

To fulfil that objective, we plan to explore in detail the considerable amount of data we have collected (around 60 hours of screen recordings, and just as much structured translation logs). The evidence gathered will enable us to answer the following questions:

- apart from definitions, do some sub-categories of KRCs play a specific role in the translation process?
- in which precise situations do translators give preference to KRCs over conventional resources such as monolingual or bilingual dictionaries?
- what is the impact of the use of KRCs and the other resources on the translation quality of the 49 final translations

The main challenge the CRISTAL project plans to address in the near future is to devise a method to automatically retrieve the ‘good contexts’ whose main features will have then been identified.

## 6 References

- Atkins, Beryl T. Sue. (1996). “Bilingual Dictionaries: Past, Present and Future.” In: Euralex’96 Proceedings, Martin Gellerstam et al., ed. Gothenburg: Gothenburg University. 515-590.
- Baker, M. (2006). Contextualization in translator- and interpreter-mediated events. *Journal of Pragmatics*, 38(3), pp. 323-337.
- Bowker, L., (1998). « Using Specialized Monolingual Native-Language Corpora as a Translation Resource: a Pilot Study ». *Meta*, 43(4), pp. 631-651.
- Bowker L., (2011). “Off the record and on the fly,” *Corpus-based Translation Studies: Research and Applications* (Eds. A. Kruger, K. Wallmach and J. Munday). London/New York: Continuum, pp. 211-236.
- Bowker L., (2012). “Meeting the needs of translators in the age of e-lexicography: Exploring the possibilities”, in S. Granger, M. Paquot (eds) *Electronic Lexicography*, Oxford University Press, pp. 379-387.
- Condamines, A., Rebeyrolle, J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results. In D. Bourigault, C. Jacquemin & M.-C. L’Homme *Recent Advances in Computational Terminology*, John Benjamins, pp. 127-148.
- Durán Muñoz, I. (2010). Specialised lexicographical Resources: a Survey of Translators’ needs. In S. Granger & M. Paquot (eds.), *eLexicography in the 21st century: New challenges, new applications, Proceedings of Elex2009. Cahiers du Cental*, vol. 7. Louvain-la-Neuve. Presses Universitaires de Louvain, pp. 55-66.
- Durán Muñoz, I. (2012). Meeting translators’ needs: translation-oriented terminological management and applications. *Journal of Specialised Translation* (18), pp. 77-92.
- Ehrensberger-Dow, M., Massey, G. (2008): Exploring Translation Competence by Triangulating Empirical Data. In *Studies in Translation*. 16, 1-20.
- Halliday, M.A.K. (1999). “The notion of Context in Language education). In M. Ghadessy (ed.) *Text and Context in Functional Linguistics*. Philadelphia: John Benjamins Publishing Company.
- Josselin-Leray, A. (2005). Place et rôle des terminologies des terminologies dans les dictionnaires unilingues et bilingues. Etude d’un domaine de spécialité : volcanologie. Unpublished PhD thesis, Université Lyon 2, France.



- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P., (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *proceedings of XIII EURALEX Congress*, E. Bernal & J. DeCesaris (Eds), Barcelona, Universitat Pompeu Fabra, pp.425-431.
- Künzli, A. (2001). Experts vs. novices. L'utilisation de sources d'information pendant le processus de traduction. In *Meta*, 46, pp. 507-523.
- Marshman, E., J., Gariépy & Harms, C. (2012). Helping translators manage terminological relations: Storing and using occurrences of terminological relations and lexical relation markers. *JoSTrans: Journal of Specialised Translation* 18, 30-56.
- Melby, A.K. (2010). Context in translation: Definition, access and teamwork. *Translation & Interpreting*, 2 (2), pp. 1-15.
- Meyer, I., (2001). Extracting Knowledge-Rich Contexts for terminography: A conceptual and methodological framework. In D. Bourigault, C. Jacquemin & M.-C. L'Homme *Recent Advances in Computational Terminology*, John Benjamins, pp. 279-302.
- Reimerink, A., Garcia de Quesada, M. & Montero-Martinez, S. (2010). Contextual information in terminological knowledge bases: a multimodal approach, *Journal of Pragmatics*, 42(7), pp. 1928-1950.
- Rey-Debove, J., (2005). Statut et fonction de l'exemple dans l'économie du dictionnaire. In M. Heinz (eds), *L'exemple lexicographique dans les dictionnaires français contemporains*. In Proceedings of the « Premières journées allemandes des dictionnaires » (Klingenberg am Main, 25-27 juin 2004). Max Niemeyer Verlag, Tübingen, pp.15-20
- Roberts, R. P. (1994). Bilingual Dictionaries Prepared in Terms of Translators' Needs, *Proceedings of CTIC 3rd Conference, Translation in the Global Village*, Banff, CTIC, pp. 51-65.
- Roberts, R.P. & Bosse-Andrieu J. (2006). Corpora and Translation. In Bowker L. (ed.), *Lexicography, Terminology, and Translation – Text-based Studies in Honour of Ingrid Meyer*. Ottawa : Presses de l'Université d'Ottawa, pp. 201-214.
- Rogers M. & Ahmad K. (1998). The Translator and the Dictionary: Beyond Words? In: B.T. Sue Atkins (ed.) *Using Dictionaries. Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Niemeyer (Lexicographica. Series Maior), pp.193-204.
- Siepman, D. (2005). Discourse markers across languages: A contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography. Routledge.
- Varantola, K. (1994). The Dictionary User as Decision Maker. Sixth Euralex conference proceedings, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands, pp. 606-611.
- Varantola, K. (1998). Translators and their Use of Dictionaries. In B.T.S. Atkins (ed), *User Needs and User Habits. Studies of Dictionary Use by Language learners and Translators*, Tübingen: Niemeyer (Lexicographica. Series Maior), pp. 179-192.
- Varantola, K. (2006). The Contextual Turn in Learning to Translate. In Bowker L. (ed.), *Lexicography, Terminology, and Translation – Text-based Studies in Honour of Ingrid Meyer*. Ottawa : Presses de l'Université d'Ottawa, pp. 215-226.



# Kontextbasierte lexikalische Kontrolle von Anforderungsdokumenten

Jennifer Krisch  
Daimler AG  
jennifer.krisch@daimler.com

## Abstract

Das Verwalten und Prüfen von Anforderungsdokumenten ist ein sehr wichtiger Bereich in den frühen Phasen von Industrieprojekten. Je früher unpräzise und unvollständige Anforderungen identifiziert werden, desto geringer sind die Folgekosten und die Überarbeitungszeit. Unpräzise und unvollständige Anforderungen resultieren häufig aus der Verwendung von Weak-Words, d.h. von unscharfen Wörtern oder Konstruktionen, die in bestimmten Kontexten Mehrdeutigkeiten auslösen. Um Fehlinterpretationen aufgrund von Weak-Words entgegenzuwirken, müssen diese Wörter oder Konstruktionen identifiziert werden. Eine reine Wortsuche reicht hierbei nicht aus, weil Weak-Words nur in bestimmten Satzkontexten Mehrdeutigkeiten auslösen. Der Beitrag beschreibt eine als Prototyp entwickelte automatisierte Weak-Word-Analysemethodik, durch welche ausgewählte Weak-Words identifiziert werden und bei der aufgrund des Satzkontexts entschieden wird, ob eine Warnung bezüglich eines Mangels an Präzision und Eindeutigkeit des Texts an den Autor zurückgegeben werden soll. Bei der Weak-Word-Analyse kommt ein Lexikon zum Einsatz, in welchem die Weak-Words selbst und deren Kontexte abgelegt sind. Die Evaluation des entwickelten Werkzeugs hat ergeben, dass die Anzahl der Warnungen, die an den Benutzer zurückgegeben werden, durch eine Kontextanalyse im Verhältnis zur wortbasierten Suche erheblich reduziert werden kann, ohne dass problematische Kontexte übersehen werden. Dies stellt eine Verbesserung gegenüber den bisher bekannten Systemen dar.

**Keywords:** Weak-Words; Kontext-Analyse; Linguistische Annotation

## 1 Einführung

Das Verfassen von Anforderungen ist in der Industrie der erste Schritt, um von einer Idee zu einem Produkt zu kommen. Die Ideen der Projektbeteiligten müssen angemessen dokumentiert werden, damit jeder weiß was im Projekt gefordert wird. Anforderungstexte sind also „Aussagen über Eigenschaften oder Leistungen eines Produktes, eines Prozesses oder der am Prozess beteiligten Personen“ (Rupp 2007).

Die Qualität einer Anforderung wird durch *Qualitätskriterien* festgelegt. Je mehr Qualitätskriterien von einer Anforderung erfüllt werden, desto höher ist deren Qualität. Anforderungen werden meist in

natürlicher Sprache verfasst, die aber das Risiko von Mehrdeutigkeit, mangelnder Präzision oder Unklarheiten über die rechtliche Verbindlichkeit birgt; daher ist es wichtig, Anforderungen vor allem auf die Qualitätskriterien *Präzision* und *Eindeutigkeit* zu prüfen. Eine Anforderung erfüllt das Qualitätskriterium *Präzision*, „wenn diese keine ungenauen Angaben enthält, wenn wo immer möglich und sinnvoll quantitative Angaben gemacht werden und wenn alle Angaben genauso präzise sind, wie es für die Problemstellung erforderlich ist“ (Daimler AG 2013). Das Qualitätskriterium *Präzision* hängt eng mit dem Qualitätskriterium *Eindeutigkeit* zusammen:

Eine eindeutige Anforderung [sollte] nur auf eine Art und Weise verstanden werden können [...]. Es darf nicht möglich sein, andere Sachverhalte hineinzuzinterpretieren. Alle Leser einer Anforderung sollten zu einer einzigen, konsequenten Interpretation der Anforderung gelangen. (Rupp 2007)

Eines der zentralen Qualitätsprobleme von Präzision und Eindeutigkeit ist (fach-) lexikographischer Natur: *Weak-Words*, auch *unscharfe Wörter* genannt, verstoßen in gewissen (Satz-)Kontexten gegen die aufgeführten Qualitätskriterien. Weak-Words sind „Wörter oder Phrasen, deren Benutzung in einem Freitext darauf schließen lässt, dass der Freitext mit hoher Wahrscheinlichkeit unpräzise ist“ (Melchisedech 2000). Ein Beispiel für ein Weak-Word ist das Adverb *lang* in der Anforderung, die in Beispiel (1) dargestellt ist.

(1) *Die Taste muss lang gedrückt werden.*

Das Wort *lang* eröffnet hier einen großen Interpretationsspielraum. Ein Projektbeteiligter wird das Wort *lang* als eine Zeitspanne von drei Sekunden interpretieren, ein anderer möglicherweise als eine Zeitspanne von 500 Millisekunden. Dieser Interpretationsspielraum kann im Produktentwicklungsprozess zu großen Problemen führen, weil ein Fehler in der Anforderungsformulierung sich durch viele Teile eines Projekts ziehen kann und möglicherweise erst zu einem sehr späten Zeitpunkt aufgedeckt wird. Aus diesem Grund ist es wichtig, Anforderungen auf das Vorkommen von Weak-Words zu prüfen. Praktisches Ziel der hier beschriebenen Arbeiten ist es, einen Teil dieser Prüfung zu automatisieren.

Die Annahme liegt nahe, dass jedes Auftreten eines Weak-Words die Qualität einer Anforderung mindert, man deshalb Weak-Words identifizieren und Anforderungsautoren bitten sollte, ihre Formulierungen nochmals zu überarbeiten. Es ist aber zu beachten, dass Weak-Words nicht immer die Qualität einer Anforderung mindern. Dies ist im hohen Maße vom Kontext abhängig, wie man in Beispiel (2) sehen kann.

(2) *Die Taste muss 2 Sekunden lang gedrückt werden.*

Hier eröffnet das Weak-Word *lang* keinen Interpretationsspielraum, sondern trägt ganz im Gegenteil dazu bei, die Anforderung zu präzisieren. Bei einer Weak-Word-Analyse muss folglich der Kontext der Weak-Words miteinbezogen werden, um entscheiden zu können, ob diese in einer Anforderung Interpretationsspielraum eröffnen oder nicht (Krisch 2013).

## 2 Stand der Technik

Es existieren bereits Prüfwerkzeuge, die eine automatisierte Weak-Word-Analyse durchführen. Das sind u.a. *ReQualize* (Heidenreich 2010), *DESIRE*<sup>®</sup> (Stöckel et al. 2009) und *QuARS* (Lami 2005). Bei allen drei Werkzeugen kann gezielt nach Wörtern gesucht werden, welche in einer Liste abgelegt sind. Wird ein solches Wort in einem Lastenheft gefunden, werden eine Warnung und ein Hinweis an den Benutzer zurückgegeben, dass er die entsprechende Anforderung nochmals überarbeiten sollte. Die Kontexte der Wörter werden bei diesen Werkzeugen nicht in die Analyse miteinbezogen. Bei einer großen Textmenge hat dies die Konsequenz, dass sehr viele Fehlermeldungen produziert werden und der Benutzer sehr oft unnötig gewarnt wird; das führt wiederum dazu, dass das System vom Benutzer nicht akzeptiert wird.

Diese unnötigen Warnungen sollen durch den hier beschriebenen Ansatz zur Erstellung kontextabhängiger Prüf-Regeln für bestimmte Weak-Words reduziert werden.

## 3 Kontextanalyse und Entwicklung eines lexikalischen Prüfwerkzeugs

### 3.1 Korpuserstellung

Um die Kontexte ausgewählter Weak-Words identifizieren zu können, musste zunächst ausreichend Textmaterial bereitgestellt werden. Hierfür wurden Lastenhefte aus einer Datenbank der potentiellen Anwender exportiert, durch ein eigens entwickeltes Werkzeug tokenisiert (Identifikation von Wort- und Satzgrenzen) und linguistisch annotiert (mit den *mate*-Werkzeugen, (Bohnet 2009)). Die tokenisierten Lastenhefte wurden zunächst lemmatisiert, anschließend mit Wortart- und Morphologie-Annotationen versehen und zuletzt noch auf der Grundlage von Dependenzsyntax analysiert (geparst). Abbildung 1 zeigt ein Beispiel eines Anforderungssatzes nach der computerlinguistischen Annotation. In schwarz ist die jeweilige Satz- bzw. Wortnummer dargestellt und in dunkelblau die Wortform (Token). Orange markiert sind die Lemmata, grün die Wortart-Tags, pink die Morphologie-Tags, rot die Abhängigkeiten zwischen den einzelnen Tokens und hellblau die syntaktischen Informationen. Das Wort *Wiedereinschalten* beispielsweise ist das zweite Token im ersten Satz mit der ID *SPM-8014*. Die Grundform, also das Lemma, ist *Wiedereinschalten*. Außerdem ist es ein Nomen, hat den Kasus Nominativ, steht im Singular und ist neutrum. *Wiedereinschalten* bezieht sich auf Token 3, also auf *erfolgt*, und ist dessen Subjekt.

Die linguistische Annotation der Anforderungen dient vor allem zur Generalisierung der in Abschnitt 3.2 entwickelten Regeln.

1_1_SPM-8014	Das	der	ART	case=nom number=sg gender=neut	2	NK
1_2_SPM-8014	Wiedereinschalten	Wiedereinschalten	NN	case=nom number=sg gender=neut	3	SB
1_3_SPM-8014	erfolgt	erfolgen	VVFIN	number=sg person=3 tense=pres mood=ind	0	--
1_4_SPM-8014	immer	immer	ADV	_	3	MO
1_5_SPM-8014	durch	durch	APPR	_	3	MO
1_6_SPM-8014	eine	ein	ART	case=acc number=sg gender=fem	8	NK
1_7_SPM-8014	erneute	erneut	ADJA	case=acc number=sg gender=fem degree=pos	8	NK
1_8_SPM-8014	Anforderung	Anforderung	NN	case=acc number=sg gender=fem	5	NK
1_9_SPM-8014	vom	von	APPRART	case=dat number=sg gender=neut	8	MNR
1_10_SPM-8014	Logikträger	Logikträger	NN	case=dat number=sg gender=neut	9	NK
1_11_SPM-8014	.	--	\$.	_	10	--

Abbildung 1: Darstellung eines Anforderungs-Satzes mit linguistischen Annotationen.

Insgesamt wurde mit diesem Vorgehen ein Korpus erzeugt, das 88 166 Sätze und 990 011 Tokens enthält. Auf Grundlage dieses Korpus wurden Experimente zur Identifikation präziser und unpräziser Kontextmuster ausgewählter Weak-Words durchgeführt. Für die Evaluation wurde ein Korpus mit neuen Lastenheften erzeugt (vgl. Kapitel 4).

### 3.2 Ermittlung der Kontexte ausgewählter Weak-Words

Die verwendete Arbeitsmethodik ist von der korpusbasierten Lexikographie inspiriert: Zur Entwicklung lexemspezifischer Regeln für die Identifikation von Weak-Words im Kontext wurde das Anforderungs-Korpus interaktiv mit dem Suchwerkzeug CQP (Evert & Hardie 2011) durchsucht. In Tabelle 1 sind Beispiele für „gute“ und „problematische“ Kontexte des potentiellen Weak-Words *mal* dargestellt.

Zusammensetzung	Beispiel
<b>CARD mal</b>	3 mal
<b>erst.+ mal</b>	das erste Mal
<b>XXX_X... mal</b>	<Parameter> mal
<b>tbdX (default CARD) mal</b>	tbd1 (default 10) mal
<b>[0-9]+ten mal</b>	nach dem 3ten Mal
<b>nächste.* mal</b>	das nächste Mal
<b>letzt.+ mal</b>	beim letzten Mal
<b>[a-z] mal</b>	x mal
<b>jedes mal wenn</b>	jedes mal wenn
<b>auch mal</b>	auch mal

**Tabelle 1: Kontextmuster des Weak-Words *mal/Mal*.**

Die grün markierten Spalten sind Kontextmuster von *mal*, in welchen die Verwendung des Wortes nicht dazu führt, dass die Anforderung unpräzise wird. Angaben wie *3 mal* oder *nach dem 3ten Mal* sind präzise und schaden der Qualität einer Anforderung nicht. Auch Kontextmuster, bei denen ein Parameter vor dem Wort *mal* steht, sind in Ordnung. Genauso verhält es sich mit Konstruktionen, die eine flektierte Form von *erst* und ein nachfolgendes *Mal* enthalten und mit Konstruktionen, die die Form *tbdX (default 3) mal*<sup>1</sup> aufweisen. Die rot markierten Muster hingegen sollten in Anforderungen vermieden werden, da sie die Verständlichkeit der betreffenden Anforderungen mindern. Ein Beispiel hierfür ist in (3) gegeben.

(3) *Die Taste kann auch mal klemmen.*

In (3) wird nicht eindeutig spezifiziert, wann dieser Fall, dass die Taste klemmen kann, genau auftreten kann. Das Weak-Word *mal* eröffnet einen großen Interpretationsspielraum und aus diesem Grund müssen Anforderungen wie diese zur nochmaligen Überarbeitung an den Benutzer zurückgeliefert werden.

Die orange markierten Muster sind Fälle, bei denen aufgrund des Satzkontexts alleine nicht sicher ist, ob sie präzise genug sind. In manchen Fällen wird es klar sein, wann *das nächste Mal* oder *das letzte Mal* ist, in anderen Fällen wiederum nicht. Hier sollte zur Sicherheit eine Warnung an den Benutzer ausgegeben werden.

Der Grundgedanke der Weak-Word-Analyse ist, dass nur die wirklich präzisen Konstruktionen mit potentiellen Weak-Words dem Benutzer nicht zurückgeliefert werden sollen. Alle anderen Konstruktionen (rot und orange markiert) sollten an den Benutzer zurückgegeben werden, damit dieser sich die betreffende Anforderung nochmals anschauen und sie überarbeiten kann. Dies kann zwar dazu

1 Konstruktionen wie diese sind Standardkonstruktionen in Lastenheften und müssen deshalb nicht als Warnung an den Benutzer zurückgegeben werden.

führen, dass weiterhin unnötige Warnungen an den Benutzer geliefert werden, dafür wird aber auch keine unpräzise Anforderung übergangen.

In gleicher Weise wurden auch die Kontextmuster für das Weak-Word *lang* identifiziert. In Tabelle 2 sind ausschließlich die präzisen Kontextmuster von *lang* angegeben. Insgesamt wurden acht unpräzise Kontextmuster und drei Kontextmuster ermittelt, bei denen nicht klar entschieden werden konnte, ob sie immer zu präzisen Anforderungen führen oder nicht. Konstruktionen wie *länger als 3 Sekunden* und *langes Drücken von 1 Sekunde* sind präzise und lösen keinen Interpretationsspielraum aus. Ein ebenfalls präzises Beispiel ist *länger als für die Nachlaufzeit*. Hier weist der bestimmte Artikel darauf hin, dass der Begriff *Nachlaufzeit* in einer vorangehenden Anforderung schon verwendet wurde und somit spezifiziert sein sollte. An dieser Stelle wird zwar keine Warnung an den Autor ausgegeben, am Schluss einer Analyse werden aber dem Autor alle Nomina zurückgegeben, denen ein bestimmter Artikel vorangeht und die mit dem Wort *lang* zusammenhängen, mit der Bitte um Prüfung, ob die Wörter eindeutig beschrieben sind. Tritt hingegen eine Konstruktion mit unbestimmtem Artikel auf wie in *länger als eine Erholungszeit*, wird in jedem Fall eine Warnung ausgegeben, da der unbestimmte Artikel darauf hinweist, dass das Wort *Erholungszeit* an dieser Stelle neu eingeführt wird und noch nicht vollständig beschrieben ist.

Für die Weak-Words *kurz* und *schnell* wurden dieselben Regeln verwendet, die auch bei *lang* zum Tragen kommen. Ziel war es zu testen, ob sich Regeln eines Weak-Words auf semantisch ähnliche Weak-Words übertragen lassen, d.h. ob die definierten Regeln generalisierbar sind. Die Weak-Words und die korrespondierenden Regeln wurden in einem Lexikon abgelegt, in einen ersten Prototypen implementiert und anschließend evaluiert.

Zusammensetzung	Beispiel
lang als ca.	länger als ca. 3 Sekunden
CARD NN/NE lang	15 Sekunden lang
lang als CARD	länger als 3 Sekunden
lang als für def.ART	länger als für die Nachlaufzeit
wie lang	wie lange
lang als def.ART	länger als die Verzögerungszeit
XXX_X... lang	<Parameter> lang
lang VERB als	nicht länger betrieben als
lang als XXX_X...	länger als <Parameter>
lang CARD NN/NE	länger 500 Millisekunden
lang NN/NE (</> ...)	langes Drücken (> 1 Sekunde)
lang NN/NE von	langes Drücken von 1 Sekunde
lang def.ART NN/NE	länger der Nachlaufzeit

**Tabelle 2: Präzise Kontextmuster des Weak-Words *lang*.**



Für jedes Weak-Word wurde ein Lexikoneintrag mit den entsprechenden Regeln erstellt. Die Kontextregeln sind als Suchmuster in der Skriptsprache Python, auf der Basis der in Abschnitt 3.1 illustrierten Annotationen, implementiert. Das Lexikon enthält auf dem Stand von Frühjahr 2014 die Weak-Words *mal*, *lang*, *kurz* und *schnell*. Neue Wörter und zugehörige Regeln können analog hinzugefügt werden. Im Lexikon stehen nur die präzisen Kontextmuster der Weak-Words. Dies gewährleistet, dass alle präzisen Kontexte vom System erkannt werden und alle anderen Kontexte, die nicht im Lexikon vermerkt sind, an den Benutzer zurückgegeben werden. Dadurch werden einerseits Fehlermeldungen, die bei anderen Tools unnötigerweise ausgegeben werden, vermieden, und andererseits keine unpräzisen Anforderungen übergangen.

## 4 Evaluation

Die Regeln wurden anhand eines Korpus von insgesamt 99 955 Sätzen und 1 160 409 Tokens evaluiert. In diesem Korpus kam das Weak-Word *mal* 44 mal vor, *lang* trat 147 mal, *kurz* 106 mal und *schnell* 207 mal auf. Es wurden alle Wortformen der vier Weak-Words analysiert, nicht nur die Grundformen.

Tabelle 3 zeigt die Auswertung der Ergebnisse für die 44 Erscheinungsstellen von *mal*. 26 Anforderungen wurden automatisch korrekt als präzise Anforderung klassifiziert. 15 unpräzise Anforderungen wurden ebenfalls richtig als unpräzise Anforderungen klassifiziert. Drei präzise Anforderungen hingegen wurden fälschlicherweise als unpräzise interpretiert, d.h. es wurden insgesamt drei überflüssige Warnungen an den Benutzer zurückgegeben. Interessant ist vor allem, dass keine unpräzise Anforderung übersehen wurde.

Das Ziel der Weak-Word-Analyse ist es, die unnötigen Warnungen an den Benutzer zu reduzieren, es soll dabei aber keine unpräzise Anforderung übergangen werden. Durch die Formulierung von Kontextregeln des Weak-Words *mal* für unproblematische Fälle konnten gegenüber einer wortbasierten Strategie 26 Warnungen entfallen, was fast 60 % der Gesamtmenge an Warnungen des wortbasierten Systems entspricht.

		automatische Analyse →	
		<b>Unpräzise Anforderung</b>	<b>Präzise Anforderung</b>
manuelle Analyse ↓	Unpräzise Anforderung	15	0
	Präzise Anforderung	3	26

**Tabelle 3: Evaluation der Ergebnisse für das Weak-Word *mal*.**

Ein Beispiel für eine falsch klassifizierte Anforderung, die das Wort *mal* enthält, ist in (4) gegeben.

(4) Die Funktion darf nicht ein viertes Mal gestartet werden.

Die beiden anderen falsch klassifizierten Anforderungen sind von derselben Art wie (4). Der POS-Tagger interpretiert das Zahlwort *viertes* nicht als Zahl (Tag = CARD), sondern als Artikel. Eine Konstruktion wie diese wird bisher nicht mit den Regeln aus Tabelle 1 abgefangen. Eine entsprechende Regel sollte in zukünftigen Arbeiten entworfen und in das Lexikon integriert werden.

In Tabelle 4 sind die entsprechenden Precision- und Recall-Werte und die daraus resultierenden F1-Measure-Werte für die Klassen *Unpräzise Anforderung* und *Präzise Anforderung* zu sehen. Der Precision-Wert gibt den Anteil der Suchergebnisse an, die tatsächlich relevant sind in Bezug auf die Gesamtgröße des Ergebnisses (Dörre, Gerstl, & Seiffert 2004). Der Precision-Wert der Klasse *Unpräzise Anforderung* liegt beispielsweise bei 83,33 %. Der Grund hierfür ist, dass zwar alle relevanten Anforderungen gefunden wurden, aber noch überflüssige Warnungen an den Benutzer geliefert werden. Dies ist bei der Klasse *Präzise Anforderung* nicht der Fall, und deshalb beträgt hier der Precision-Wert 100 %.

Der Recall-Wert bezeichnet den Anteil der relevanten Suchergebnisse in Bezug auf die Menge aller für diese Suche relevanten Ergebnisse (Dörre, Gerstl, & Seiffert 2004). Für die Klasse *Präzise Anforderung* werden nicht alle relevanten Instanzen gefunden, für die Klasse *Unpräzise Anforderung* allerdings schon, was die Werte von 89,66 % und 100 % erklärt.

	Klasse: Präzise Anforderung	Klasse: Unpräzise Anforderung
Precision	100 %	83,33 %
Recall	89,66 %	100 %
F <sub>1</sub> -Measure	94,55 %	90,91 %

**Tabelle 4: Precision, Recall und F<sub>1</sub>-Measure: mal.**

Das F1-Maß ist eine Kombination aus Precision und Recall (geometrisches Mittel) und gibt einen Gesamtüberblick über die Performanz des entwickelten Werkzeugs in Bezug auf die einzelnen Klassen. Tabelle 5 zeigt die Auswertung der Ergebnisse für das Weak-Word *lang*. 129 Kontexte von *lang* konnten korrekt klassifiziert werden. Es gab nur 18 Fehlklassifizierungen. Insgesamt konnten 69 unnötige Warnungen vermieden werden, was in der Gesamtmenge knapp 50 % entspricht.

		automatische Analyse →	
		Unpräzise Anforderung	Präzise Anforderung
manuelle Analyse ↓	Unpräzise Anforderung	60	0
	Präzise Anforderung	18	69

**Tabelle 5: Evaluation der Ergebnisse für das Weak-Word lang.**

Die Fehlanalysen beruhen vor allem auf Klammer- und Adjunkt-Konstruktionen, die vom System bisher nicht behandelt werden können. Ein Beispiel für eine Klammerkonstruktion, die falsch klassifiziert wurde, ist in (5) zu sehen.

(5) *Die Eingriffe sollen nicht innerhalb einer applizierbaren langen Vergangenheit (5 Sekunden) liegen.*

Für Konstruktionen wie diese wurden bisher keine Regeln formuliert. Inwiefern hier eine Regel entworfen werden kann, soll in weiteren Arbeiten untersucht werden.

Weak-Word	lang		kurz		schnell	
	Prüz. Anf.	Unpräz. Anf.	Prüz. Anf.	Unpräz. Anf.	Prüz. Anf.	Unpräz. Anf.
<b>Precision</b>	100 %	76,92 %	100 %	82,83 %	100 %	98,47 %
<b>Recall</b>	79,31 %	100 %	29,12 %	100 %	78,57 %	100 %
<b>F<sub>1</sub>-Measure</b>	88,46 %	86,95 %	45,11 %	90,61 %	88 %	99,23 %

**Tabelle 6: Precision, Recall und F<sub>1</sub>-Measure: lang, kurz, schnell.**

Eine Adjunkt-Konstruktion, die ebenfalls nicht korrekt klassifiziert werden konnte, ist in Beispiel (6) gegeben.

(6) *War die Strecke bis zu Position A zu lang, d.h. es wurden mehr als drei Hallimpulse gezählt, wird das Glasdach geschlossen.*

Bei Beispiel (6) handelt es sich um eine linguistisch sehr komplexe Anforderung. Regeln für eine Konstruktion wie diese zu definieren, ist maschinell nur schwer umsetzbar. Man sollte sich hier auch die Frage stellen, ob sich der Aufwand eine Regel zu finden lohnt, oder ob es nicht doch sinnvoll ist, den Autor auf eine so komplizierte Formulierung wie in (6) hinzuweisen, damit er die Anforderung nochmals in eine verständlichere Konstruktion umwandelt.

Tabelle 6 zeigt die Auswertung der Ergebnisse der Weak-Words *lang*, *kurz* und *schnell*. Für die Analyse von *kurz* und *schnell* wurden dieselben Regeln verwendet, die auch bei *lang* zum Tragen kommen. Die Ergebnisse von *lang* und *schnell* sind solide. Bei *kurz* kann man einen deutlichen Einbruch erkennen. Zwar generalisieren die Kontexttypen von *lang* gut auf *schnell* und ordentlich auf *kurz*, aber die drei Wörter werden durchaus unterschiedlich verwendet. Regeln, die für jedes einzelne Wort separat festgelegt werden, könnten also noch bessere Ergebnisse liefern. Auch hier resultieren die Fehlanalysen des Systems vor allem aus Klammer- und Adjunkt-Konstruktionen.

## 5 Zusammenfassung und Ausblick

Es wurde ein sehr stark auf Kontext-Typen aus Anforderungstexten zugeschnittenes Lexikon von potentiell unspezifischen Wörtern erstellt, welches automatisch anwendbar ist. Das Lexikon enthält Wörter und deren Kontextmuster; die Arbeitsmethodik ist von der korpusbasierten Lexikographie inspiriert, zielt aber auf sehr präzise, automatisch überprüfbare Muster von Kontexttypen. Die entwickelte Weak-Word-Analyse, die dieses Lexikon als Quelle heranzieht, liefert solide und über semantisch verwandte Wörter generalisierbare Ergebnisse.

In weiteren Arbeiten sollen weitere Weak-Words in die Analyse integriert und neue Lexikoneinträge mit zugehörigen Regeln erzeugt werden. Für zukünftige Arbeiten wäre es auch interessant zu untersuchen, welche anderen linguistischen Probleme, neben den Weak-Words, gegen die Qualitätskriterien verstoßen und ob auch diese automatisiert behandelt werden können.

## 6 References

- Bohnet, B. (2009). Efficient Parsing of Syntactic and Semantic Dependency Structures. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009*, Boulder, Colorado, US.
- Daimler AG (2013). Qualitätskriterien, Daimler AG, Böblingen, DE.
- Dörre, J., Gerstl, P., & Seiffert, R. (2004). Volltextsuche und Text Mining. In K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde, & H. Langer, *Computerlinguistik und Sprachtechnologie: Eine Einführung*. München: Spektrum, pp. 479-495.
- Evert, S., Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. University of Birmingham, UK.
- Heidenreich, M. (2010). Metriken und Werkzeugunterstützung zur Überprüfung von Anforderungen. OBJEKTSpektrum, Ausgabe RE/2010.
- Krisch, J. (2013). Identifikation kritischer Weak-Words aufgrund ihres Satzkontextes in Anforderungsdokumenten. Diplomarbeit. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart, DE.
- Lami, G. (2005). QuARS: A Tool for Analyzing Requirements. Carnegie-Mellon Software Engineering Institute.
- Melchisedech, R. (2000). Verwaltung und Prüfung natürlichsprachlicher Spezifikationen. Dissertation. Universität Stuttgart, Stuttgart, DE.
- Rupp, C. (2007). Requirements-Engineering und -Management: professionelle, iterative Anforderungsanalyse für die Praxis. München: Hanser Verlag.
- Stöckel, F., Stolz, P., Uddin, I., & Endriss, L. (2009). DESIRE® Dynamic Expert System for Improving Requirements. HOOD Group.

# From Term Dynamics to Concept Dynamics: Term Variation and Multidimensionality in the Psychiatric Domain

Pilar León-Araúz, Arianne Reimerink  
Department of Translation and Interpreting, University of Granada  
pleon@ugr.es, arianne@ugr.es

## Abstract

Medical terminology is one of the most dynamic terminological domains, and the choice of one term instead of the other is not random, but the result of different perspectives towards reality. VariMed is a research project on medical term variants and its overall objective is to generate a multifunctional resource on the medical domain for linguistic research, translation and technical writing. In this paper, we propose a systematic way of extracting term variants from large corpora within the subdomain of Psychiatry and how to represent them according to cognitive and communicative parameters. Our aim is to discover if different conceptualizations, or different conceptually motivated term variants, of the same concept are preferred in expert or semi-specialized communication. A corpus on Psychiatry was compiled and classified according to user types. A grammar was designed in NooJ (Silberstein, 2003) in order to extract term variants based on the usual lexico-syntactic patterns accompanying synonyms (*also known as; commonly referred to as*, etc.). Corpus analysis results indicate that, from a cognitive perspective, term variants reflect the prototypical dimensions in which psychiatric disorders may be classified. From a communicative perspective, terms and dimensions can also be associated with user-based parameters.

**Keywords:** term variation; cognitive motivation; communicative motivation

## 1 Introduction

The medical domain has over 25 centuries of history and involves numerous disciplines which affect all human beings to some extent. Therefore, medical terminology is one of the most dynamic terminological domains, and the use of one term instead of the other implies perceiving and conceptualizing aspects of reality from different perspectives (Prieto Velasco et al 2013: 168). VariMed is a research project on medical term variants and its overall objective is to generate a multifunctional resource on the medical domain for linguistic research, translation and technical writing. In this paper, we propose a systematic way of extracting term variants from large corpora within the subdomain of Psychiatry and how to represent them according to cognitive and communicative parameters. The relationship of specialized communications between a terminological resource and its user implies a

prototypical discursive positioning (Harré and Langenhove 1999), which is reflected in a specialized text sender and receivers with a different background knowledge level. Terminological resources should provide adequate terminological units and an adequate knowledge load (Tarp 2005: 8-9), always according to their potential users' continuum of general-specific language and knowledge purposes (León-Araúz et al 2013: 33).

This is in consonance with the Functional Theory of Lexicography (FTL; Bergenholtz and Tarp 1995, 2003). According to the FTL, there are two main types of lexicographic functions that cover *use situations* and different *user needs* (Wiegand 1989). These functions are cognition and communication-oriented (Bergenholtz and Tarp 2003; Bergenholtz and Nielsen 2006). In cognition-oriented situations, users seek additional information to widen their knowledge about the conceptual structure of a particular subject-field (psychiatry, neurology, oncology, etc.). Bergenholtz and Nielsen (2006: 286) explain that in these situations, the only communicative act taking place is between the terminographer and the users of the resource. The users want knowledge and the lexicographers provide it at a cognitive level, nothing more. The most difficult task is then, for the terminographer to decide *how much* information is to be included and how to represent its underlying structure to make the dictionary suitable to meet users' needs. On the other hand, in communication-oriented situations, two or more persons are engaged in producing or receiving a piece of language. This is the case of a translator who receives and must subsequently produce a text, as well as scientific writers, proofreaders, etc. Here the terminographer acts as a kind of mediator who helps to solve communication problems. We believe any terminological resource should satisfy both (León-Araúz et al 2013: 34).

In section 2, we give a brief overview on term variation. In section 3, we present how term variants are extracted from a specialized corpus on Psychiatry with a pattern-based grammar in NooJ, an NLP application (Silberztein, 2003). In section 4, a selection of the extracted variants is classified according to dimensional features and the results are compared across three subcorpora in order to see if certain dimensions are preferred in one discourse or the other. Finally, section 5 covers the conclusions and further research.

## 2 Term Variation

Although specialized language initially aspired to having one linguistic designation for each concept for greater precision, it is true that the same concept can often have many different types of linguistic designations. In the same way as in general language, there is terminological variation based on user-based parameters of geographic, temporal or social variation or usage-based parameters of tenor, field, and mode (Gregory and Carroll 1978). However, terminological variation also occurs for reasons that are often considerably more complex and difficult to explain. Freixa (2006: 52) classifies the causes for terminological variation in the following categories: (1) dialectal, caused by different origins of the authors; (2) functional, caused by different communicative registers; (3) discursive, caused by dif-

ferent stylistic and expressive needs of the authors; (4) interlinguistic, caused by contact between languages; (5) cognitive, caused by different conceptualizations and motivations. According to Freixa (2002), certain term variants are not only formally different, but also semantically diverse, as they give a particular vision of the concept. In this sense, Fernández-Silva et al (2011) describe this phenomenon as the linguistic reflection of conceptual *multidimensionality*. Multidimensionality has been defined by many authors (Bowker 1997, Kageura 1997, Wright 1997, Rogers 2004) as the phenomenon in which certain concepts can be classified according to different points of view or conceptual facets. This has important consequences in how domains are categorized and modelled. According to Picht and Draskau (1985, 48 *apud* Rogers 2004, 219), multidimensionality depends on who is the classifier as well as the different knowledge sources that may reflect different criteria when organizing the same domain or knowledge node. For example, botanists would classify roses different from rose growers. However, multidimensionality has also an impact on term variation, since concepts can be designated in more than one way based on the different characteristics that it possesses (Fernández-Silva et al 2011). Thus term variation should not be regarded as a linguistic phenomenon isolated from conceptual representations, since it is one of the manifestations of the dynamicity of categorization and expression of specialized knowledge (Fernández Silva et al in press).

Fernández-Silva (2010: 60-71) classifies the cognitive factors involved in term variation, based on numerous authors, according to two criteria. Firstly, the first category division depends on whether the cognitive factor refers to the conceptual organization or to its usage. Secondly, within the usage category, the categorization of the factor depends on how reality is conceptualized by certain groups of people or individuals or how reality is conceptualized according to the specific context in which the concept is used (Table 1).

Conceptual organization	
	Conceptual system
	Conceptual class
	Multidimensionality of the conceptual system
	Flexibility of the concept
	Linguistic system/culture
Usage	
Different conceptualizations	Knowledge evolution
	Dialects/cultures
	Thematic areas in interdisciplinary contexts
	Schools of knowledge/Ideologies
	Socio-professional groups
	Individual/individual point of view
Adaptation to specific context use	Adaptation to level of expertise of the receiver
	Intention/Aim/Point of view

**Table 1: Cognitive factors of term variation (adapted from Fernández-Silva 2010: 61).**

As can be inferred by Table 1, all Freixa’s causes for term variation can be approached from a cognitive perspective according to Fernández-Silva. This study combines Freixa’s second (functional) and fifth (cognitive) causes for term variation with Fernández-Silva’s perspective, since it analyses the multidimensionality of the conceptual system and how the different conceptual dimensions correlate with the adaptation to the level of expertise of the receiver. Our aim is to discover if different conceptualizations, or different conceptually motivated term variants, of the same concept, are preferred in expert communication or semi-specialized communication.

### 3 Extracting Term Variants

A specialized corpus was compiled on the Psychiatry domain, which has more than 10 million tokens, and it was divided according to user and genre types: expert, semi-specialized and encyclopaedic. The expert corpus contains specialized books and journal papers written by experts for experts, such as the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2013). The semi-specialized corpus consists of web pages and brochures written by experts from Medline Plus and the National Institute of Mental Health, which combines basic and clinical research with information for patients, or their relatives, suffering from any kind of mental disorder. Finally, the encyclopaedic corpus consists of a Wikipedia dump which was automatically collected through categories such as *Psychiatry, Syndromes, Disorders*, etc. We considered that Wikipedia should belong to this corpus because, being an encyclopaedic resource, it usually contains metalinguistic information on synonyms and variants that could be useful in our research.

Once the corpus was compiled and classified, a NooJ local grammar was designed in order to extract term variants (Figure 1). The grammar is based on the usual lexico-syntactic patterns accompanying synonyms (*also known as; commonly referred to as*, etc.) combined with specialized terms, namely syndromes, disorders and diseases.

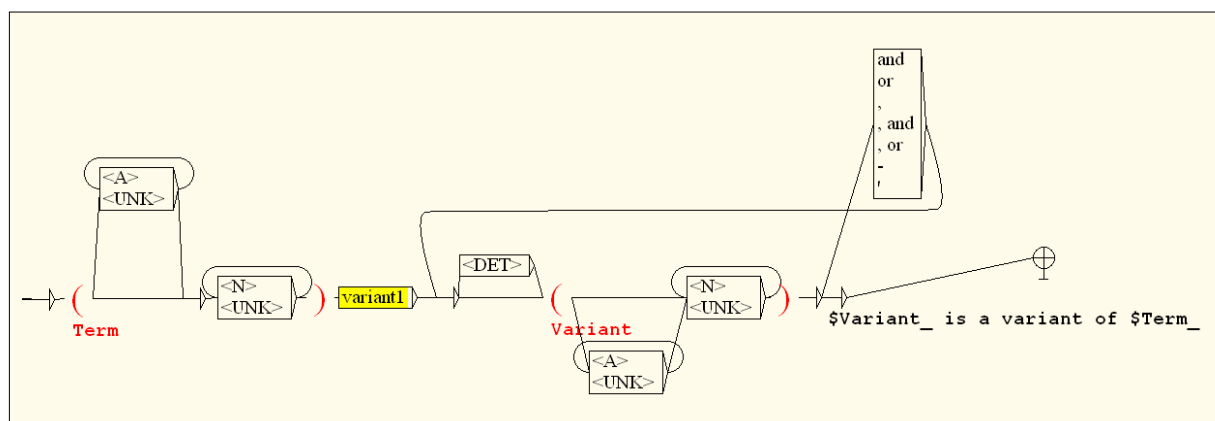


Figure 1: Pattern-based grammar for term variants’ extraction.





de that, not surprisingly, term variance structures are most often lexicalized in the Wikipedia corpus, then in the semi-specialized corpus, and finally in the expert corpus (Figure 4).

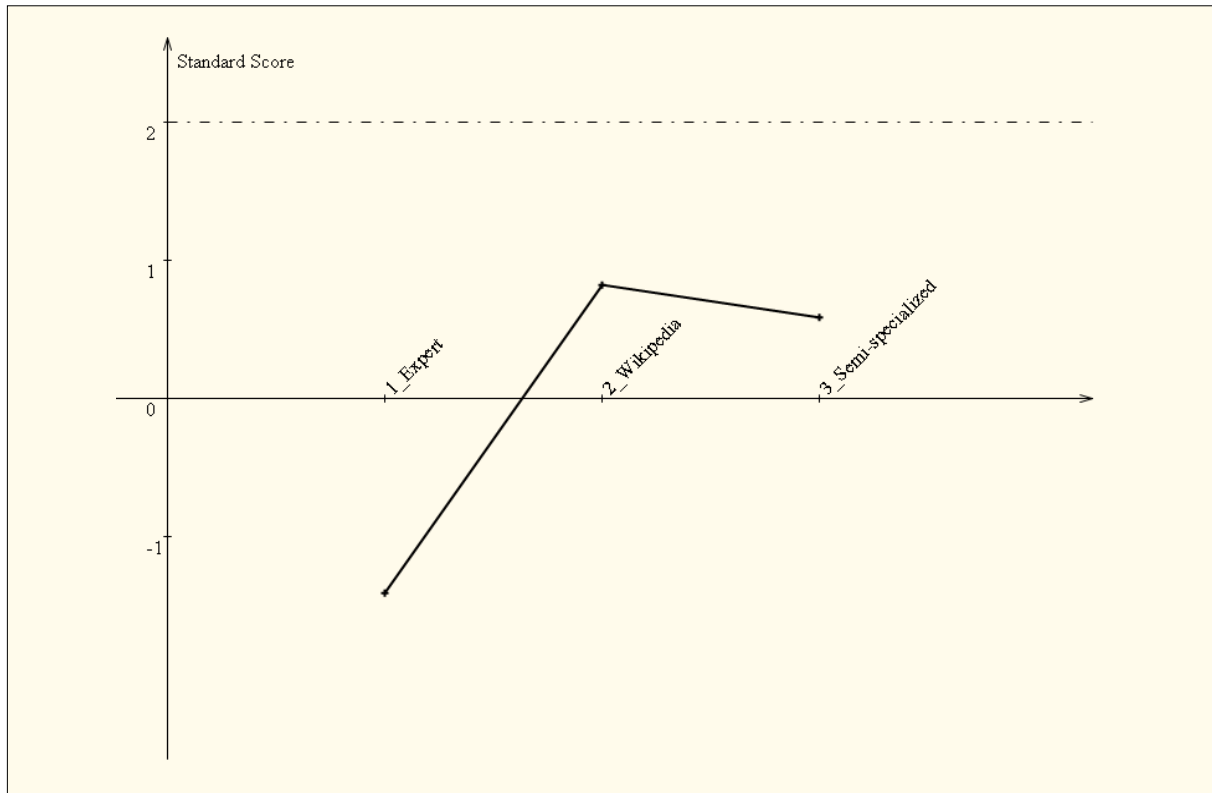


Figure 4: Term variance structures in the three corpora.

## 4 Representing Term Variants

In terminological resources, users are often confronted with a vast array of variants with no other information on how term variation arises and how their use may be constrained. However, they need to know when to use each of the variants and the conceptual connotations they imply, since this will affect the receiver's interpretation of the message.

In our study, we have found many different types of term variants for the same concept. Some of them were just acronyms, dialectal, orthographic or morphological variants, which, of course, need to be stored in any terminological resource, but their impact on communication is obvious, and their use does not usually need any further explanation. In this paper, however, we focus on dimensional variant types, which need more in-depth study, since they affect both cognitive and communicative situations. Dimensional variants show different conceptualizations of the same concept according to different facets and are usually conveyed by multi-word terms. For instance, *Ganser syndrome*, *nonsense syndrome* and *prison psychosis* are all variants of the same concept, but the first one highlights a DISCO-

VERER dimension (Sigbert Ganser was the first to describe the syndrome), the second one focuses on the SYMPTOM dimension (saying nonsense is one of them) and the third one on the LOCATION dimension (it often takes place in prisons, since it affects inmates). We collected from the corpus all the concepts that showed more than one variant type and classified their corresponding variants according to the dimension conveyed. In Table 2, we show the dimensions we found with an illustrative example.

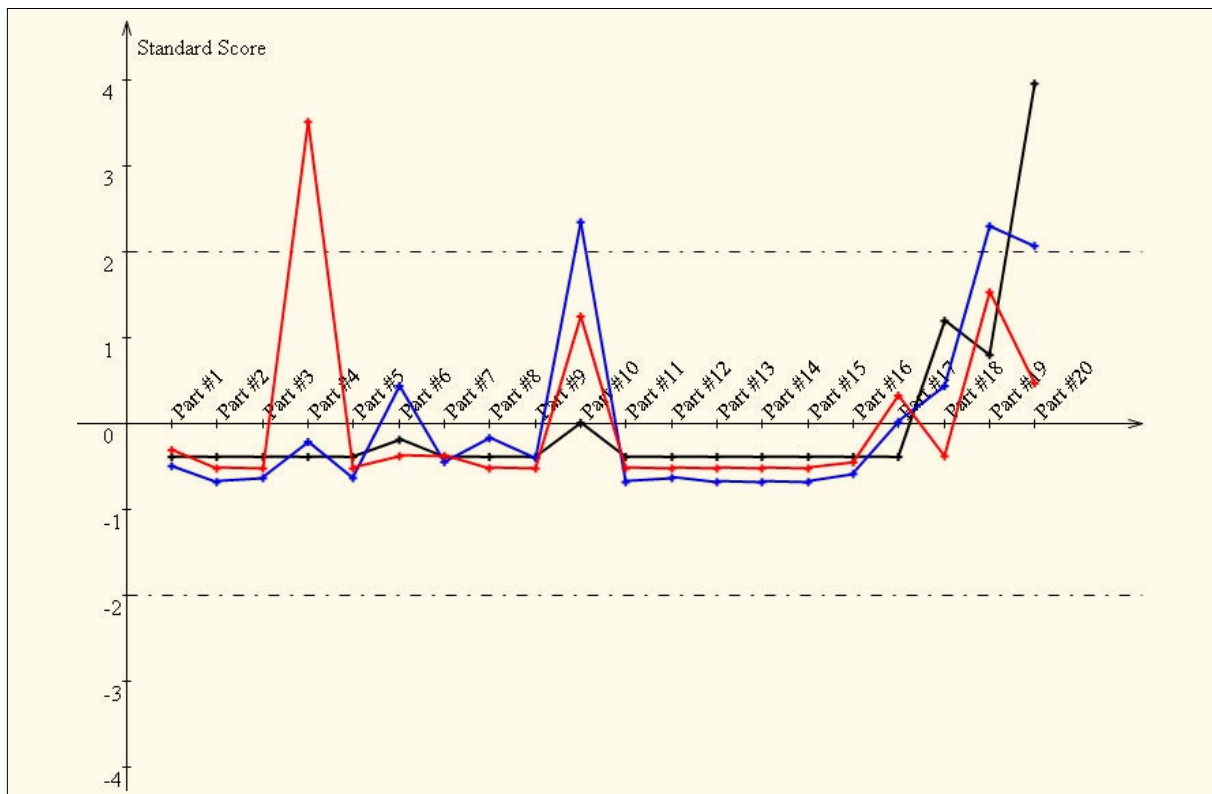
Dimension	Term variant
+Discoverer	Korsakoff's psychosis
+Symptom	burning-mouth syndrome
+Cause	alcohol-induced amnesic disorder
+Body_part	Broca's aphasia
+Patient	boxer's dementia
+Result	bedwetting
+Intensity	mild cognitive impairment
+Time	short-term insomnia
+Location	prison psychosis

**Table 2: Dimensional variant types found in the corpus.**

Of course, there are variants that may show different dimensions at the same time, such as *chronic* [+Time] *traumatic* [+Cause] *brain* [+Body\_part] *injury associated with boxing* [+Cause] or *alcohol-induced* [+Cause] *amnesic* [+Symptom] *disorder*; and different variants for the same concept that highlight the same dimension, such as *Alice in Wonderland* [+Symptom] *syndrome* and *Lilliputian hallucination* [+Symptom], which refer to the same symptom of the disorder, or *nonsense* [+Symptom] *syndrome* and *.syndrome of approximate answers* [+Symptom], which convey the same dimension but refer two different symptoms.

Terminological resources should add the conceptual dimension conveyed by each variant so that users can make a cognitively sound choice, but as previously stated, term choice also depends on communicative situations, namely, the expert-lay continuum. Therefore, term entries should also add use-related information. When querying the corpus, we see that the variant *dementia pugilistica* is much more often used in the expert corpus as compared to *punchdrunk syndrome* or *boxer's dementia*. In this case, the latin origin of the term points to the usual preference in expert settings. However, this preference is not always as straight forward. In such cases, users should have this information at hand. For instance, in Figure 5 three different variants for the same concept are represented according to usage-based preferences: *postnatal depression* (red), *postpartum depression* (blue), and *baby blues* (black). For this analysis, the three corpora were merged into a single file in order to observe these preferences as a continuum. When a single file is loaded in NooJ, its statistical module automatically splits it into 20 parts. Thus, in Figure 5, the first third of the graph represents the expert corpus, the second the Wi-

ikipedia corpus and the third the semi-specialized corpus. Not surprisingly, *baby blues* is the preferred term in semi-specialized communication, although the other two variants are also commonly found. At first, *postnatal* or *postpartum* would seem interchangeable choices, but corpus analysis indicates that the use of one term or the other imposes a strong constraint on the communicative situation. *Postpartum depression* seems to be a more neutral term most often found in the Wikipedia corpus and *postnatal depression* is the preferred term in expert texts.



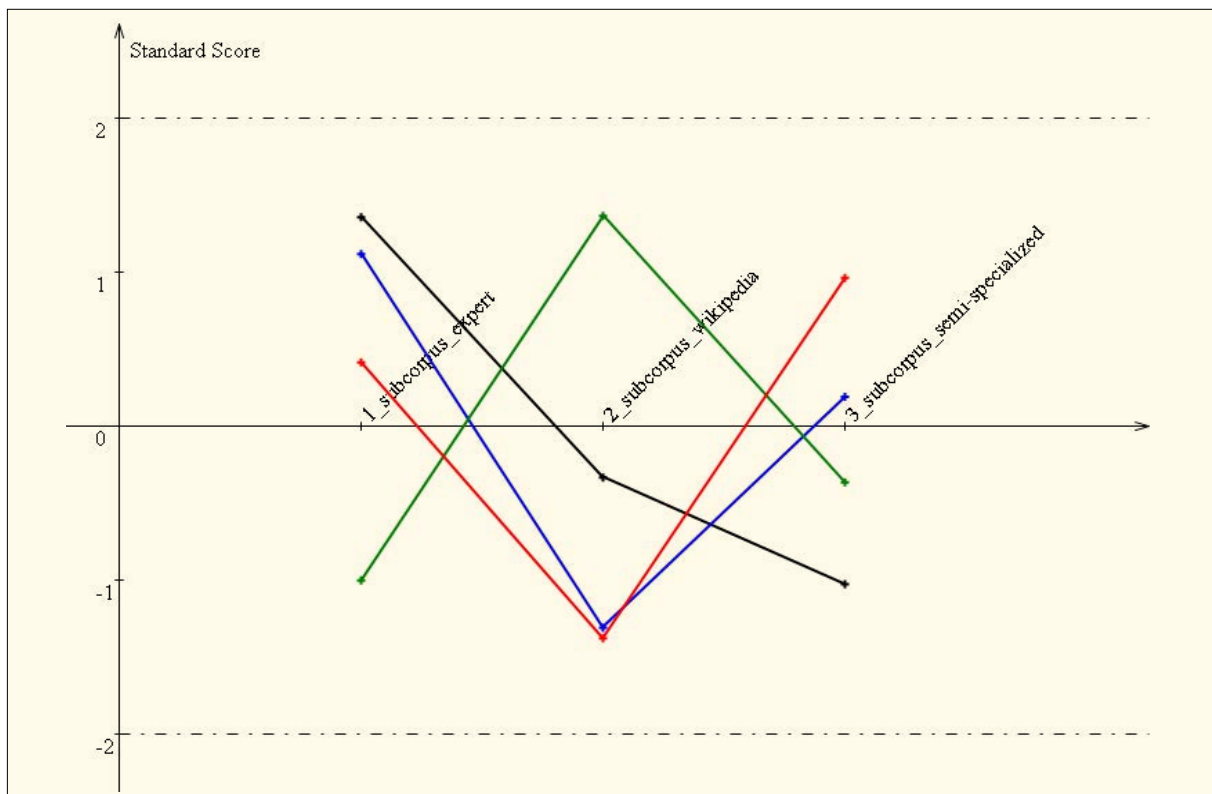
**Figure 5: Usage-based preferences for *baby blues* (black); *postpartum depression* (blue); *postnatal depression* (red).**

Nevertheless, apart from showing usage information related to individual terms, we believe that more generalized patterns can be found in the comparison of the three corpora with regards to the prevalence of dimensions. For this reason, as a further step, we built a new dictionary with all the dimensional variants related to the same entry in order to automatically annotate them as such in the corpus. So far, the dictionary has 73 concept entries associated with 326 dimensional variants. Figure 6 shows an example of a NooJ dictionary entry, where N means noun, FLX=TABLE refers to the inflectional paradigm, and UNAMB is a special code that lets us deal with multi-word terms as a single unit. Apart from NooJ codes, we can also add any other semantic feature to each entry (e.g. Term, Symptom).

```
ganser syndrome, N+FLX=TABLE+Term+Discoverer+UNAMB  
ganser's syndrome, ganser syndrome, N+FLX=TABL+Term+Discoverer+UNAMB|  
nonsense syndrome, ganser syndrome, N+FLX=TABLE+Term+Symptom+UNAMB  
balderdash syndrome, ganser syndrome, N+FLX=TABLE+Term+Symptom+UNAMB  
syndrome of approximate answers, ganser syndrome, N+FLX=TERMOFTERM+Term+Symptom+UNAMB  
pseudodementia, ganser syndrome, N+FLX=TABLE+Term+Intensity+Old+UNAMB  
prison psychosis, ganser syndrome, N+Term+Location+UNAMB
```

**Figure 6: NooJ dictionary entry for *Ganser syndrome* and its variants.**

Once we had our own dictionary, we applied it to the main corpus in order to build three new subcorpora, which would only contain sentences where our previously extracted term variants occurred. This was done for a better performance of the system and especially in order to eliminate any bias towards lexical diversity or distribution differences in the three corpora. After that, we performed new queries based on four of the most relevant dimensions found in this domain (Figure 7): +Discoverer (black), +Cause (red), +Symptom (blue), +Body\_part (green).



**Figure 7: Correlation of dimensional variants with the three corpora.**

According to a patient-oriented approach, one might think at first that the +Symptom and +Body\_part dimensions could be more semi-specialized than expert-related. However, surprisingly enough, it seems that the +Body\_part dimension is only significant in the Wikipedia subcorpus, whereas the +Symptom dimension is most often found in the expert subcorpus. In the expert subcorpus +Discover-

rer is the most prevalent dimension, but +Cause is also significantly represented. As for the semi-specialized subcorpus, +Cause and +Symptom dimensions are most prototypical, but in an inverted way as compared to the expert subcorpus.

This is only a first approach to the study of cognitive and communicative correlation that should be further extended to a higher number of variants and other medical domains, since results can change dramatically when comparing domain-based differences. In this sense, Tsuji and Kageura (1998) observed in a medical corpus that person or virus names were more dominant than other variant types.

## 5 Conclusion

In this paper we have shown how term variants in the psychiatric domain are cognitive and communicatively motivated. From a cognitive perspective, variants reflect the prototypical dimensions in which psychiatric disorders may be classified. From a communicative perspective, terms and dimensions can also be associated with user-based parameters. However, further studies need to be done with regards to the correlation between the cognitive and communicative factors underlying term variation, especially from a cross-linguistic perspective, since not all cultures conceptualize specialized domains in the same way and nor do they address their audience in the same manner.

## 6 References

- American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- Bergenholtz, H. and S. Nielsen. 2006. Subject-Field Components as Integrated Parts of LSP Dictionaries. *Terminology* 12 (2): 281-303.
- Bergenholtz, H. and S. Tarp. 2003. Two Opposing Theories: On H. E. Wiegand's Recent Discovery of Lexicographic Functions. *Hermes: Journal of Linguistics* 31: 171-196.
- Bergenholtz, H. and S. Tarp (eds.) 1995. *Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries*. Amsterdam: John Benjamins.
- Bowker, L. 1997. Multidimensional Classification of Concepts and Terms. In Wright, S. E. and G. Budin (eds.) *Handbook of Terminology Management: Basic Aspects of Terminology Management*. Amsterdam: John Benjamins, 133-143.
- Cruse, D. A. 2002. Hyponymy and Its Varieties. In Green, R., C. A. Bean and S. H. Myang (eds.) *The Semantics of Relationships*. Dordrecht: Kluwer, 3-21.
- Cruse, D. A. 1995. Polysemy and Related Phenomena from a Cognitive Linguistic Viewpoint. In Dizier, P. St. and E. Viegas (eds.) *Computational Lexical Semantics*. Cambridge: Cambridge University Press, 33-39.
- Fernández-Silva, S. 2010. Variación Terminológica y Cognición. Factores Cognitivos en la Denominación del Concepto Especializado. PhD Thesis. Universitat Pompeu Fabra.
- Fernández-Silva, S., J. Freixa and M. T. Cabré. 2011. A Proposed Method for Analysing the Dynamics of Cognition through Term Variation. *Terminology* 17 (1): 49-73.

- Fernández-Silva, S., Freixa, J., Cabré, M.T. (in press). "A proposed method for analysing the dynamics of naming from a intralingual and interlingual perspective". In: Temmerman, R. and Van Campenhoudt, M. (eds.) *The dynamics of culture-bound terminology in monolingual and multilingual communication*. Amsterdam/Philadelphia: John Benjamins.
- Freixa, J. 2006. Causes of Denominative Variation in Terminology: A Typology Proposal. *Terminology* 12 (1): 51-77.
- Freixa, J. 2002. Reflexiones acerca de las causas de la variación denominativa en terminología. In Guerrero, G. and M. F. Pérez (coords.) *Panorama actual de la terminología*. Granada: Editorial Comares-Interlingua, 107-115.
- Gregory, M. and S. Carroll. 1978. *Language and situation: Language varieties and their social contexts*. London: Routledge and Kegan Paul.
- Harré, R. and L. van Langenhove (eds.) 1999. *Positioning Theory: Moral Contexts of Intentional Action*. Malden: Blackwell.
- Kageura, K. 1997. Multifaceted/Multidimensional Concept Systems. In Wright, S. E. and G. Budin (eds.) *Handbook of Terminology Management: Basic Aspects of Terminology Management*. Amsterdam: John Benjamins, 119-132.
- León Araúz, Pilar, Arianne Reimerink, and Alejandro García Aragón. 2013. Dynamism and Context in Specialized Knowledge. *Terminology* 19 (1): 31-61. doi:10.1075/term.19.1.02leo.
- Picht, H. and J. Draskau. 1985. *Terminology: An Introduction*. Guildford: University of Surrey.
- Prieto Velasco, Juan Antonio, Maribel Tercedor, and Clara Inés López Rodríguez. 2013. La multidimensionalidad conceptual en la traducción médica. *Shopos. Revista Internacional de Traducción e Interpretación* 2: 167-183.
- Rogers, M. 2004. Multidimensionality in Concepts Systems: A Bilingual Textual Perspective. *Terminology* 10 (2): 215-240.
- Tarp, S. 2005. The Pedagogical Dimension of the Well-Conceived Specialised Dictionary. *Iberica* 10: 7-21.
- Tsuji, K, Kageura, K. 1998. An analysis of medical synonyms: the word-structure of preferred terms. *Terminology*, 5:2, 229-249
- Wiegand, H. E. (ed.) 1989. *Wörterbücher in der Diskussion* [I] (Lexicographica. Sieres Maior 27). Tübingen: M. Niemeyer.
- Wright, S. E. 1997. Representation of Concept Systems. *Handbook of Terminology Management: Basic Aspects of Terminology Management*. Amsterdam: John Benjamins, 89-97.

## Acknowledgements

This study was carried out in the framework of the R&D project VARIMED (FFI2011-23120) financed by the Spanish Ministry of Education, Culture and Sports.





# ***Bon usage* vs. Fachliches: Fachsprache in der Geschichte der französischen Sprachpflege und Lexikographie**

Martina Mayer  
Universität Innsbruck  
contact@martina-mayer.com

## **Abstract**

Frankreich ist seit dem 16. Jahrhundert dafür bekannt, sich besonders intensiv der Pflege seiner Nationalsprache zu widmen. Seit dem 17. Jahrhundert ist die französische Sprachpflege und Sprachnormierung institutionalisiert und wird durch entsprechende lexikographische Aktivitäten gestützt. Allerdings zielten all diese Bemühungen bis zur Mitte des 20. Jahrhunderts in erster Linie auf die Gemeinsprache ab – ein Umstand, der vergessen lässt, dass auch die Pflege der Fachsprachen in Frankreich eine nähere Betrachtung verdient. Dies bezieht sich nicht nur auf die modernen sprachpflegerischen Bemühungen an den Fachsprachen, sondern hat durchaus ab dem 16. bzw. 17. Jahrhundert Gültigkeit. Der vorliegende Beitrag wird sich dem Aspekt der Pflege der Fachsprachen und ihrer Berücksichtigung in der französischen Lexikographie daher aus einer historischen Perspektive widmen. Der Schwerpunkt liegt dabei vor allem auf entsprechenden Aktivitäten im *Grand Siècle*, das mit seinen Sprachpflegern, dem *bon usage* des *honnête homme*, der Gründung der *Académie française* und seinen lexikographischen Referenzwerken bis in die Gegenwart maßgeblich auf die französische Sprache einwirkte.

**Keywords:** Fachsprachen; Sprachpolitik; Sprachpflege; Lexikographie; Frankreich;

## **1 Einleitendes zur Sprachpflege und Sprachnormierung in Frankreich<sup>1</sup>**

Die beiden Schlagwörter *Sprachpflege* und *Sprachnormierung* begleiten Europa seit der Renaissance, in der viele europäische Sprechergemeinschaften mit einer intensiven Auseinandersetzung mit ihren noch jungen Nationalsprachen begannen. Das Ziel dieser von humanistischem Gedankengut beeinflussten Aktivität: die Nationalsprachen zunächst in möglichst vielen Domänen gemein- und fachsprachlicher Natur zu effizient einsetzbaren Kommunikationswerkzeugen auszubauen, um ihnen im Weiteren eindeutige Regeln zu verleihen. Das traf auch auf Frankreich zu: Nach dem von der Transiti-

---

1 Der vorliegende Beitrag beruht zum Teil auf der 2013 veröffentlichten Diplomarbeit *Sprachpflege und Sprachnormierung in Frankreich am Beispiel der Fachsprachen vom 16. Jahrhundert bis in die Gegenwart* (Mayer 2013), die in chronologischer Reihenfolge entsprechende Schlüsselmomente beleuchtet.

on vom Lateinischen zum Französischen sowie entsprechenden Elaborierungsbestrebungen geprägten 16. Jahrhundert begannen die französischen Sprachnormierer und Lexikographen des *Grand Siècle*, ihr sprachpflegerisches Augenmerk vor allem auf die präskriptive Definition des *bon usage* zu richten. Damit wurde im 17. Jahrhundert ein sprachpuristischer Kurs eingeschlagen, mit dem eine strikte Separation von Gemein- und Fachsprache einherging. Es sollten unbedingt Regelmäßigkeit und Reinheit in der Gemeinsprache gewährleistet werden, um einerseits die Schönheit sowie das Prestige des Französischen zu steigern und seine Tauglichkeit als das Konversationsmedium des Hofes sicherzustellen: Der Sprachnutzer bei Hof und dessen Sprachgebrauch standen im Mittelpunkt dieser Bemühungen; der fachlich gebildete *pédant* hingegen, mit seinen spezialisierten Kenntnissen und der entsprechenden Art zu kommunizieren, repräsentierte einen als geistlos geltenden Antityp, der bei Adel und Dichtern kaum Ansehen genoss. Andererseits sollte über den Umweg der sprachlichen Vereinheitlichung bzw. der Ausgrenzung unschicklicher Wörter sowie fachlichen Wortschatzes letztlich auch die politische Stabilität des Landes gefördert werden. Zugleich setzten intensive lexikographische Aktivitäten ein, die im Sinne der sprachpuristischen Standards der offiziellen französischen Sprachpolitik zumeist auf eine Beschneidung des Französischen ausgerichtet waren. Als maßgebliche normierende Instanz sei hier die *Académie française* genannt, der bis in die Gegenwart der Ruf einer altmodischen bzw. konservativen Einrichtung anhaftet. Diese sprachnormativen Bemühungen halten in Frankreich schon seit dem 17. Jahrhundert an. Damit kann die seit über vierhundert Jahren betriebene staatlich koordinierte Sprachpflege mit ihren lexikographischen Bestrebungen als ein französisches Phänomen bezeichnet werden, das eine gewisse Faszination auf die Sprachwissenschaft vor allem anderer Länder auszuüben scheint.

Der Forschungsbereich der Sprachpflege und Sprachnormierung in Frankreich wurde vor allem in Bezug auf die Gemeinsprache und die entsprechenden lexikographischen Referenzwerke umfassend analysiert. Weniger häufig konzentrierte sich die Forschung in diesem Zusammenhang bisher jedoch auf die fachsprachliche Komponente: Auf den ersten Blick mag es gar scheinen, als hätte in Frankreich vor dem 20. Jahrhundert kaum eine strukturierte Fachsprachenreflexion stattgefunden, als seien die Fachsprachen immer nur verdrängt bzw. ignoriert anstatt bewusst ins sprachpflegerische Denken einbezogen worden. Die aktuellen Versuche eines Ausbaus der originär französischen Fachlexik im Kontext einer *crise du français* durch die ministeriellen Terminologiekommissionen beispielsweise könnten in Bezug auf Fachsprache nahezu als einzige Manifestation sprachpflegerischen und -normativen Wirkens in der Geschichte Frankreichs wahrgenommen werden, was allerdings der Realität nicht gerecht würde. Dieser Beitrag wird daher nicht der ohnehin häufig besprochenen Gegenwart, sondern dem historischen Aspekt besondere Aufmerksamkeit widmen: Im Mittelpunkt steht das noch für die heutige französische Sprachpolitik und Lexikographie prägende 17. Jahrhundert mit seinen Sprachnormierern, dem Sprachideal des *honnête homme*, seinen lexikographischen Referenzwerken und der Haltung der *Académie française*. Damit verbunden ist natürlich auch die Frage, wie es überhaupt dazu kommen konnte, dass die offizielle Sprachpolitik im *Grand Siècle* kontinuierlich die

Fachsprachen zurückgedrängt hat. Die Ursache dieses Phänomens ist chronologisch früher anzusetzen.

## 2 Französische Fachsprachenreflexion vor dem 17. Jahrhundert

Reflektiert man die Entstehung von Fachsprache, so ist von einer Herauskristallisierung auf Basis eines spezialisierten Kommunikationsbedarfs durch menschliche Arbeitsteilung auszugehen. Eine solche bringt unabhängig von der Domäne menschlicher Tätigkeit notwendigerweise Fachkommunikation hervor, sei es in Form einer Werkstatt- oder einer Wissenschaftssprache, sei es auf mündlichem oder schriftlichem Wege. Mündliche Fachkommunikation prägte Europa bis in die frühe Neuzeit; dann stieg die Anzahl schriftlicher Zeugnisse von Fachkommunikation aus den verschiedensten Bereichen. Entsprechende Textbelege weisen sowohl theoretisch-wissenschaftlich orientierte (z.B. Chemie, Medizin, Philosophie, Recht) als auch praktisch-handwerklich ausgerichtete Fachsprachen auf, wobei das kommunikative Ziel jeweils ein fachinterner, also auf das jeweilige Kollektiv begrenzter Austausch bzw. eine entsprechende Konservierung von Inhalten war (Fluck 1996: 27). Interessant ist hierbei die sprachliche Aufteilung: Während sich wissenschaftliche Fachkommunikation nach Tradition der *septem artes liberales*, der Sieben Freien Künste, damals natürlich noch in erster Linie in lateinischer Sprache vollzog, lässt sich für die Fachkommunikation der *artes mechanicae*, also der handwerklichen Künste, schon von Anfang an eine Konzentration auf die Volkssprachen feststellen, was auch verständlich ist, mussten die fachsprachlichen Inhalte doch der richtigen Zielgruppe zugänglich sein – in Bezug auf Werkstattsprachen nun eben dem „einfachen Volk“, sofern es überhaupt lesen konnte (Berschlin/Felixberger/Goebel 2008 198; Roelcke 2005: 162ff).

Mit dem 16. Jahrhundert erfuhr nun Frankreich im Bereich der Sprachpolitik einen ersten entscheidenden Umbruch bezüglich dieser Zielgruppenorientierung bzw. Vulgarisierung: Mit der *Ordonnance de Villers-Cotterêts* vom 25. August 1539 wurde der „*langage maternel françois*“ durch François I<sup>er</sup> als offizielle Gerichts-, Urkunden- und damit Amtssprache festgesetzt, wodurch der breiten Bevölkerung ganz neue Inhalte zugänglich gemacht wurden.<sup>2</sup> Damit war der sich in den verschiedensten fachlichen Domänen fortan zügig vollziehende Ersatz des Lateinischen durch das Französische besiegelt (Müller 1975: 36). Pécheur (2001: 51) spricht hier bildhaft von einem „[...] embryon de politique linguistique [...]“, dessen Ziel eine Festschreibung des verpflichtenden Gebrauchs der französischen Sprache im Sinne einer erstmaligen Normierung als Standard war. Im Folgenden wurde Joachim du Bellay (1522-1560) zum ersten Verfechter der französischen Sprachpflege. Der Titel seiner 1549 erschienenen *Deffen-*

---

2 Die *Ordonnance de Villers-Cotterêts* umfasst 192 Artikel, von denen vor allem die Artikel 110 und 111 maßgeblichen Einfluss auf die Entwicklung des Französischen hatten. Nachzulesen sind sie im vollständigen Text der Verordnung auf der Website der *Assemblée Nationale* (<http://www.assemblee-nationale.fr/histoire/villers-cotterets.asp>, Stand: 5.4.2014).

*ce et Illustration de la Langue Francoyse* ist programmatisch: Das Französische sei dem Lateinischen gegenüber von der Basis her gleichwertig und müsse gegen den Vorwurf eines mangelnden Kommunikationspotenzials verteidigt werden – eine entsprechende Pflege im Sinne eines umfassenden Sprachausbaus sowohl der Literatur- als auch der Fachsprache sei allerdings notwendig und käme den Sprachnutzern selber zu (du Bellay 1969: 47f). Zu du Bellays Zukunftszielen zählte außerdem eine kontinuierliche Etablierung des Französischen in weiteren, über die Fachsprache des Rechts hinausgehenden spezialisierten Kontexten der Wissenschaften und des Handwerks, sodass über kurz oder lang alle Inhalte allen Sprachnutzern gegenüber demokratisiert werden sollten (du Bellay 1969:133ff). Du Bellays Vision einer üppigen französischen Lexik der vielseitigen kommunikativen Möglichkeiten wurde Realität; außerdem hatte das 16. Jahrhundert bereits eine ganze Reihe von Glossaren und frühen Wörterbüchern zu bieten, wie Lindemanns (1994) umfassende Bestandsaufnahme *Die französischen Wörterbücher von den Anfängen bis 1600* eindrucksvoll zeigt. Allerdings hatte es die französische Sprechergemeinschaft durch den sehr freizügigen Sprachausbau der Renaissance selbst trotz der bereits bestehenden lexikographischen Aktivitäten in Folge mit einem fast überbordenden und uneinheitlich verwendeten Wortschatz zu tun.

### 3 Der Sprachpurismus des *Grand Siècle*

Dieser nicht unproblematischen sprachlichen Situation gegenüber fanden sich nun die Sprachnutzer wie auch die Sprachnormierer des *Grand Siècle* wieder. Vor allem Letztere reagierten darauf mit einer puristischen Grundhaltung sowie stark normativen Bestrebungen, was der Stellung der Fachsprachen innerhalb der französischen Sprachpflege und Sprachpolitik nicht unbedingt zuträglich war, zumindest was die offizielle Lexikographie betraf. Dies erklärt sich allerdings schlichtweg aus den sprachpolitischen Bedürfnissen eines zunehmend zentralistisch organisierten Staates, der eine einheitliche, universell einsetzbare und starke Sprache benötigte, um sich von innen heraus zu stabilisieren.

#### 3.1 François de Malherbe

Die wohl erste und eine der wichtigsten unter diesen normierend eingreifenden Persönlichkeiten war der Hofdichter François de Malherbe (1555-1628) mit seiner Suche nach sprachlicher Reinheit und seiner Kategorisierung der Wörter in drei Gruppen: Er klassifizierte zunächst die schicklichen *mots nobles*, weiters die aus diversen Gründen (z.B. aufgrund eines Konnexes zur Welt der Arbeit) zumindest für den Adel unverwendbaren *mots bas* und schließlich die tabuisierten, da auf Körperfunktionen oder den Bereich der Sexualität bezogenen *mots sales*. Damit schloss er im Grunde sämtliche fachsprachliche Begriffe, die großteils natürlich als *mots bas* zu betrachten waren, von einem literarischen Gebrauch aus, erklärte sie als nicht salonfähig und verbannte sie damit kollektiv aus dem Sprachgebrauch der führenden Schicht wie auch der Autoren, worin Brunot (1930: 3-9) ganz eindeutig einen Rückschritt für die französische Sprache

erkannte: Schließlich gingen dadurch in dieser Zeit, die auch noch nicht von umfassenden lexikographischen Bemühungen durchdrungen war, zahlreiche Wörter verloren, die nirgends verzeichnet waren und somit tatsächlich aus dem Französischen verschwanden. Eines ist Malherbe diesbezüglich allerdings nicht vorzuwerfen: Er versuchte zwar, die Literatursprache und die Kommunikation in nicht-fachsprachlichen Kontexten von Terminologie und anderen „Störfaktoren“ der Lexik zu reinigen, verfolgte aber nicht das Ziel, normativ in den Bereich der Fachkommunikation selber vorzudringen.

Die von Malherbes Bestrebungen direkt angesprochenen Gruppen, die *noblesse* und das literarische Milieu, akzeptierten jedenfalls seine linguistischen Entscheidungen großteils – bzw. hätte es anders betrachtet auch gar keine Möglichkeit gegeben, in irgendeiner Weise dagegen vorzugehen: Malherbes normatives Wirken war schließlich durch den König gedeckt und war somit stillschweigend zu akzeptieren. Inwiefern? Während es dem Landadel ohnehin an Mitteln und an Einfluss fehlte, so war der Hochadel durch den absolutistischen Willen des Königs an ein starres Gerüst aus Regeln gebunden, welches das Leben bei Hofe dominierte und als obligatorisch zu betrachten war; bei Missachtung drohte laut Elias (1983: 353f) eine „Einschränkung oder Verminderung der königlichen Gunst“ und damit „eine schwer zu ertragende Gefahr“, die ganz präzise in einem Verlust an sozialem Prestige und folglich einer Verschlechterung der Zukunftschancen, der Lebensverhältnisse usw. bestand. Das Konzept des höfischen Menschen sah es nicht vor, dass ein *honnête homme* jemals Unabhängigkeit vom Hof erlangen sollte: Er musste dem König unterworfen bleiben, damit dessen absolute Macht keinesfalls angetastet werden konnte. Diese Unterwerfung des Adels wiederum konnte nur solange funktionieren, als dieser keine Möglichkeit hatte, sein Leben dank der Anwendung etwaiger nicht ausschließlich höfischer Kompetenzen oder Fachkenntnisse in finanzieller Hinsicht eigenständig zu bestreiten. Während also die fachliche und damit auch finanzielle Unselbständigkeit des *honnête homme* als grundlegende Bedingung des Erhalts der königlichen Macht bezeichnet werden kann, war für den Adel umgekehrt der Erhalt der königlichen Gunst durch eine demonstrative Anpassung an das Regelkorsett des Hofes die grundlegende Bedingung für den Fortbestand seines Lebensstils. Ein guter Grund für jeden *honnête homme*, zum eigenen Schutz nicht einmal verbal auf das Vorhandensein etwaiger spezialisierter Kenntnisse hinzuweisen und sich präventiv dem präskriptiven Sprachgebrauch des *bon usage* zu beugen, der vollkommen auf die mündliche Konversation am Hof bzw. in den Pariser Salons ausgerichtet war, das gesellschaftliche und zugleich schöngeistige Leben durchdrang und alles Fachliche ausgrenzte (Grimm 2005: 145; Hassler 1998: 323f). Da sich die Malherbesche Beschneidung der Sprache in ausgerechnet diesem historischen Kontext situierte, der stark von sozialen Gegebenheiten und Machtverhältnissen determiniert und vom Einsetzen der französischen Klassik durchdrungen war, fiel sie auf fruchtbaren Boden und fand eine entsprechende Nachfolge.

### 3.2 Claude Favre de Vaugelas

Claude Favre, baron de Pérouges et seigneur de Vaugelas (1585-1650), der sich nicht nur als Mitglied der *Académie française*, sondern auch individuell sprachnormativ betätigte, stellte sein Wirken eben-

falls in das Zeichen des normativen Purismus, wenngleich unter anderen Vorzeichen als Malherbe. Im Rahmen seiner *Remarques sur la langue française, utiles à ceux qui veulent bien parler et bien écrire* aus dem Jahr 1647, die man unter Rückgriff auf die Kategorien der heutigen französischen Lexikographie am ehesten mit einem *dictionnaire des difficultés* vergleichen kann, prägte Vaugelas (1880: 12f) die Begriffe des *bon usage* sowie des *mauvais usage* erstmals in einer schriftlichen Definition, die keinen Zweifel darüber ließ, dass *bon* und *mauvais* für ihn von einem elitären Sprachverständnis und sozialer Exklusion ausgehend zu deuten waren. Wie bei Malherbe stand auch für Vaugelas eine Politik der Ab- und Ausgrenzung im Mittelpunkt, allerdings war diese anders ausgerichtet: Während Malherbe die Reinheit und Klarheit der Sprache in ganz bestimmten Bereichen sprachlicher Äußerung gewährleisten wollte, so ging es Vaugelas darum, den sprachlich-modellhaften Charakter einer Elite von Sprachnutzern als Optimum darzustellen. Vaugelas konnte sich darin natürlich auf das Resultat von Malherbes Vorarbeit stützen – den bereits reinen und schönen Sprachgebrauch des Hofes, dem von nun an unumstrittene Vorbildhaftigkeit zugeschrieben wurde und der bereits ohne fachsprachliche Elemente auskam, allerdings bis auf eine große Ausnahme: den höfisch-fachlichen Wortschatz, der sich auf den typischen Zeitvertreib des *honnête homme* und das Leben bei Hofe bezog.

Dementsprechend finden sich in den *Remarques* nur vereinzelt fachsprachliche Elemente. Im Allgemeinen zögerte Vaugelas kaum, sich in Hinblick auf den Gebrauch einzelner Wörter – auch in Hinblick auf Fachsprache – sprachnormativ ganz eindeutig zu positionieren, um Vorgaben durchzusetzen, die er vom Sprachgebrauch der „*plus saine partie de la Cour*“ (Vaugelas 1880:12f) ableitete: beispielsweise dann, wenn Vaugelas (1880: 144) dem auch heute noch verwendeten Begriff *naviguer* aus dem Technolekt der Seeleute zu Gunsten der vom Hof präferierten phonetisch-orthographischen Variante *naviger* ohne objektive Begründung seine Richtigkeit abspricht; wenn er sich ganz offensichtlich über sämtliche etymologische Grundlagen hinwegsetzt, indem er aus *tempe*, der Schläfe, ohne entsprechende Argumentation *temple* macht (Vaugelas 1880: 266) oder seiner sonstigen Haltung gegenüber völlig konträr den fachsprachlichen Neologismus (für Vaugelas‘ übliche Maßstäbe eigentlich zwei Ablehnungskriterien in einem) *conjoncture* in seine *Remarques* aufnimmt und sich diesem Begriff gegenüber noch dazu höchst positiv gestimmt zeigt (Vaugelas 1880: 345). Vaugelas‘ Arbeit diente wiederum als Grundlage weiterer normativer Ansätze.

### 3.3 Dominique Bouhours

Auch Dominique Bouhours (1628-1702) war sprachnormativ tätig, entfernte sich zunächst in seiner schriftlichen Befassung mit dem Thema allerdings von der durch Vaugelas im weitesten Sinne vorgegebenen Form und bediente sich in *Entretiens d'Ariste et Eugène* aus dem Jahr 1671 des Lehrdialogs, um zwei sprachpflegerische Grundhaltungen einander gegenüber zu stellen: einerseits das Bewusstsein, dass der Bedarf einer Emanzipation des Französischen anderen Sprachen gegenüber (Spanisch, Italienisch und Latein) bestünde, die durch Perfektion erreichbar sei, und andererseits den Gedanken, dass bei aller Hinwendung zu ebendieser Perfektion aber nicht die kommunikativen Bedürfnisse der Spra-

che ignoriert werden dürfen; fachsprachlicher Wortschatz sei demnach keinesfalls zu vernachlässigen – schon allein deshalb, weil sogar die für das Leben bei Hof wichtigen Sprachen der Jagd, der Falknerie oder der Kunst auch nicht ohne Fachlichkeit auskommen konnten (Bouhours 1962: 47ff). Bouhours bewies damit also durchaus ein realistisches Gespür für die Notwendigkeiten einer lebenden Sprache, was sich auch an seinen sprachnormativen Folgewerken feststellen lässt: Die *Remarques nouvelles sur la Langue française* aus dem Jahr 1675 wie auch die *Suite des Remarques nouvelles sur la Langue française* von 1687 folgten sowohl dem Titel nach als auch von der Struktur her betrachtet Vaugelas' direktem Vorbild, beriefen sich auch auf den *bon usage*, konzentrierten sich aber unter anderem mit Rückgriff auf Autorenzitate vermehrt auf die Sprache der schöngeistigen Literatur und hatten nicht den Anspruch, eine elitäre Sprachauffassung unter Abwertung bestimmter sozialer Gruppen bzw. unter absolutem Ausschluss der von diesen Gruppen verwendeten Technolekte zu propagieren.

### 3.4 Die Académie française

Der 1635 gegründeten *Académie française* hingegen wurde und wird immer wieder unterstellt, sie verfolge ein derartiges Ziel, habe sich in ihrem sprachnormativen und lexikographischen Wirken der systematischen Bekämpfung allen fachlichen Wortschatzes verschrieben und vertrete damit eine Grundhaltung, die einer modernen Sprache gegenüber nicht angemessen sei. Nun mag es zwar stimmen, dass die *Académie française* seit der Definition des guten Sprachgebrauchs des *honnête homme* durch eine entsprechende Gestaltung ihrer Wörterbücher fachsprachliche Elemente konsequent aus der französischen Sprache zurückdrängt; jedoch wäre es wünschenswert, hier zu einem etwas differenzierteren Bild zu finden. Einen ersten Ansatzpunkt dafür bietet ein Blick in den Artikel XXIV der am 22. Februar 1635 beschlossenen Gründungsstatuten der Akademie:

La principale fonction de l'Académie sera de travailler avec tout le soin et toute la diligence possibles à donner des règles certaines à notre langue et à la rendre pure, éloquente et capable de traiter les arts et les sciences<sup>1</sup>.

<sup>[1]</sup> Article essentiel qui formule la raison d'être de l'Académie, lui prescrit sa mission et fonde son autorité.

(Académie française s.a.: [sites/academie-francaise.fr/files/statuts\\_af.pdf](http://sites/academie-francaise.fr/files/statuts_af.pdf), Hervorhebung des Absatzes weggelassen, MM)

Obleich eine intensive Pflege der Fachsprachen aus den im Kapitel 3.1 bereits dargelegten machtpolitischen Gründen nicht wirklich mit dem höfischen Sprachideal zu vereinbaren und die *Académie française* diesbezüglich ganz sicher Kompromissen unterworfen war, stellen also zumindest die Gründungsdokumente der Akademie eine gewisse Sensibilität für die kommunikativen Anforderungen unter Beweis, die sich einer in Entwicklung befindlichen Sprache und Sprechergemeinschaft stellen. Eine völlige Abkoppelung fachsprachlicher Inhalte war nie von Grund auf intendiert – ganz im Ge-



genteil, was die Statuten auch beweisen –, ergab sich aber wohl in Folge aus dem von Vaugelas postulierten Sprachideal. Das 1694 erstmalig erschienene Akademiewörterbuch, der *Dictionnaire de l'Académie française*, erwies sich dementsprechend auch als rein gemeinsprachlich ausgerichtete lexikographische Werk, was der Akademie allerdings kaum vorzuwerfen ist: Schließlich steht für ein Wörterbuch wie für jedes andere auf ein gewisses Ziel ausgerichtete Werk der Rezipient, oder in diesem Falle präziser der Nutzer, im Fokus – und die Zielgruppe bestand in diesem Fall nun einmal aus dem Adel und den Dichtern jener Zeit, für die ein klar prädefinierter Normgebrauch galt, dem der *Dictionnaire* wiederum gerecht werden musste. Mit dieser erstmaligen, zumindest aus nachvollziehbaren Gründen den Fachsprachen gegenüber restriktiven Positionierung ihres lexikographischen Wirkens stellte nun die *Académie française* zwar die Weichen für ihre gesamte weitere Redaktion von Wörterbüchern, blieb allerdings im Lauf der Zeit in Bezug auf ihre Konzeption der Sprachpflege bei den Fachsprachen auch nicht so vollkommen statisch wie es ihr regelmäßig vorgeworfen wird. Dies lässt sich anhand einer diachronischen Analyse der Paratexte der Akademiewörterbücher nachvollziehen.

Das Vorwort der ersten Ausgabe des Akademiewörterbuchs von 1694 zeugt ganz klar von der Zielgruppenorientierung des Werkes (wenn im folgenden Zitat vom *Discours* die Rede ist, so bezieht sich dies natürlich exklusiv auf jenen der *honnestes gens*) und außerdem vom Anspruch, als ein Medium des kollektiven linguistischen Gedächtnisses die französische Sprache in ihrem Istzustand zu illustrieren:

[...] C'est dans cet estat où la Langue Françoisse se trouve aujourd'huy qu'a esté composé ce Dictionnaire; & pour la représenter dans ce mesme estat, l'Académie a jugé qu'elle ne devoit pas y mettre les vieux mots qui sont entierement hors d'usage, ni les termes des Arts & des Sciences qui entrent rarement dans le Discours; Elle s'est retranchée à la Langue commune, telle qu'elle est dans le commerce ordinaire des honnestes gens, & telle que les Orateurs & les Poëtes l'employent; Ce qui comprend tout ce qui peut servir à la Noblesse & à l'Elegance du discours. Elle a donné la Definition de tous les mots communs de la Langue dont les Idées sont fort simples; & cela est beaucoup plus mal-aisé que de définir les mots des Arts & des Sciences dont les Idées sont fort composées; Car il est bien plus aisé, par exemple, de définir le mot de Telescope, qui est une Lunette à voir de loin, que de définir le mot de voir; Et l'on esprouve mesme en definissant ces termes des Arts & des Sciences, que la Definition est toujours plus claire que la chose definie; au lieu qu'en definissant les termes communs, la chose definie est toujours plus claire que la Definition. [...] (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 1<sup>re</sup> préface)

Vor dem Hintergrund, dass die Gemeinsprache der Fachsprache als ein Subsystem ihrer selbst die notwendigen grammatischen und lexikalischen Rahmenbedingungen vorgibt, ist auch der Verweis auf definatorische Schwierigkeiten bei gemeinsprachlichen Lemmata interessant, deren lexikographische Erfassung laut dem obigen Auszug eine größere Leistung darstellt als jene von Terminologie: Fachsprache kann ohne Gemeinsprache nicht existieren; Gemeinsprache hingegen wird durch Fach-



sprache maximal durch die Übernahme neuen Wortgutes aus der fachlichen Domäne in die gemeinsprachliche beeinflusst. Dies hat auch die *Académie française* indirekt berücksichtigt:

L'Académie en bannissant de son Dictionnaire les termes des Arts & des Sciences, n'a pas creu devoir estendre cette exclusion jusques sur ceux qui sont devenus fort communs, ou qui ayant passé dans le discours ordinaire, ont formé des façons de parler figurées; comme celles-cy, Je luy ay porté une botte franche. Ce jeune homme a pris l'Essor, qui sont façons de parler tirées, l'une de l'Art de l'Escrime, l'autre de la Fauconnerie. On en a usé de mesme à l'esgard des autres Arts & de quelques expressions tant du style Dogmatique, que de la Pratique du Palais ou des Finances, parce qu'elles entrent quelquefois dans la conversation. (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 1<sup>re</sup> préface)

Das Vorwort der zweiten Auflage des Wörterbuchs von 1718 zeigte ein ähnliches Bild (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 2<sup>e</sup> préface); mit der dritten Auflage ab 1740 hingegen begann die *Académie française* auf eine zusätzliche Öffnung den Fachsprachen gegenüber hinzuweisen, die mit dem nun schon sehr weit fortgeschrittenen Ersatz des Lateinischen durch das Französische in Zusammenhang stand: Die französische Sprache hatte zahlreiche neue Fachbereiche erobert, und entsprechendes Vokabular war von diesen Fachsprachen aus in die Gemeinsprache übergegangen – ein Phänomen, dem auch die ausdrücklich gemeinsprachlich ausgerichtete Lexikographie gerecht werden musste (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 3<sup>e</sup> préface). Die vierte Auflage von 1762 bekräftigte dies und räumte einen weiteren Aspekt ein: Da die schöngeistige Literatur inzwischen auch wissenschaftliche und fachliche Themen behandle, sei der *honnête homme* in seinem Zeitvertreib zunehmend mit entsprechender Lexik konfrontiert. Im Sinne einer Ausrichtung des Werkzeugs Wörterbuch auf die Bedürfnisse des Nutzers schien es damit wiederum geboten, zusätzliche fachsprachliche Elemente in den *Dictionnaire* zu übernehmen. Diese Aussage zeigt doch, dass die *Académie française* dem tatsächlichen Sprachgebrauch gegenüber nicht vollkommen unsensibel war; weiters zeugt sie von einem bedarfsbedingten Konnex zwischen lexikographischen Bestrebungen und schöngeistiger Literatur (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 4<sup>e</sup> préface). Die Préface der fünften Auflage von 1789 brachte diesbezüglich wieder keine Neuerungen ein (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 5<sup>e</sup> préface); die sechste Auflage hingegen bewies mit ihrem Vorwort, dass die Akademie durchaus auch zu Selbstreflexion in der Lage war (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 6<sup>e</sup> préface): 1835 stellte die *Académie française* fest, dass ihr *Dictionnaire* nur über ein sehr begrenztes fachsprachliches Repertoire verfügte und dass dies ein Versäumnis sei – allerdings ein zu entschuldigendes, seien doch Nomenklaturen ohnehin schnell veraltet und habe nur die Literatursprache wirklich Bestand. Demnach wäre es sinnlos gewesen, Energie auf die Integration fachlicher Lemmata zu verschwenden. Zugleich sei die den Fachsprachen gegenüber restriktive Haltung aber doch zu stark beibehalten worden, außer in Bereichen wie Wappenkunde und Jagdwesen, die dem *Dictionnaire* einzelne Termini wie auch stehende Wendungen beschert hätten. Mit der siebten Auflage des Wörterbuchs von 1878 zeigte die

Akademie sich schließlich den Fachsprachen gegenüber noch offener: 2200 neue Lemmata seien zum Wörterbuch hinzugekommen, wobei auch den Fachsprachen ein gewisser Raum zugestanden worden sei. Vor allem Felder wie die Philosophie, die Archäologie, die Philologie, die Politikwissenschaften, die Industrie und die Landwirtschaft seien nun berücksichtigt worden, wobei diese lexikographische Leistung aber nicht ohne Vorbehalte der eigenen Haltungsänderung gegenüber zustande gekommen sei:

Naturellement la part des sciences et des inventions nouvelles a été grande dans les deux mille mots ajoutés. Les chemins de fer, la navigation à vapeur, le télégraphe électrique ont fait irruption dans notre bon vieux français, avec leurs dénomination [sic] d'une forme souvent bizarre ou étrangère; force a été d'admettre: *un télégramme, un steamer, un tunnel, des tramways*: l'ombre de nos prédécesseurs a dû plus d'une fois en frémir. (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 7<sup>e</sup> préface).

Das Vorwort der achten Auflage von 1935 unterstrich dies, indem es zwar über die Aufnahme zahlreicher spezialisierter Lemmata berichtete, sich allerdings negativ zu den Fachsprachen äußerte, die sich kontinuierlich und maßlos ausbreiten würden. Die neu aufgenommenen Termini seien daher sehr bewusst ausgewählt worden, als Kriterium habe die vermutete Dauer ihres Bestehens gegolten. Außerdem sei es natürlich nach wie vor Aufgabe der Akademie, ein Wörterbuch des *bon usage* zu redigieren (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 8<sup>e</sup> préface). Die neunte Auflage des Akademiewörterbuches ist noch im Entstehen begriffen, liegt derzeit bis zum Eintrag *réglage* vor und umfasst im Vergleich zur vorherigen Auflage etwa 10 000 zusätzliche Lemmata (Académie française s.a.: Le Dictionnaire – La 9<sup>e</sup> édition). Die Aufnahmekriterien sind, laut eigenen Angaben, wieder höchst pragmatisch angelegt; sie orientieren sich erneut daran, welche fachsprachlichen Elemente bereits in die Gemeinsprache übergegangen sind:

Nous ne donnons entrée, parmi les termes techniques, qu'à ceux qui, du langage du spécialiste, sont passés par nécessité dans le langage courant, et peuvent donc être tenus pour réellement usuels. (Académie française s.a.: Le Dictionnaire – Les neuf préfaces – 9<sup>e</sup> préface)

In Summe zeigt sich also, dass die Akademie zwar einerseits an einem konservativen Konzept von Sprachpflege festhält, dass sie sich andererseits aber auch nicht vollständig der Notwendigkeit einer Aktualisierung ihres "allgemeingültigen" Wortschatzes verschließt. Eine Institution wie die *Académie française* kann sich letztlich nicht über die Jahrhunderte hinweg völlig wirklichkeitsfremd zeigen, indem sie einen tatsächlichen Kommunikationsbedarf ignoriert.

### 3.5 Die Fachwörterbücher des 17. Jahrhunderts

Dennoch fand Fachlexik damals zunächst in lexikographischen Werken Beachtung, die nicht auf Initiative der staatlichen Sprachpolitik zurückgingen. Zu nennen ist beispielsweise César-Pierre Richelets *Dictionnaire François, contenant les Mots et les Choses* von 1680 und außerdem Antoine Furetières *Dic-*

*tionnaire universel, contenant généralement tous les Mots François* von 1690. Vor allem das lexikographische Werk des Letzteren wurde von der *Académie française* als Konkurrenz wahrgenommen und löste die sogenannte *Bataille des dictionnaires* aus, die schließlich indirekt in der Redaktion eines dritten großen fachsprachlichen Wörterbuches des 17. Jahrhunderts gipfelte, des *Dictionnaire des Arts et des Sciences* von Thomas Corneille.

Während die Akademie sich darauf konzentrierte, 1694 in ihrem ersten Wörterbuch den tatsächlichen Sprachgebrauch ihrer primären Zielgruppe abzubilden, veröffentlichte César-Pierre Richelet schon 14 Jahre vorher ein zweibändiges Fachwörterbuch mit dem Titel *Dictionnaire François, contenant les Mots et les Choses : Plusieurs nouvelles Remarques sur la Langue Française : Ses Expressions Propres, Figurées & Burlesques, la Prononciation des Mots les plus difficiles, le Genre des Noms, le Regime des Verbes : Avec les Termes les plus connus des Arts & des Sciences. Le tout tiré de l'Usage et des bons Auteurs de la Langue française*. Dieses lexikographische Werk war vom *bon usage* losgelöst, obgleich es derselben Zielgruppe gewidmet war wie die Akademiewörterbücher. Richelet verzeichnete darin auch *mots sales* bzw. *mots bas* und legte ein besonderes Augenmerk auf die Fachlexik, dies laut eigener Auskunft im Vorwort, um der Zielgruppe ein noch nützlicheres Hilfsmittel an die Hand zu geben (Richelet 1680: Avertissement, ohne Paginierung). So bot Richelet ein im Vergleich zu den Akademiewörterbüchern vollständigeres Bild der damaligen französischen Sprache und erlebte in Frankreich auch einen gewissen Erfolg; vor allem konnte er aber von der *Académie française* unbehelligt arbeiten.

Antoine Furetière erregte mit seinem 1684 teilweise, 1690 postum vollständig veröffentlichten *Dictionnaire universel, contenant généralement tous les Mots François, tant vieux que modernes, & les Termes de toutes les Sciences et des Arts* hingegen den Unmut der Akademie: All ihren Grundhaltungen entgegen hatte das Akademiemitglied Furetière die Absicht, die französische Sprache mit ca. 40 000 Lemmata im Sinne der Vermittlung von enzyklopädischem Wissen so vollständig wie möglich zu illustrieren, was natürlich auch Fachsprache einschloss. Furetière erhielt für sein Projekt 1684 tatsächlich eine königliche Druckerlaubnis und publizierte daraufhin eine Vorschau auf sein Wörterbuch. Damit überholte er sozusagen die *Académie française*, die ihrerseits seit 1674 ein königliches Druckprivileg für die Erstellung eines französischen Wörterbuches innehatte, aber für die Redaktion ihres Werkes sehr viel Zeit benötigte. Die Akademie erhob Furetière gegenüber nun Plagiatsvorwürfe, schloss ihn aus ihren Reihen aus und erreichte auch die Aufhebung seines Druckprivilegs. Der *Dictionnaire universel* erschien trotzdem, allerdings in Den Haag und erst nach dem Tod seines Autors (Chassagne et al. 1994:63).

Diese Geschehnisse brachten allerdings die *Académie française* in Zugzwang: Sie musste den Fachsprachen daraufhin etwas mehr Raum geben und beauftragte Thomas Corneille 1690 mit der Redaktion eines separaten *Dictionnaire des Arts et des Sciences*, das als Ergänzung zum damals noch in Ausarbeitung befindlichen fachsprachenrestriktiven Akademiewörterbuch gedacht war und ebenfalls 1694 erschien. Bemerkenswert ist die extrem kurze Redaktionszeit dieses weniger enzyklopädisch, sondern hauptsächlich definitorisch angelegten Wörterbuches – ein Umstand, der nun wiederum Corneille den Vorwurf eintrug, Richelets Wörterbuch plagiiert zu haben (Chassagne et al. 1994: 58). Das Vorwort des *Dictionnaire des Arts et des Sciences* befasste sich wiederum in Form einer harschen Kritik intensiv

mit dem Wörterbuch von Furetière und einer ganzen Reihe an Defekten, die selbiges nach Meinung der Akademie aufwies – ganz anders als das *Dictionnaire des Arts et des Sciences* selber, das laut Aussage seines Autors mit höchster Sorgfalt erstellt worden war (Corneille 1694: Préface, ohne Paginierung).

## 4 Fachsprachen im Frankreich des 18. Jahrhunderts und danach

Wie bereits unter 3.4 festgestellt, passte die *Académie française* ihre lexikographische Ausrichtung in den Folgejahren aber jeweils nur so geringfügig wie möglich an die sich ändernde Situation an und nahm im 18. Jahrhundert auch nicht aktiv am Umbruch hin zum *âge encyclopédique* teil: Im Zuge einer generellen Hinwendung der Öffentlichkeit zu Fachlichem und Fachsprachlichem bildete sich damals das Bewusstsein heraus, dass Fachsprache und Gemeinsprache nicht strikt voneinander getrennt werden können. Der bis dahin vorherrschende, fachsprachenfreie *style noble* wich nun einem neuen Sprecherverhalten, das sich durch die zunehmende Verwendung von Fachsprache in nicht-fachlichen Kommunikationssituationen der gehobenen Schichten auszeichnete. Damit ging wiederum ein Aufschwung der Fachwörterbücher einher: Das für das 18. Jahrhundert wichtigste fachsprachliche lexikographische Werk war zweifelsohne Diderots und d’Alemberts vielimitierte *Encyclopédie*, deren Ziel es war, sämtliches Wissen ihrer Zeit zu erfassen.

Etwas später, während der Französischen Revolution, entstand dann aus einer politischen Motivation heraus ein starkes wissenschaftliches Interesse für die sprachlichen Gegebenheiten in Frankreich, das in der großen Sprachumfrage des Abbé Henri Grégoire seine Verkörperung fand: Im Hinblick auf die Fachsprachen versuchte er vor allem herauszufinden, in welchen Bereichen die französische Standardvarietät ein zu schwaches Kommunikationspotenzial aufwies – es waren schließlich wie bereits dargestellt auch aufgrund der sprachpuristischen Bestrebungen des 17. Jahrhunderts sprachliche Lücken entstanden. Nach einer vier Jahre andauernden Datenerhebung zeigte sein *Rapport sur la nécessité et les moyens d’anéantir les patois et d’universaliser l’usage de la langue française*, den er am 4. Juni 1794 der Nationalversammlung vorlegte, nicht unerhebliche Problemstellen im Bereich der Fachsprachen: Die Sprechergemeinschaft musste zur Gewährleistung einer effizienten Kommunikation beispielsweise in der Handwerker- und Bauernsprache bzw. in Fächern, die Berührungspunkte mit dem Alltag der einfachen Bevölkerung aufwiesen, häufig auf Begriffe aus Regiolekten zurückgreifen, da es schlichtweg keine standardfranzösischen Ausdrucksmittel gab. Diese Tendenz machte Lösungsstrategien zugunsten einer einheitlichen und starken Sprache im Sinne einer geeinten und starken Nation notwendig. Der Wandel weg von regional geprägten Termini, hin zu einer panfranzösisch einheitlichen Fachsprachverwendung, sollte mittels einer kostenlosen Grundschulausbildung für alle Kinder in französischer Sprache, der Übersetzung fachsprachlichen Wortgutes aus den Patois ins Normfranzösische und durch die Verbreitung entsprechender fachsprachlicher Texte mit Relevanz für das all-

tägliche Leben erreicht werden (Bochmann 1993: 64ff; ib.: 88; ib.: 139). Die offizielle französische Lexikographie trug allerdings nur wenig dazu bei.

Nicht nur für das 18. Jahrhundert, sondern auch noch bis in die Gegenwart lässt sich feststellen, dass mehr als vier Jahrhunderte fachsprachenrestriktiver Sprachpolitik in Frankreich Spuren in Form von sprachlichen Lücken hinterlassen haben. Obgleich sich heute neue Werke der französischen Lexikographie den Fachsprachen gegenüber aufgeschlossener zeigen und auch die institutionalisierte Sprachpflege sehr bemüht ist, vorhandenen Problemen entgegenzuwirken, strömen etwa seit Ende des 19. Jahrhunderts zahlreiche Fremdwörter – vornehmlich aus dem Englischen – ins Französische, um dort vor allem fachsprachliche Defizite auszugleichen. Von dort aus finden sie anschließend häufig Eingang in die Gemeinsprache, die sie immer stärker durchsetzen. Dies hat zu einem Aufruf zur „Verteidigung der Sprache“ geführt. Dieser Kampf hat bis heute nichts von seiner Vehemenz verloren: Eine moderne Sprachgesetzgebung sowie eine beachtliche Anzahl staatlicher und privater Sprachpflegeorganisationen stehen im Dienst des Schutzes des Französischen, was unter anderem einen koordinierten Ausbau der französischen Fachsprachen und ein entsprechendes lexikographisches Wirken impliziert.

## 5 Resümee

Zusammenfassend lässt sich feststellen, dass die französische Sprache im europäischen Vergleich durchaus eine Art Sonderstatus innehat. In kaum einem anderen Land wird die Tradition der Sprachpflege so großgeschrieben und so konsequent betrieben; in kaum einem anderen Land bringen die Sprachnutzer ihrer Nationalsprache so viel Aufmerksamkeit entgegen. Zugleich war die französische Sprachpflege allerdings lange Zeit über stark auf die Gemeinsprache fokussiert und vernachlässigte die Fachsprachen. Deshalb wird ihr in ihrer institutionalisierten Form, der *Académie française*, auch immer wieder vorgeworfen, sie würde den Anforderungen einer modernen Sprache im Wandel nicht gerecht; sie praktiziere eine Art der lexikalischen Zensur, die der Sprache ernst- und dauerhaft Schaden zufügen könne. In der Tat schlug sich das sprachpolitische Eingreifen der Akademie seit dem 17. Jahrhundert mit immer noch anhaltender Wirkung im tatsächlichen Sprachgebrauch nieder: Die französische Sprache ist eine vereinheitlichte Sprache, eine wahre Standardvarietät, die allerdings gerade ob der zur Vereinheitlichung und ästhetischen Perfektionierung gesetzten restriktiven Maßnahmen teils nach wie vor unter Defiziten in der Lexik leidet. Für das Heute gilt, dass sich der für das Französische typische, lebhaft sprachpolitische Diskurs im Lauf der Zeit inhaltlich gewandelt hat, aber immer noch von der immensen Bedeutung zeugt, die der Pflege und Wertschätzung der Nationalsprache in Frankreich zugesprochen wird.

## 6 Primär- und Sekundärliteratur

- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *1re préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neuf-prefaces/preface-de-la-premiere-edition-1694> [06/04/2014].
- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *2e préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neuf-prefaces/preface-de-la-deuxieme-edition-1718> [08/04/2014].
- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *3e préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-troisieme-edition-1740> [08/04/2014].
- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *4e préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-quatrieme-edition-1762> [08/04/2014].
- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *5e préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-cinquieme-edition-1798> [08/04/2014].
- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *6e préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neuf-prefaces/sixieme-preface> [08/04/2014].
- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *7e préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-septieme-edition-1877> [08/04/2014].
- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *8e préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neufs-prefaces/preface-de-la-huitieme-edition-1932-1935> [08/04/2014].
- Académie française* > *Le Dictionnaire* > *La neuvième édition*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire-les-neuf-prefaces/preface-la-neuvieme-edition> [08/04/2014].
- Académie française* > *Le Dictionnaire* > *Les neuf préfaces* > *9e préface*. Zugriff unter: <http://academie-francaise.fr/le-dictionnaire/la-9e-edition> [08/04/2014].
- Académie française* > *L'institution* > *Les statuts*. Zugriff unter: [http://www.academie-francaise.fr/sites/academie-francaise.fr/files/statuts\\_af.pdf](http://www.academie-francaise.fr/sites/academie-francaise.fr/files/statuts_af.pdf) [06/04/2014].
- Assemblée nationale* > *Accueil* > *Histoire et Patrimoine* > *Ordonnance de Villers-Cotterêts*. Zugriff unter: <http://www.assemblee-nationale.fr/histoire/villers-cotterets.asp> [05/04/2014].
- Berschin, H./Felixberger, J./Goebel, H. (2008) *Französische Sprachgeschichte. 2., überarbeitete und ergänzte Auflage*. Hildesheim: Georg Olms Verlag.
- Bochmann, K. (1993) *Sprachpolitik in der Romania. Zur Geschichte sprachpolitischen Denkens und Handelns von der Französischen Revolution bis zur Gegenwart*. Berlin/New York: Walter de Gruyter.
- Bouhours, D. (1962) *Les Entretiens d'Ariste et d'Eugène. Présentation de Ferdinand Brunot*. Paris: Armand Colin.
- Bouhours, D. (1973) *Remarques nouvelles sur la Langue française (1675). Suite des Remarques nouvelles sur la Langue française (1687)*. Genf: Slatkine Reprints.
- Brunot, F. (1930) *Histoire de la Langue française des Origines à 1900, Tome III: La Formation de la Langue classique. Première Partie, Édition revue et corrigée*. Paris: Librairie Armand Colin.
- Chassagne, A./Gasnault, P./Pastoureau, M./Service du Dictionnaire de l'Académie française (1994) *Le Dictionnaire de l'Académie française: 1694-1994 – sa naissance et son actualité*. Paris: Institut de France.
- Corneille, T. (1694) *Dictionnaire des Arts et des Sciences. Par M. D. C. de l'Académie française. Tome Premier. A-L*. Paris: Chez la Veuve de Jean Baptiste Coignard et chez Jean Baptiste Coignard. Zugriff unter: <http://gallica.bnf.fr/ark:/12148/bpt6k50507s> [09/04/2014].
- Du Bellay, Joachim (1969) *La Deffence et Illustration de la Langue francoyse. Edition critique par Henri Chamard*. Genf: Slatkine Reprints.
- Ehlich, K./Osner, J./Stammerjohann, H. (eds.) *Hochsprachen in Europa. Entstehung, Geltung, Zukunft*. Freiburg im Breisgau: Fillibach Verlag.
- Elias, N. (1983) *Die höfische Gesellschaft. Untersuchungen zur Soziologie des Königtums und der höfischen Aristokratie*. Frankfurt/Main: Suhrkamp (Taschenbuch Wissenschaft 423).



- Fluck, H. (1996) *Fachsprachen. Einführung und Bibliographie*. Fünfte, überarbeitete und erweiterte Auflage. Basel/Tübingen: A. Francke Verlag (UTB 483).
- Grimm, Jürgen (2005) *Französische Klassik. Lehrbuch Romanistik*. Stuttgart/Weimar: Verlag J. B. Metzler.
- Hassler, G. (1998) Anfänge der europäischen Fachsprachenforschung im 17. und 18. Jahrhundert. In: Hoffmann, L./Kalverkämper, H./Wiegand, H. (eds.) *Fachsprachen. Languages for Special Purposes. 1. Halbband*. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 14.1), 322-326.
- Hoffmann, L./Kalverkämper, H./Wiegand, H. (eds.) *Fachsprachen. Languages for Special Purposes. 1. Halbband*. Berlin/New York: Walter de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 14.1).
- Lindemann, Margarete (1994) Die französischen Wörterbücher von den Anfängen bis 1600. Entstehung und typologische Beschreibung. Tübingen: Max Niemeyer Verlag (Lexicographica Series Maior 54).
- Mayer, M. (2013) *Sprachpflege und Sprachnormierung in Frankreich am Beispiel der Fachsprachen vom 16. Jahrhundert bis in die Gegenwart*. Innsbruck: Innsbruck University Press (Studien des Interdisziplinären Frankreich-Schwerpunkts der Universität Innsbruck 5).
- Müller, B. (1975) *Das Französische der Gegenwart. Varietäten, Strukturen, Tendenzen*. Heidelberg: Carl Winter Universitätsverlag.
- Pécheur, J. (2001) Nouveaux espaces pour le français. In: Ehlich, K./Osnner, J./Stammerjohann, H. (eds.) *Hochsprachen in Europa. Entstehung, Geltung, Zukunft*. Freiburg im Breisgau: Fillibach Verlag, pp. 47-72.
- Richelet, C. (1680) *Dictionnaire françois, contenant les Mots et les Choses, plusieurs nouvelles Remarques sur la Langue françoise*. Genf: Chez Jean Herman Widerhold. Zugriff unter: <http://gallica.bnf.fr/ark:/12148/bpt6k509323/f6.image> [09/04/2014].
- Roelcke, T. (2005) *Fachsprachen. 2., durchgesehene Auflage*. Berlin: Erich Schmidt Verlag (Grundlagen der Germanistik).
- Vaugelas, C. (1880) *Remarques sur la Langue françoise, par Vaugelas. Nouvelle Édition, par A. Chassang. Tome Premier*. Paris: Léopold Cerf.





# EU-Terminologie in den einsprachigen Wörterbüchern des Deutschen

Diana Stantcheva  
American University in Bulgaria  
dstantcheva@aubg.bg

## Abstract

Die heutige Europäische Union blickt nun schon auf eine über 50 jährige Geschichte zurück. In dieser Zeit ist die europäische Staatengemeinschaft von ursprünglich 6 auf derzeit 28 Länder gewachsen. Deutschland gehörte zu den Gründungsmitgliedern der Europäischen Gemeinschaft für Kohle und Stahl, der Europäischen Wirtschaftsgemeinschaft und der Europäischen Atomgemeinschaft, der Organisationen, die später in die Europäische Union zusammenflossen. Österreich trat 1995 der Europäischen Union bei. Diese Fakten zum Anlass nehmend, untersucht der vorliegende Beitrag die lexikographische Kodifizierung EU-bezogener Terminologie in ausgewählten allgemeinen einsprachigen Wörterbüchern des Deutschen von 1961 bis 2009.

**Keywords:** terminology; terminography; lexicography; specialised languages; monolingual German dictionaries; Terminologie der Europäischen Union, Deutsch-als-Fremdsprache-Unterricht; Europa-Studien; einsprachige Wörterbücher

## 1 Einleitung

Mit dem Begriff *Terminus* und synonym dazu *Fachwort* bzw. *Fachwendung* wird im vorliegenden Beitrag zusammen mit WENDT (1997: 63) „[...] eine Funktion beschrieben, die jede Einheit einer natürlichen und künstlichen Sprache (Wort, Wortverbindung, Abkürzung, Symbol) unter bestimmten Bedingungen übernehmen kann“. Im Falle der EU-Terminologie wären diese Bedingungen das Auftreten einer sprachlichen Einheit in der fachlichen Kommunikation über das vereinte Europa und die Europäische Union mit ihren Institutionen und Organen oder die Zugehörigkeit dieser Einheit zum fachlichen Gegenstand *Amtssprache und Arbeitssprache der Europäischen Union*. Ausgehend von dieser Definition verstehe ich im Folgenden unter EU-Terminologie nicht nur genuine EU-Termini der „supranationalen Varietät ‚EU-Deutsch‘“, wie *Euro*, *Europäische Union*, *EU*, *EU-Richtlinie*, *Eurobarometer*, *Europäisches Parlament*, *Europäische Kommission*, *Europäische Zentralbank*, *Europäische Integration*, sondern auch Fachwörter und -wendungen aus den Sachgebieten Wirtschaft, Handel, Politik, Recht, wie z.B.

---

1 Zum Begriff „supranationale Varietät ‚EU-Deutsch‘“ vgl. MARKHARDT (2004).

*Supranationalität, Sicherheitspolitik, Binnenmarkt, Mehrheitsentscheidung, Minderheitenschutz, Verordnung*, die in Bezug auf das Zusammenwachsen Europas und/oder die Europäische Union verwendet werden. Die Basis der vorliegenden empirischen Analyse der lexikographischen Kodifizierung und Darstellung von EU-bezogenen Termini bildet die Wortstrecke *eu – ev* in einsprachigen Wörterbüchern des Deutschen. Es handelt sich dabei um eine verhältnismäßig kurze Wortstrecke mit einer großen Konzentration an EU-relevanten Wörterbucheinträgen. Die ausgewählten Wörterbücher decken den Zeitraum von 1961 bis 2009 ab und lassen damit Veränderungen in der Terminologie und deren Darstellung im Wörterverzeichnis festhalten.

Der Untersuchung liegen folgende allgemeine einsprachige Wörterbücher des Deutschen zugrunde:

*Wörterbuch der deutschen Gegenwartssprache* (1961-1977) [WDG]

*Ullstein Lexikon der deutschen Sprache* (1969) [ULLSTEIN-LdS]

*Der Sprachbrockhaus* (1972) [SBH<sub>3</sub>]

*Duden. Das große Wörterbuch der deutschen Sprache* (1976-1981) [DUDEN-GWB<sub>1</sub>]

*Brockhaus-Wahrig. Deutsches Wörterbuch* (1980-1984) [BW]

*Duden. Deutsches Universalwörterbuch* (1983) [DUDEN-UW<sub>1</sub>]

*Knaurs Großes Wörterbuch der deutschen Sprache* (1985) [KNAURS-GW]

*Gerhard Wahrig. Deutsches Wörterbuch* (1986) [WAHRIG-DW<sub>4</sub>]

*Duden. Das große Wörterbuch der deutschen Sprache* (1993-1995) [DUDEN-GWB<sub>2</sub>]

*Duden. Das große Wörterbuch der deutschen Sprache* (1999) [DUDEN-GWB<sub>3</sub>]

*Gerhard Wahrig. Deutsches Wörterbuch* (2000) [WAHRIG-DW<sub>7</sub>]

*Duden. Deutsches Universalwörterbuch* (2003) [DUDEN-UW<sub>5</sub>]

*Duden. Deutsches Universalwörterbuch* (2007) [DUDEN-UW<sub>6</sub>]

*Gerhard Wahrig. Deutsches Wörterbuch* (2008) [WAHRIG-DW<sub>8</sub>]

Darüber hinaus werden auch drei Lernerwörterbücher, ein Rechtschreibwörterbuch und ein historisches Wörterbuch des Deutschen herangezogen, um ein umfassenderes Bild von der Behandlung der EU-Terminologie in den Wörterbüchern des Deutschen in der untersuchten Zeitspanne zu bekommen:

*Langenscheidt GWDaF* (1998) [LGWDaF<sub>1</sub>]

*Grimm. Deutsches Wörterbuch* (Neubearbeitung, Band 8, 1999) [GRIMM-DW]

*Pons GWDaF* (2004) [PGWDaF]

*Duden. Die deutsche Rechtschreibung* (2006) [DUDEN-DR<sub>24</sub>]

*Langenscheidt GWDaF* (2008) [LGWDaF<sub>2</sub>]

*Gerhard Wahrig. Großwörterbuch DaF* (2008) [WAHRIG-GDaF<sub>1</sub>]

*Duden. Die deutsche Rechtschreibung* (2009) [DUDEN-DR<sub>25</sub>]

Die vorliegende Untersuchung ist im Zusammenhang mit der studienbegleitenden fachsprachlichen Deutschausbildung im Rahmen des Bachelorstudienganges Europa-Studien<sup>2</sup> an der American University in Bulgaria (= AUBG) entstanden und ist Teil eines größeren Projekts, das zum Ziel hat, das Curriculum für diesen Unterrichtstyp zu optimieren, Übungsmaterialien mithilfe von Korpustexten geschriebener Sprache zu erstellen sowie ein Minimum an deutschen EU-Termini für die Zwecke der studienbegleitenden Fremdsprachenausbildung zu ermitteln.

## 2 Lexikographische Kodifizierung EU-bezogener Terminologie in den untersuchten Wörterbüchern

Die EU-Terminologie zeichnet sich durch Breite und Dynamik aus. Das macht sich auch in den durchgesehenen Nachschlagewerken der letzten zwei Jahrzehnte bemerkbar. Jedes neue Wörterbuch und jede weitere Wörterbuchauflage verzeichnen neue Termini. Die Anzahl der aufgenommenen Lemmata der Wortstrecke *eu* – *ev* ist in den herangezogenen Nachschlagewerken kontinuierlich gewachsen: Von fünf Begriffen (*Europa*, *Europäer*, *europäisch*, *europäisieren*, *Europäisierung*) zum damaligen Zeitpunkt noch ohne gemeinschaftseuropäischen Bezug im mehrbändigen WDG 1961-1977 (2. Band, 1964-1967) bis hin zu 47 Termini mit gemeinschaftseuropäischem Bezug im einbändigen DUDEN-DR<sub>25</sub> 2009. Auffällig viel sind dabei Neulexeme mit den Komponenten *EU*-, *Europa*- und *Euro*- als Erstglied, die vor allem politische und wirtschaftliche Neuentwicklungen innerhalb der Europäischen Union bezeichnen, wie z.B. *EU-Bürger*, *EU-Erweiterung*, *EU-Staat*, *Eurocent*, *Eurozone* sowie terminologische Wortverbindungen mit dem Attribut *europäisch*, wie z.B. *Europäischer Gerichtshof*, *Europäische Union*, *Europäische Währungseinheit*, *Europäisches Parlament*, *Europäische Zentralbank*, *Europäische Kommission*. Der Entwicklung der EU-Terminologie tragen auch die von manchen Lernerwörterbüchern angebotenen kurzen enzyklopädischen Zusatzinformationen in Form eines Informationskastens außerhalb des Wörterbuchartikels Rechnung, wie z.B. zu den Begriffen *Euro* und *Europäische Union* im PGWDaF 2004 und zu *EU* und *Euro* im WAHRIG-GDaF<sub>1</sub> 2008.

Das Veralten von EU-Termini lässt sich anhand der durchgesehenen Wörterbücher ebenfalls beobachten, auch wenn die Wörterbücher dies nicht immer explizit deutlich machen: Bestimmte EU-Termini, die mittlerweile zu Historismen geworden sind, weil das Denotat außer Gebrauch kam, sind in manchen der Wörterbücher noch vorhanden, allerdings nicht immer entsprechend markiert, wie z.B.: *Europäische Gemeinschaft für Kohle und Stahl* im WAHRIG-DW<sub>8</sub> 2008 (unter *europäisch*). Dieser Begriff ist nicht als Historismus markiert und nur durch die Zeitangaben in der semantischen Paraphrase

---

2 Den Bachelorstudiengang Europa-Studien gibt es an der AUBG seit 1999 als Nebenfach (*European Studies Minor*) und seit 2001 als Hauptfach (*European Studies Major*). Im Rahmen des Hauptfaches ist seit 2004 eine obligatorische studienbegleitende Fremdsprachenausbildung in einer der Sprachen Deutsch, Französisch oder Spanisch vorgesehen. Näheres zum Studiengang Europa-Studien an der AUBG siehe unter <http://www.aubg.bg/template5.aspx?page=4419&menu=001001002003> und <http://www.aubg.bg/template5.aspx?page=4428&menu=001001003003>.

wird signalisiert, dass die Europäische Gemeinschaft für Kohle und Stahl nicht mehr existiert: „1952 gegründete u. 2002 wieder aufgelöste Gemeinschaft europäischer Länder mit dem Ziel eines gemeinsamen Marktes für Kohle u. Eisen, Montanunion“.

Insgesamt unklar bleiben die Kriterien für die Aufnahme der EU-Termini in die Wörterverzeichnisse, denn alle Wörterbücher weisen Lemmalücken auf. So verzeichnen z.B. WAHRIG-DW<sub>7</sub> 2000, WAHRIG-DW<sub>8</sub> 2008 und Grimm-DW 1999 den Begriff *Euro* als Lemma nicht. *Eurogeld* ist gebucht nur im DUDEN-DR<sub>24</sub> 2006 und DUDEN-DR<sub>25</sub> 2009, *Euromünze* nur im DUDEN-DR<sub>24</sub> 2006, DUDEN-DR<sub>25</sub> 2009 und WAHRIG-DW<sub>8</sub> 2008, *Eurowährung* nur im WAHRIG-DW<sub>8</sub> 2008.

Eine Recherche in den Textkorpora des Deutschen<sup>3</sup> bringt darüber hinaus eine Reihe von Termini zu Tage, die allesamt als Lemma nicht in den untersuchten Wörterbüchern zu finden sind: *Europarecht*, *europarechtlich*, *Europahymne*, *Europabegeisterung*, *Europamüdigkeit*, *Europapolitiker*, *europapolitisch*, *Europa-Studien*, *europakritisch*, *Europawahlkampf*, *Eurokratie*, *EU-Ebene*, *EU-Haushalt* usw., um hier nur einige zu nennen.

### 3 Lexikographische Defizite bei der Darstellung von EU-Termini in den untersuchten Wörterbüchern

In diesem Abschnitt soll auf die Darstellung der EU-Terminologie in den Wörterverzeichnissen der untersuchten Wörterbücher eingegangen werden. Exemplarisch soll dabei die lexikographische Darstellung folgender EU-Begriffe näher betrachtet werden: *Europa*, *Europäer*, *Europäerin*, *Euro*, *Euroskeptiker*, *Europäische Zentralbank*, *EU*, *Eurokrat*, *EU-Erweiterung* und *Euroland*.

Bei *Europa*, *Europäer*, *Europäerin* und *Euro* handelt es sich einerseits um Termini, die eine Bedeutungserweiterung im Rahmen der EU-Terminologie erfahren haben, wie *Europa*, *Europäer*, *Europäerin*, und andererseits um ein Neulexem *Euro*, dessen Denotat genuin mit der Europäischen Union verbunden ist. Im Folgenden soll anhand der lexikographischen Darstellung dieser Begriffe deutlich gemacht werden, dass die durchgesehenen Nachschlagewerke, von wenigen Ausnahmen abgesehen, den Bedeutungswandel der Begriffe *Europa*, *Europäer*, *Europäerin* nicht reflektieren und den terminologischen Neologismus *Euro* uneinheitlich beschreiben.

In sechs der durchgesehenen Wörterbücher wird *Europa* gemäß der lexikographischen Tradition, Eigennamen nicht zu verzeichnen, als Lemma nicht aufgenommen: So im BW 1980-1984, Knaurs-GW 1985, WAHRIG-DW<sub>4</sub> 1986, WAHRIG-DW<sub>7</sub> 2000, WAHRIG-DW<sub>8</sub> 2008 und Grimm-DW 1999. In acht Wörterbüchern wird *Europa* als monosemes Lexem dargestellt: WDG 1961-1977 (2. Band, 1964-1967), Ullstein-LdS 1969, LGWDaF<sub>1</sub> 1998, LGWDaF<sub>2</sub> 2008 verzeichnen nur die Bedeutung ‚Kontinent‘. Der Sprachbrockhaus 1972, DUDEN-DR<sub>24</sub> 2006 und DUDEN-DR<sub>25</sub> 2009 listen lediglich die Bedeutung ‚griechische

3 Z.B. im ZEIT-Korpus (<http://www.dwds.de>), im Korpus der Berliner Zeitung (<http://www.dwds.de>) oder in den Korpora geschriebener Sprache des Instituts für Deutsche Sprache in Mannheim (<http://www.ids-mannheim.de/cosmas2/>).

weibliche Sagengestalt' auf und PGWDaF 2004 nur die Bedeutung ‚Staatenkomplex‘. DUDEN-GWB<sub>1</sub> 1976-1981 ist das erste der durchgesehenen Wörterbücher, das *Europa* als polysemes Wort darstellt und die Unterbedeutung ‚Staatenkomplex‘ auflistet:

Staatenkomplex, der durch einen Zusammenschluß der europäischen Staaten entstehen soll: Konrad Adenauer will ein E[uropa], das sich in eine größere Atlantische Gemeinschaft einfügt (Dönhoff, Ära 128); sich für E[uropa] (einen Zusammenschluß der europäischen Staaten) begeistern, einsetzen.

Diese Zeilen sind in die neueren DUDEN-Wörterbücher (DUDEN-GWB<sub>2</sub> 1993-1995, DUDEN-GWB<sub>3</sub> 1999, DUDEN-UW<sub>5</sub> 2003 und DUDEN-UW<sub>6</sub> 2007) mit Ausnahme der Tilgung des Dönhoffs Zitats und des Verbs „begeistern“ unverändert übernommen, was insbesondere im Hinblick auf die Formulierung, dass der Staatenkomplex auch nach einer über 50 jährigen Geschichte noch „entstehen soll“, problematisch ist. Im DUDEN-GWB<sub>2</sub> 1993-1995, DUDEN-GWB<sub>3</sub> 1999, DUDEN-UW<sub>1</sub> 1983, DUDEN-UW<sub>5</sub> 2003 und DUDEN-UW<sub>6</sub> 2007 sind auch je zwei Lemmata *Europa* verzeichnet: Das erste ist polysem und hat die Unterbedeutungen ‚Kontinent‘ und ‚Staatenkomplex‘, das zweite monosem mit der Bedeutung ‚weibliche Gestalt der griechischen Mythologie‘.

LGWDaF<sub>1</sub> 1998, LGWDaF<sub>2</sub> 2008 und GRIMM-DW 1999 listen eine Reihe von Zusammensetzungen mit *Europa* als Erst- oder Zweitglied auf, allerdings ohne eine Bedeutungserläuterung oder Illustrationsbeispiele. Dabei wird auch kein Unterschied zwischen *Europa* als ‚Kontinent‘ und *Europa* als ‚Staaten-gemeinschaft‘ gemacht, sodass beispielsweise *Europa-Parlamentarier*, *Europareise*, *Europa-Währung* im GRIMM-DW 1999 und *Nordeuropa*, *Westeuropa*, *Europapolitik* im LGWDaF<sub>1</sub> 1998 und LGWDaF<sub>2</sub> 2008 nebeneinander und ohne Bedeutungs-differenzierung stehen.

Keines der untersuchten Wörterbücher führt phraseologische Wortverbindungen mit dem Grundelement *Europa* als ‚Staatengemeinschaft‘ auf, wie z.B. *(die) Festung Europa*, *das offizielle Europa*, *das gemeinsame Haus Europa*, *Europa à la carte*, die in Textkorpora des Deutschen mehrfach belegt sind.

Noch weniger reflektieren die durchgesehenen Nachschlagewerke den Bedeutungswandel bei den Termini *Europäer* und *Europäerin*. Der Begriff *Europäer* ist im GRIMM-DW 1999 als Lemma nicht gebucht. WDG 1961-1977 (2. Band, 1964-1967), DUDEN-DR<sub>24</sub> 2006 und DUDEN-DR<sub>25</sub> 2009 verzeichnen *Europäer* ohne eine Bedeutungserläuterung. Als monosemes Wort mit der Bedeutung ‚Bewohner des europäischen Kontinents‘ findet man *Europäer* im Sprachbrockhaus 1972, Ullstein-Lds 1969, WAHRIG-DW<sub>4</sub> 1986, WAHRIG-DW<sub>7</sub> 2000 und WAHRIG-DW<sub>8</sub> 2008. PGWDaF 2004 hat den Begriff ebenfalls nur mit einer Bedeutung aufgenommen, nämlich mit „jmd., der Bürger eines europäischen Staates ist“. Der erste Eintrag von *Europäer* als polysemes Wort in den durchgesehenen Wörterbüchern ist erneut im DUDEN-GWB<sub>1</sub> 1976-1981 zu finden, mit den Unterbedeutungen „Vertreter des abendländischen Kulturkreises“ und „Politiker, der für einen Zusammenschluß der europäischen Staaten eintritt“. Seitdem ist *Europäer* als polysemes Wort mit variierenden Bedeutungserläuterungen verzeichnet, wie:

1. Einwohner Europas
2. Anhänger einer Politik, die den Interessen der Europäischen Gemeinschaft Vorrang einräumt (BW 1980-1984)

1. Bewohner Europas
2. jmd., der für den Zusammenschluß der Staaten Europas ist (KNAURS-GW 1985)
1. j-d, der in Europa geboren ist u. zu e-r europäischen Nation gehört
2. verwendet für j-n, dessen Denken u. Handeln die Einheit Europas zum Ziel hat (LGWDaF1 1998, LGWDaF2 2008)

DUDEN-UW<sub>1</sub> 1983 ist das erste der untersuchten Wörterbücher, das *Europäer* auch als 'Einwohner des europäischen Staatenkomplexes' definiert, auch wenn diese Definition durch einen Verweis auf das Lemma *Europa* zustande kommt:

1. Ew. zu 'Europa [d.i. 'Einwohner des europäischen Kontinents' und 'Einwohner des Staatenkomplexes' – meine Anmerkung D.S.]
2. Politiker, der für einen Zusammenschluß/ss der Staaten Europas eintritt.

DUDEN-GWB<sub>2</sub> 1993-1995, DUDEN-GWB<sub>3</sub> 1999, DUDEN-UW<sub>5</sub> 2003 und DUDEN-UW<sub>6</sub> 2007 haben diese Artikelstruktur und diese Bedeutungserläuterung ohne Veränderungen übernommen. Abgesehen von den fünf DUDEN-Wörterbüchern erwähnen die anderen durchgesehenen Nachschlagewerke die Unterbedeutung 'Einwohner des Staatenkomplexes in Europa' bzw. 'EU-Bürger' zu *Europäer* nicht. In Textkorpora des Deutschen ist diese Verwendung von *Europäer* dagegen etwa seit den 70er Jahren des 20. Jahrhunderts belegt.

Die Movierung *Europäerin* ist in den durchgesehenen Wörterbüchern zum ersten Mal im DUDEN-GWB<sub>2</sub> 1993-1995 als Lemma mit der Bedeutungserläuterung „w[eibliche] Form zu ↑ Europäer“ gebucht (so auch im DUDEN-GWB<sub>3</sub> 1999, DUDEN-UW<sub>5</sub> 2003, DUDEN-UW<sub>6</sub> 2007). GRIMM-DW 1999, WAHRIG-DW<sub>7</sub> 2000, LGWDaF<sub>1</sub> 1998 verzeichnen *Europäerin* nicht, LGWDaF<sub>2</sub> 2008, DUDEN-DR<sub>24</sub> 2006 und DUDEN-DR<sub>25</sub> 2009 haben *Europäerin* ohne eine semantische Paraphrase aufgenommen. PGWDaF 2004 listet *Europäer* und *Europäerin* nebeneinander mit dem bestimmten Artikel, der Genitivendung und der Pluralform nur des männlichen Substantivs auf<sup>4</sup> und definiert beide Begriffe zusammen als „jmd., der Bürger eines europäischen Staates ist“. Im WAHRIG-DW<sub>8</sub> 2008 und im WAHRIG-GDaF<sub>1</sub> 2008 ist die Definition „weibl[icher] Europäer“ zu finden. Mit anderen Worten: Die Unterbedeutung 'EU-Bürgerin' ist ebenfalls nur in den DUDEN-Wörterbüchern (DUDEN-GWB<sub>2</sub> 1993-1995, DUDEN-GWB<sub>3</sub> 1999, DUDEN-UW<sub>5</sub> 2003, DUDEN-UW<sub>6</sub> 2007) vorhanden, auch wenn sie sich auch hier durch eine Zirkeldefinition erschließen lässt: *Europäerin* → *Europäer* → *Europa*.

Die Währungsbezeichnung *Euro* ist in den untersuchten Wörterbüchern erst seit 1998 gebucht, allerdings nicht in allen, denn GRIMM-DW 1999, WAHRIG-DW<sub>7</sub> 2000 und WAHRIG-DW<sub>8</sub> 2008 verzeichnen den *Euro* als Lemma nicht. Die Wörterbücher, die den Begriff aufgenommen haben, geben voneinander abweichende, unpräzise und zuweilen unkorrekte Definitionen an. Das LGWDaF<sub>1</sub> 1998 erläutert

---

4 Diese Darstellungsform ist im gesamten Wörterverzeichnis bei der Bildung einer weiblichen Personenbezeichnung mit einem Suffix von einer männlichen Form zu finden, z.B. auch „Europagegner, Europagegnerin der <-s, ->“. So eine Darstellung mag platzökonomisch sein, ist aber höchst problematisch besonders in Anbetracht der Tatsache, dass es sich bei PGWDaF 2004 um ein Lernerwörterbuch handelt.

z.B. *Euro* als „Bezeichnung für die gemeinsame Währung in den Staaten der Europäischen Union (ab 1999)“. Diese Definition ist nicht korrekt, denn nicht alle EU-Länder haben den Euro als Währung. Im LGWDaF<sub>2</sub> 2008 ist diese Bedeutungserläuterung korrigiert in: „Bezeichnung für die gemeinsame Währung in den Staaten der Europäischen Währungsunion“. DUDEN-UW<sub>5</sub> 2003, DUDEN-UW<sub>6</sub> 2007, DUDEN-GWB<sub>3</sub> 1999 beschreiben den Euro als „Währungseinheit der Europäischen Währungsunion“, DUDEN-DR<sub>24</sub> 2006 und DUDEN-DR<sub>25</sub> 2009 als „europ[äische] Währungseinheit“, und PGWDaF 2004 als „die europäische Währungseinheit“.

Wenn man die Wörterbucheinträge zum Lemma *Euro* vergleicht, fällt nicht nur die unterschiedliche Bedeutungserläuterung auf, sondern auch die variable Darstellung der Formen des Plurals und des Genitivs in den einzelnen Wörterbüchern:<sup>5</sup>

Eu|ro, der; -[s], -s <aber: 10 Euro> (DUDEN-UW<sub>5</sub> 2003, DUDEN-UW<sub>6</sub> 2007, DUDEN-GWB<sub>3</sub> 1999)

Eu|ro, der; -[s], -s ...30 Euro (DUDEN-DR<sub>24</sub> 2006 und DUDEN-DR<sub>25</sub> 2009)

Eu·ro *der*; -(s), -(s); (LGWDaF<sub>1</sub> 1998)

Eu·ro *der*; -(s), -(s); ... *Dieses Buch kostet 10 Euro* (LGWDaF<sub>2</sub> 2008)

Eu·ro *der* <-s, -s> (PGWDaF 2004)

Keines der untersuchten Wörterbücher verzeichnet auch phraseologische Wortverbindungen zum Lemma *Euro*, wie z.B. *der schnelle Euro*, *jeden Euro zweimal umdrehen*, die in Textkorpora des Deutschen mehrfach belegt sind.

Der Begriff *Euroskeptiker* wird im DUDEN-UW<sub>6</sub> 2007, DUDEN-GWB<sub>2</sub> 1993-1995 und im DUDEN-GWB<sub>3</sub> 1999 als „Politiker, der dem Europagedanken, der Politik der Europäischen Gemeinschaften skeptisch gegenübersteht“ umschrieben. Warum ein Euroskeptiker nur ein Politiker sein soll, leuchtet dabei nicht ein. Nicht ganz unproblematisch in dieser Definition ist auch die Fachwendung „Europäische Gemeinschaften“, denn die drei Gemeinschaften (EGKS, EURATOM und EG) sind seit 1993 in der Europäischen Union zusammengeführt. In den untersuchten Wörterbüchern herrscht darüber hinaus Unsicherheit im Hinblick auf die Markierung des Begriffs: *Euroskeptiker* wird z.B. im PGWDaF 2004 mit dem Marker „POL“ (für „Politik“), im DUDEN-GWB<sub>2</sub> 1993-1995, DUDEN-GWB<sub>3</sub> 1999 und DUDEN-UW<sub>6</sub> 2007 mit „Politik Jargon“ versehen. Im DUDEN-DR<sub>25</sub> 2009 ist der Begriff dagegen nicht markiert und dementsprechend als neutral anzusehen.

Die Paraphrase des Terminus *Europäische Zentralbank* (unter *europäisch*) ist im WAHRIG-DW<sub>7</sub> 2000 gänzlich unkorrekt: „1950 von den Mitgliedstaaten der OEEC gegründete Union zur multilateralen Verrechnung von Zahlungen, die 1958 durch das europäische Währungsabkommen abgelöst wurde“. Diese Definition wurde höchstwahrscheinlich vom BW 1980-1984 übernommen, allerdings nicht richtig eingesetzt, denn die gleiche Definition steht im BW unter der terminologischen Wortverbindung *Europäische Zahlungsunion*.

---

5 Die fehlende Eindeutigkeit in der Darstellung der flektierten Formen dieses Neulemmas hat HERBERG (2001) bereits in sechs anderen Wörterbüchern des Deutschen festgestellt, die im Zeitraum 1996-1999 erschienen sind.



Eine sachliche Ungenauigkeit enthält auch der folgende Text im Infokasten zu *EU* im WAHRIG-GDaF<sub>1</sub> 2008: „Mit der Abkürzung *EU* wird der große europäische Staatenbund, die *Europäische Union*, bezeichnet. Hervorgegangen ist diese aus der im Jahr 1952 gegründeten Europäischen Wirtschaftsgemeinschaft (EWG)“. Die Europäische Wirtschaftsgemeinschaft wurde 1957 und nicht 1952 gegründet und die *EU* ist, wie unter *Euroskeptiker* bereits erwähnt, nicht einzig und allein aus der Europäischen Wirtschaftsgemeinschaft hervorgegangen.

Problematisch ist auch die Definition des Lemmas *EU* im LGWDaF<sub>1</sub> 1998 und LGWDaF<sub>2</sub> 2008: „e-e Union von europäischen Staaten, die in allen politischen Bereichen eng zusammenarbeiten u. e-e politische Einheit Europas wollen“. Diese Definition lässt Bereiche wie Wirtschaft und Kultur außer Acht und reduziert die Rolle der *EU* auf den politischen Bereich.

Das Lemma *Eurokrat* im DUDEN-GWB<sub>2</sub> 1993-1995, DUDEN-GWB<sub>3</sub> 1999 und im DUDEN-UW<sub>1</sub> 1983, DUDEN-UW<sub>5</sub> 2003 und DUDEN-UW<sub>6</sub> 2007 wird erläutert als „Politiker, der den Interessen der Europäischen Gemeinschaft (besonders gegenüber den USA) Vorrang einräumt“. DUDEN-GWB<sub>2</sub> 1993-1995 und DUDEN-GWB<sub>3</sub> 1999 führen zusätzlich zu dieser irreführenden Definition auch das folgende Illustrationsbeispiel an, das in keinem Zusammenhang zur angegebenen Bedeutungserläuterung steht: „Schon jetzt scheint klar, dass die Brüsseler [Eurokrat]en mit ihrem Haushaltsgeld ... nicht auskommen (Spiegel 41, 1980, 22)“. Dieses Illustrationsbeispiel würde eher zur Bedeutung ‚EU-Beamter/EU-Angestellter‘ oder gar ‚EU-Bürokrat‘ passen, welche auch Textkorpora des Deutschen belegen. Man vergleiche die nachfolgenden Textausschnitte:

Plötzlich scheint ein Tabu zu brechen: [...] Ausgerechnet in Belgien, wo die imposanten Gebäude der EU-Institutionen in den Himmel ragen, wo Eurokraten still und geduldig ihre bürokratischen Fesseln über den Erdteil legen und Politiker aus ganz Europa den Abgesang aufs Nationale anstimmen. (Die ZEIT 47/2007)

Immer wieder wundern sich Politiker und Beamte in Brüssel, warum sie als Eurokraten verschrien sind. (Die ZEIT 47/2007, Wirtschaft)

Unter *EU-Erweiterung* findet man im PGWDaF 2004 die Definition: „Ausdehnung der EU-Mitgliedschaft auf weitere Länder (im Osten)“. Der Zusatz „im Osten“ mag den vermehrten Gebrauch des Begriffs in den Medien im Zusammenhang mit der EU-Osterweiterung in den letzten Jahren reflektieren, ignoriert aber die Tatsache, dass man auch über eine EU-Erweiterung nach Norden oder nach Süden sprechen kann (1995 traten z.B. Finnland, Österreich und Schweden der EU bei, und diese drei Länder liegen keineswegs im Osten).

*Euroland* wird in den untersuchten Wörterbüchern unterschiedlich behandelt: Die drei Lernerwörterbücher und WAHRIG-DW<sub>8</sub> 2008 stellen den Terminus als monosemes Lexem dar, wobei jedes Wörterbuch eine andere Bedeutungserläuterung angibt: Laut LGWDaF<sub>2</sub> 2008 versteht man unter *Euroland* „ein Land mit dem Euro als Währung“ (keine Information über eine mögliche Pluralform), laut PGWDaF 2004 ist damit das „Gesamtgebiet der EU-Länder“ (ohne Plural) gemeint und laut WAHRIG-GDaF<sub>1</sub> 2008 und WAHRIG-DW<sub>8</sub> 2008 ist *Euroland* die „Gesamtheit der europäischen Ländern [SIC!], in



denen die europäische Währungseinheit gilt“ (ohne Plural). DUDEN-DR<sub>24</sub> 2006, DUDEN-DR<sub>25</sub> 2009, DUDEN-GWB<sub>3</sub> 1999, DUDEN-UW<sub>5</sub> 2003 und DUDEN-UW<sub>6</sub> 2007 verzeichnen *Euroland* mit zwei Unterbedeutungen: „an der Europäischen Währungsunion teilnehmende Staatengruppe“ (ohne Plural) und „Staat, der an der Europäischen Währungsunion teilnimmt“. Bei dieser zweiten Unterbedeutung ist im DUDEN-UW<sub>5</sub> 2003 und DUDEN-UW<sub>6</sub> 2007 keine Pluralform angegeben. DUDEN-DR<sub>24</sub> 2006, DUDEN-DR<sub>25</sub> 2009 und DUDEN-GWB<sub>3</sub> 1999 listen dagegen Verwendungsbeispiele mit der Pluralform „Euroländer“ bei dieser Unterbedeutung auf. Diese Uneinheitlichkeit in der Beschreibung des Terminus führt zur Verunsicherung des Benutzers. Abhilfe können in solchen Situationen die Textkorpora des Deutschen schaffen.<sup>6</sup>

## 4 Fazit

Die vorstehende kurze Analyse der Darstellung von EU-Termini der Buchstabenstrecke *eu* – *ev* in ausgewählten allgemeinen einsprachigen Wörterbüchern des Deutschen brachte Lemmalücken, Bedeutungslücken, fehlende feste Wortverbindungen, sachliche Ungenauigkeiten in den semantischen Paraphrasen und uneinheitliche Angaben zu Tage. Dabei kann man keinen Unterschied zwischen den älteren und den neueren der durchgesehenen Wörterbücher ausmachen. Man kann sich nur vorstellen, was eine Durchsicht der gesamten Wörterverzeichnisse in Bezug auf EU-relevante Informationen für Funde bringen würde. Für die Zwecke der studienbegleitenden Fremdsprachenausbildung lässt sich in Sachen EU-Terminologie keines der untersuchten Wörterbücher des Deutschen empfehlen. Man sollte vielmehr verschiedene allgemeinsprachliche Wörterbücher gleichzeitig konsultieren und vor allem Korpusbelege sowie Sachlexika der deutschen Bundeszentrale für politische Bildung und/oder der Europäischen Union zum Vergleich heranziehen. Im Hinblick auf die Tatsachen, dass Deutschland zu den Gründerländern der heutigen EU gehörte, dass Österreich seit 1995 Mitglied der EU ist und dass das Deutsche offiziell als Arbeitssprache der EU gehandelt wird, auch wenn mit einem gewissen Abstand hinter dem Englischen und Französischen, erstaunt der mangelhafte Niederschlag, den die EU-Terminologie in den Wörterbüchern des Deutschen findet. In diesem Zusammenhang würde sich auch eine vergleichende Untersuchung der Darstellung der EU-Terminologie in einsprachigen Wörterbüchern des Englischen und Französischen lohnen.

## 5 Literatur

Brockhaus-Wahrig (1980-1984). Deutsches Wörterbuch in sechs Bänden. Hrsg. von Gerhard Wahrig, Hildegard Krämer, Harald Zimmermann. Wiesbaden/Stuttgart: F. A. Brockhaus/Deutsche Verlags-Anstalt.

---

6 Textkorpora des Deutschen belegen beide Bedeutungen von *Euroland* sowie eine weitere Schreibweise *Euro-Land*, die in den Wörterbüchern nicht lemmatisiert ist.

- Der Duden: in 12 Bänden (242006). Die deutsche Rechtschreibung. 24., völlig neu bearbeitete und erweiterte Auflage, hg. von der Dudenredaktion. Mannheim u. a.: Dudenverlag. (= Der Duden; Bd. 1).
- Der Duden: in 12 Bänden (252009). Die deutsche Rechtschreibung. 25. Auflage, hg. von der Dudenredaktion. Mannheim u. a.: Dudenverlag. (= Der Duden; Bd. 1).
- Duden-GWB (11976-1981). Duden Das große Wörterbuch der deutschen Sprache. In sechs Bänden. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter Leitung von Günther Drosdowski. Mannheim u. a.: Bibliographisches Institut.
- Duden-GWB (21993-1995). Duden Das große Wörterbuch der deutschen Sprache. In acht Bänden. 2., völlig neu bearb. und stark erw. Aufl. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter Leitung von Günther Drosdowski. Mannheim u. a.: Dudenverlag.
- Duden-GWB (31999). Duden Das große Wörterbuch der deutschen Sprache. In zehn Bänden. 3., völlig neu bearb. und erw. Aufl. Hrsg. vom Wissenschaftlichen Rat der Dudenredaktion. Mannheim u. a.: Dudenverlag.
- Duden-UW (11983). Duden Deutsches Universalwörterbuch. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion unter Leitung von Günther Drosdowski. Mannheim u. a.: Bibliographisches Institut.
- Duden-UW (52003). Duden Deutsches Universalwörterbuch. 5., neu bearb. und erw. Aufl. Hrsg. und bearb. vom Wissenschaftlichen Rat und den Mitarbeitern der Dudenredaktion. Mannheim u. a.: Dudenverlag.
- Duden-UW (62007). Duden Deutsches Universalwörterbuch. 6. überarbeitete und erw. Aufl. Hrsg. von der Dudenredaktion. Mannheim u. a.: Dudenverlag.
- Grimm, Jacob/ Grimm, Wilhelm (1999). Deutsches Wörterbuch. Neubearbeitung. 8. Band, emporerheben-exzitieren. Stuttgart: S. Hirzel Verlag.
- Herberg, Dieter (2001). Euro. Zur Karriere eines europäischen Neologismus in deutschen Presstexten (1995-1999). <http://www1.ids-mannheim.de/fileadmin/lexik/LexikalischeInnovationen/pdf/euroanalysis.pdf> [10/04/2014].
- Knaurs-GW (1985). Knaurs Großes Wörterbuch der deutschen Sprache. Der große Störig. Erarb. v. U. Hermann unter Mitarbeit v. H. Leisering und H. Hellerer. München: Droemer Knaur.
- LGWDaF (11998). Langenscheidts Großwörterbuch Deutsch als Fremdsprache. Hrsg. von Dieter Götz, Günther Haensch, Hans Wellmann. Neubearbeitung. In der neuen deutschen Rechtschreibung. Berlin u. a.: Langenscheidt.
- LGWDaF (22008). Langenscheidt Großwörterbuch Deutsch als Fremdsprache. Neubearbeitung. Hg. von D. Götz/G. Haensch/H. Wellmann. Berlin u. a. Langenscheidt.
- Markhardt, Heidemarie (2004). Das österreichische Deutsch im Rahmen der Europäischen Union. In: Lebende Sprachen 1, S. 15-22.
- Pons GWDaF (2004). Großwörterbuch Deutsch als Fremdsprache. Barcelona u. a. Klett Sprachen.
- SBH (1972). Der Sprachbrockhaus. Deutsches Bildwörterbuch. 8., völlig Neubearb. u. erw. Aufl. Wiesbaden: Brockhaus.
- Ullstein-Lds (1969). Ullstein Lexikon der deutschen Sprache. Wörterbuch für Rechtschreibung, Silbentrennung, Aussprache, Bedeutungen, Synonyme, Phraseologie, Etymologie. Hrsg. u. bearb. von R. Köster unter Mitarbeit v. H. Hahmann, H. Hartmann u. F. Mehling. Frankfurt am Main/Berlin: Ullstein.
- Wahrig-DW (41986). Gerhard Wahrig. Deutsches Wörterbuch. Hrsg. in Zusammenarbeit mit zahlreichen Wissenschaftlern und anderen Fachleuten. Völlig überarbeitete Neuausgabe. München: Mosaik.
- Wahrig-DW (72000). Gerhard Wahrig. Deutsches Wörterbuch. Mit einem „Lexikon der deutschen Sprachlehre“. Neu hrsg. von Renate Wahrig-Burfeind. Gütersloh/München: Bertelsmann.
- Wahrig-DW (82008). Gerhard Wahrig. Deutsches Wörterbuch. 8., grundlegend bearb. und erweiterte Auflage, Nachdruck. Rechtschreibung 2006. Hrsg. von Renate Wahrig-Burfeind. Gütersloh: Bertelsmann.
- Wahrig-GDaF (12008). Wahrig-Burfeind, Renate: Wahrig Großwörterbuch Deutsch als Fremdsprache. 1. Aufl. Gütersloh: Wissen Media Verlag.

- WDG (1961-1977). Wörterbuch der deutschen Gegenwartssprache. Hrsg. von Ruth Klappenbach und Wolfgang Steinitz. 6 Bände. Berlin: Akademie-Verlag.
- Wendt, Susanne (1997). Terminus – Thesaurus – Text: Theorie und Praxis von Fachbegriffssystemen und ihrer Repräsentation in Fachtexten. Tübingen: Narr (= Forum für Fachsprachen-Forschung; Bd. 37).



# Text Boxes as Lexicographic Device in LSP Dictionaries

Elsabé Taljard\*, Danie J. Prinsloo\*, Rufus H. Gouws\*\*

\*University of Pretoria, \*\*University of Stellenbosch

elsabe.taljard@up.ac.za, danie.prinsloo@up.ac.za, rhg@sun.ac.za

## Abstract

It is the duty of the lexicographer to maximally utilise all lexicographic devices at his/her disposal in the compilation of a dictionary. In modern-day dictionaries lexicographers often rely on lexicographic text boxes as a means to present lexicographic data, especially where stronger emphasis or additional data regarding a specific entry is needed. It could be argued that the use of text boxes as a lexicographic device is underutilised in general dictionaries and even more so in LSP dictionaries. They should however not be used in a haphazard way but rather according to a well-devised system. Looking at existing dictionaries one often realises that a random use of text boxes has a detrimental influence on successful data retrieval. Text boxes should be used for data that need to be presented in a position of salience. The aim of this paper is first to give an overview of the current status of the use of text boxes in LSP dictionaries; secondly, to offer some suggestions as to the effective use of text boxes in these dictionaries by drawing on an analysis of the current use of text boxes in LGP dictionaries.

**Keywords:** Text boxes; LSP dictionaries; LGP dictionaries

## 1 Introduction

As pointed out by Gouws and Prinsloo (2010) the function of text boxes as lexicographic device is to place more than the default focus on a specific data item. They are typically used for data that need to be presented in a position of salience. One of the challenges for lexicographers is to make an informed decision as to what kind of data should be presented in textboxes in order to maximize data transfer and where to include the text box in order to ensure optimal access and a successful consultation by the user. Such a decision needs to be informed by a clear understanding of the user and his / her particular lexicographic needs and reference skills.

As pointed out by Tarp (2012), modern society is characterised by an ever-growing need for lexicographical tools that provide quick and easy access to data which have been prepared and selected in such a way as to enable the prospective user to retrieve the punctual information required to satisfy his / her lexicographic needs in a specific user situation. This need has resulted in an ever-increasing number of specialized dictionaries being published; however, many of these are of dubious quality, due to a lack of relevant lexicographic data. This paper consequently investigates the role that the in-

novative use of text boxes can potentially play in producing better quality specialized dictionaries. The aim of this paper is first to give an overview of the current status of the use of text boxes in dictionaries dealing with languages for special purposes, i.e. LSP dictionaries; secondly, to offer some suggestions as to the effective use of text boxes in these dictionaries by drawing on an analysis of the current use of text boxes in dictionaries dealing with language for general purposes, i.e. LGP dictionaries. For the purpose of this paper, the focus of the second point will be on LSP dictionaries for South African school learners. Within the current South African educational context, this will imply a perspective that includes the South African Bantu languages, since by far the majority of school learners and therefore of prospective dictionary users have a Bantu language as home language. The paper thus has both a contemplative and a transformative approach with regard to the use of text boxes.

## 2 The Nature of Text Boxes

Text boxes are typically included within the article structure of a dictionary, albeit that text boxes could also be included as article-external entries and immediate constituents of an article stretch. As article-internal entries they can prevail in a variety of article positions, cf. Wiegand and Gouws (2011). The article structure typically consists of the microstructural items and various indicators to identify article slots and item types. Text boxes are not part of these default constituents of dictionary articles. Like items they can be functionally positionally segmented but as another type of entry, mostly an inserted inner text. In this regard they usually are immediate constituents of the article but they could also be immediate constituents of a specific part of the article, e.g. the comment on semantics. They can be positioned where the lexicographer deems it necessary and can have any part of the article in their scope. They don't need to be addressed at an item in the article but can function as non-addressed entries. The nature of inserted inner texts gives lexicographers the freedom to utilize text boxes in different ways to contain whatever text the lexicographer regards as of enough importance to be allocated to a text box.

## 3 Text Boxes in LSP Dictionaries

It would seem that, generally speaking, text boxes are extremely rarely used in LSP dictionaries. Perusal of a random sample of five monolingual standard English LSP paper dictionaries revealed that none of the selected dictionaries utilizes this device. These dictionaries include the *Oxford Dictionary of Science*, *Oxford Dictionary of Chemistry*, *Oxford Dictionary of Economics*, *Concise Oxford Dictionary of Music* and *Stedman's Medical Dictionary*. One can only speculate as to what the possible reasons for this may be. It could be that the kind of information typically provided in text boxes in LGP dictionaries, is not deemed to be important by LSP lexicographers. In fact, provision of this kind of data in LSP dictionary-

es, even as items in the normal microstructure, is very rare, since these dictionaries do not normally intend to inform about issues such as grammar, pronunciation, spelling, etc. Another possible reason for the sparse provision of this kind of data could be that the dictionary has been designed by experts in the subject field in question, possibly with less than adequate lexicographic background, especially regarding article structures. In the third instance, the rudimentary treatment of comment on form could be a space-saving consideration, especially in the case of paper dictionaries. One should also consider the possibility that many changes introduced in LGP dictionaries might not have been reflected in LSP dictionaries because theoretical lexicographers have paid too little attention to LSP lexicography. LSP lexicographers might have used existing LSP dictionaries as their only model. A feature of the current development in theoretical lexicography is a stronger theoretical focus on LSP lexicography, cf. publications like Fuertes-Olivera (2010), Jesenšek (2013) and Tarp and Fuertes-Olivera (2014). This should result in positive changes in the planning and compilation of LSP dictionaries.

An informed decision as to whether text boxes could be a useful lexicographic device in LSP dictionaries can only be made if the skills, needs and knowledge of the target user and the function and genuine purpose of the dictionary are taken into consideration. When deciding on the possible use of text boxes in LSP dictionaries one should negotiate the subtypological diversity based on target users in the field of LSP lexicography. Three distinct user groups can be distinguished, i.e. experts, semi-experts and laypeople. Text boxes could be extremely helpful for users introduced into a new subject field who rely on their LSP dictionary as an important source of guidance. Consequently, guidance by means of text boxes is particularly important for users of school dictionaries. This information would also be decisive when deciding on the nature of the content of text boxes. We return to this issue below.

## 4 Typical Content of Text Boxes in LGP Dictionaries

As pointed out by Gouws and Prinsloo (2010), care should be taken as to what kind of data should be presented in a text box, and what data should form part of the normal microstructure of an article, since overuse of this device can detract from its usefulness. An analysis of a number of LGP dictionaries reveals that the typical data types found in text boxes include (but are not restricted to) contrasting related words, application range, pronunciation, register, grammar, spelling, collocations, metaphor, syntactic restrictions and contrasting British English and American English.

Obviously, some of the types of data provided in text boxes in LGP dictionaries are indeed irrelevant for LSP dictionaries: it is for example highly unlikely that data on metaphorical use of lemmata in an LSP dictionary would be needed, as would data on offensive use be. However, other data types can indeed be relevant for LSP dictionaries as well, and the possibility that additional data types might be necessary specifically for LSP dictionaries in specific usage situations by specific users needs to be investigated. Since the focus of our investigation is on LSP dictionaries for South African school learn-

ers, our assumption is that an analysis of the use of text boxes in LGP school dictionaries could provide useful information with regard to the potential use of this device in LSP dictionaries.

## 5 Use of Text Boxes in South African LGP School Dictionaries

It seems that the utilization of text boxes is gaining in popularity in the compilation of South African school dictionaries. In the most recent editions of three flagship school dictionaries, the *HAT Afrikaanse Skoolwoordeboek*, the *OXFORD Bilingual School Dictionary (English-Northern Sotho)*, and the *Oxford Bilingual School Dictionary (Afrikaans - English)* text boxes are used to provide additional data to the user. Compare the following examples from the *Oxford Bilingual School Dictionary (Afrikaans - English)* and *HAT Afrikaanse Skoolwoordeboek* in figures 1 and 2 respectively:



Figure 1: *Lekker* in *Oxford Bilingual School Dictionary (Afrikaans – English)*.



Figure 2: *Laf* in *HAT Afrikaanse Skoolwoordeboek*.

Of the three dictionaries mentioned above, the *HAT Afrikaanse Skoolwoordeboek* is the only one in which the text boxes are explicitly labelled in order to provide the user with some guidance as to the nature of the data provided in the text box. Typical labels are *Tesourus* 'Thesaurus', *Gebruik* 'Use', *Spelling* 'Spelling', *Skryfwyse* 'Conjunctive and/or disjunctive writing' and *Uitdrukking* '(Fixed) expression'. This



could be regarded as labels but also as topics or titles for the text boxes. Such an approach is supported by the role of text boxes as inserted inner texts, occupying an article slot and not only a micro-structural slot.

An analysis of the contents of text boxes found in the randomly selected alphabetical stretch 'L' in the *HAT Afrikaanse Skoolwoordeboek* is provided in table 1.

Related words and their meaning (Thesaurus boxes)	7
Spelling	4
Usage	2
Expression	1
Approximate no. of lemma signs in alphabetical stretch 'L'	321
Average of 1 text box per 23 lemma signs	

**Table 1: Text Boxes for the Alphabetical Stretch 'L'.**

Compare the results of a similar analysis for the *OXFORD Bilingual School Dictionary (English-Northern Sotho)* in table 2:

<b>Northern Sotho - English</b>		<b>English - Northern Sotho</b>	
Translation / non-translation of function words	2	Register	1
Composition of multiword lemmas	1	Pronunciation	4
Morphologically shortened forms	12	Contrast related words	2
Range of application	1	Correct usage	1
Discourse pragmatic information	5	Cross reference to another note	1
Offensive use	1		
Tense	1		
Additional information on part of speech	2		
Syntactic information	1		
Approximate no. of lemma signs in alphabetical stretch 'L'	416	Approximate no. of lemma signs in alphabetical stretch 'L'	178
Average of 1 text box per 16 lemma signs		Average of 1 text box per 19 lemma signs	

**Table 2: Text Boxes for the Alphabetical Stretch 'L' in the Oxford Bilingual School Dictionary.**

Table 3 presents the results for *The Oxford Bilingual School Dictionary (Afrikaans – English)*:

Afrikaans - English		English - Afrikaans	
Attributive vs. predicate use of adjectives	2	Register	1
Conjunctive vs. disjunctive writing	1	Pronunciation	6
Singular vs. plural use	2	Contrast related words	1
		Cross reference to another note	1
		Correct usage	1
		Singular vs. plural use	4
Approximate no. of lemma signs in alphabetical stretch 'L'	171	Approximate no. of lemma signs in alphabetical stretch 'L'	176
Average of 1 text box per 34 lemma signs		Average of 1 text box per 12 lemma signs	

**Table 3: The Oxford Bilingual School Dictionary (Afrikaans – English).**

## 6 Text Boxes in South African LSP School Dictionaries: Some Suggestions

A logical point of departure for the LSP lexicographer would be to ascertain which of the data types provided in text boxes in LGP dictionaries could also be utilized in LSP dictionaries. Perusal of existing LSP dictionaries reveals that none of the currently available ones makes use of text boxes. A lexicographic device which seems popular is the use of shaded blocks, especially in math and science dictionaries. Formulae, equations and chemical reactions mostly make up the contents of these blocks. The data provided in these blocks are however not additional to the comments on form and semantics, but form an essential and inherent part of it. The function of these shaded blocks is mainly to act as typographical markers of a specific data type. Compare the following example from the *Oxford Physical Sciences Dictionary Grades 10-12* in figure 3:

**pressure in liquids** Gravity acting on a liquid causes pressure to be exerted on the walls of its container. *This, like air pressure, acts equally in all directions but does not depend on the shape of the container. It does depend on the density of the liquid and it increases with depth. The pressure at a point in a liquid is given by:*

pressure in liquid at a point	=	height ( <i>h</i> ) of liquid at a point	×	density ( <i>ρ</i> ) of liquid	×	gravitational constant ( <i>g</i> )
-------------------------------------	---	--	---	--------------------------------------	---	---

*This is in addition to the atmospheric pressure above the liquid. In water, the pressure increases by approximately one atmosphere for every 10 metres of depth.*

**Figure 3: Pressure in Liquids in Oxford Physical Sciences Dictionary Grades 10-12.**

These shaded blocks contain microstructural data and are not inserted inner texts. Taking the function of text boxes as explained by Gouws and Prinsloo (2010) into consideration, there seems to be no reason why they cannot be utilized as a lexicographic device to assist with enhanced data retrieval, within an extended obligatory microstructure, from an LSP dictionary. Further motivation for the use of text boxes is the fact that they are increasingly utilized in LSP school dictionaries. It can therefore be assumed that the target user is already familiar with text boxes and knows what kind of data he / she can expect to find in them.

When deciding on the data types to be provided in text boxes for LSP dictionaries for South African school learners, care should be taken not to provide data that should form part of the default obligatory microstructure in text boxes. The fact that comments on form are generally inadequate in these dictionaries does not automatically qualify them as good candidates for presentation in a text box. Useful data on, for example, (regular) singular / plural forms, morphological derivations and pronunciation would indeed add value to LSP dictionaries as good lexicographic products, but these data categories should rather form part of the obligatory microstructure.

One data type that appears in text boxes in all three dictionaries mentioned above is information on related words, which is specifically aimed at distinguishing between words which are semantically and conceptually related. LSP dictionaries essentially deal with terminology in that the lemma signs are in actual fact terms. Apart from their function as subject specific linguistic labels for concepts in LSP, terms are often also used by laypersons as words in LGP, in which case the conceptual distinction between related concepts is often fuzzy. Words referring to related but distinct concepts are often used interchangeably, precisely because they are used by lay persons. Examples include 'temperature' and 'heat', 'weight' and 'mass', and 'power' and 'force'. It is likely that learners' first acquaintance with these concepts would have been in the non-technical sense. Using a text box in an LSP dictionary to explicate the distinction between conceptually related terms would add to the cognitive unravelling of subject specific concepts. A typical example of a text box in an LSP school dictionary would highlight the difference between the related concepts 'weight' as 'force experienced by an object due to gravity' and 'mass' as 'the amount of matter in an object'.

South African English is strongly based on British English. However, an increasing use of American English, also in subject field literature, is noticeable in South African English. American and British English do not always use the same terms. Lexicographers will do well to make provision for this situation by discussing the British and American use where applicable in the text boxes. Besides British and American English it is sometimes also necessary to refer to International English because these terms may differ from those found in both British and American English.

LSP lexicographers furthermore need to distinguish between culture-dependent and culture-independent terms. In South African dictionaries terms, e.g. in the field of medicine where terms from traditional healers still prevail, need additional treatment which falls outside the default microstructural slots. Text boxes could assist the lexicographer in this regard to ensure an optimal retrieval of information by the user.

An issue which is particularly relevant to the South African situation is standardization of terminology, especially with regard to the nine official South African Bantu languages. Official structures which are responsible for development, standardization and validation of terminology are largely non-functional, particularly those dealing with the Bantu languages. Consequently, one finds an almost unchecked proliferation of terminology, resulting in multiple terms for a single concept. Compare for example table 4:

	<b>Northern Sotho equivalents</b>	<b>Zulu equivalents</b>
	lerathana	inhlayiya
particle	karolonyana	intwanyana
	seripana	iphathikili
	sekgawana	

**Table 4: Northern Sotho and Zulu Equivalents for *Particle*.**

Apart from standard treatment of the lemma ‘particle’ in the dictionary, which may include listing and labelling of variant equivalent forms, data in the text box should provide the user with adequate guidance with regard to the use of standardized versus non-standardized forms and also with regard to the status of a specific term as having been standardized or not.

## 7 Conclusion

The use of text boxes as a lexicographic device in LSP dictionaries is under-utilized, even more so than in LGP dictionaries. Text boxes can potentially add the same value to LSP dictionaries as to LGP dictionaries, although the nature and content of these boxes will differ. The function of text boxes, i.e. to place more than the default focus on a specific data item, should be the main consideration when deciding on data inclusion in or exclusion from text boxes.

## 8 References

### 8.1 Dictionaries

- HAT. Afrikaanse Skoolwoordeboek.* Cape Town: Pearson Education. 2009.  
*Oxford Dictionary of Chemistry.* Oxford: Oxford University Press. 2008  
*Oxford Dictionary of Economics.* Oxford: Oxford University Press. 2009.  
*Oxford Dictionary of Music.* Oxford: Oxford University Press. 2007.  
*Oxford Dictionary of Science.* Oxford: Oxford University Press. 2010.  
*Oxford School Dictionary Afrikaans-Engels / English-Afrikaans.* Cape Town: Oxford University Press. 2007.

*Oxford School Dictionary Northern Sotho – English / English – Northern Sotho*. Cape Town: Oxford University Press. 2007.

*Oxford Physical Sciences Dictionary Grades 10-12*. Cape Town: Oxford University Press. 2013.

*Stedman's Medical Dictionary*. Philadelphia, Pa.: Wolters Kluwer Health/Lippincott Williams & Wilkins. 2012.

## 8.2 Other References

Fuertes-Oliviera, P. (ed.) 2010. *Specialised Dictionaries for Learners*. Berlin: De Gruyter.

Gouws, R H and Prinsloo, D J. 2010. *Thinking out of the box – perspectives on the use of text boxes*. Euralex Conference proceedings, 14th Euralex Conference, Leeuwarden, 6 -10 July, 2010.

Jesenšek, V. (ed.) 2013. *Specialised Lexicography*. Berlin: De Gruyter

Tarp, S. 2012. Specialised lexicography: 20 years in slow motion. *Iberica* 24, 117-128.

Tarp, S. & Fuertes-Oliviera, P. 2014. *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. Berlin: De Gruyter.

Wiegand, H.E. and Gouws, R.H. 2011. Theoriebedingte Wörterbuchformprobleme und wörterbuchformbedingte Benutzerprobleme I: Ein Beitrag zur Wörterbuchkritik und zur Erweiterung der Theorie der Wörterbuchform. *Lexikos* 21: 232-297



# Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources

Izabella Thomas<sup>1</sup>, Blandine Plaisantin Alecu<sup>2</sup>, Bérenger Germain<sup>3</sup>, Marie-Laure Betbeder<sup>4</sup>

<sup>1</sup>Centre L.Tesnière, Université de Franche-Comté

<sup>2</sup>Prolipsia, France,

<sup>3</sup>Share and Move Solutions, France

<sup>4</sup>Institut Femto-ST, Université de Franche-Comté

izabella.thomas@univ-fcomte.fr, blandine.alecu@prolipsia.com,

berenger.germain@shareandmove.fr, marie-laure.betbeder@univ-fcomte.fr

## Résumé

Dans cet article nous décrivons l'architecture générale d'une plateforme web paramétrable, modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non-terminologiques : la Station Sensunique. Conçue dans l'objectif initial de faciliter et d'accélérer le processus de constitution du lexique d'une Langue Contrôlée (LC), son champ d'application peut être élargi à l'acquisition de tout type de ressources termino-ontologiques. La Station s'articule autour de deux points de vue du processus d'acquisition de ressources : (1) chronologique (centré processus) : import des textes d'entrées, analyse automatique, analyse manuelle approfondie, validation, et enfin export; (2) ergonomique (centré utilisateur-analyste) : mise en adéquation de l'analyse selon le corpus et l'application visée, visualisation et gestion des unités lexicales candidates (ULC), recherches complexes en corpus ou dans la liste d'ULC, modification ou enrichissement des descriptions ou relations des ULC, validation progressive, demande de validation par l'expert-métier, etc. La Station dispose de deux interfaces utilisateurs faciles à manipuler ; son utilisation se fait sans aucune contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées.

**Mots-clés** : lexique; langue contrôlée; ressource terminologique; extraction des termes; acquisition des termes; plateforme terminologique

## 1 Introduction

La Station Sensunique est une plateforme web paramétrable, modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non-terminologiques créée à l'Université de Franche-Comté dans le cadre du projet ANR Sensunique<sup>1</sup>. Conçue dans l'objectif initial d'assister le

---

1 Projet ANR -EMMA-2010-039 (2010-12), <http://tesniere.univ-fcomte.fr/sensunique.html> [08/04/2014].

processus de constitution du lexique d'une Langue Contrôlée (LC), telle que définie par (Renahy et al. 2011 ; Renahy et al. 2009 ; Vuitton et al. 2009), son objectif premier est de diminuer le temps (et donc le coût) nécessaire à la conception d'une LC. Automatiser ce processus implique deux types de contraintes : (1) liées au travail terminologique et (2) liées à la conception d'une LC (Renahy et al. 2009), à savoir recenser l'ensemble du lexique d'une LC (qu'il soit terminologique ou non), gérer les constructions particulières, notamment les structures lexicales, et respecter les principes communs à toute LC (non-ambiguïté et non-redondance). Ceci présuppose la gestion de relations entre les unités lexicales (UL), qu'elles soient lexico-sémantiques (synonymie, antonymie), morphologique (flexionnelle, dérivationnelle, de variation morphosyntaxique faible, etc.) ou syntaxico-lexicales (grâce à la recherche de collocations par patterns prédictifs) (Plaisantin Alecu et al. 2012).

L'implémentation logicielle de ce processus, i.e. la Station Sensunique, automatise l'extraction d'UL candidates (ULC) à partir de corpus. Elle est configurable en finesse pour répondre aux multiples contextes d'utilisation possibles des ressources à construire, en termes de : domaines, types de textes, publics cibles des textes rédigés en une LC, ressources terminologiques préexistantes, ou plus généralement ressources linguistiques existantes et accessibles (notamment avec le courant fortement émergent des linked open linguistics data<sup>2</sup> (Chiarcos et al. 2012)). Elle offre aussi les fonctionnalités adéquates aux étapes suivantes du processus, à savoir, premières sélection et validation des UL par un analyste, seconde validation par l'expert métier et export de la ressource finale exploitable. Les interfaces utilisateurs (*interface de gestion* et *interface de travail*), très ergonomiques, ne nécessitent aucun savoir-faire technique et sont faciles à prendre en main et à explorer. L'application exploitant les lexiques d'une LC à concevoir (besoin initial de la Station) est un logiciel d'aide à la rédaction de textes techniques en LC sur mesure. La Station a été évaluée et validée dans ce cadre précis, sur l'intérêt du procédé de multi-extraction, implémenté dans la Station, pour le recensement du lexique d'une LC (Plaisantin Alecu et al. 2012).

Le processus métier d'acquisition du lexique d'une LC est très proche de l'acquisition de ressources termino-ontologiques (RTO) tel que décrit par Bourigault (2003). L'acquisition de dictionnaires, glossaires, lexiques, thesaurus<sup>3</sup> à partir de corpus doit répondre à une double contrainte de pertinence, vis-à-vis du corpus et de l'application visée, e.g. aide à la traduction, extraction d'information, indexation, etc. (Bourigault 2003).

La Station répondant à ces contraintes, son champ d'application initial (lexique d'une LC) peut être élargi à l'ensemble des RTO. Ses fondements méthodologiques et son architecture logicielle donne à la Station le potentiel d'un outil générique pouvant produire des ressources variées tout en étant fonction de l'application visée. Dans ce sens, elle suit le principe d'adéquation de Slodzian (2003) : *Qu'il s'agisse d'indexation, de mémoires de traduction bi- ou multilingues, d'aide à la rédaction de docu-*

---

2 Qu'il faudra toutefois télécharger puis convertir au format des ressources intégrables compatibles à la Station.

3 Pour la liste complète des RTO, voir Bourigault (2003).



*ments experts, les outils proposés doivent présenter un degré d'adéquation suffisant avec le problème que l'utilisateur cherche à résoudre.*

Dans la suite de cet article, nous allons tout d'abord situer notre travail (&1) dans le contexte des LC (&1.1) pour ensuite définir les besoins qui ont guidé la conception de la Station Sensunique (&1.2). Nous présenterons ensuite l'architecture générale de la Station Sensunique (&2), en mettant l'accent sur ses aspects modulaire et paramétrable. Nous décrirons les multiples possibilités de paramétrage de l'analyse automatique, en fonction du corpus d'entrée et de l'application visée. Puis nous présenterons le module de gestion qui, à partir de la liste des ULC issues du module d'analyse automatique, offre un ensemble de manipulations visant à faciliter les processus de sélection et de validation manuelle de ces ULC. Nous finirons par la présentation du module d'export qui permet également un paramétrage fin des ressources à constituer en fonction de l'application visée. Dans la conclusion, nous présenterons les possibles évolutions de la Station Sensunique.

Nous nommons « analyste » tout linguiste, terminologue, ingénieur des connaissances ou autre utilisateur de la Station et « ressources » tout type de RTO et lexique d'une LC.

## 1.1 Contexte

Les recherches concernant les Langues Contrôlées, sous leur multiples dénominations (langues simplifiées, langues construites, etc.) ne sont pas nouvelles, même si largement sur l'anglais : Kuhn (2013) recense plus de 100 LC conçues dans cette langue. Les travaux portant sur le français restent extrêmement rares : on peut citer l'initiative du COSLA (Comité de Simplification de la Langue Administrative ou le Français Rationalisé du GIFAS (Groupement d'Industries Françaises Aéronautiques et spatiales mais qui avait pour bases les règles du Simplified English de l'AECMA (Association Européenne des Constructeurs de Matériel aérospace)) (GIFAS, 1990).

Pour situer notre approche sur le panorama des travaux entrepris sur les LC, nous reprendrons une récente enquête sur l'ensemble des LC anglaises, dans laquelle Kuhn (2013 : 3) propose la définition suivante : *A controlled natural language is a constructed language that is based on a certain natural language, being more restrictive concerning lexicon, syntax, and/or semantics while preserving most of its natural properties.*

Nous sommes en accord avec cette définition, dans le sens où elle met l'accent sur les caractéristiques essentielles des LC telles que nous les concevons : une LC est toujours construite à partir d'une langue naturelle et en conserve les propriétés. Si l'on adopte la classification des LC proposée par Kuhn (2013), les LC que nous concevons sont de type CTWDAI, puisque : elles sont conçues dans l'objectif d'améliorer la compréhensibilité (*Comprehensibility*) ; elles augmentent la traductibilité (*Translability*) ; elles sont destinées à être écrites (*Written*) ; elles sont spécifiques à un domaine (*Domain-dependent*) ; elles sont initiées par une recherche académique (*Academic*) ; elles sont aussi industrielles (*Industrial*) dans la mesure où l'applicabilité des LC en industrie est un critère prépondérant des travaux présentés dans cet article.

Plus qu'une LC, nous cherchons à mettre en place un cadre méthodologique de conception assistée de LC *sur mesure*. Nous appelons LC sur mesure une LC reposant sur les besoins précis d'une structure particulière ayant pour objectifs l'amélioration de la qualité de son système documentaire et l'amélioration de la fiabilité de ses textes afin de diminuer les risques liés à leur mauvaise interprétation/application. Une telle LC est donc circonscrite à un domaine et à un environnement de rédaction précis, c'est-à-dire une activité précise, un public défini et un type de textes particulier (Renahy et al. 2009). Elle repose sur une analyse de corpus délimité, lequel doit recenser l'ensemble des textes en vigueur pour l'activité et le public concernés<sup>4</sup> (Plaisantin Alecu et al. 2012). Enfin, la LC conçue doit permettre aux personnes en charge de la rédaction technique au sein d'une structure de rédiger des documents en conformité avec les principes de cette LC.

Le cadre méthodologique d'établissement d'une LC sur mesure doit intégrer une collaboration étroite entre les linguistes et les experts du domaine et de l'activité concernés (experts métier). Il doit, de plus, prendre en considération le coût et temps de conception. Ce temps a été estimé par Jeff Allen (2005) à 5 à 10 ans dans un contexte industriel. Jusqu'à aujourd'hui, ce coût de conception était un frein à l'exploitation de LC par des structures autres que les grandes industries<sup>5</sup>. L'accessibilité des LC à des structures plus modestes par la diminution de leur coût de conception a orienté nos travaux sur le cadre méthodologique de conception des LC.

Les travaux présentés ici concernent le premier verrou que nous avons souhaité lever, à savoir le recensement du lexique d'une LC sur mesure. La spécificité du recensement de ce lexique par rapport à la conception de ressources terminologiques est qu'il se doit d'être exhaustif : toutes les unités lexicales nécessaires lors de l'écriture effective de documents, qu'elles soient ou non terminologiques, doivent être encodées (pour être utilisables) dans le dictionnaire de la LC. Cependant, ce critère d'exhaustivité lexicale ne doit pas permettre n'importe quel emploi des unités recensées par le futur rédacteur : l'emploi de certaines unités lexicales doit être contraint. Ceci implique de distinguer au moins deux types de dictionnaires pour chaque LC sur mesure : un dictionnaire des unités lexicales d'une LC et un dictionnaire des structures lexicales, deux notions que nous précisons par la suite.

Le développement de la Station Sensunique s'inscrit dans cet objectif précis : accélérer l'acquisition du lexique pour la construction du Lexique d'une Langue Contrôlée (LLC).

## 1.2 Lexique d'une Langue Contrôlée: recensement des besoins

La notion du lexique d'une LC telle que nous la considérons mérite quelques précisions dans la mesure où elle ne correspond pas à la définition du 'vocabulaire contrôlé' ('lexique contrôlé', 'termes normalisés') communément employée dans la communauté scientifique :

---

4 Pour donner un exemple de corpus délimité, le système documentaire du laboratoire d'immunologie avec lequel nous avons travaillé comporte environ 40 modes opératoires.

5 Une liste récente des entreprises ayant investi dans la création d'une LC est donnée par Uwe Muegge et disponible ici : [http://www.tekom.de/upload/2750/tcw02\\_2009.pdf](http://www.tekom.de/upload/2750/tcw02_2009.pdf) [03/04/2014].

Vocabulaire contrôlé (ou lexique contrôlé ou liste de termes normalisés) : Un vocabulaire contrôlé est un ensemble de termes reconnus, fixés, inaltérables, normalisés et validés par un groupe (une communauté de pratiques) utilisés pour indexer ou analyser le contenu et pour rechercher de l'information dans un domaine d'information défini (...).<sup>6</sup>

La première différence entre un LLC et un vocabulaire contrôlé vient de son objectif. Comme le montre la définition précédente, un vocabulaire contrôlé est dans la majorité des cas défini pour l'indexation de documents dans le but d'en faciliter la recherche. Par exemple, le MeSH considéré comme vocabulaire contrôlé (Névéol, 2004) sert à l'indexation de ressources de santé.

La deuxième différence vient de leur périmètre respectif. L'ensemble des unités composant un vocabulaire contrôlé renvoie uniquement aux termes spécifiques d'un domaine. Le LLC, quant à lui, doit permettre la rédaction d'un texte technique dans sa globalité tout en respectant l'ensemble des contraintes d'une LC. Il devra donc recenser bien plus que les termes afin de pouvoir rédiger un texte en entier. En ce sens, (Møller et al. 2006) parle de « mots » (référant alors à des unités monolexémiques comme multilexémiques) afin de ne pas confondre les unités d'un LLC avec des unités terminologiques. Nous choisissons, quant à nous, de considérer comme *unités lexicales*<sup>7</sup> (UL) toutes les unités d'un LLC.

Pour recenser l'ensemble du vocabulaire contenu dans une collection de textes techniques, plusieurs types de vocabulaires sont nécessaires, comme le souligne également Camlong (1996). Ensemble, ils constituent un continuum allant du vocabulaire terminologique du domaine jusqu'au vocabulaire général. En effet, pour écrire un protocole dans le domaine d'immunobiologie<sup>8</sup>, par exemple, plusieurs types de vocabulaire sont nécessaires :

- les termes du domaine (simples et complexes) : nominaux (*anticorps monoclonaux, réactif de lyse, tampon de fixation*), verbaux (*numéroter (les cellules), centrifuger (la suspension cellulaire)*) et adjectivaux (*aneuploïde, mononucléé*) ;
- les termes d'un autre domaine (*fenêtres informatiques, répartitions gaussiennes*) ;
- les unités du lexique général: soit entrant dans la composition des termes (*anticorps de souris*) ; soit 'autonomes' (*échantillons, divers, en particulier, étude, etc.*) ; soit potentiellement ambiguës, puisque possédant un sens spécifique dans le domaine traité (*solution, population* dans le domaine de l'immunobiologie, par exemple).

Une LC exige le respect, entre autres, des principes de non-ambiguïté et de non-redondance : une unité lexicale ne peut avoir qu'une seule définition ; et une définition ne peut correspondre qu'à une seule unité lexicale dans un domaine choisi. Pour être en conformité avec ces exigences, il s'avère

---

6 [http://appui.upmf-grenoble.fr/wiki/index.php/Vocabulaire\\_contrôlé](http://appui.upmf-grenoble.fr/wiki/index.php/Vocabulaire_contrôlé) [03/04/2014].

7 Nous reprenons ici la notion d'unité lexicale telle que définie par L'Homme (2005).

8 Tous nos exemples se basent sur le corpus-test établi dans le domaine d'immunobiologie, composé de 14 protocoles, pour un total de 10 064 mots. Ce corpus a été soumis à des extracteurs de termes, ce qui a permis de produire une liste de 2945 unités lexicales candidates (ULC), parmi lesquelles 1512 unités lexicales ont été finalement validées par un analyste.

nécessaire de contrôler l'ensemble du lexique utilisé pour la rédaction de la documentation dans un domaine. Pour exemple, il est indispensable d'éviter d'employer le mot *solution* au sens général (*Ensemble des opérations mentales, intellectuelles susceptibles de fournir une réponse théorique ou pratiques visant à la résolution, l'analyse, la compréhension d'un problème (...), TLFi<sup>9</sup>*) dans les protocoles d'immunobiologie, dans lesquels *solution* prend un sens très spécifique (*Liquide formé par la dissolution d'une substance solide (p. ex. médicament) dans un solvant, GDT<sup>10</sup>*). Il est également nécessaire d'identifier de multiples relations entre les unités lexicales ou leurs formes telles que :

- relations morphologiques (relation flexionnelle, relation dérivationnelle, relation de variation morphosyntaxique faible, etc.) ;
- relations lexico-syntaxiques (grâce à la recherche de collocations par patterns prédictifs) ;
- relations lexico-sémantiques (par exemple, relations de synonymie, homonymie).

### 1.2.1 Structures lexicales d'une Langue Contrôlée

Nous introduisons la notion de structure lexicale pour répondre au critère de non-ambiguïté tout en conservant le caractère exhaustif du lexique et la nécessité de restriction d'emploi selon le contexte. La notion de structure lexicale dépasse la définition d'unité lexicale à strictement parler puisqu'elle s'appuie sur la combinatoire lexico-syntaxique entre plusieurs unités lexicales, se situant ainsi à la frontière du lexique et de la syntaxe. Cette notion est à rapprocher de celles de classes de sélection distributionnelles, classes d'objets, fonctions lexicales, cadres prédictifs, pour ne citer que quelques unes des dénominations décrivant ces types de construction dans différentes théories linguistiques. On définit une Structure Lexicale (SL) comme un patron morphosyntaxique imposé et contrôlé par un lexème, souvent prédictif, composée d'une partie figée (lexicalisée, variable uniquement en flexion) et d'une partie variable (mais contrainte par des traits morphosyntaxiques et sémantiques). Par exemple, *marquage* est le lexème prédictif dans *marquage des cellules*, *marquage des cellules leucocytaires*, *marquage des cellules endothéliales vasculaires animales*, *marquage des cellules en suspension*. Le besoin de définir des structures lexicales vient, d'une part, de l'impossibilité d'encoder ces constructions dans un dictionnaire de termes (puisque ce ne sont pas des UL) et, d'autre part, de la nécessité de contrôler leur distribution et leur variabilité dans un environnement de rédaction d'une LC. C'est pour ces raisons que nous proposons de les recenser dans un dictionnaire spécifique, sous un format décrivant leurs principales caractéristiques :

Exemple

*marquage de* < NOM : CELLULE >

La partie variable, introduite par les chevrons (<>), est généralement définie par sa catégorie fonctionnelle (ici : NOM), qui peut être en plus caractérisée par son appartenance à une classe sémantique (ici : CELLULE).

9 Trésor de la Langue Française Informatisé, atilf.atilf.fr [03/04/2014].

10 Grand Dictionnaire Terminologique, <http://gdt.oqlf.gouv.qc.ca/Resultat.aspx> [03/04/2014].

La notion de structure lexicale est primordiale lorsque, nous éloignant de la théorie terminologique classique, nous considérons comme termes des syntagmes autres que les syntagmes nominaux. En effet, des verbes ou des adjectifs peuvent renvoyer à des concepts bien spécifiques dans des domaines précis. Certains dictionnaires terminologiques recensent d'ores et déjà des termes de nature verbale. Par exemple, on trouve aussi bien le nom 'centrifugation' que le verbe 'centrifuger' dans Le GDT. Simplement, la description de ce verbe, en s'arrêtant à l'identification de sa catégorie verbale, ne nous renseigne ni sur la présence ni sur la nature de ses compléments : pourtant, on *centrifuge* toujours *quelque chose, du sang total, du plasma sanguin* etc. Nous proposons donc de recenser ce verbe dans un dictionnaire de structures, en indiquant clairement qu'il doit être accompagné de compléments d'une certaine classe fonctionnelle et sémantique : centrifuger <NOM : SANG> .

Un autre avantage concernant l'identification des structures lexicales est l'établissement des relations entre des UL dérivées et la vérification de la cohérence du recensement du vocabulaire. En théorie, les UL prédicatives en relation de dérivation ne peuvent introduire dans leurs structures que des compléments appartenant à des classes sémantiques identiques :

#### Exemple

*numéroter* < NOM : CELLULE > ; < NOM : CELLULE > *numéroté(es)* ; *numération de* < NOM : CELLULE >

Pour rédiger : *numération des populations leucocytaires, numéroter les lymphocytes T, B et NK*

L'avantage du recensement de ces structures est double : d'une part, cela permet de contrôler que *populations leucocytaires* et *lymphocytes T, B et NK* portent bien la contrainte sémantique *CELLULE* et que *numéroter*, *numération* (voire le participe passé adjectival *numéroté*) renvoient toujours à la même classe sémantique.

En résumé, le recensement du LLC implique la création de dictionnaires pour quatre types de données : les unités lexicales terminologiques, les unités lexicales non-terminologiques, les structures lexicales terminologiques et les structures lexicales non-terminologiques.

## 2 Architecture générale de la Station Sensunique

A notre connaissance, il n'existe pas d'outil spécifique dédié à l'aide au recensement du LLC. Par contre, il existe de nombreux outils d'extraction de termes, tâche à laquelle s'apparente l'établissement du LLC. Par conséquent, après avoir testé l'hypothèse que les extracteurs de termes peuvent aider au recensement d'un LLC (Plaisantin Alecu et al. 2012), nous avons décidé de les intégrer à la Station Sensunique.

Comme toute plateforme terminologique (par exemple : HyperTerm<sup>11</sup>, Terminae<sup>12</sup>, Terminus<sup>13</sup>), la Station intègre la mise en séquençement de plusieurs outils TAL (étiquetage, lemmatisation et extraction de termes). Sa spécificité repose sur ses autres fondements méthodologiques. Le premier est la multi-extraction ou coopération de plusieurs extracteurs. Ce procédé donne des résultats significativement meilleurs que l'utilisation d'un seul extracteur et il permet de réduire le silence et filtrer automatiquement le bruit. Plus précisément, cumuler les résultats de 3 extracteurs de termes permet de couvrir 79 % des termes (par opposition à 58% de rappel pour le meilleur extracteur), et le meilleur moyen d'aider à déterminer le statut terminologique d'une ULC est de se baser sur les résultats communs aux 2 extracteurs (Yatea et Termostat dans l'étude) avec une précision de 37 % par opposition à 28% d'un seul extracteur (Plasantin Alecu et al. 2012). Ce procédé reprend celui des systèmes à base de vote (Fiscus 1997 ; Brunet-Manquat 2004 ; Matusov 2007 ; Serp et al. 2008), mais n'a jamais été employé avant nos travaux pour l'acquisition de ressources.

La seconde spécificité de la Station est le recouplement des résultats d'extraction avec des ressources lexicales et terminologiques existantes interrogées automatiquement. Ceci permet, d'une part, d'augmenter le potentiel terminologique d'une ULC déjà recensée comme terme dans une ressource externe, et d'autre part d'attribuer un statut non-terminologique à des ULC présentes dans les ressources lexicales intégrées à la Station.

Le dernier fondement méthodologique est le calcul de trois pondérations, en fonction de diverses informations recueillies automatiquement par la Station : (1) le Poids Terminologique (PT) ou potentiel d'une ULC à être un terme ; (2) le Poids de Structure Lexicale (PSL) ou potentiel d'une ULC à être transformée en une structure lexicale ; et (3) le Poids d'Unité Lexicale (PUL) ou potentiel d'une ULC à être une unité lexicale bien formée. Le calcul de ces pondérations organisent le travail de validation et facilitent la prise de décision et l'établissement de consensus entre plusieurs analystes ou entre l'analyste et l'expert métier.

Bien que chacun de ces procédés (multi-extraction, interrogation des ressources existantes, pondération) ne soient pas nouveaux, ils n'ont jamais été combinés, à notre connaissance, pour cumuler leurs bénéfices au sein d'une seule et même plateforme de recensement de ressources terminologiques ou non terminologiques.

La Station s'articule sur deux points de vue du processus d'acquisition de ressources : (1) chronologique (centré processus) : import des textes d'entrées, analyse automatique<sup>14</sup>, validation, et enfin export ; (2) ergonomique (centré analyste) : mise en adéquation de l'analyse selon le corpus et l'application visée par la ressource, visualisation des ULC (fiche lexicale et contextes d'occurrence), analyse d'un groupe d'ULC (pour l'organisation du travail ou pour des actions en masse pertinentes), re-

---

11 <http://www.tedopres.com/hyperterm-terminology-management> [03/04/2014].

12 [http://lipn.univ-paris13.fr/terminae/index.php/Main\\_Page](http://lipn.univ-paris13.fr/terminae/index.php/Main_Page) [03/04/2014].

13 <http://terminus.iula.upf.edu/cgi-bin/terminus2.0/terminus.pl> [03/04/2014].

14 L'analyse automatique comprend : étiquetage, lemmatisation, racinisation, extraction des ULC, interrogation des ressources externes et internes, calcul des pondérations.

cherches complexes en corpus ou dans la liste d'ULC, modification ou enrichissement des descriptions ou relations des ULC, validation progressive, demande de validation par l'expert-métier, etc.

De cette double articulation résultent 4 modules distincts : (1) Module de configuration de l'analyse automatique, (2) Module d'analyse automatique, (3) Module de gestion des ULC et (4) Module d'export, ainsi que 3 étapes successives d'établissement du lexique (Figure 1) :

- Etape 1 : Analyse automatique, qui extrait, à partir d'un corpus textuel, une liste composée d'unités terminologiques et non-terminologiques classées en fonction de leur statut et de leur potentiel terminologique ;
- Etape 2 : Analyse manuelle approfondie, qui consiste en un premier filtrage de la liste opéré par l'analyste pour ne retenir que les unités potentiellement valables et un second filtrage réalisé avec l'aide de l'expert métier aboutissant à des ressources validées ;
- Etape 3 : Etablissement et export paramétré des ressources établies.

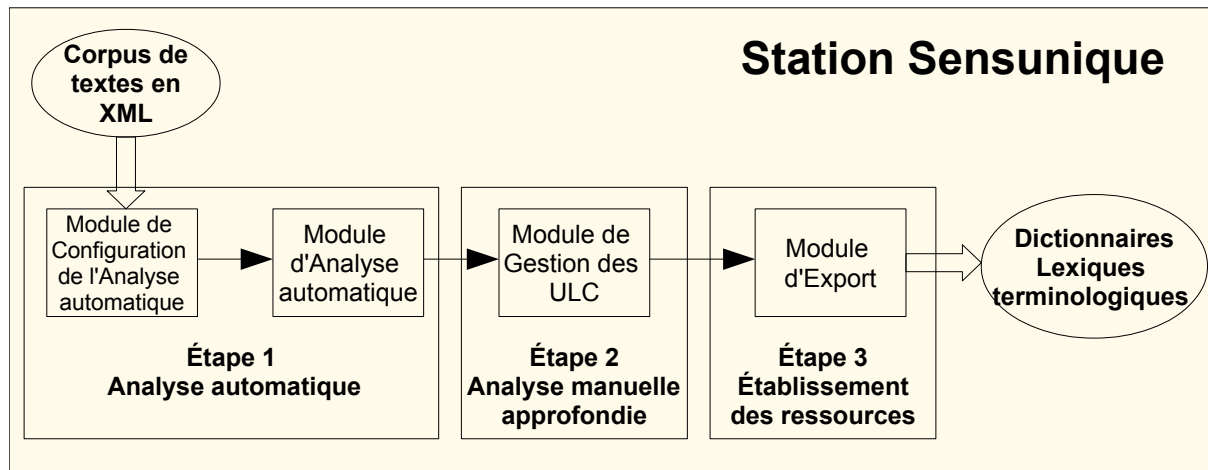


Figure 1: Schéma de la Station Sensunique.

## 2.1 Module Configuration de l'analyse automatique: Paramétrer l'analyse en fonction de la ressource visée

L'analyse automatique doit être configurée en fonction de l'application visée. L'analyste peut choisir ce qu'il souhaite exploiter comme types de corpus, outils, ressources et valeurs initiales de l'algorithme de pondération, selon leur adéquation au corpus et à la ressource visée. La qualité des résultats de l'analyse-extraction dépend de ces paramètres.



### 2.1.1 Sélection de corpus

Pour le même projet, trois types de corpus textuels<sup>15</sup> peuvent être simultanément analysés par la Station :

- (1) le Corpus d'Analyse (CA) : c'est un corpus obligatoire duquel sont extraites les ULC à analyser ;
- (2) le Corpus Support (CS) : c'est un corpus facultatif, du même domaine que le CA. En recoupant les ULC retrouvées dans les deux corpus (CA et CS), l'algorithme de pondération renforce leur potentiel terminologique. Ce procédé est inspiré de l'hypothèse de Drouin (2003) prouvant qu'une UL extraite de deux corpus différents du même domaine a plus de probabilité d'être un terme du domaine ;
- (3) le Corpus Contrastif (CC) : c'est un corpus facultatif, contenant des textes généralistes, non relatifs au domaine analysé. L'exploitation d'un CC permet à l'algorithme de pondération d'augmenter la qualité des résultats en diminuant le potentiel terminologique des ULC issues du CA et du CC à la fois. De nouveau, ce procédé est inspiré de Drouin (2003) qui prouve qu'une UL extraite d'un corpus de domaine et d'un corpus généraliste a plus de probabilité d'être une unité du lexique général qu'un terme du domaine.

Les corpus sont (ré)utilisables dans plusieurs projets. En outre, un corpus n'est pas intrinsèquement lié à un statut particulier (CA, CS ou CC) : ce statut lui est attribué en fonction du projet, par un analyste. Par conséquent, le même corpus peut être utilisé comme un CA dans un projet particulier et comme un CC dans un autre projet. Ceci permet une meilleure exploitation de différents corpus constitués dans un groupe de travail ayant des projets différents.

### 2.1.2 Sélection des outils

Pour effectuer une analyse automatique, la Station intègre un certain nombre d'outils, à savoir :

- les étiqueteurs morphosyntaxiques : statistique Treetagger (Schmid, 1994) et à base de règles Brill<sup>16</sup> (Brill 1992) ; l'annotation de chaque forme fléchie du corpus par sa catégorie morphosyntaxique et ses traits morphosyntaxiques est utile non seulement à l'analyse flexionnelle et à l'extraction de termes mais également aux diverses recherches en corpus que l'analyste peut effectuer via le concordancier intégré à la Station Sensunique ;
- l'analyseur flexionnel du français Flemm v2 et v3 (Namer, 2000) : l'annotation de chaque forme fléchie du corpus par sa forme lemmatisée est utile non seulement aux extracteurs mais également aux diverses recherches en corpus que l'analyste peut effectuer via le concordancier intégré à la Station Sensunique ;
- les extracteurs de termes Acabit (Daille, 1994), TermoStat (Drouin, 2003) et YaTeA (Aubin et al., 2006) : les extracteurs de termes fournissent chacun des propositions de termes assortis d'une ma-

---

15 Mis au préalable au format XML TEI P5, [http://www.tei-c.org/Guidelines/Customization/Lite/teiu5\\_fr.html](http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_fr.html) [04/04/2014].

16 Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.



trice morphosyntaxique ; de plus, Acabit regroupe des variantes du même terme ; Acabit et YaTeA découpent les termes composés en tête et expansion ; enfin, d'autres informations fournies par les extracteurs permettent de calculer certaines types de collocations (ULC incluses, composées et associées, cf. & 2.2) ;

- le racinisateur *Lingua::Stem*<sup>17</sup>: les racines ajoutées grâce à cet outil permettent d'identifier les relations dérivationnelles entre les ULC et sont également exploitées pour une recherche en corpus via le concordancier.

Les outils sont reliés en chaînes de travail indépendantes et parallèles. L'analyste peut sélectionner de 1 à 3 chaînes d'outils parmi : (1) TreeTagger - Termostat ; (2) Brill - Flemm v2 - Acabit ; (3) TreeTagger - Flemm v3 - YaTeA. Bien que la sélection d'une seule chaîne suffise pour lancer une analyse automatique, la Station est optimisée lors de l'emploi des 3 chaînes grâce au procédé de multi-extraction. Les résultats d'analyse de toutes les chaînes sélectionnées sont cumulés et recoupés et les informations obtenues affichées dans la liste des ULC résultant de l'analyse.

### 2.1.3 Sélection de ressources terminologiques externes (prédéfinies)

Deux ressources externes sont actuellement prédéfinies dans la Station :

- TermSciences<sup>18</sup>, portail terminologique multidisciplinaire développé par CNRS-INIST ;
- IATE<sup>19</sup>, base de données terminologique de l'Union Européenne.

L'interrogation automatique par web service de ces deux ressources externes permet de vérifier si une ULC proposée par les extracteurs est déjà recensée en tant que terme. Pour IATE, l'interrogation peut être restreinte à un domaine ou un sous-domaine précis (selon le référencement en domaines et sous-domaines EuroVoc<sup>20</sup>). Seuls les termes qui atteignent une certaine fiabilité (selon le paramètre «reliability» défini par IATE) sont retenus. Pour TermSciences, l'interrogation permet de vérifier si les constituants d'une ULC composée (sa tête ou son expansion) sont recensés indépendamment comme terme.

L'interrogation des ressources externes influe sur les pondérations, en renforçant le potentiel terminologique d'une ULC attestée dans une (ou plusieurs) ressource(s), renforcement plus ou moins fort selon si l'ULC est attestée dans sa globalité, ou si sa tête et /ou son expansion sont attestés. Elle permet ainsi de structurer le processus de validation des ULC. De plus, elle participe à l'enrichissement des informations rattachées à chaque ULC, puisque sont importées dans la Station des informations supplémentaires telles que définitions, synonymes et classes sémantiques/conceptuelles auxquelles appartient le terme attesté.

L'analyste peut choisir d'intégrer ou non l'interrogation automatique des ressources à l'analyse.

---

17 <http://search.cpan.org/~sdp/Lingua-Stem-Fr0.02/lib/Lingua/Stem/Fr.pm> [04/12/2011].

18 <http://www.termosciences.fr/> [03/04/2014].

19 <http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load> [03/04/2011].

20 [eurovoc.europa.eu](http://eurovoc.europa.eu) [03/04/2011].

#### 2.1.4 Intégration de nouvelles ressources (dites internes)

En plus de ressources externes prédéfinies, la Station permet d'intégrer à chaque nouveau projet d'autres ressources spécifiques, moyennant leur mise au format prédéfini dans la Station. Il peut s'agir aussi bien de ressources terminologiques (e.g. des dictionnaires spécialisés) qui augmentent le potentiel terminologique des ULC, que des ressources non-terminologiques (e.g. Morphalou 2.0<sup>21</sup>) qui augmentent le poids d'unité lexicale d'une ULC tout en diminuant son potentiel terminologique. Par ailleurs, des ressources constituées au préalable dans la Station, résultant d'autres projets, peuvent aussi être intégrées en tant que ressources internes.

Du fait de l'intégration dynamique des ressources, la Station peut être considérée comme évolutive, puisque chaque analyse peut être enrichie grâce à un ensemble de ressources spécifiques et appropriées.

#### 2.1.5 Paramétrage des pondérations

Trois pondérations servent à faire ressortir la fiabilité des ULC et à les classer en vue d'organiser le travail de filtrage et de validation :

(1) *Poids Terminologique (PT)* : potentiel terminologique d'une ULC calculé selon 7 critères :

- le nombre des extracteurs ayant proposé l'ULC ;
- le seuil du statut terminologique, c'est-à-dire la valeur à partir de laquelle les ULC sont considérées comme termes ;
- présence dans le CS ou le CC (cf &2.1.1) ;
- le nombre des ressources choisies ayant attesté l'ULC ;
- le type d'attestation dans une ressource (l'attestation d'ULC globale ayant plus de poids que l'attestation de la tête et/ou l'expansion seulement) ;
- la fiabilité de la ressource externe (TermSciences ou IATE)<sup>22</sup> dans le domaine analysé ;
- et la présence d'une ULC dans une ressource terminologique interne (cf. & 2.1.4).

(2) *Poids de Structure Lexicale (PSL)* : potentiel d'une ULC à être transformée en une structure lexicale, calculé selon 8 critères dont :

- l'attestation d'une ULC globale dans une ressource terminologique (qui influe négativement sur sa possibilité d'être une structure lexicale) ;
- la matrice morphosyntaxique d'une ULC (les verbes et les participes ayant plus de probabilité de constituer les structures lexicales) ;
- le nombre de dérivées et/ou de collocations construites autour d'une ULC.

---

21 Lexique de formes fléchies du français développé par ATILE, <http://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php> [03/04/2014].

22 Estimée par l'analyste.

(3) *Poids d'Unité Lexicale (PUL)* : potentiel d'une ULC à être une unité lexicale bien formée, calculé selon 2 critères :

- le nombre d'extracteurs l'ayant proposé ;
- la présence d'une ULC dans une ressource interne non-terminologique (cf. & 2.1.4).

A chacun de ces critères correspond une valeur jouant dans le calcul global de chacune des 3 pondérations. Des valeurs préexistent par défaut, mais sont ajustables par l'utilisateur.

## 2.2 Module d'Analyse automatique

Ce module a deux fonctions. La première fonction est d'annoter linguistiquement le corpus d'analyse par incorporation des résultats des étiqueteurs, lemmatiseurs et racinisateur intégrés. Sa deuxième fonction est d'extraire de ce corpus des ULC (par multi-extraction), de les décrire (résultat des extracteurs et de l'interrogation des ressources définies) et de les pondérer (résultat de l'algorithme de pondération de la Station).

Les informations issues de l'analyse automatique sont, pour chaque ULC :

- **Forme canonique** : correspond, la plupart de temps, à la suite de lemmes de chaque élément d'une ULC, ex. *membrane cellulaire* ;
- **Statut lexical** : terminologique ou non, selon le seuil du PT paramétré par l'analyste ;
  - > **Domaine(s)** (uniquement si le statut est terminologique ; correspond dans ce cas au domaine renseigné par l'analyste dans le descriptif du projet ; ex. *immunobiologie*) ;
- **Usage** : "préconisé" ou "interdit", selon les spécifications d'une LC ;
- **Catégorie(s) sémantique(s)** : proposée(s) par les ressources externes (ex. *Structures cellulaires*, d'après TermSciences) ;
- **Fréquence** : nombre d'occurrences des formes fléchies de l'ULC en corpus ;
- **Indices de confiance** :
  - > **Pondérations internes** : PT, PSL, PUL (cf & 2.1.5) ;
  - > **Indices des extracteurs externes** : indices de confiance fournis par les extracteurs, ex. *loglike* pour Acabit ;
- **Tête** : régisseur syntaxique d'une ULC, ex. *membrane* ;
- **Expansion** : complément/modifieur d'une Tête, ex. *cellulaire* ;
- **Catégorie morphosyntaxique fonctionnelle**: en général, catégorie de la Tête d'une ULC, ex. *NOM* ;
- **Matrice morphosyntaxique** : suite des catégories morphosyntaxiques de chaque élément de l'ULC., ex. *Adj Nom* ;
- **Formes fléchies** : si trouvées en corpus, assorties des traits morphosyntaxiques et fréquence ;
- **Variantes** : provenant soit du corpus analysé, soit des ressources externes, ex. *membrane plasmique* ;
- **ULC dérivées**: ULC dont un des composants appartient à la même famille dérivationnelle, ex. *membrane cellulaire, marquage de cellule* ;

- ULC homonymes: ULC homographes d'une autre catégorie morphosyntaxique que l'ULC analysée;
- Collocations (ULC liées) ;
  - > ULC incluses : une ULC incluse est une ULC dont l'intégralité se retrouve dans l'ULC analysée ; par exemple, pour l'ULC *anticorps monoclonal de souris*, les ULC incluses sont : *anticorps monoclonal*, *anticorps* ;
  - > ULC composées : une UL composée est une ULC contenant plus que l'intégralité de la ULC analysée ; par exemple pour l'ULC *anticorps monoclonal*, les ULC composées sont *anticorps monoclonal conjugué*, *anticorps monoclonal de souris*, *anticorps monoclonal HLA-B27* <sup>23</sup>;
  - > ULC associées : une ULC associée est une ULC non incluse et non composée contenant un même lemme que l'ULC analysée ; exemple : pour l'ULC *anticorps monoclonaux*, ULC associée est *solution d'anticorps* ;
- Sources :
  - > Outil(s) ayant proposée une ULC (exemple : Termostat, Acabit) ;
  - > Ressource(s) externe(s) l'attestant (exemple : TermSciences) ;
- Définition(s) (provenant de ressources externes).

Partant du principe que chaque proposition faite lors d'une analyse automatique peut être modifiée, tous les résultats du module d'analyse (excepté les indices de confiance calculés par les extracteurs et les sources) sont éditables dans le module de Gestion des ULC.

### 2.3 Module de Gestion des ULC: Faciliter le processus de sélection et de validation

Le module de gestion des ULC rassemble des fonctionnalités facilitant la seconde phase du processus d'acquisition des ressources, à savoir l'analyse manuelle approfondie. Elle consiste en un premier filtrage des ULC par un analyste et en l'établissement du consensus final avec les experts métier (Fig.1). Le parti pris fondamental de la Station est que l'analyste peut effectuer tout changement nécessaire concernant l'ensemble de résultats proposés par l'analyse automatique. Un espace dédié, appelé *interface de travail* lui sert à visualiser, à approfondir et à élargir (si besoin) les résultats afin de les valider pour construire la ressource finale.

Dans l'interface de travail, les résultats de l'analyse automatique peuvent être visualisés sous 3 modes :

- liste des ULC contenant des informations utiles pour trier et filtrer les résultats ;
- fiche lexicale de chaque ULC détaillant toutes les informations ;

---

23 Les UL incluses et composées fonctionnent de manière symétrique : si une ULC1 est ULC incluse d'une ULC2, alors l'ULC2 sera ULC composée de l'ULC1.

- fiches de relations, détaillant l'ensemble de relations entre l'ULC analysée et d'autres ULC (telles que variantes, collocations, homonymes, ULC appartenant à la même famille dérivationnelle).

L'analyste peut ajouter, modifier, compléter, valider ou supprimer toute ULC ou information à partir d'un mode de visualisation approprié. Chaque proposition/modification de données est toujours tracée, c'est-à-dire, assortie du nom de son auteur (qu'il soit analyste, outil ou ressource).

Ce module réunit également des fonctionnalités d'exploration (des ULC et de leurs informations descriptives) et d'aide à la décision (aux rejet, modification, enrichissement, validation):

- **tri et filtre** sur la liste des ULC selon 21 paramètres différents, dont fréquence, PT, PUL, extracteur(s) d'origine, ressources attestant l'ULC, matrice morphosyntaxique, catégorie sémantique etc. ; les filtres sont cumulatifs, c'est-à-dire qu'on peut filtrer les ULC selon plusieurs paramètres à la fois (par exemple, ULC proposées par Termostat, ayant atteint un certain seuil de PT et d'une matrice morphosyntaxique particulière) ;
- **projection** pour visualiser une ou plusieurs ULC en contexte d'origine (en corpus ou par phrases) ;
- **regroupement** d'ULC dans les fiches de relations ; certaines ULC sont regroupées automatiquement, mais l'analyste peut aussi établir de nouvelles relations ;
- **concordancier évolué** offrant différents types de recherche sur le corpus<sup>24</sup> : (a) simple : sur une chaîne de caractères ; (b) morphologique simple : sur un (ou une suite de) lemme(s) permettant d'identifier toutes ses formes fléchies d'une ULC ; (c) morphologique complexe : sur un (ou une suite de) radical(aux) permettant d'identifier les familles dérivationnelles ; (d) morphosyntaxique : sur une suite d'étiquettes morphosyntaxiques ; (e) recherche dite combinée permettant de coupler les types de recherches précédents. Combiner des critères appartenant à différents niveaux d'analyse linguistique permet d'imposer des contraintes plus ou moins fortes sur les motifs recherchés, et ainsi cibler ou, au contraire, élargir le champ des résultats. Par exemple, la recherche '[e]Nom [c] de [l] cellule' (exprimée sous forme d'expression régulière Sensunique) permet de cibler les groupes dont le premier élément est le Nom suivi de la préposition 'de' et d'une forme fléchie du mot 'cellule' (ex. *nombre de cellules, greffon de cellules, analyse de cellules* etc.).

L'établissement des SL se fait manuellement, à partir du regroupement de plusieurs ULC. La fonctionnalité de dégradation permet de définir une nouvelle SL (et ses différentes informations associées, telles que statut lexical, catégorie sémantique, catégorie fonctionnelle etc.) et de l'ajouter à une liste des SL. Les opérations de tri et de filtrage peuvent être effectuées sur la liste des SL comme sur la liste des ULC.

Enfin, 7 statuts de validation, correspondant à différentes étapes d'analyse ('Non validé', 'En cours d'analyse', 'A valider par les experts', 'Invalidée par les experts', 'Validé par les experts', 'Validée', 'Invalidé') permettent de suivre le processus d'établissement du lexique.

---

24 Sous forme d'Expressions Régulières (selon <http://fr2.php.net/manual/fr/book.pcre.php>) adaptées à la Station Sensunique.

## 2.4 Module d'Export: Paramétrer les ressources produites en fonction d'une application

Ce module permet d'exporter en dictionnaires les données recensées dans la station au format XML afin de :

- créer des ressources terminologiques diverses ;
- exploiter les données dans d'autres applications ;
- durant l'analyse, valider les données nécessitant des compétences spécifiques par des experts métiers.

En fonction de son objectif, l'utilisateur peut paramétrer les dictionnaires de sortie, en choisissant le(s) type(s) d'informations qu'il souhaite exporter. Toute la finesse de description d'une ressource produite dans la Station n'est pas forcément utile à l'application qui va exploiter cette ressource. De même, on peut n'être intéressé que par un périmètre restreint des UL recensées.

La sélection s'effectue à l'aide des filtres cumulatifs servant à restreindre le périmètre des données exportées selon deux axes :

- sélection des propriétés des ULC (parmi les 17 propriétés proposées, telles que définition, synonymes, matrice morphosyntaxique, catégorie sémantique, collocations, statut de validation, etc.) :

Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Définition et Variantes

Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Matrice morphosyntaxique et Fréquence

- sélection des propriétés des ULC et des valeurs de propriétés :

Exemple : dictionnaire d'UL contenant seulement : Forme canonique, Classe Sémantique, Définition, Statut de Validation ; ET le Statut de Validation est « Validée »

Le même projet permet de créer plusieurs ressources en fonction d'une application visée. Le principe est le même pour les dictionnaires de SL.

## 3 Conclusion

Conçue dans l'objectif d'optimiser (en termes de qualité et de coût) l'acquisition du lexique d'une langue contrôlée, les possibilités d'exploitation de la Station Sensunique dépassent considérablement ce champ d'action. En effet, l'éventail des configurations d'analyse (choix des outils et ressources, intégration de nouvelles ressources, personnalisation des bases de calcul des pondérations, paramétrage de l'export) en fonction de nombreux contextes d'utilisation, fait d'elle non seulement un outil d'acquisition du lexique d'une langue contrôlée, mais aussi une plateforme pertinente pour tout travail de constitution de RTO à partir de corpus.

Sur le plan méthodologique, la multi-extraction permet à la Station Sensunique d'offrir à ses utilisateurs les points forts de chaque extracteur de termes. Renforcée par l'interrogation des ressources existantes et par le principe des 3 types de corpus, la Station pondère ses résultats et permet ainsi d'organiser le processus de validation des ULC. L'interrogation des ressources existantes permet d'enrichir automatiquement la description morphologique, syntaxique et sémantique des ULC. Par ailleurs, la Station est conçue pour respecter et faciliter le processus métier d'acquisition de ressources : elle prend en compte les différentes phases de ce processus et modélise l'implication de plusieurs acteurs, y compris la validation finale par un expert-métier. De plus, l'utilisation de la Station se fait sans aucune contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées par la Station. Enfin, la Station Sensunique est dotée d'une interface utilisateur facile à manier et à explorer.

En ce qui concerne les futurs développements de la Station, plusieurs directions sont envisagées. Premièrement, nous considérons l'ajout d'autres chaînes d'outils ou le développement d'outils propres pour améliorer les performances de la Station, ainsi que l'intégration d'autres ressources externes, bien que ceci pose problème concernant les licences d'utilisation des fois difficiles à obtenir : d'où l'importance d'interagir avec les courants tels que linked open data. Deuxièmement, nous souhaitons améliorer le traitement du contenu sémantique des textes, principalement la détection des relations sémantiques et conceptuelles entre les unités lexicales. Une autre direction de recherche est l'exploitation de la plateforme pour la construction d'ontologies de domaine.

## 4 Références bibliographiques

- Allen J. (2005). How are we responding to industrial and business needs for Controlled Language and Machine Translation, *Journées Linguistiques – Langues contrôlées, traduction automatique et langues spécialisées : 5-6 May 2004 Besançon, France*, <http://web.science.mq.edu.au/~rolfs/controlled-natural-languages/papers/Jeff-Allen.pdf> [08/04/2014].
- Aubin S. et Hamon, T. (2006). Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing*, 5th International Conference on NLP (FinTAL'2006), Springer, 2006, p. 380-387.
- Brill E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing* (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155.
- Camlong (1996). Méthode d'analyse lexicale textuelle et discursive, Paris, Orphrys.
- Bourigault, D., & Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes. In *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues* (pp. 27-50).
- Brunet-Manquat F. (2004). Fusionner pour mieux analyser : Conception et évaluation de la plate-forme de combinaison. In *Actes de TALN-2004*. Fez, Maroc, 19-22 avril 2004. vol. 1/1, pp. 111-120.
- Chiarcos Ch., Hellmann S. and Nordhoff S. (2012). Linking linguistic resources: Examples from the Open Linguistics Working Group, In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), *Linked Data in Linguistics. Representing Language Data and Metadata*, Springer, Heidelberg, p. 201-216.



- Daille B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Workshop at the 32nd Annual Meeting of the ACL (ACL'94), Las Cruces, New Mexico, USA.
- Drouin P. (2003). Term Extraction Using non-Technical Corpora as Point of Leverage. In *Terminology*, vol.9, n°1, John Benjamins Publishing Company: Amsterdam/Philadelphia, p. 99-115.
- Fiscus J.G.(1997). A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), Automatic Speech Recognition and Understanding, 1997. In *Proceedings IEEE Workshop*, pp.347-354.
- GIFAS, *Guide du rédacteur*. Groupement des Industries Françaises Aéronautiques et Spatiales, Paris, France, 1990.
- Kuhn T. (2013) A Principled Approach to Grammars for Controlled Natural Languages and Predictive Editors. *Journal of Logic, Language and Information*, 22(1).
- Kuhn T. (2014) A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1), 2014.
- L'Homme M-C. (2005). Sur la notion de terme. In *Meta: journal des traducteurs / Meta: Translators' Journal*, vol. 50, n° 4, p. 1112-1132. <http://id.erudit.org/iderudit/012064>.
- Matusov E. et al. (2007). System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7):1222-237.
- Møller M.H., Christoffersen E., Hansen M. (2006). Building a Controlled Language Lexicon for Danish. In *LSP and Professional Communication*, vol. 6, Nr. 1, p. 12-38.
- Namer, F. (2000). FLEMM: un analyseur flexionnel du français à base de règles. In *Traitement Automatique des Langues*; vol. 41/2, p. 523-547.
- Névéal A. (2004). Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé. In *RECITAL 2004*, Fès.
- Plaisantin Alecu B., Thomas I., Renahy J. (2012). La « multi-extraction » comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques. In Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN, ATALA/AFCP, pp.511-518, <http://www.aclweb.org/anthology/F/F12/F12-2047>.
- Renahy J., Devitre D., Thomas I., Dziadkiewicz A. (2009). Controlled language norms for the redaction of security protocols: finding the median between system needs and user acceptability. In *Proceedings of the 11th International Symposium on Social Communication*, Santiago de Cuba, Cuba, 19-23 January 2009, pp. 289-293.
- Renahy J., Thomas I., Chippeaux G., Germain B., Petiaux X., Rath B., De Grivel V., Cardey S., Vuitton DA. (2011). La langue contrôlée et l'informatisation de son utilisation au service de la qualité des textes médicaux et de la sécurité dans le domaine de la santé. In P. Staccini, A. Harmel, S. Darmoni, R. Gouider, *Systèmes d'information pour l'amélioration de la qualité en santé*, Comptes rendus des quatorzièmes Journées francophones d'informatique médicale (JFIM'2011), Tunis, 23-24 septembre 2011, Springer-Verlag.
- Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing* (Vol. 12, pp. 44-49).
- Thomas I., Betbeder M.L., Renahy J., Vuitton DA. (2012). *Optimisation d'un logiciel pour la rédaction de textes techniques de qualité : application-pilote au domaine de la santé*. Projet ANR -EMMA-2010-039 (2010-12), rapport final (non-publié).
- Serp C., Cazal E., Laurent A., Roche M. (2008). TERVOTIQ : un système de vote pour l'extraction de la terminologie d'un corpus en français médiéval. In *9èmes journées internationales d'analyse statistique de données textuelles (JADT'2008)*, Lyon, 2008.
- Slodzian, M. (2000). L'émergence d'une terminologie textuelle et le retour du sens. Le sens en terminologie, 61-85.
- Vuitton DA., Aishan A., Renahy J., Jin G., Wu X., De Grivel V., Cardey S. (2009). Controlled language: a Linguistic Concept to Improve Health Care Safety in a "Globalised" World? Application to Medical Proto-



cols Written within the Hospital Accreditation/Certification Framework in France and China. In *ISMTCL Proceedings, International Review BULAG*, PUFC, ISBN 978-2-84867-261-8, pp. 260-268.

### **Remerciements**

Nos travaux ont été financés par l'Agence Nationale de la Recherche, programme Emergence 2010. Nous remercions toute l'équipe du projet Sensunique, les auteurs des outils intégrés et les organismes gérant les ressources terminologiques prédéfinies dans la Station Sensunique.



# Station Sensunique: une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non terminologiques (orientée Langues Contrôlées)

Izabella Thomas<sup>1</sup>, Blandine Plaisantin Alecu<sup>2</sup>, Bérenger Germain<sup>3</sup>, Marie-Laure Betbeder<sup>4</sup>

<sup>1</sup>Centre L. Tesnière, Université de Franche-Comté

<sup>2</sup>Prolipsia, France

<sup>3</sup>Share and Move Solutions, France

<sup>4</sup>Institut Femto-ST, Université de Franche-Comté

izabella.thomas@univ-fcomte.fr, blandine.alecu@prolipsia.com,

berenger.germain@shareandmove.fr, marie-laure.betbeder@univ-fcomte.fr

## Résumé

Dans cet article, nous présentons le fonctionnement et les services proposés par la Station Sensunique, une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et du lexique d'une Langue Contrôlée. La Station prend en charge et facilite l'ensemble de ce processus, à travers une analyse automatique du corpus, puis la possibilité d'approfondir l'analyse manuellement, pour aboutir à la création des fiches terminologiques/lexicales et des ressources exportables. La Station est un outil prêt à l'emploi, ergonomique, facile à prendre en main et à exploiter à travers ses différentes interfaces. Elle allie les avantages d'une analyse automatique (rapidité, coût) avec l'exactitude et la fiabilité d'une analyse humaine.

**Mots-clés:** lexique; langue contrôlée; ressource terminologique; extraction des termes; acquisition des termes; plateforme terminologique

## 1 Description générale

La Station Sensunique est une plateforme Web modulaire, collaborative et évolutive d'acquisition assistée de ressources terminologiques et non terminologiques. Elle a été conçue à l'Université de Franche-Comté durant le projet ANR-EMMA-2010-039 intitulé Sensunique<sup>1</sup> (2011-2012) dans l'objectif d'accélérer le processus d'établissement du lexique d'un domaine ou d'une Langue Contrôlée (LC). Elle prend en entrée un corpus de textes et produit en sortie des diverses ressources (dictionnaires, lexiques, glossaires) enrichies de multiples informations linguistiques. Elle s'appuie sur une analyse automatisée de corpus, dont les résultats sont la base d'une phase de validation manuelle effectuée par un analyste, puis par un expert d'un domaine. La spécificité de cette station par rapport à d'autres

---

1 <http://tesniere.univ-fcomte.fr/sensunique.html> [08/04/2014].

plateformes de travail terminologique (HyperTerm<sup>2</sup>, Terminae<sup>3</sup>, Terminus<sup>4</sup>), repose (Thomas et al. 2014) :

- d'une part, sur les choix méthodologiques sur lesquels elle est fondée : la collaboration de plusieurs outils TAL (Plaisantin Alecu et al. 2012), l'interrogation automatique des ressources terminologiques existantes, l'intégration et l'interrogation des ressources terminologiques ou lexicales propres; ceci en vue de faciliter le travail de l'analyste en lui proposant une liste d'Unités Lexicales Candidates (ULC) pondérées et enrichies de multiples informations linguistiques acquises automatiquement ;
- d'autre part, sur les objectifs spécifiques desquels elle découle, dont notamment le recensement du Lexique d'une Langue Contrôlée (LLC), définie comme une *Langue Contrôlée sur mesure* (Renahy et al. 2009, 2011).

La spécificité d'un tel lexique est qu'il se doit d'être exhaustif : toutes les unités nécessaires lors de l'écriture de documents, qu'elles soient ou non terminologiques, doivent être encodées (pour être utilisables) dans le dictionnaire d'une LC. De plus, cette contrainte d'exhaustivité du niveau lexical d'une LC implique de distinguer au moins deux types de dictionnaires : un dictionnaire du lexique d'une LC et un dictionnaire des structures lexicales (Thomas et al. 2014), chacun pouvant être soit terminologique soit général. Une autre contrainte liée à la conception de LLC provient des principes d'une LC : de non-ambiguïté (à une Unité Lexicale (UL) ne correspond qu'un sens) et, inversement, de non-rendance (à un sens correspond une et une seule UL). Ceci présuppose la gestion de la synonymie, et plus généralement la gestion des relations entre plusieurs unités lexicales (telles que homonymie, dérivation, collocations etc.).

La station est orientée analyste - utilisateur. Tous les résultats (ULC ou informations associées) sont des propositions que l'utilisateur peut modifier (ajouter, modifier, compléter, valider ou invalider). Il est assisté dans ce processus par un ensemble de fonctionnalités d'exploration des résultats, à savoir : visualisation des ULC sous plusieurs modes, tris et filtres sur la liste des ULC, projection des ULC sur le corpus d'origine, regroupement de différentes ULC, recherches sur le corpus à l'aide d'un concordancier avancé, etc.

L'utilisation de la Station se fait, sans contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisés par la Station. Enfin, la Station Sensunique est dotée d'une interface utilisateur facile à manier et à explorer.

---

2 <http://www.tedopres.com/hyperterm-terminology-management> [08/04/2014].

3 [http://lipn.univ-paris13.fr/terminae/index.php/Main\\_Page](http://lipn.univ-paris13.fr/terminae/index.php/Main_Page) [08/04/2014]

4 <http://terminus.iula.upf.edu/cgi-bin/terminus2.0/terminus.pl> [08/04/2014].

## 2 Architecture et services

La station Sensunique fonctionne de façon modulaire, chaque module proposant à l'utilisateur plusieurs services. Les modules sont organisés pour correspondre au processus d'acquisition de ressources, divisé en plusieurs étapes (représenté par la Figure 1).

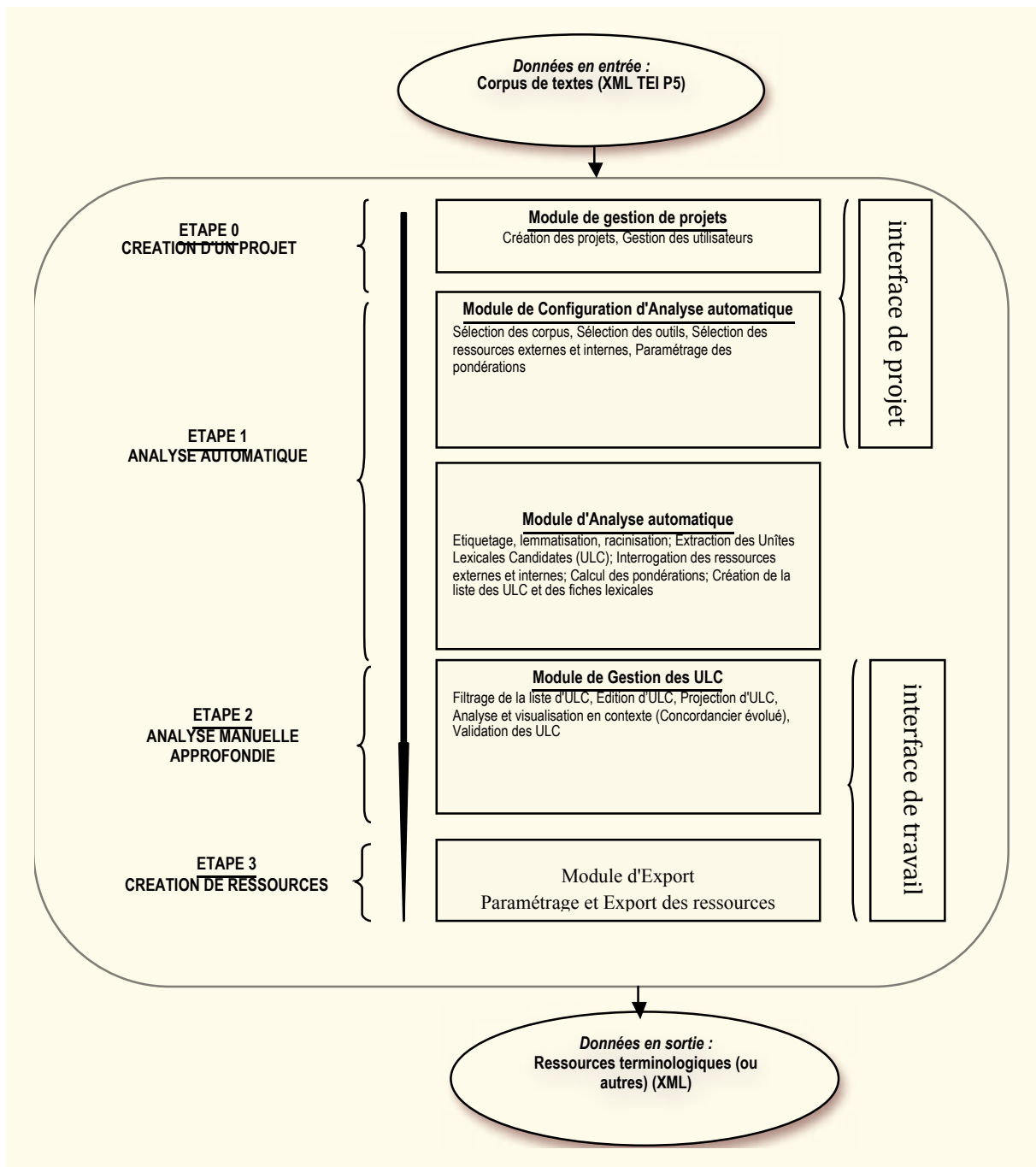


Figure 1: Schéma de la Station Sensunique.

L'utilisateur peut interagir avec la Station à partir de deux interfaces : (1) *l'interface de projet*, qui sert à définir et paramétrer un projet en termes d'utilisateurs, de corpus, d'outils et de ressources utilisés pour l'analyse automatique et (2) *l'interface de travail*, qui permet d'explorer les résultats d'analyse automatique en vue de l'établissement d'une ressource finale. Nous présentons les diverses fonctionnalités de la Station en suivant le processus chronologique d'un utilisateur souhaitant implémenter un nouveau projet.

## 2.1 Etape 0: Création d'un projet

La création d'un projet commence par l'établissement d'un groupe de travail, c'est-à-dire par la déclaration d'un ou plusieurs utilisateurs ayant le droit de travailler sur le projet. En effet, la Station est collaborative : elle permet à plusieurs utilisateurs d'interagir sur la même tâche. Elle assure aussi la traçabilité de toute modification (correction, modification ou complétion des données) grâce à une étiquette portant le nom de l'utilisateur concerné. Ces étiquettes de traçabilité permettent également de distinguer les données obtenues de façon automatique des données créées ou modifiées par un utilisateur.

Un groupe de travail peut créer plusieurs projets (Figure 2) ; chaque projet, en plus du nom et de sa date de création, est caractérisé par son domaine et le public auquel il est destiné. Un projet ne peut contenir qu'un corpus pour chaque type de corpus permis : Corpus d'Analyse, Corpus Support, Corpus Contrastif (Thomas et al., 2014). Les corpus doivent être chargés dans la Station au format XML TEI P5<sup>5</sup>; la conversion de tout document vers ce format doit être faite au préalable en utilisant, par exemple Oxgarage<sup>6</sup>, un convertisseur automatique de format de documents en ligne.

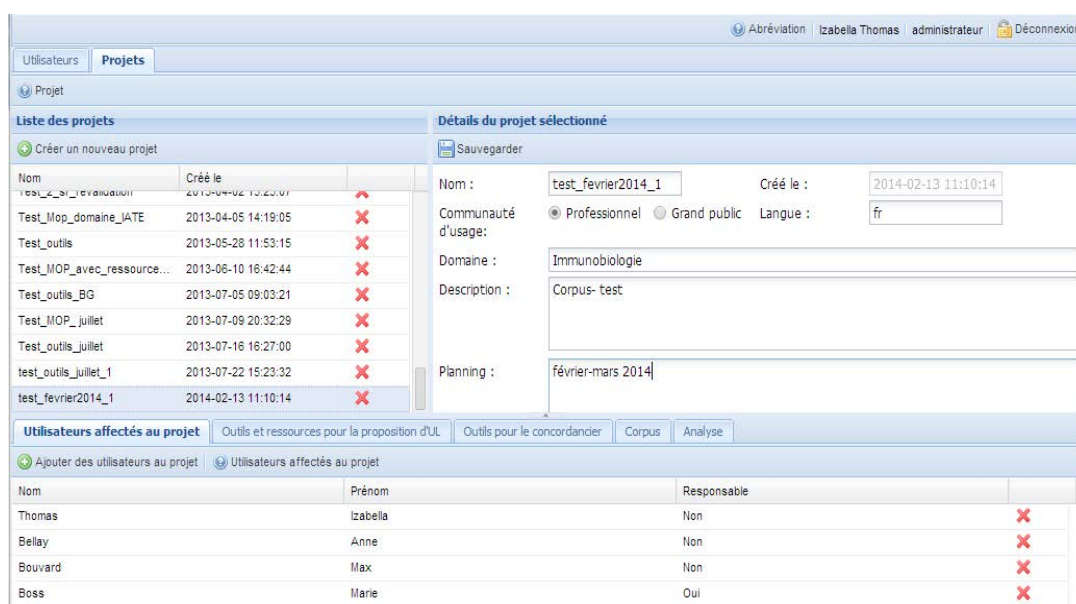


Figure 2: Déclaration d'un projet (capture d'écran).

5 [http://www.tei-c.org/Guidelines/Customization/Lite/tei5\\_fr.html](http://www.tei-c.org/Guidelines/Customization/Lite/tei5_fr.html) [04/04/2014].

6 <http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient> [accédé le 08/04/2014].

## 2.2 Etape 1: Analyse automatique

Cette étape est composée de deux sous-étapes : (1) la configuration de l'analyse automatique par l'utilisateur et (2) l'analyse automatique.

La Station Sensunique est hautement paramétrable (Thomas et al. 2014), dans l'objectif d'assurer l'adéquation de l'analyse avec la ressource à construire. Les paramétrages se font à partir de l'interface de projet (Figure 3): en fonction de son objectif, l'utilisateur peut choisir les chaînes d'outils et les ressources externes à interroger, paramétrer l'algorithme de pondération et incorporer des ressources internes (moyennant leur mise en format appropriée). Le choix de ressources internes n'est pas restreint, ce qui assure à la Station son caractère évolutif.

La qualité de l'analyse et le nombre d'informations recueillies par la Station Sensunique à partir de diverses ressources dépend du paramétrage effectué par l'utilisateur.

Actuellement, les outils et les ressources intégrés à la Station Sensunique sont les suivants :

- étiqueteurs morphosyntaxiques : TreeTagger<sup>7</sup> et Brill<sup>8</sup> (Brill 1992) ;
- analyseur flexionnel du français : Flemm v2 et v3 (Namer, 2000) ;
- extracteurs de termes : Acabit (Daille 1994), TermoStat (Drouin 2003) et YaTeA (Aubin et al. 2006) ;
- racinisateur Lingua::Stem<sup>9</sup> ;
- TermSciences<sup>10</sup>, portail terminologique multidisciplinaire développé par CNRS-INIST ;
- IATE<sup>11</sup>, base de données terminologique de l'Union Européenne.



Figure 3: Sélection des outils et des ressources (capture d'écran).

En ce qui concerne la durée de l'analyse automatique, elle dépend de la taille du corpus et du paramétrage utilisateur. Pour un corpus de 50 fichiers représentant un volume de 507 Ko, soumis aux outils TreeTagger et Flemm v3, jugés représentatifs du comportement global des outils lors d'une analyse, le temps d'exécution s'élève à 7,735 s. Il n'augmente que modérément avec l'augmentation du volume

7 <http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schmid/Tagger-Licence> [08/04/2014].

8 Avec le lexique et le fichier de règles fournis par l'ATILF-CNRS, de Nancy.

9 <http://search.cpan.org/~sdp/Lingua-Stem-Fr0.02/lib/Lingua/Stem/Fr.pm> [04/12/2011].

10 <http://www.termosciences.fr/> [08/04/2014].

11 <http://iate.europa.eu/iatediff/SearchByQueryLoad.do?method=load> [08/04/2014].

d'informations à traiter (3,486 s pour un 1 corpus de 20 fichiers représentant un volume de 214 Ko). Par contre, l'interrogation des ressources externes (par web services) peut rallonger considérablement le temps d'exécution (jusqu'à plusieurs heures).

## 2.3 Etape 2: Analyse manuelle approfondie

Les résultats de l'analyse automatique sont affichés dans l'interface de travail. Cette interface est divisée en 4 espaces (Figure 4) que l'on peut repositionner, redimensionner, afficher ou cacher.

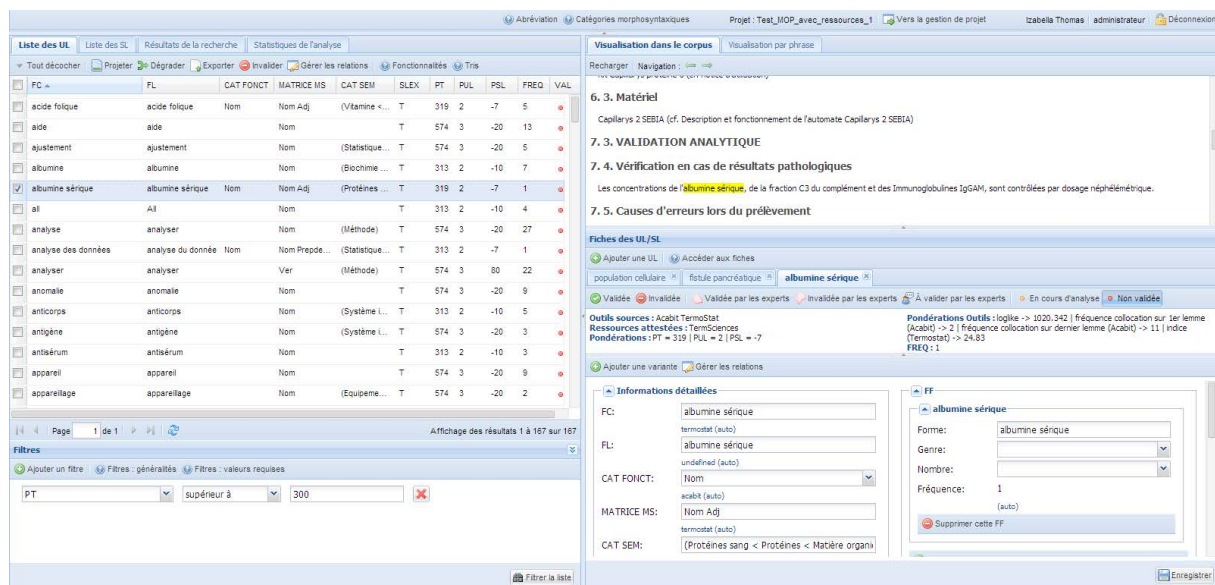


Figure 4: Interface de travail, vue globale (capture d'écran).

Dans l'espace 1, l'analyste visualise la liste des ULC assorties d'informations servant aux tris (par simple clic sur la colonne correspondante) et filtres des résultats. L'espace 2 permet de filtrer la liste des ULC à l'aide de 21 paramètres, telles que la fréquence, l'extracteur-source ou la classe sémantique. Les filtres sont cumulatifs : il est possible de filtrer la liste selon plusieurs paramètres simultanément ; par exemple, les ULC de matrice morphosyntaxique Nom Adj, proposées par TermoStat, avec au moins 20 occurrences dans le corpus. L'espace 3 sert à visualiser les ULC dans leur contexte initial, en projetant une ou plusieurs ULC sélectionnée(s) dans la liste, en corpus ou par phrase (cf. exemple de «albumine sérique» sur la Figure 4). L'espace 4 sert à afficher les fiches lexicales des ULC sélectionnées dans la liste. Une fiche lexicale comporte l'ensemble d'informations concernant l'ULC, c'est-à-dire une description complète de la forme canonique d'une ULC avec des spécifications sur ses formes fléchies. L'analyste peut modifier, ajouter, valider ou enlever les informations.

Chaque ULC est aussi assortie d'une *fiche de relations* qui permet de visualiser et/ou de définir un réseau de relations qu'elle entretient avec d'autres formes recensées. Il s'agit de :

- relations morphologiques (recensement de formes fléchies (FF) d'une ULC, recensement de formes en relation de dérivation avec une (partie de) ULC (cf. UL dérivées sur la Figure 5) ;



- relations lexico-syntaxiques (cf. UL incluses, composées et associées sur la Figure 5) ;
- relations lexico- sémantiques (cf. UL homonymes, UL variantes sur la Figure 5).

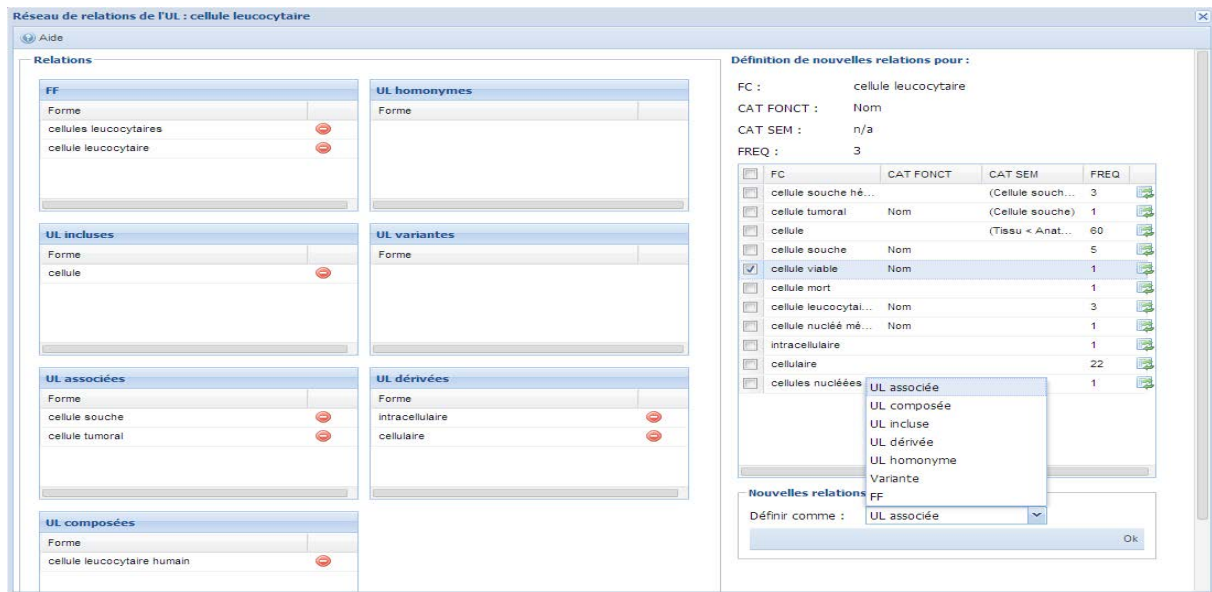


Figure 5: Exemple d'une fiche de relations pour UL *cellule leucocytaire* (capture d'écran).

Certaines relations entre les ULC sont proposées automatiquement, soit par les outils/ ressources, soit à partir de calculs effectués par la station. L'analyste peut les valider/invalider, mais aussi établir des nouvelles relations (cf. Figure 5, «*cellule viable*» sélectionnée dans la liste de gauche est définie comme une UL associée à «*cellule leucocytaire*»).

Afin d'approfondir l'analyse en explorant le corpus, la Station Sensunique propose un concordancier évolué, qui permet de visualiser les occurrences d'une forme en contexte, dans le corpus ou dans les phrases isolées (Figure 6). La recherche s'effectue soit directement à partir des informations saisies par l'utilisateur, soit en demandant à la Station de calculer les informations linguistiquement plus complexes (lemmes, racines ou catégories morphosyntaxiques) concernant les formes à rechercher. Le concordancier est dit 'évolué' au sens où il permet différents types de recherche, allant d'une simple recherche sur une chaîne de caractères jusqu'à une recherche impliquant la combinaison de différents critères linguistiques (lemmes, racines, catégories morphosyntaxiques d'une ou plusieurs unités lexicales). Combiner des critères appartenant à différents niveaux d'analyse permet d'imposer des contraintes plus ou moins fortes sur les motifs recherchés, et ainsi cibler (ou, au contraire, élargir) le champ des résultats. Par exemple, la recherche '[l]cellule [e]Adj' (exprimée sous forme d'expressions régulières Sensunique) permet de cibler les groupes composés d'une des formes fléchies du mot «*cellule*» suivie d'une forme à fonction adjectivale (ex. «*cellules nucléées*», «*cellules totales*», «*cellules mortes*» etc.) (cf. Figure 5).

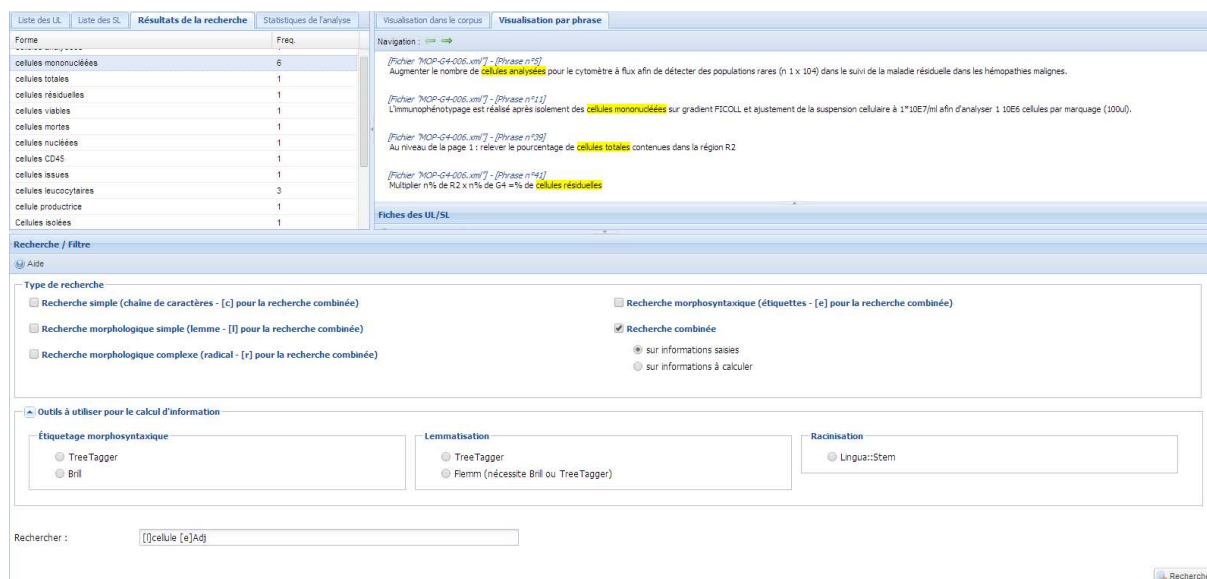


Figure 6: Concordancier évolué (capture d'écran).

## 2.4 Etape 3 : Création de ressources

A tout moment de son travail, l'utilisateur peut exporter les données recensées dans la station afin de les exploiter dans d'autres applications, les valider par un expert-métier ou simplement, créer une ressource terminologique finale. Les données à exporter, au format XML, peuvent être sélectionnées et restreintes à certaines valeurs grâce à un système de filtres cumulatifs (Figure 7).

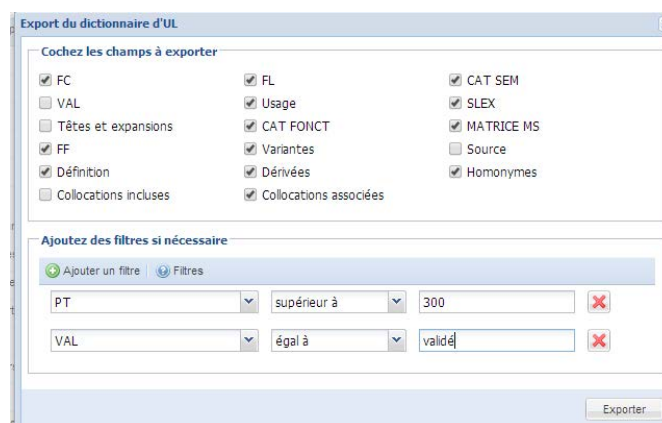


Figure 7: Paramétrage de l'export (capture d'écran).

## 3 Conclusion

Les fondements méthodologiques et l'architecture logicielle de la Station Sensunique lui permettent de dépasser son objectif initial (établissement du lexique d'une LC) et lui donne le potentiel d'être un

outil générique assistant l'établissement de diverses ressources terminologiques : glossaires, dictionnaires, bases de données, thesaurus, index, termino-ontologies etc., aussi bien pour une consultation directe que comme entrées pour d'autres applications en TAL (recherche et extraction d'information, systèmes d'indexation, acquisition et représentation des connaissances etc.). La facilité d'utilisation de la Station Sensunique nous semble un véritable avantage : c'est un outil prêt à l'emploi, ergonomique, facile à prendre en main et à exploiter. Elle allie les avantages d'une analyse automatique (rapidité, coût) et l'exactitude et la fiabilité d'une analyse humaine. L'utilisation de la Station se fait sans aucune contrainte technique ni installation préalable, à partir d'une interface web qui intègre l'ensemble des outils et ressources utilisées par la Station.

Nous projetons d'intégrer de nouvelles fonctionnalités à la Station Sensunique, concernant le traitement du contenu sémantique des textes, principalement l'amélioration de la détection des relations sémantiques et conceptuelles entre les unités lexicales. Une autre direction de recherche inclut la construction d'ontologies de domaine.

## 4 Références bibliographiques

- Aubin S. et Hamon, T. (2006). Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processing*, 5th International Conference on NLP (FinTAL'2006), Springer, p. 380-387.
- Brill E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing* (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155.
- Daille B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Workshop at the 32nd Annual Meeting of the ACL (ACL'94), Las Cruces, New Mexico, USA.
- Drouin P. (2003). Term Extraction Using non-Technical Corpora as Point of Leverage. In *Terminology*, vol.9, n°1, John Benjamins Publishing Company: Amsterdam/Philadelphia, p. 99-115.
- Namer, F. (2000). FLEMM: un analyseur flexionnel du français à base de règles. In *Traitement Automatique des Langues*; vol. 41/2, p. 523-547.
- Plaisantin Alecu B., Thomas I., Renahy J. (2012). La « multi-extraction » comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques, In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, volume 2 : TALN, ATALA/AFCP, pp.511 518, <http://www.aclweb.org/anthology/F/F12/F12-2047>.
- Renahy J., Devitre D., Thomas I., Dziadkiewicz A. (2009). Controlled language norms for the redaction of security protocols: finding the median between system needs and user acceptability, in *Proceedings of the 11th International Symposium on Social Communication*, Santiago de Cuba, Cuba, 19-23 January 2009, pp. 289-293.
- Renahy J., Thomas I., Chippeaux G., Germain B., Petiaux X., Rath B., De Grivel V., Cardey S., Vuitton DA. (2011). La langue contrôlée et l'informatisation de son utilisation au service de la qualité des textes médicaux et de la sécurité dans le domaine de la santé, In P. Staccini, A. Harmel, S. Darmoni, R. Gouider, *Systèmes d'information pour l'amélioration de la qualité en santé*, Comptes rendus des quatorzièmes Journées francophones d'informatique médicale (JFIM'2011), Tunis, 23-24 septembre 2011, Springer-Verlag.
- Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing* (Vol. 12, pp. 44-49).

Thomas I., Plaisantin Alecu B., Germain B., Betbeder M.L. (2014). Station Sensunique: Architecture générale d'une plateforme web paramétrable, modulaire et évolutive d'acquisition assistée de ressources. In *Proceedings of Euralex 2014*, EURAC, Institute for Specialised Communication and Multilingualism.

# **Lexicography of Lesser Used Languages**



# Towards an Integrated E-Dictionary Application – The Case of an English to Zulu Dictionary of Possessives

Sonja Bosch\*, Gertrud Faaß<sup>1</sup>

\*University of South Africa, University of Hildesheim

E-mail: boschse@unisa.ac.za, faassg@uni-hildesheim.de

## Abstract

This paper describes a first version of an integrated e-dictionary translating possessive constructions from English to Zulu. Zulu possessive constructions are difficult to learn for non-mother tongue speakers. When translating from English into Zulu, a speaker needs to be acquainted with the nominal classification of nouns indicating possession and possessor. Furthermore, (s)he needs to be informed about the morpho-syntactic rules associated with certain combinations of noun classes. Lastly, knowledge of morpho-phonetic changes is also required, because these influence the orthography of the output word forms. Our approach is a novel one in that we combine e-lexicography and natural language processing by developing a (web) interface supporting learners, as well as other users of the dictionary to produce Zulu possessive constructions. The final dictionary that we intend to develop will contain several thousand nouns which users can combine as they wish. It will also translate single words and frequently used multiword expressions, and allow users to test their own translations. On request, information about the morpho-syntactic and morpho-phonetic rules applied by the system are displayed together with the translation. Our approach follows the function theory: the dictionary supports users in text production, at the same time fulfilling a cognitive function.

**Keywords:** Zulu; less-resourced languages; learner's dictionary, extralexigraphic features; integrated e-dictionary; possessives

## 1 Introduction

This paper describes an integrated approach to e-dictionaries as learning tools informed by the Function Theory of Lexicography. Issues raised by information science as well as learners' levels of knowledge in text production (cf. Bothma and Tarp 2012) are taken into consideration. The case of possessive constructions in Zulu will be used as example. We will explore ways of using an e-dictionary integrated in a learning environment to translate possessives from English to Zulu. We aim to enhance the language learning experience of users in a didactically valuable manner.

---

1 This project is part of the activities of Gertrud Faaß as an affiliated research fellow of the University of South Africa.

Our users in focus are not only language learners, but also others who do not strictly fall into one of the categories defined by the function theory, e.g. advanced speakers of the language possibly requiring knowledge about correct grammatical constructions, in this case possessives. We combine lexicographic and morpho-syntactic information and take particular extra-lexicographical situations and user needs into consideration.

To our knowledge this research on an integrated e-dictionary for Zulu is novel work. While e.g. Radtke and Heid (2012:798) propose an approach to address the needs of text production by making use of corpus data, and Prinsloo et al. (2012:292) offer graphical decision trees as a solution for explaining and translating Northern Sotho kinship terminology and copulative constellations to users, we go a step further by integrating morpho-syntactic and lexicographical information as well. We address text production by showing users how to form possessives and by offering information about how these are formed. In addition, we address the cognitive function by assisting users to understand the morphology of the other language (Gouws 2007:85). Though the e-dictionary described here in the long term may be developed into an intelligent computer assisted language learning (ICALL) system, this e-dictionary is not as yet, comparable to what is already developed for other Bantu languages, as e.g. described by Katushemerewe and Nerbonne (2013) for Runyakitara or Hurskainen (2009) for Swahili.

In the next section, a brief background on the morphological structure of Zulu will be given, followed by an exposition of the complexities encountered with the translation of possessives from English into Zulu. Thereafter the implementation of the integrated e-dictionary application will be discussed with detailed exemplification, followed by a conclusion and notes on future work.

## **2 Background of the morphological structure of Zulu**

Zulu [zul] belongs to the Bantu languages which have a rich agglutinating morphological structure, based on two principles, namely the nominal classification system, and the concordial agreement system. According to the nominal classification system, nouns are categorized by prefixal morphemes, which for analysis purposes have been assigned class numbers. These noun class prefixes use concordial agreement to link the noun to other words in the sentence such as verbs, adjectives, pronouns, possessives etc. (cf. Poulos and Msimang 1998).

Some degree of morpho-phonological complexity occurs which can mainly be ascribed to the phonological phenomena at morpheme boundaries resulting from the conjunctive orthography of the Zulu language. These phenomena are mostly predictable and rule-based. Because of its complexity, the construction of the possessive constitutes a particular challenge for language learners. First, information on the noun class of the possession and the possessor is required in order to determine the connecting element: it can either be a “regular” class specific possessive concord containing the possessive marker *a*, or if the possessor is a noun belonging to class 1a, then a special possessive marker



*ka* is used (cf. Poulos and Msimang 1998:146). After selecting the appropriate connecting element, the speaker has to know the morpho-phonetic rules for cases where the connecting element and the possession are to be fused. See some examples in the following table:

English possessive construction	Zulu possession (class)	connecting element	Zulu possessor (class)	Translation	Rule
Sipho's cats lit. 'cat's of Sipho'	amakati (6)	ka-	uSipho (1a)	amakati kaSipho	fuse connecting element with possessor by deleting initial vowel <i>u-</i>
old women's box lit. 'box of old women'	ibhokisi (5)	la-	izalukazi (8)	ibhokisi lezalukazi	fuse connecting element with possessor by means of vowel coalition <i>a+i &gt; e</i>
mother's bread lit. 'bread of mother'	isinkwa (7)	sika-	<i>umama</i> (1a)	isinkwa sikamama	fuse connecting element with possessor by deleting initial vowel <i>u-</i>

**Table 1: Examples of Zulu possessive constructions.**

When a possessor is to be replaced by a pronoun, its noun class again plays an important role. For instance when translating the phrase 'their box' (lit. 'box of them') referring to 'old women' as possessor, the connecting element or possessive concord of the possession ('box') class 5 *la-* is prefixed to the (abbreviated) pronoun of class 8 *zo*, resulting in *ibhokisi lazo*.

Examples of even higher complexities are highlighted in the implementation section below.

### 3 Challenges for learners

The microstructure of bilingual dictionaries usually contains source language lemmas and inter alia their respective translation equivalent in the target language. The user needs in terms of the function theory could be on the level of text reception<sup>2</sup>, text production or cognitive needs in order to understand the construction. Occasionally a language learner might want to query short phrases, such as possessive structures e.g. "Sipho's food", or "(the) doctors' medicine" without having to look up the possessor, the possession as well as the grammatical rules on how to formulate the construction.

Although Computer Aided Language Learning (CALL-) applications offer such possibilities, these applications usually only have restricted dictionaries that contain solely what is described in the learning material accompanying them. This is usually sufficient, as the learner can learn about the rules

<sup>2</sup> The work described here, however, does not cater for receptive user needs as this issue is foreseen to be taken into account at a later stage.

for producing the constellation in the target language and only has to use a bilingual dictionary to look up translations for the words (s)he needs to know.

But what if even simple constellations show a huge variety in forms? In languages like German, English or French, there are only three parameters to take into account when formulating a possessive: person, number, and gender, the patterns are not too difficult to learn and there are few exceptions. For Zulu, a fourth parameter replaces the 3 choices of gender with up to 16 choices of noun class, some of which are illustrated in Table 1. Language learners hence struggle for quite a while to learn full paradigms for each of the day-to-day constellations. In addition, the above mentioned morpho-phonetic phenomena need to be memorized and kept track of as well. One of the constellations that is very challenging to absorb is the possessive construction, as knowledge of the following facts is required: (i) class of the possessor; (ii) class of the possession; (iii) the concord that is to be used; (iv) generic morpho-phonological rules.

To give an idea of the complexity of possessive constructions, in some Zulu learner guides, the description of the possessive construction is spread over an average of four pages which “can be very user-unfriendly or time consuming for him/her to find the right information by having to read entire sections of grammatical descriptions.” (Prinsloo and Bosch 2012:300).

## 4 Current State of Implementation

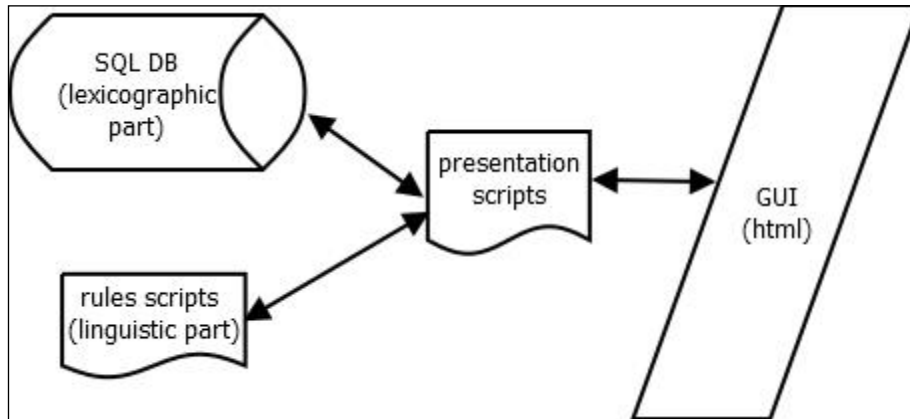
Our system architecture comprises three parts: (1) an sql-database containing the dictionary (lexicographic content part); (2) morpho-phonological rules (linguistic part) and, lastly, (3) programming scripts that build the Graphical User Interface (GUI) and build the web page, i.e. the interface between database and user (the presentation part), cf. figure 1.

We opted for using the XAMPP<sup>3</sup>-environment for developing the sql-database, because it assists when designing and filling databases with data and also simplifies the creation of websites. *php* is a dynamic scripting language specifically designed to develop websites and is part of the XAMPP environment. It offers phpMyAdmin, an sql-interface, i.e. sql-commands can be executed on a respective database with *php* and allows the results to be written into variables that can then be displayed on a website. After the development has been finalized, the php scripts and the database can easily be ported onto an apache webserver. Our test interface can be found on [www.uni-hildesheim.de/iwist-cl/projects/eZulu/login.php](http://www.uni-hildesheim.de/iwist-cl/projects/eZulu/login.php), access is currently restricted to known users<sup>4</sup>.

---

3 <http://www.apachefriends.org/de/xampp.html>

4 Please contact G. Faaß if you should like to try out the test interface



**Figure 1: System Architecture.**

For a start, our database is kept simple and currently contains representative data of all noun classes necessary to perform the given task. We are still working on the design of the website, as it should fulfil the users' needs in terms of the relevant issues of the function theory described above. Currently, we focus on an accurate implementation of the rules and a user-friendly way to explain their application to those users who request such an explanation by ticking the checkbox.

Though the website design is preliminary, we show screenshots in Figures 2, 3 and 4 in order to demonstrate our implementation principle: keep it simple and intuitive. In the header, the user is asked what kind of information (s)he requires over and above the translation which is done in all cases. Checkbox 1 on the left asks whether the user would like to have a pronominal version of the translation (i.e. the possessor noun being replaced by the pronoun belonging to the same noun class) – in addition to the full translation as Zulu uses different pronouns dependent on the noun class they belong to, hence the pronouns “it” or “they” has several translation equivalents, dependent of the noun class of the noun that it stands for. If we ask the user what possessor the pronoun stands for, it will be possible to identify the noun class of this noun so that the tool can choose the right pronoun.



**Figure 2: Start page: user requires translation of “food of person” plus rules plus pronominal translation of the possessor.**

Ticking checkbox 2 on the right hand side will lead to an additional explanation how the resulting Zulu translation was formed concerning the elements and the rules applied. These additional explanations are currently still kept brief, but we plan to link these to other pages where the applied rules will be described in more detail and with additional examples.

Figure 3 demonstrates the result of the query shown in Figure 2: A dynamic webpage which is divided into 4 blocks is visibly distinguishable: The 1<sup>st</sup> block, i.e. the header consists of static information on the page as such (“Zulu e-Dict test version”), while the 2<sup>nd</sup> block – which appears in all cases of use – repeats the English expression entered (“food of person”<sup>5</sup>). It then shows the word-by-word translation of possession and possessor (*ukudla* and *umuntu*), and also the possessive concord that is to be used (*kwa*). Since the checkbox requiring the pronominal form was ticked, it appears as well (*ukudla kwakhe*).

Ticking the checkbox “Do you require an explanation of the rules applied” results in the popping up of a 3<sup>rd</sup> block appears (identified by a darker blue background). Here, the English expression is repeated, however now, possession and possessor are clearly categorized as such. This 3<sup>rd</sup> block then also informs the user that the elements of the English construction “food of person” are *ukudla kwa*, and *umuntu* in Zulu. In addition to the translation of each of the elements, the rules forming the correct translation equivalent, *ukudla komuntu*, are briefly explained. A second paragraph describing the rules of forming the pronominal form is added, because the respective checkbox was ticked.

**Zulu e-Dict test version**  
Translation of English possessives into Zulu

food of person is translated as *ukudla komuntu*      pronominal form: *ukudla kwakhe*

	possession	possessive concord	possessor
English	food		person
Zulu	ukudla (class 15)	kwa	umuntu (class 01)

rules applied:  
possession is of class 11 or 15 and possessor is not of classes 01a/02a  
rule: *vowel coalescence* (.a + u.. becomes .o..)  
intermediate result: ukudla kwomuntu  
rule 2: *semi vowel elision* (.w+vowel becomes .vowel)

rule for forming the pronominal version:  
(*kwa* merges with *khe* (possessive pronoun of class 01));

Next query? Go to [query](#) or [logout](#).

© 2013 2014 University of Hildesheim (UHI) and University of South Africa (UNISA)

**Figure 3: Webpage resulting from query shown in Figure 2.**

We make use of the checkboxes mainly to show only the relevant data on the resulting page, i.e. the required units of information as they were defined by the user (information on demand, cf. Bothma and Tarp 2012:90). This is in line with Tarp’s (2009:26) claim that “The study of access routes, i.e. the rapid and easy access to the relevant needs-adapted data, is of utmost importance to lexicography.” Figure 4 demonstrates what happens if both checkboxes remained unticked.

5 We also consider offering the shorter form, i.e. “person’s food” with the final design of the website.

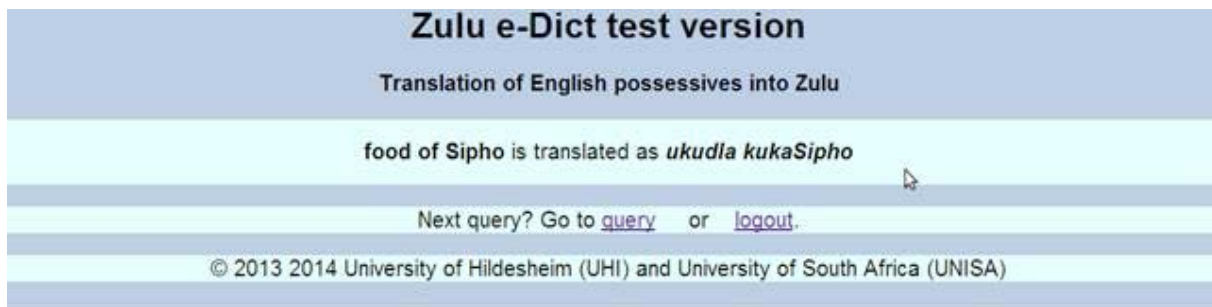


Figure 4: Webpage resulting from the query: “food of Siphoh”, no checkboxes ticked.

In this version of the website, no linking information to study guide pages/chapters is given, however as it is planned to become part of a free online course (cf. University of South Africa, 2010), this will be added then. As mentioned earlier on, the lexicon presently only contains a limited number of nouns (approximately 500 entries) that may be entered. However, we are currently working towards a second version of the database that will contain the vocabulary described in the study material used in the online course. The third version is then planned to become part of an e-dictionary of the African languages currently developed in the framework of the SeLA<sup>6</sup> project which will translate single words and other morpho-syntactic constructions in both directions

## 5 Challenges and Plans for further Implementation Steps

What we still consider a problem is the case where a possessive concord ends in *-a* while the possessor begins with *-o*. In this case – restricted to class 2a noun possessors – there are two ways of forming the Zulu possessive, as demonstrated by Figure 5: The first rule, when applied, merges the concord with the possessor and thereby deletes the vowel (“vowel elision”) leading to *yodohotela* while the second, alternative rule adds *-w-* (“semi-vowel insertion”) resulting in the constellation *yawodohotela*. As both constellations occur frequently in a Zulu corpus<sup>7</sup>, we cannot decide prescriptively which of the constellations should be used. This is a problem in the light of the productive function that this e-dictionary should fulfil: the user might be confused not knowing which of the two constellations is “better”. Currently, we need to offer both alternatives, without distinctive features, as more corpus based research will be necessary to determine whether there are contextual factors that give preference to one of the constellations above the other.

An open challenge is the didactic element that is foreseen: we would like give users the opportunity to test their knowledge of Zulu, i.e. we will allow them to enter a translation equivalent that they

6 Scientific e-Lexicography for Africa, cf. Heid (2012) and <http://www.uni-hildesheim.de/iwist-cl/projects/sela/>

7 We are very thankful to Prof. DJ Prinsloo for giving us access to the Zulu corpus at the University of Pretoria, cf. Prinsloo and de Schryver (2005).

deem correct in one of the next versions of the dictionary. The system is then to compare the user’s input with the correct output and, in a first stage, it will inform the user whether his/her input was correct or not. However, we will ask each user whether (s)he will give his/her consent to store the given data. If this consent is given, the input plus the attempted and the correct output will both be stored in a log file. We hope that these data will help to investigate typical learners’ mistakes and incorrect generalizations. Depending on the output, we hope that in future the system can be enabled to dynamically offer appropriate feedback. These feedback messages, like the log file, are not planned to be personalized, but will only depend on the given inputs.



**Figure 5: Webpage resulting from a query with a noun of class 2a as possessor (usually beginning with o–) and a noun of a class where the possessive concord ends in –a as possession.**

## 6 Conclusion and Future Work

We believe that this prototype e-dictionary with integrated morpho-syntactic information will go a long way to address user needs in specific user situations, in particular with respect to complicated issues such as the translation of possessives from English to Zulu. A next stage of development will allow for a bi-directional translation, where Zulu possessives given as input will be translated into English. This method will serve as a starting point for adding further constructions, while the vocabulary contained will be extended continuously. A parallel activity will be a thorough evaluation of the system; so far, the performed random sample tests have not discovered any errors.

At a later point, we also foresee an expansion of the prototype to closely related Bantu languages such as Xhosa, Swati, Southern Ndebele and Zimbabwe Ndebele. The prototype will also be testing for usability and user preferences, and therefore a test access has been made available for interested users.

## 7 References

- Bothma, T. and Tarp, S. (2012). Lexicography and the Relevance Criterion. *Lexikos* 22, pp. 86-108.
- Gouws, R. (2007). A Transtextual Approach to Lexicographic Functions. *Lexikos* 17, pp. 77-87.
- Heid, U. (2012). SeLA – a new project on electronic lexicography. *Lexicographica* 28(1), pp. 437-440.
- Hurskainen, A. (2009). Intelligent computer-assisted language learning: Implementation to Swahili. Technical Reports in Language Technology Report No 3, University of Helsinki, Finland. <http://www.njas.helsinki.fi/salama>. [09/04/2014]
- Katushemererwe, F. and Nerbonne, J. (2013). Computer-Assisted Language Learning (CALL) in support of (re-) learning native languages: The case of Runyakitara. *Computer-Assisted Language Learning*, 1-18 DOI: 10.1080/09588221.2013.792842 [09/04/2014]
- Poulos, G. and Msimang, C.T. (1998). *A linguistic analysis of Zulu*. Pretoria: Via Afrika.
- Prinsloo, D.J. and Bosch, S. (2012). Kinship terminology in English-Zulu/Northern Sotho dictionaries - a challenge for the Bantu lexicographer. In: *Proceedings of the 15th EURALEX International Congress*. 7-11 August (2012), Oslo, pp. 296-303. Oslo: Representralen, UiO. ISBN 978-82-303-2095-2
- Prinsloo, D.J. and de Schryver, G-M. (2005). Managing eleven parallel corpora and the extraction of data in all official South African languages. In W. Daelemans, T. du Plessis, C. Snyman & L. Teck (eds.) *Multilingualism and Electronic Language Management*. Proceedings of the 4th International MIDP Colloquium, 22-23 September 2003, Bloemfontein, South Africa (Studies in Language Policy in South Africa 4), pp. 100–122. Pretoria: Van Schaik Publishers. Available: <http://tshwanedje.com/publications/11Par-Cor.pdf> [28/10/2013]
- Prinsloo, D.J., Heid, U., Bothma, T., and Faaß, G. (2012). Devices for Information Presentation in Electronic Dictionaries. *Lexikos* 22 pp. 290-320
- Radtke, J. and Heid, U. (2012). Word formation in electronic language resources: state of the art analysis and requirements for the future. In: *Proceedings of the 15th EURALEX International Congress*. 7-11 August 2012, Oslo, pp. 794-802. Oslo: Representralen, UiO. ISBN 978-82-303-2095-2
- Tarp, S. 2009. Beyond Lexicography: New Visions and Challenges in the Information Age. In Bergenholtz, H., Nielsen, S. & Tarp, S. (eds), *Lexicography at a Crossroads Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Linguistic Insights 90, pp.17-32. ISBN 978-3-03911-799-4 br
- University of South Africa Free Online Course. 2010. Available: [http://www.unisa.ac.za/free\\_online\\_course/](http://www.unisa.ac.za/free_online_course/) [26/08/2013].





# Zur (Vor-)Geschichte der saamischen Lexikografie: ein lateinisch-saamisches Wörterverzeichnis aus dem 17. Jahrhundert

Eino Koponen  
Institut für die Landessprachen Finnlands  
eino.koponen@kotus.fi

## Zusammenfassung

Die Saamen (früher Lappen) sind ein nordisches Minoritätsvolk, das auf dem Gebiet von vier Staaten (Norwegen, Schweden, Finnland, Russland) lebt. Die Wurzeln der auf Saamisch gedruckten Literatur reichen ins 17. Jahrhundert zurück, und aus demselben Jahrhundert datiert der erste Versuch eines saamischen Lexikons. Dabei handelt es sich um ein von dem schwedischen Pastor Z. Plantinus verfasstes handschriftliches Wörterverzeichnis von ca. 850 lateinischen Wörtern mit saamischen Äquivalenten, das 1888 in Stockholm wiederentdeckt und später publiziert wurde. Das uns überlieferte Dokument scheint eine korrumpierte Abschrift des Originals zu sein, deren zahlreiche Fehler das Erkennen der saamischen (und bisweilen auch lateinischen) Wörter stellenweise (fast) unmöglich machen. Hier wird das Wörterverzeichnis einer genaueren Analyse unterzogen. Die Arbeit wurde im Rahmen eines größeren Forschungsprojekts, das eine etymologische Datenbank der saamischen Sprachen und der verwandten uralischen Sprachen darstellt, am Institut für die Landessprachen Finnlands durchgeführt. Alle saamischen Wörter des Verzeichnisses sind jetzt in die Datenbank eingespeist. Auch die auf den ersten Blick obskuren Wörter werden dadurch der künftigen Forschung zugänglicher. Bei der Präsentation werden die Resultate der Analyse anhand der Álgu-Datenbank veranschaulicht.

**Stichworte:** Geschichte der Lexikografie; etymologische Datenbank; Saamisch; Latein

## 1 Einleitung

### 1.1 Die saamische Literatur im 17. Jahrhundert

Die ersten auf Saamisch gedruckten Bücher, ein ABC-Buch und ein Gesangbüchlein, erschienen 1619 in Schweden. Der Verfasser war Nicolaus Andreae, Pastor der Gemeinde Piteå. Auch wenn man alle Schwierigkeiten in Acht nimmt, die der Bahnbrecher zu beseitigen hatte, zeugen diese Werke nicht über besonders gute Sprachkenntnisse des Verfassers. Noch bescheidener war die Beherrschung des Saamischen bei dem Fortsetzer seiner Arbeit, dem in Umeå ansässigen Pastor Olaus Petri Niurenius. Der von ihm 1633 herausgegebene Katechismus ist, was die Sprache anbetrifft, als das erbärmlichste

je gedruckte saamische Werk beurteilt worden (Wiklund 1922: 22). Im Gegensatz zu den oben genannten, deren Sprache als entstelltes Süd- oder Umesaamisch mit finnischer Mischung charakterisiert werden kann, repräsentiert das nächste Druckwerk, das ABC-Buch eines unbekanntes Autors (1638), einen archaischen lulesaamischen Dialekt (Bergsland 1982).

Der Höhepunkt der saamischen Literatur im 17. Jahrhundert, wenigstens was den Umfang betrifft, wurde 1648 erreicht, als das 950-seitige Manuale Lapponicum von Johannes Tornaëus (Pfarrer von Nieder-Torneå) herauskam. Tornaëus hatte zur Aufgabe eine Schriftsprache zu erschaffen, die allen Ansprüchen der Saamen im schwedischen Reich genügen würde. Aus dem Blickwinkel der modernen Sprachwissenschaft kann das Resultat unterschiedlich beurteilt werden (s. zuletzt Koponen 2010 mit weiterer Literatur). Auf jeden Fall befriedigte es nicht alle Zeitgenossen, sodass Olaus Stephani Graan, Pastor in Lycksele, das Manuale (allerdings in engerem Umfang) 1669 erneut in seine eigene umesaamische Muttersprache übersetzte. Da die nächsten auf Saamisch gedruckten Bücher schon dem folgenden Jahrhundert angehören, kann dieses Werk neben einem von demselben Autor 1668 herausgegebenen Katechismus als der Endpunkt der saamischen Literatur im 17. Jahrhundert betrachtet werden. (Qvigstad & Wiklund 1899: 22-23.) Sprachlich werden die Werke von Graan dem südlichen Teil der saamischen Bevölkerung (d. h. den Süd- und Umesaamen) gefallen haben, für die anderen (d. h. für die Pite-, Lule- und Tornesaamen; von den östlichen Kemi- und Inarisaamen ganz zu schweigen) waren sie sicher mehr oder weniger unverständlich.

Wenigstens was die Lautbezeichnung anbetrifft, bedeuteten die Arbeiten von Graan einen Schritt rückwärts. Obgleich die Orthografie von Tornaëus unvollständig und inkonsequent ist, sind die Wörter bei ihm meistens auch ohne Kontext (oder Übersetzung) zu erkennen. Bei Graan (und seinen Vorgängern) hingegen ist das Lesen der Wörter nur aufgrund des Schriftbilds öfters unmöglich. Besonders markant kommt das zum Vorschein bei den Affrikaten [ts] und [tš], die bei Tornaëus (meist) durch *z* bezeichnet werden, während die übrigen Druckwerke des 17. Jahrhunderts für sie (oft) *ti*, *hi*, *gi* (oder aber auch nur *t*, *h*, *g*) benutzen. Da dieselben Buchstaben (vor allem wenn kein *i* folgt aber nicht nur) für normale Klusile stehen können, ist die Lautung der Wörter nur aufgrund der Bedeutung zu erschließen. Neben den Affrikaten sei bemerkt, dass der Sibilant [š] bei Tornaëus (meist) *sch*, bei den übrigen Autoren des 17. Jhs (meist) *hi* geschrieben wird. (Qvigstad 1899: 13; Qvigstad 1947: 19; Bergsland 1982; Sköld 1986)

## 1.2 Die saamischen Wörterbücher des 18. Jahrhunderts

Das erste saamische Wörterbuch erschien 1738. Der Autor war der als Schullehrer (später als Pfarrer) in Lycksele ansässige Petrus Fjellström (geb. 1697 im schwedischen Lappland), der in demselben Jahr auch die erste saamische Grammatik herausgab. Das Wörterbuch umfasst 190 Seiten und ca. 7.500 schwedische Stichwörter mit saamischen Übersetzungen. Da einerseits ein schwedisches Stichwort mehrere saamische Wörter als Übersetzungen haben, andererseits aber ein und dasselbe saamische Wort als Übersetzung für mehrere schwedische Wörter auftreten kann, beläuft sich die Anzahl der

saamischen Lexeme in diesem Wörterbuch auf 6.000 bis 9.000. Mit diesen und anderen Arbeiten von demselben Autor und seinen Zeitgenossen (s. genauer Qvigstad & Wiklund 1899: 26-) wurde in der Mitte des Jahrhunderts die erste schwedischsaamische normierte Schriftsprache entwickelt, die sich im *Lexicon Lapponicum* (LL 1780) von Ericus Lindahl und Johannes Öhrling manifestierte. Dieses Großwerk besteht aus einem 584-seitigen saamisch-lateinisch-schwedischen Teil mit reichlichen Beispielen aus der saamischen Phraseologie und aus einem 130-seitigen schwedisch-saamischen Index, der als ein selbständiges kleineres Wörterbuch betrachtet werden kann (Larsson 1997: 107).

Während die Schriftsprache sich in Schweden auf der Basis der südlichen Dialekte entwickelte, basierte die Schriftsprache in Norwegen auf einem nördlichen (dem schwedischen Tornesaamisch nahestehenden) finnmarksaamischen Dialekt. Obgleich das Saamische auf der norwegischen Seite erst seit dem 18. Jahrhundert als Schriftsprache benutzt wurde, wurde der erste Teil eines norwegischsaamischen *Lexicon Lapponicum bipartitum* von Knud Leem schon 1768 herausgegeben. Dieses Werk, das mehr mit dem Wortreichtum als mit systematischer Darstellung imponiert, umfasst über 1600 Seiten und gibt die Bedeutungen der saamischen Stichwörter auf Dänisch und Lateinisch an. Der zweite Teil mit dänischen Stichwörtern und mit einem lateinischen Index erschien erst 1781, sieben Jahre nach Leems Tod, und wurde von seinem Kollegen Gerhard Sandberg herausgegeben. (Larsson 1997: 106-107; zu den alten und neueren saamischen Wörterbüchern s. auch Magga 2012.)

### **1.3 Die lateinisch-schwedischen gedruckten Wörterverzeichnisse des 17. Jahrhunderts**

Da die Bildungssprache am Anfang der Neuzeit auch im schwedischen Reich Lateinisch war, wurden nach dem mitteleuropäischen (vor allem deutschen) Vorbild Wörterverzeichnisse als Hilfsmittel für das Studium des Lateinischen herausgegeben. Die bekanntesten sind das 4-sprachige (lateinisch-schwedisch-deutsch-finnische) *Lexicon Latino-Scondicum* von Ericus Schroderus (1637) und das „*Variarum Rerum Vocabula Latina*“, das im 16. und 17. Jahrhundert in mehreren Auflagen (zuerst Lateinisch-Schwedisch, später unter Hinzufügung des Finnischen) herausgegeben wurde. Für unser Thema ist nicht ohne Interesse, dass man ein handschriftliches lateinisch-finnisches Wörterverzeichnis aus dem Jahre 1669 kennt, das eine Abschrift des finnischen Materials aus der letzten Auflage des letztgenannten Werkes (VR 1668) darstellt. Es handelt sich um eine Arbeit, die der deutsche Arzt und Gelehrte Martinus Fogel auf die Aufforderung des Großfürsten von Toscana Cosimo III., ihm Materialien über die finnische Sprache zu verschaffen, durchführen ließ. (Hierzu genauer Stipa 1990: 78-80.)

## 2 Die etymologische Datenbank Álgú

Am Institut für die Landessprachen Finnlands wurde im Laufe der letzten 12 Jahre eine etymologische Datenbank der saamischen Sprachen aufgebaut, die vor allem der Erforschung des Ursprungs und der Geschichte des saamischen Wortschatzes dienen soll. Die Datenbank namens Álgú (nordsaamisch: ‚Beginn, Herkunft‘) hat neben einer finnischen und nordsaamischen Benutzeroberfläche auch eine auf Deutsch und Englisch. Die Datenbank steht im Internet allen Interessierten frei zur Verfügung. Alle saamischen Wörter aus dem hier zu besprechenden Wörterverzeichnis sind jetzt in Álgú eingespeist und dort (unter der Sprache Schwedischlappisch) leicht aufzufinden. Aus der Datenbank geht hervor, in welcher Relation (Äquivalenz, Ableitung usw.) jedes Wort zu dem uns aus anderen Quellen bekannten saamischen Wortmaterial steht. Auch die auf den ersten Blick obskuren Wörter des Verzeichnisses werden dadurch einleuchtend und der künftigen Forschung zugänglicher, vorausgesetzt natürlich, dass ihre Rätsel richtig gelöst sind. Bei Wörtern, die vorläufig nicht oder nur mit Vorbehalt identifiziert worden sind, erfährt der Benutzer auch diese Information. (Zum Aufbau der Datenbank s. genauer Aapala et al. 2010.)

## 3 Das Wörterverzeichnis von Zacharias Plantinus

### 3.1 Die Entdeckung, Entstehung und Herausgebung des Verzeichnisses

Das Wörterverzeichnis, das hier einer genaueren Analyse unterzogen wird, wurde 1888 im Nachlass des schwedischen Schriftstellers und Sprachforschers Georg Stiernhielm (1598–1672) in der Königlichen Bibliothek zu Stockholm wiederentdeckt. Die Entdeckung wurde von dem finnischen Finnougristen E. N. Setälä gemacht, der das Verzeichnis etwas später auch herausgab (Setälä 1890). Der Verfasser des Wörterverzeichnisses ist Zacharias Olai Plantinus, geboren in den zwanziger Jahren des 17. Jahrhunderts in Umeå als Sohn des oben erwähnten Olaus Petri Niurenius (1580–1645). Über den späteren Lebenslauf von Z. Plantinus kann erwähnt werden, dass er nach seinem Studium in den fünfziger und sechziger Jahren u. a. als Lektor der griechischen Sprache am Hernösander Gymnasium arbeitete und zweimal auch als Rektor des Gymnasiums tätig war. 1672 wurde er zum Pastor des Kirchspiels Offerdal und Probst über ganz Jämtland ernannt. Plantinus starb 1688. (Setälä 1890: 86–87) Vater und Sohn waren Pastoren in Gemeinden mit (auch) saamischer Bevölkerung und engagierten sich beide in der Erforschung der Saamen und ihrer Sprache. Wie mehrere seiner Kollegen hat Z. Plantinus Johannes Schefferus bei seinem berühmten Werk „Lapponia“ (1673) Hilfe geleistet. Neben einem von Vater Niurenius verfassten und von Sohn Plantinus ergänzten handschriftlichen Lappmarkbericht hat Schefferus das Vorwort eines ungedruckt gebliebenen saamischen Lexikons von Plantinus benutzt (Löw 1956: 14, 409). Auch dieses handgeschriebene Vorwort wurde von Setälä gefunden, allerdings an einer ganz anderen Stelle, sodass es seiner Meinung nach unsicher sei, ob das Vorwort

und das Wörterverzeichnis zusammengehören. Auf jeden Fall kennt man kein anderes lexikalisches Werk von Plantinus, zu dem dieses Vorwort gehören könnte (Setälä 1890: 90).

Was den Zeitpunkt der Entstehung betrifft, stellt Setälä (1890: 86) fest, dass Stiernhielm bereits 1672 starb, sodass das Wörterverzeichnis nicht aus einer späteren Zeit stammen kann. Setälä verbindet die Entstehung des Wörterverzeichnisses einerseits mit der Ernennung von Plantinus zum Pastor in Ofverdal und andererseits mit der Aufforderung des schwedischen Antiquitätsarchivs, Schefferus für sein Werk Nachrichten mitzuteilen, und hält deswegen für am wahrscheinlichsten, dass das Wörterverzeichnis auch nicht früher sondern eben um 1672 geschrieben wurde. G. J. Stipa (1990: 83, 145) hingegen verlegt den Zeitpunkt der Entstehung in das zweite Drittel des Jahrhunderts, d. h. zeitlich wenigstens sechs Jahre früher. Laut Stipa wurde das Wörterverzeichnis von Stiernhielm, Schefferus und späteren Forschern benutzt. Falls Stipa recht hat, muss das Wörterverzeichnis früher in mehreren Exemplaren existiert haben. Dies ist an sich sehr gut möglich, zumal das einzige uns bekannte Exemplar mit seinen zahlreichen Schreibfehlern eine korrumpierte Abschrift des Originals zu sein scheint. Im Gegensatz zu dem eigenhändigen Begleitschreiben ist das an Stiernhielm versandte Exemplar des Wörterverzeichnisses nicht von Plantinus selbst, sondern von unbekannter Hand „ins Reine geschrieben“ worden (Setälä 1890: 86). Löw (1956: 434) zufolge hätte Setälä es „nicht ganz adäquat“ herausgegeben, was bedeuten würde, dass die Schreibfehler wenigstens zum Teil erst dann entstanden wären. Setälä (1890: 90) hingegen versichert, dass er und sein Mitarbeiter, der schwedische Lappologe K. B. Wiklund, keine Mühen beim gewissenhaften Dechiffrieren der Handschrift gescheut hätten, die Handschrift allerdings oft beinahe unlesbar gewesen sei, sodass er nicht zu behaupten wage, dass alle schwierigen Stellen richtig interpretiert worden seien.

### 3.2 Bekannte und rätselhafte Wörter

Das Verzeichnis enthält ca. 850 alphabetisch geordnete lateinische Stichwörter mit saamischen Übersetzungen. Für ca. 50 lateinische Wörter werden zwei (oder drei) saamische Übersetzungen aufgeführt, sodass die Anzahl der saamischen Wörter ca. 900 ist. Etwa 60 saamische Wörter begegnen im Verzeichnis als Übersetzungen für zwei (oder drei) lateinische Wörter. Wenn man ein und dasselbe Wort nur einmal mitrechnet, beläuft sich die Anzahl der saamischen Lexeme im Verzeichnis auf etwa 840.<sup>1</sup> In Wortklassen teilen sie sich folgendermaßen auf: ca. 370 Substantive, ca. 270 Verben, ca. 140 Adjektive, ca. 40 Partikeln, ca. 10 Numeralien und ca. 10 Pronomina.

Etwa 720 der saamischen Wörter sind mit Sicherheit oder mit einigem Vorbehalt mit Lexemen in LL (und im heutigen Schwedischsaamisch) gleichzusetzen oder enthalten wenigstens einen in LL belegten Wortstamm, d. h. stehen in einem (korrelativen) Ableitungsverhältnis zu einem dort vorkommenden Lexem. Etwa 40 Wörter sind mit größerem Vorbehalt mit einem in LL belegten Wort(stamm) zu vergleichen, und etwa 140 Wörter haben keine Vergleichspunkte in LL. Einige obsoletere Wörter sind

1 Laut Setälä (1890: 90) ist die Anzahl der saamischen Wörter 820; er erklärt nicht, wie er zu diesem Ergebnis gekommen ist.

mir bei Tornaus oder Graan aufgefallen.<sup>2</sup> In Fjellströms Wörterbuch habe ich passende Wörter aufgrund ihrer anzunehmenden schwedischen Bedeutung gesucht,<sup>3</sup> und in gleicher Weise wurden von mir alle Wörterbücher der heutigen saamischen Sprachen sowie der lateinische Index in Leems Wörterbuch durchgearbeitet. Die meisten und sichersten Entsprechungen für in LL unbelegte Wörter habe ich im Südsaamischen gefunden.<sup>4</sup> Im heutigen Lulesaamisch (sowie im Nord- und/oder Ostsaamischen, nicht aber im Südsaamischen) begegnende Entsprechungen haben *sábge* ‚radix‘ (*suobde* ‚Wurzel eines gefällten Baumes‘) und *loivijth* ‚frangere‘ (*lajggit* ‚losreißen, abziehen, abreißen‘), und zwei Wörter wären vielleicht mit dem Kildinsaamischen zu vergleichen: *wout* ‚divitiae‘ (*vāptōk* ‚reich; Reichtum‘) und *vetzeth* ‚currere‘ (*viđže* ‚laufen‘); das letztere könnte aber gleich gut mit LL *qwotset* ‚currere‘ identisch sein. Auch wenn die Vergleichen mit den kildinsaamischen Wörtern zutreffend sind, beweisen sie natürlich nicht, dass Plantinus auch ostsaamische Wörter in seine Liste aufgenommen hätte, sondern, dass diese Wörter früher auch im Schwedischsaamischen (Süd- oder Lulesaamischen) existiert haben. Gleichfalls kann es sich bei einigen dunkel gebliebenen um alte Wörter handeln, die später spurlos verlorengegangen sind, bei anderen wiederum um Wörter, deren Erkennung aus verschiedenen Gründen (noch) nicht gelungen ist.

Wie aus meiner Darstellung zu ersehen ist, ist es meistens zwecklos, Plantinus' Wörter in den alphabetisch geordneten saamischen Wörterbüchern dem Schriftbild nach aufzusuchen. Am leichtesten findet man sie aufgrund der Bedeutung. Der lateinische Index in Leem 1781 gibt z. B. für ‚aqua‘ finnmarksaamisch *zhjatze* an, welches im heutigen Nordsaamisch *čáhci* geschrieben wird. Dasselbe Wort begegnet in verschiedenen phonetischen und orthografischen Varianten in allen saamischen Sprachen, im heutigen Südsaamisch *tjaetsie*, Lulesaamisch *tjáhstse* usw.; LL *tjatse* ‚aqua; vatten‘. Bei Plantinus steht für ‚aqua‘ *kiatie*, welches ohne Zweifel dasselbe Wort ist. Mehr Mühe erfordert die Identifizierung des nächsten Wortes *hiergitt* für ‚aquila‘. Leem gibt für dieses Wort *guasskem* an, ein im Saamischen weit verbreitetes Wort für Adler, offensichtlich aber nicht das, was Plantinus gemeint hat. Der schwedische Index in LL hilft nicht weiter, denn dort steht sub *örn* (= ‚aquila‘) *arnes* und *kåskem*. In dem norwegisch-südsaamischen Wörterbuch findet man endlich (s. v. *ørn*) ein passendes Wort *giergehtse* und erfährt dazu, dass *h(i)* hier nicht für eine Affrikate sondern für einen Klusil steht. Es stellt sich heraus, dass dieses Wort auch in LL als *hergits* ‚aquila; örn‘ belegt ist, obgleich im Index darauf nicht verwiesen wird.

2 *håividhith* ‚verberare‘ (Tornaus 1649); *raik* ‚gaudium; laetitia‘, *hagge* ‚ira‘ (Graan 1668).

3 Das Resultat: *ächte* ‚genus‘, *laiketh* ‚deserere‘, *tiackohieth* ‚placare‘, *karrath* ‚ornare‘ (Fjellström 1738: *ächt* ‚slächt‘, *lacketet* ‚öfvergifva‘, *tiotzkotet* ‚stilla wreden‘, *garfwot* ‚bepryda‘, *garfwet* ‚pryda‘).

4 *voszes* ‚animosus‘ (Südsaamisch *vuasehks* ‚dreist, kühn, zudringlich‘), *pullames* ‚bulla‘ (*bollenjes* ‚Wasserblase‘), *strappo* ‚cadaver‘ (*straahpoe* ‚die Leiche eines vom Raubtier getöteten Rentiers‘), *sveibul* ‚flamma‘ (*svääjpele* ‚Flamme‘), *kasnåth* ‚fomes‘ (*gasnege* ‚Zunder‘), *girbe* ‚pudenda‘ (*girpie* ‚Penis eines Hundes‘), *vaknerteth* ‚luctari‘ (*viengerdidh* ‚sich mit etwas sehr anstrengen‘), *miacht* ‚res‘ (*mij-akt* ‚irgendwas‘), *kaijeh* ‚sonus‘ (*gaajege* ‚Widerhall‘), *harithe* ‚celer‘ (*haerrehtje* ‚schleunig, rasch, flink‘; hierzu wohl auch *varitha* ‚celeriter‘), *botnadt* ‚custodi-re‘ (*båtnodh* ‚verbergen, verstecken, verwahren, aufbewahren, sparen‘), *beendeth* ‚tumescere‘ (*bonhtegidh* ‚anschwellen‘), *hokedeth* ‚allicere‘ (*gåvhkohtidh* ‚locken‘), *jaebeteth* ‚aperire‘ (*gaehpididh* ‚öffnen, eine Öffnung, einen Riss machen‘), *veividh* ‚cogere‘ (*meejvedh* ‚(Rentiere) zusammentreiben‘), *galdij* ‚fons‘ (*gaaltije* ‚Quelle‘); alle bis auf das letzte mit enger südsaamischer Verbreitung.

Wörter wie *aqua* und *aquila* sind in dem Sinne unproblematisch, dass ihre Bedeutungsäquivalente im Schwedischen, Norwegischen, Finnischen und Deutschen bei Bedarf anhand der zweisprachigen lateinischen Wörterbücher leicht herauszufinden sind, was die Voraussetzung für die Benutzung der existierenden zweisprachigen saamischen Wörterbücher ist. Die meisten lateinischen Wörterbücher orientieren sich zunächst nach dem klassischen Latein, Plantinus wiederum nach dem des 17. Jahrhunderts. So ist es nicht immer eindeutig, welche Bedeutung eines polysemantischen lateinischen Wortes Plantinus gemeint hat. Das erste Wort des Verzeichnisses ist *abies*, was im Lateinischen die Benennung eines Nadelbaums ist. Laut den lateinisch-deutschen Wörterbüchern bedeute es eine Tanne oder eine Fichte, von welchen nur die letztere in Frage kommen kann, denn die Tanne wächst im schwedischen Lappland nicht. Die Fichte heißt im Saamischen aber südsaamisch *goese*, nordsaamisch *guossa* usw. Dieses Wort begegnet bei Plantinus (*håsze* geschrieben) als Äquivalent für ‚pinus‘, welches laut den lateinisch-deutschen Wörterbüchern ‚Kiefer‘ bedeuten soll. Auch in LL steht für *kuosse* lateinisch ‚pinus‘, obgleich aus der schwedischen Bedeutung („gran“) zu ersehen ist, dass nicht die Kiefer sondern die Fichte gemeint ist. Die Kiefer heißt südsaamisch *bietsie*, nordsaamisch *beahci* usw. In LL ist *petse* ‚abies; tall, furu‘ belegt, woraus eindeutig hervorgeht, dass *abies* hier für ‚Kiefer‘ steht. Der Baumname *lätha* bei Plantinus bezieht sich also zweifelsohne auf die Kiefer, nicht die Fichte. Eine Behauptung, woran man wenigstens auf den ersten Blick viel mehr zweifeln kann, ist, dass es sich bei *lätha* und *petse* um ein und dasselbe Wort handelt. Nicht nur der Umstand, dass für dieses Wort im Saamischen sonst keine Vergleichspunkte zu finden wären, sondern die Beobachtung, dass das Verzeichnis auch andere Wörter mit dem Kopierfehler *l* pro *p* enthält (*langsing* ‚labium‘, *loide* ‚adeps‘; LL *pankse* id., *puoite* ‚pingvitulo, lardum‘), erlauben uns hier von einer ursprünglichen Form *\*pätha* (?\**pätia*, *\*pätza*) auszugehen. Auf die Problematik des Vokalismus lohnt es sich hier nicht näher einzugehen (siehe Fußnote 7), es sei nur kurz darauf verwiesen, dass die Diminutivform des Wortes im heutigen Lulesaamisch *bätsasj* lautet.

Als saamisches Bedeutungsäquivalent für lat. *insidiae* ‚Hinterhalt, Versteck zum Auflauern; Falle, Fallstrick‘ steht bei Plantinus *biutas*. Dieses gehört ohne Bedenken zusammen mit LL *piwto(sma)* ‚tendiculla, laqueus aut alius eius generis, quo captura sit‘, *piwtosmab takket* ‚retia ponere aliasve insidias feris struere‘, nordsaamisch *bivddus* ‚Fanggerät‘ (zu LL *piwtet* ‚capturae avium vel piscium etc. operam navare‘). Ein ähnlich aussehendes Wort steht einige Zeilen weiter unten für lat. *institutum* ‚Einsetzung, Einrichtung, Sitte, Gewohnheit, Unterweisung‘. Hier handelt es sich jedoch wohl um ein ganz anderes Wort, und zwar LL *piejates*, *piejetus* ‚mandatum; befallning, förordnande‘.<sup>5</sup> Das Stammwort von *piejetus* ist *piejet*, das laut LL ‚ponere; lägga, sätta‘ aber auch ‚ordinare; befalla‘ (*naute le Jubmel piejam* ‚ita Deus ordinavit‘) und ‚instituere; instikta‘ (*Lådnesteje le kastateseb sisa piejam* ‚Salvator Sacramentum Baptismi instituit‘) bedeutet. Dieses Verb begegnet bei Plantinus als *paijeth* ‚ponere‘ und *paieth* ‚iubere‘. Eine Ableitung von demselben Verb ist auch Plantinus *paijemat* ‚jussum‘. Das Suffix *-mat* ist für das Saami-

5 Die Verwendung des Wortes wird in LL mit folgenden Beispielen beleuchtet: *kalka målsotet tait piejetusit* ‚mutabit instituta illa‘ (ein Bibelzitat Act 6:14, in der Vulgata allerdings: *mutabit traditiones*) und *Jubmelen piejetusen melte wiesot* ‚ita vivere ut jussit Deus; lefva efter Guds befallning‘.



sche befremdend, sodass es nicht ausgeschlossen ist, dass es sich hierbei um eine Entstellung von *pä-ijetom*, *piejetum* o. ä. handelt. Eine solche Ableitung, die von demselben Typus wie *wärnotom* ‚juratus‘ (worüber unten) ist, begegnet bei Graan (1668: 60): [Weralden pijemusij pargån ja] *piejetumen piira* ‚Om [thet Werdlsliga Regementet och] Instichtelse‘. LL gibt als lateinisches Bedeutungsäquivalent für *piejetus* nur *mandatum* an, aber offensichtlich wären auch *institutum* und *jussum* (sowie die Nebenform *piejetum*) hinzuzufügen. Plantinus hat auch ein drittes zu diesem Begriffskreis gehörendes Wort: *vijeh-tes* (*vijchtes*) ‚lex‘. LL gibt für schwedisch *lag* ‚Gesetz‘ auch *piejetus* an. Es stellt sich die Frage, ob nicht auch *vijeh-tes* auf dieses zurückzuführen wäre.<sup>6</sup>

Wenn auch die meisten lateinischen Wörter in den mir zur Verfügung stehenden Wörterbüchern aufzufinden sind, gibt es bei Platinus einige, die ich in keiner anderen Quelle habe belegen können. Möglicherweise handelt es hierbei um falsch geschriebene (und/oder kopierte) Wörter oder Neologismen mit enger Verbreitung. Zu den erst genannten scheint *cherus* für saamisch *kierk* (vgl. LL *keres*, *kerok* ‚carus, dilectus; kär‘, auch VR und Schroderus *charus* ‚käär‘) zu gehören, zu den letzteren *insipidus* (auch in VR *insipidum* ‚osmakande‘), worüber weiter unten. Bei einigen Fällen hat sich das lateinische Wort durch Abschreibfehler geändert: so steht bei saamisch *vösze* und *sacke* lateinisch *sanus* pro *saccus* (vgl. LL *wuoss* ‚saccus‘ und südsaamisch *siehke* ‚Sack‘), bei saamisch *kijrz* lateinisch *guttus* pro *guttur* (vgl. LL *haras*, *kirs* ‚gula, cartilaginosa pars colli‘), bei saamisch *brades* lateinisch *festivus* pro *festinus* (vgl. LL *brad(es)* ‚praeruptus, acclivis, celer‘), bei saamisch *kiomatz* lateinisch *aridus* (?) pro *armus* (vgl. LL *tjå-motes* ‚quadrupedum armus‘), wahrscheinlich auch bei saamisch *hiäg* lateinisch *angelus* pro *angulus* (vgl. LL *tjåk* ‚angulus‘; *ängel* ‚angelus‘; hier muss also auch das saamische Wort beim Kopieren entstellt worden sein). Wie bei *sanus/saccus* spricht auch bei *angelus/angulus* für die Richtigkeit des letzteren Wortes der Umstand, dass es (im Gegensatz zu dem ersteren) alphabetisch an der richtigen Stelle stehen würde. Es ist erwartungsgemäß, dass die lateinischen Wörter, auch wenn sie falsch gelesen und kopiert worden sind, immerhin lateinische Wörter geblieben sind, während aus den – dem Kopierer wohl unbekannt – saamischen Wörtern Kauderwelsch geworden ist. Es ist möglich, dass es unter den bisweilen dunkel gebliebenen Wörtern noch mehrere Fälle gibt, in denen vor allem das geänderte lateinische Bedeutungsäquivalent das Problem ist. Ohne Kontext und ohne (richtige) Information über die Bedeutung sind die saamischen Wörter wegen der ungenauen und launischen Schreibweise meist unerkennbar, auch wenn sie „richtig“ geschrieben (und kopiert) sind.

Beim Vergleich von Plantinus’ Wörterverzeichnis mit den zeitgenössischen lateinisch-schwedischen Vokabularien ist die relativ große Anzahl der Verben (ca. 190 von Plantinus’ 290 lateinischen Verben begegnen weder in VR noch bei Schroderus) auffällig. Sowohl die lateinischen als auch (mit wenigen Ausnahmen) die saamischen Verben stehen in der Infinitivform, deren Endung bei den meisten saamischen Wörtern *-th* geschrieben wird (bei 8 Verben jedoch *-dh*, bei 5 *-dt*, bei 4 *-d* und zweimal *t*). Auch

6 Bei Fjellström steht für schwedisch *lag* saamisch *lag*, *räcktie*, *biäjetz* und weiter *lagenbiäjertia* ‚lagstiftare‘ (= Gesetzgeber) sowie *biäjjet* ‚stiffta‘. Es ist auch möglich (und vielleicht sogar wahrscheinlicher), dass *vijchtes* eine Entstellung von *räcktie*, LL *rektas* ‚ius, forum; rätt, domstol‘, *laga ja rektas* ‚lag och rätt‘ ist (v pro r auch in *vöxe* ‚femur‘, LL *ruoksje* id.).



alle oben erwähnten in Schweden gedruckten Wörterbücher (VR, Schroderus, Fjellström, LL) führen die Verben in der Infinitivform an, während sie bei Leem in der Form der 1. Person Präsens stehen. Bei 30 saamischen Verben scheint die Form bei Plantinus jedoch nicht der Infinitiv zu sein. Die Formen *matta* ‚intelligere‘, *kauna* ‚reperire‘, *sava* ‚optari [!]‘, *sijta* ‚velle‘, *ähta* ‚amare‘ *bijla* ‚timere‘, *billa* ‚metuere‘, *gåive* ‚haurire‘ sind eindeutig 3. Person Singular (die ersten zwei begegnen im Verzeichnis dazu im Infinitiv: *mattith* ‚posse‘, *kaunath* ‚invenire‘), mit Vorbehalt auch *jore* ‚rotari‘ (= *joreth* ‚volvere‘) und *bårre* ‚comedere‘.<sup>7</sup> Derselben Kategorie dürften weiter die Verbformen *kiotha* ‚concitare‘, *nålla* ‚abigere‘, *laggå* ‚irasci‘ und *naldne* ‚titubare‘ angehören, aber bei ihnen ist nicht klar, um welches Wort es sich überhaupt handelt.

Im Verb *tabran* ‚hærere‘ ist das auslautende *n* ein Ableitungssuffix. Es handelt sich um den dreisilbigen Stamm LL *tabranet* ‚adhærere‘; die Form dürfte 3. Sg. sein. Ein ähnlicher Fall ist möglicherweise *hiekin* ‚fremere‘, wobei das ganze Wort dunkel ist. Das Verb *kanest* ‚oscitare‘ ist ohne Bedenken eine (3. P. Sg.) Form des Verbs LL *kawestet* id., das Verb *mijhes* ‚singultire‘ wohl eine des Verbs LL *niakkestet* id. Bei *kattest* ‚pendere‘ handelt es sich entweder um LL *katsostet* ‚suspendere‘ oder LL *katsahet* ‚pendere‘ (die Bedeutung deutet auf ersteres, das Schriftbild auf letzteres, das im Verzeichnis auch als *kittesteth* ‚suspendere‘ begegnet). Das Wort *sinerves* ‚ringi‘ gehört mit den gleichbedeutenden Verben südsaamisch *snorvet*, lulesaamisch (*s*)*nirvvit* (vgl. auch LL *snerdset* id.) zusammen. Als Verb würde ihm südsaamisch *snjerviestidh* am nächsten stehen, das jedoch semantisch (‚brünstig sein‘) weiter entfernt ist. Es wäre vielleicht nicht ausgeschlossen, dass es sich bei *sinerves* nicht um ein Verb sondern um ein Adjektiv handelte, vgl. nordsaamisch *snirvvas* ‚grinsend‘. Bei *sadgus* ‚mussitare‘ und *sumdkus* ‚fremere‘ handelt es möglicherweise um ein und dasselbe Wort, das eventuell mit *sabkasath* ‚sibilare‘ identisch ist (vergleiche jedoch auch südsaamisch *sjodkesjidh* ‚pfeifen (Wind)‘). Der Stamm von *sabkasath* ist mit dem Stamm der im LL (und anderswo) belegten Ableitung *samkelet* ‚insusurrare, in aures dicere‘ identisch (-*bk*- ~ -*mk*- ist ein interdialektaler Wechsel), die ein anderes Suffix enthält. Ein suffixales -*l*- steckt auch in *kalmell* ‚frigere‘ (vgl. LL *kalmet* ‚gelari‘, *kalmes* ‚frigidus‘), sowie im ersten Teil der Konnexion *gioggel pijas* ‚surge‘. Hier handelt es sich also sowohl auf Lateinisch als auch auf Saamisch um eine Imperativform (*pijas* ‚auf‘). Daneben enthält das Wörterverzeichnis auch die Infinitivform *gioggeleth* ‚surgere‘ (sowie *kioggeleth* ‚erigere‘ pro ‚erigi‘; vgl. LL *tjuodtjelet* ‚surgere‘, *tjuodtjel paijas* ‚erige te‘, *tjuodtjaldattet* ‚erigere, exsurgere facere‘). – Aus zwei Wörtern bestehende Konnexionen sind weiter

7 Der Grund für den Vorbehalt ist der Auslautvokal *e* statt des zu erwartenden *a*, vgl. modernes Südsaamisch *jårra* (3. P. Sg. aus *jarredh* ‚rollen, sich drehen‘), *bårra* (3. P. Sg. aus *bårredh* ‚fressen‘). Einem einzelnen Buchstaben, insbesondere einem Vokal kann jedoch keine große Bedeutung beigemessen werden. Erstens sind die Vokale im Südsaamischen paradigmatischen Wechseln unterworfen (*billedh* – *bælla* – *billieh*; *gåajvodh* – *gååjve* – *gåajvoeh*; *iehtsedh* – *eahtsa* – *iehtsieh*). Zweitens weisen sie schon innerhalb des Südsaamischen interdialektale Variation auf (z. B. *jarredh*, *barredh* neben *jarredh*, *bårredh*), und das Bild wird noch bunter, wenn man auch andere saamische Sprachvarietäten (Ume-, Pite- und Lulesaamisch) berücksichtigt. Drittens begegnen im Saamischen Vokale (vor allem Diphtonge), bei denen es gar nicht evident ist, mit welchen lateinischen (oder schwedischen) Buchstaben sie zu bezeichnen wären. Und viertens muss man immer auch damit rechnen, dass das Schriftbild beim Kopieren entstellt worden ist (so erklären sich wohl die nicht seltenen Fälle, wo anstelle eines zu erwartenden *ä* ein *ä* oder *a*, anstelle eines *å* ein *å* oder *a* und anstelle eines *a* ein *ä* oder *å* steht).

*raszataieth* ‚rumpere‘ (LL *rasta tåjet* ‚confringere‘ s. v. *rasta* ‚trans‘), *palaidienames* (LL *ädnamī palet* s. v. *be-grafva*), *sorritrånchelet* ‚deprecari‘ (vgl. LL *särrit* = *erit* ‚in alium locum [= weg]‘, Leem *erit rokkadalam* ‚deprecor‘) und *postäithpotheth* ‚redire‘ (vgl. südsaamisch *bååstede* ‚zurück‘).

Bei einigen Wörtern stellt sich die Frage, ob es sich bei dem lateinischen und dem saamischen Wort um ein und dieselbe Wortklasse handelt. Das Wort *svolateth* steht im Verzeichnis als Äquivalent für lat. *furax*, obgleich es sich dabei wohl nur um das Verb LL *suoladet* ‚furari‘ handeln kann. Das ursprüngliche lateinische Wort wird also *furari* gewesen und *furax* ein Fehler beim Abschreiben sein. Ein analoger Fehler scheint bei *holgahieth* ‚concitatus‘ vorzuliegen: das saamische Wort ist offenbar mit dem südsaamischen Verb *holkesidh* ‚nachfolgen‘ identisch. Bei genauerer Betrachtung stellt sich weiter heraus, dass das ursprüngliche lateinische Wort wohl nicht *concitare* ‚(an)treiben, jagen, aufhetzen‘ sondern *comitari* (od. *comitare*) ‚begleiten‘ gewesen ist, nicht nur weil dieses semantisch besser passt, sondern auch, weil es im Verzeichnis zwischen *comedere* und *comperdere* steht.<sup>8</sup> Das Verb *concitare* begegnet außerdem weiter unten (an der richtigen Stelle), wo ihm, wie oben festgestellt, ein dunkles *kiiotha* als saamisches Äquivalent zugeordnet wird. Dem saamischen Wort *plaidgaseth* sollte dem Verzeichnis zufolge lateinisch *fulgus* entsprechen. Ein solches Wort ist in Wörterbüchern unbelegt; die vom Schriftbild her am nächsten gelegenen Wörter, die ich finden kann, wären *fulgor* oder *fulgur* ‚Glanz, Blitz‘ und *fulgere* ‚glänzen, blitzen, funkeln‘. Das saamische Wort ist anscheinend ein Verb und zunächst mit südsaamisch *pleajhkasjidh* ‚glänzen, funkeln, blinken‘ zu vergleichen.

Während die oben aufgezählten saamischen Wörter auf *-th* trotz ihrer angeblichen nominalen Bedeutung höchst wahrscheinlich Verben sind, gibt es andere, die mehr (*tiabuth* ‚collum‘, LL *tjåpot* id.; *eckith* ‚vespera‘, LL *ekked* ‚vesper‘) oder weniger eindeutige Substantive sind. Zu den letzteren gehören *måneth* ‚exitus‘, *kolath* ‚examen‘, *jelith* ‚sumptus‘ und *jelijth* ‚expensa‘, deren Stämme zwar jeweils mit einem verbalen Stamm gleichzusetzen sind (vgl. LL *mannat* ‚proficisci, iter facere‘, *kullet* ‚audire‘, *jelet* ‚vivere‘), deren *-th* aber ein deverbales nominales Ableitungssuffix sein wird. Die zwei letzten sind vielleicht mit der nordsaamischen Ableitung *ealádat* ‚Nahrung‘ (oder *ealáhat* ‚Lebensunterhalt‘) identisch. Nicht ausgeschlossen ist, dass es sich in dem einen oder anderen Wort um dasselbe Suffix handelt wie bei *iamate* ‚mors‘ (vgl. südsaamisch *jaemedē* ‚Tod‘) und *jelidh* ‚vita‘ (vgl. südsaamisch *jielede* ‚Leben‘).<sup>9</sup> Zu den Verbalsubstantiven gehört anscheinend auch *måreth* ‚cura‘, vgl. LL *mårraha*, *mårrek* id., *mårrahet* ‚curare, curam gerere‘, *mårretet* id. Zu dem letztgenannten Verb würde dem Schriftbild nach ideal *morreteth* passen, das im Verzeichnis jedoch nicht für ‚curare‘ sondern ‚conari‘ steht. Um einen Fehler beim Kopieren scheint es sich nicht zu handeln, denn *conari* steht hier alphabetisch an der

8 Lat. *comperdere* ist in den mir zur Verfügung stehenden Wörterbüchern unbelegt und dunkel ist auch das entsprechende saamische Wort *kieketeth*. Dem Schriftbild nach würde es ideal zu LL *tjåketet*, frequ. zu *tjåket* ‚abscondere, occultare‘ (lulesaam. *tjiekhat* ‚verstecken, verbergen; geheimhalten‘, skoltsaam. *čiekhâd* ‚verbergen; vergraben‘) passen. Möglicherweise ist das ursprüngliche lateinische Wort *compercere* ‚ersparen, zusammensparen‘ gewesen. Laut Lagercrantz (1939: 86) kann nordsaam. *čiekhat* auch in der Bedeutung ‚sparen, aufbewahren‘ gebraucht werden. Die semantische Zusammengehörigkeit von ‚verbergen, verstecken, vergraben‘ und ‚aufheben, zurücklegen, für die Zukunft aufbewahren, sparen‘ ist naheliegend.

9 Im Verzeichnis steht an dieser Stelle „Vita *hågke* vivere *jelidh*“, wobei nicht eindeutig ist, ob *jelidh* hier ‚vita‘ oder ‚vivere‘ bedeuten soll. Auf jeden Fall steht weiter unten noch einmal „Vivere *Jälith*“.

richtigen Stelle, sowie auch *curare*, dessen saamisches Äquivalent *kattith* mit LL *kattet* ‚cavere, custodi- re; curam gerere‘ identisch ist. Auf das rätselhafte *morreteth* wird am Ende des Beitrags zurückgekommen.

Einige Wörter, die im Verzeichnis die Endung *-th* haben, gehören zu den sog. Karitivadjektiven, d. h. sie enthalten ein Suffix, das dem deutschen *-los* entspricht: *kialmeth* ‚cæcus‘, *peliet* ‚surdus‘, *maimath* ‚liber‘ und (trotz der angeblichen substantivischen Bedeutung wohl auch) *jwrmath* ‚imprudencia‘ (also ‚augenlos‘, ‚ohrlos‘, ‚schuldlos‘, ‚vernunftlos‘, vgl. LL *tjalmete/bme*, *-s* ‚cæcus‘, *peljete/bme*, *-s* ‚surdus‘, *mainete/bme*, *-s*, *-k* ‚innocens‘, *jerbmete/bme*, *-s* ‚insipiens, imprudens‘ zu *tjalme* ‚oculus‘, *pelje* ‚auris‘, *maine* ‚culpa, vitium; morbus; delictum‘, *jerbme* ‚cerebrum; intellectus‘). Möglicherweise begegnet dasselbe Suffix (allerdings anders geschrieben) auch im Adjektiv *pakott* ‚rudis‘. Es könnte mit LL *lakhot(e)s* ‚dimidius‘, *lakhots pargo* ‚opus cuius non nisi dimida pars absoluta est‘ identisch sein (*p* pro *l* falsch abgeschrieben). Dieses Wort wird ein Karitivadjektiv zu LL *lakke* ‚pars dimidia‘ sein, d. h. ‚dem die Hälfte fehlt‘. Das Karitivsuffix hat im Saamischen (und im LL) mehrere Varianten. Die oben aufgezählten Formen auf *-th* dürften der Variante *-tek* entsprechen, das Verzeichnis enthält aber auch der Variante *-tebme* entsprechende Formen: *almatijm* ‚insipidus‘ (wohl: ‚geschmacklos‘, vgl. (neu)lat. *insipiditas* ‚Geschmacklosigkeit‘ zu LL *almes* ‚dulcis‘, südsaamisch *aelmie* ‚würziger Geschmack‘), *maimetijm* ‚insons‘ und möglicherweise noch *iwösetim* ‚rarus‘ und/oder *jeszetijm* ‚tepidus‘, die eventuell mit LL *assetes* ‚tenuis‘, südsaamisch *jissehts*, inarisaamisch *asettim* id. zu vergleichen wären. Hierbei handelt es sich um ein Karitivadjektiv zu LL *asse* ‚cutis, cuticula‘, südsaamisch *jissie* ‚Innenseite des Felles‘ mit der ursprünglichen Bedeutung ‚das eine dünne Innenseite hat (vom Fell)‘. Das letztgenannte Wort würde lautlich fast ideal zu dem südsaamischen Wort passen, lateinisch *tepidus* könnte ein Abschreibfehler (pro *tenuis*) sein. Bei dem erstgenannten Wort, das mit größerem Vorbehalt hierzu gehört, wäre von einem entstellten Schriftbild und einer ungenauen Wiedergabe der Bedeutung (*rarus* ‚spärlich, nicht dicht (z. B. vom Haar eines Felles)‘) auszugehen.<sup>10</sup>

10 Vgl. LL *niarbe* ‚rarus, gles; tenuis, tunn‘, *niarbot* ‚rareferi, rarescere, blifva tunn och gles‘, *qwolga niarbo* ‚pili rarescunt‘. – Neben *maimath* ‚liber‘ und *maimetijm* ‚insons‘ begegnet im Verzeichnis noch *maimetime*, dessen Bedeutung mit ‚liberare‘ angegeben wird. Möglicherweise hat hier ursprünglich *liberatus* gestanden. Die Bedeutung wäre hier (wie bei *maimath*) als ‚von einer Schuld freigesprochen‘ o. ä. zu verstehen. – Anscheinend identisch mit *jwrmath* ist *jermet* ‚stultitia‘ (gleichfalls mit substantivischer Bedeutung!); vgl. LL *jerbmetes-wuot* ‚dementia, imprudencia‘ mit einem Ableitungssuffix *-wuot*, das im Saamischen abstrakte Substantive aus Adjektiven bildet. Dieses Suffix ist häufig auch bei Plantinus: *irmes vâth* ‚prudencia‘ (LL *jerbmeswuot* id.), *pahas vâth* ‚inimicitia‘ (LL *pahaswuot* ‚malitia‘, *paha(s)* ‚malus‘), *varres vâth* ‚sanitas‘ (LL *warreswuot* ‚valetudo, integra‘; zu *varres* ‚sanus‘, LL *warres* id.), *sudes vâth* ‚quies‘ (LL *sâddoswuot* id.; zu *sieddâs* ‚tutus‘ und ‚securus‘, LL *sâddos* ‚tranquillus‘), *bâreswothe* ‚senectus‘ (nordsaamisch *boaris* ‚alt‘, *boarisvuohta* ‚Alter‘), *vöridvôt* ‚sobrietas‘ (LL *wuorredeswuot* id.; zu *vörid* ‚sobrius‘, LL *wuorredes* id.), *kalgos vâth* ‚negligentia‘ und *kolgos vâth* ‚pigritia‘ (LL *kâlkoswuot* ‚pigritia, ignavia‘; zu *kalgos* ‚negligens‘ und *kolgos* ‚piger‘, LL *kâlgos* ‚tardus, segnis‘), *arges vâth* ‚timor‘ (LL *argeswuot* ‚timiditas‘; zu *arges* ‚timidus‘, LL *arge(s)* id.), *arves vâtes* ‚terror‘ (zu *arvos* ‚pavidus‘; vgl. südsaamisch *asvoeh* ‚entsetzlich, schrecklich‘?), *skamoszvâth* ‚verecundus‘ (mit einer angeblichen adjektivischen Bedeutung ‚schamhaft, schüchtern‘ statt der zu erwartenden *\*verecundia* ‚Scham(gefühl), Schüchternheit‘; falsch kopiert?; vgl. LL *skabmokes* ‚pudendus‘, *skabmokeswuot* ‚turpitudus‘). Trotz der substantivischen Bedeutung stehen in einer adjektivischen Form anscheinend *sadnes* ‚veritas‘ (LL *sadneswuot* id., *sadnes* ‚verus‘), eventuell auch *jallo* ‚invidia‘ (falls es sich hierbei um dasselbe Wort handelt, wie *jallos* ‚inimicus‘; zu der Wortfamilie des erstgenannten gehört jedenfalls *jallolit* ‚invidere‘, unklar ist hingegen, ob diese Wörter irgendwie mit LL *jall(a)* ‚stultus, stolidus‘, *jalla-wuot* ‚stultitia‘, *jallastet* ‚stultum agere‘ zu verbinden sind).

Während *rasatijm* ‚instructus‘ und *podetijm* ‚nuntius‘ mit keinem saamischen Wort(stamm) zu identifizieren sind, scheint *puđetijm* ‚tabes‘ mit folgenden Wörtern zu verbinden zu sein: LL *påd(w)o* ‚inquinamentum, sordes, omne istud, quod inquinat‘, *påd(w)(ot)et* ‚contaminare, inquinare‘. Hier handelt es sich nicht um ein Karitivadjektiv, sondern um ein Verbalsubstantiv auf *-m* (zunächst des dreisilbigen Stammes *pådwoťet*). Zu derselben Kategorie gehören die zwei ‚biblischen‘ Wörter *hiettetalim* ‚similitudo‘ (zu LL *sjättetallet* ‚applicare, congruere facere, aptare‘) und *kegge labma* ‚tentatio‘ (LL *kättjelem* ‚tentatio, periculum‘ zu *kättjelet* ‚periclitari, tentare‘). Das letztere (fälschlich in zwei Wörtern geschriebene) Wort ist anscheinend eine Illativform und entstammt dem Vaterunser (et ne nos inducas in tentationem; im ABC-Buch 1638 *kiäggielebma*, s. genauer Bergsland 1982: 15). Das erstere begegnet (allerdings in einer kürzeren Form *hiettedem*) bei Graan 1669, nicht aber in LL, wo für ‚similitudo‘ *muotolwas* angegeben wird. Zwei Verben stehen in einer auf *-m* auslautenden Form, die die 1. Person Sg. sein könnte: *munnum* ‚lactere‘ und *vtigem* ‚minuere‘.<sup>11</sup> Möglich wäre aber, dass es sich auch hierbei um Verbalsubstantive handelt, vgl. LL *utsanem* ‚ipse actus, dum quid minuitur, sive minus sit, imminutio‘ (zu *utsanet* ‚minui‘).<sup>12</sup> Neben den Verbalsubstantiven haben die (aktiven und passiven) Partizipien des Präteritums eine Endung auf *-m*. Hierher scheinen folgende Wörter zu gehören: *vaddijm* ‚donum‘, *valtuin* (val-tum?) ‚conjugium‘, *sijlome* ‚lassus‘ (vgl. LL *waddes* ‚donum‘, *waldom* ‚nuptus‘, *sillom* ‚fessus‘ zu *waddet* ‚dare‘, *waldot* ‚uxorem ducere, nubere‘, *sillot* ‚defatigari‘), sowie *pådohijm* ‚furibundus‘ zu *pådohieth* ‚fure-re‘ (vgl. LL *piädat* ‚insanire, stultum esse‘, *piädatet* ‚injicere alicui amentiam‘) und *wårnotom* ‚juratus‘ (vgl. LL *wuordnes* ‚jus jurandum‘, *wuordnom* id., *wuordnot* ‚jurare‘, *wuordnotet* ‚jus jurandum imponere‘).

### 3.3 Zur Semantik

Der Hauptteil der konkreten Substantive in Plantinus’ Wörterverzeichnis kann in folgende semantische Gruppen eingeteilt werden: (i) leblose Natur (Zeit, Licht, Wetter, Gelände u. dgl.), (ii) lebendige Natur (Pflanzen, Tiere), (iii) Anatomie (Organe, Körperteile u. dgl.), (iv) Physiologie (Nahrung, Krankheiten, Sekrete), (v) Mensch (Alter, Geschlecht, Verwandtschaft, Beruf u. dgl.), (vi) Materielle Kultur (Bauten, Fahrzeuge, Kleider, Instrumente u. dgl.). Bei den Substantiven mit abstrakter Bedeutung handelt es sich zum großen Teil um Ableitungen von Adjektiven und Verben. Beispiele dafür sind

- 
- 11 Bei dem ersteren ist trotz des entstellten Schriftbildes (süd)saamisch *njammedh* (1. P. Sg. *njammem*) ‚saugen‘ zu erkennen, das letztere (oder wenigstens sein Stamm) dürfte mit LL *utsetet* ‚minuere‘, (süd)saamisch *uhtjiedidh* (1. P. Sg. *uhtjedem*) ‚vermindern‘ zu verbinden sein.
- 12 Mehr oder weniger sicher gehören zu den Verbalsubstantiven auf *-m* weiter folgende Wörter: *koikolim* ‚sitis‘ (zu *koikolith* ‚sitere‘; LL *käikeles* ~ *käikelwas* ‚sitis‘, *käikelet* ‚celeriter arescere‘, *käikeluet* ‚sitere‘), *jäkijm* ‚haustus‘ (LL *jukkem* ‚potatio‘ zu *jukket* ‚bibere‘), *höfvidjan* ‚laus‘ (zu *höffvidh* ‚laudare‘, südsaamisch *heevvedh* ‚beloben‘; LL *hewetem* ~ *hewetes* ‚laus‘, *hewetet* ‚laudare‘), *sijtom* ‚voluntas‘ (zu *sijta* ‚velle‘; LL *sit(t)em* ~ *situd* ‚voluntas‘, *sit(t)et* ‚velle‘), *kailekothin* ‚jocus‘ (zu *kailakotith* ‚jocari‘; vgl. LL *skalkestallet* ‚nequitiam perpetrare‘, *skalk* ‚nequam, nebulo‘, nordsaamisch *skälkkaštit* ~ *skälkkošit* ‚Späße machen‘, *skälka* ‚Spaßmacher, Schelm‘), *spallim* ‚alapa‘ (LL *spekkestem* ‚plaga, alapa‘, *niärab spekket* ‚alepam impingere‘ s. v. *spekket* ‚vola pulsare sive percutere‘; *njär* ‚genæ, mala‘), *kiäres potin* ‚acceptus‘ (vgl. LL *påtem* ‚adventus‘, *heres* ‚carus, dilectus‘, lulesaamisch *buorisboahthem* ‚willkommen‘), *maitim* ‚sapor‘ (zu *maitzateth* ‚gustare‘, LL *maistet* id.), *siöggiom* ‚reditus‘ (dunkel).

oben (Fußnoten 10 und 12) angeführt. Besondere Aufmerksamkeit unter den abstrakten Substantiven verdienen Wörter, die zum Begriffskreis (vii) der geistigen Kultur (Religion, Gesellschaft) gehören. (Appendix 1)

Obgleich die meisten Verben einen Menschen oder wenigstens ein Lebewesen als prototypisches Subjekt voraussetzen, kann das Subjekt bei einem Teil der Verben auch ein Ding, eine Naturerscheinung usw. sein. Zwei Gruppen zeichnen sich ab: (viii) intransitive Verben und (ix) aktive, kausative und transitive Verben, die (auch) mit einem nicht belebten Subjekt stehen können. Ein belebtes (nicht unbedingt menschliches) Subjekt voraussetzende Verben gehören zu folgenden semantischen Gruppen: (x) Verben für physiologische Tätigkeiten, (xi) spatiale Verben (Bewegung, Dasein, Orientierung u. dgl.), (xii) faktitive, kausative und transitive Verben mit einem prototypisch belebten Subjekt. Die Verben, die einen Menschen als prototypisches Subjekt voraussetzen, bilden folgende Gruppen: (xiii) faktitive und kausative Verben (arbeiten, sich bemühen u. dgl.), (xiv) possessive und habitive Verben (besitzen, bekommen, verlieren, dienen u. dgl.), (xv) mentale und kognitive Verben (denken, fühlen, ausdrücken u. dgl.). (Appendix 2)

Die Adjektive bei Plantinus können je nach dem, ob sie sich (xvi) auf eine Eigenschaft eines Menschen (oder eines Tieres) beziehen oder (xvii) auch in Bezug auf nicht belebte Gegenstände benutzt werden können, in zwei Gruppen eingeteilt werden. (Appendix 3). Bei den Numeralien handelt es sich um die Kardinalzahlen von 1 bis 10, wobei für lateinisch *duo* saamisch *wēstes* steht, welches wohl eher die Ordinalzahl ‚primus‘ sein wird, vgl. LL *wuostes* ‚primus‘, *qwekt(e)* ‚duo‘ (aber auch *qwektes* id.). Bei den Partikeln handelt es sich um Adverben für Raum und Zeit, grammatikalische Wörter, Präpositionen u. dgl.

## 4 Schlussfolgerungen

Das Wörterverzeichnis von Z. Plantinus fällt zeitlich mit dem Katechismus und dem Manuale von O. S. Graan, der Lapponia von J. Schefferus und dem Verzeichnis der finnischen Wörter von M. Fogel zusammen. Dem letztgenannten und Plantinus' Verzeichnis ist gemeinsam, dass es sich bei ihnen um eine handgeschriebene Wortliste handelt, die auf Verlangen einer bestimmten Person für wissenschaftliche Zwecke erstellt wurde. Im Gegensatz zu Fogel, der sein Material direkt aus VR entnommen hat, hat Plantinus, was das lateinische Material anbetrifft, weder dieses noch das Lexikon von E. Schroderus als Vorlage benutzt. Von den ca. 850 lateinischen Wörtern bei Plantinus sind nämlich nur ca. 360 in VR und ca. 430 bei Schroderus belegt; ca. 300 Wörter sind sowohl in VR als auch bei Schroderus unbelegt.

Eine unerklärte Frage ist, wie sich das hier analysierte Wörterverzeichnis zu der Handschrift eines *Lexicon Lapponicum* von Plantinus verhält, dessen Vorrede von Schefferus zitiert wird. Wie oben (Abschnitt 3.1) erwähnt, hat Setälä auch die von Schefferus benutzte Vorrede gefunden, allerdings an einer anderen Stelle, sodass er es für unsicher ansieht, ob sie zu diesem Wörterverzeichnis oder zu einer



verlorengegangenen größeren Arbeit gehört. Wie das Wörterverzeichnis ist auch die Vorrede nicht von Plantinus selbst sondern von einer fremden Hand niedergeschrieben. Obgleich Setälä das nicht ausdrücklich sagt, muss die Handschrift der Vorrede wohl eine andere als die des Verzeichnisses sein, denn sonst hätte man ja keinen Grund, an ihrer Zusammengehörigkeit zu zweifeln. Eine Möglichkeit ist, dass die erhalten gebliebenen Dokumente zwei zu verschiedenen Zeiten und für verschiedene Zwecke angefertigte partielle Kopien (die eine für Schefferus und die andere für Stiernehielm) aus der Originalhandschrift von Plantinus darstellen. Eine andere Möglichkeit ist, dass es sich bei der Vorrede, die auch Plantinus' eigenhändige Ergänzungen und Berichtigungen enthält, um den Anfang eines für die Druckerei bestimmten Dokuments handelt; ungewiss ist dabei, ob es ursprünglich länger gewesen ist und auch den Wörterbucheil umfasst hat; falls ja, so könnte das uns überlieferte Verzeichnis daraus kopiert worden sein. Als Vorlage der Kopie hat natürlich auch irgendeine frühere oder spätere Arbeitsphase dienen können. Und wie es damit auch bestellt sein mag, ist es möglich, dass es sich bei der an Stiernehielm versandten Wortliste um einen Auszug (oder eine Auslese) aus einem größeren Wörterbuchmanuskript handelt.

Im 15. Kapitel der Lapponia (*De Lingva & Sermone Lapponum*) führt Schefferus (1673: 178-182) drei Wortlisten mit 52 saamischen Wörtern an. Von ihnen kommen 11 bei Plantinus nicht vor, und auch die dort begegnenden 41 Wörter sind insofern anders geschrieben, dass ihre primäre Quelle augenscheinlich nicht Plantinus ist. Zu den Informanten von Schefferus gehörten auch saamische Muttersprachler, darunter der schon erwähnte Olaus Stephani Graan (Pastor in Lycksele) und sein Namensvetter Olaus Graan (Pastor in Piteå) sowie Olaus Sirma (Student und später Komminister in Enontekis), von dem die zwei durch die Lapponia später weltberühmt gewordenen saamischen Lieder stammen. Da die von Schefferus angeführten Wörter nördliche und östliche Merkmale aufweisen (*ännä* ‚mater‘, *Immel* ‚Deus‘, *riemnes* ‚vulpes‘, vgl. Plantinus *ädne*, *Jubmel*, LL *edne*, *Jubmel*, *repe*), ist es wahrscheinlich, dass Schefferus auch die Wörter von dem in Kemi-Lappland aufgewachsenen Olaus Sirma Angaben bekommen hat. (Zur Heimat von Sirma s. auch Itkonen 1940.)

Wie gut Plantinus selbst das Saamische beherrschte, wissen wir nicht. Muttersprachler war er nicht, möglicherweise hatte er aber schon von Kindheit an Kontakt mit der Sprache. Es ist auch möglich, dass das Verzeichnis in der einen oder anderen Weise eine Kollektivarbeit ist. Wie schon Qvigstad (1947: 39) bemerkt, steht die Sprache in Plantinus' Wörterverzeichnis der in den Werken von O. S. Graan nahe, was sowohl aus zeitlichen als auch geografischen Gründen natürlich ist. Dies gilt vor allem für die Schreibweise der Wörter. Über das lexikalische Verhältnis von Plantinus' und Graans Arbeiten kann nichts Absolutes gesagt werden, weil Graans Wortschatz (besonders das Manuale) noch nicht systematisch erforscht ist. Da das Hauptgewicht in Plantinus' Verzeichnis deutlich auf der profanen (nicht kirchlichen, religiösen oder biblischen) Lexik liegt, ist anzunehmen, dass die Wörter nicht aus der gedruckten Literatur exzerpiert wurden, obgleich die zeitgenössischen Druckwerke dem Zusammensteller und seinen Informanten natürlich bekannt gewesen sind. Falls Plantinus seine Arbeit an dem Wörterverzeichnis schon in Hernösand angefangen hat, könnte ein Teil des Wortschatzes eventuell von saamischen Schülern in Hernösand herkommen. Auf den heterogenen Ursprung des Mate-

rials deutet, dass ein Teil der Verben nicht im Infinitiv angeführt sind, und vielleicht auch, dass viele saamische Wörter im Verzeichnis verschieden geschrieben als Bedeutungsäquivalente für mehr als ein lateinisches Wort begegnen.

Es ist möglich, dass das Verzeichnis (wenigstens zum Teil) auf Material basiert, in dem die Metasprache ursprünglich nicht Lateinisch, sondern Schwedisch gewesen ist. Dadurch würden zwei rätselhafte Wörter einleuchtend werden. Für lateinisch *brevis* steht im Verzeichnis sowohl *aniki* (LL *āne(hes)* ‚brevis; kort‘) als auch *spelek*. Letzteres widersetzt sich allen Vergleichen mit Wörtern mit der Bedeutung ‚kurz‘, ihm kommt aber sehr nahe LL *spiäl*, *spjåla* ‚chartae lusoriae; kort‘, also ‚Spielkarte(n)‘. Wie ersichtlich, werden diese zwei Wörter im Schwedischen in gleicher Weise geschrieben, sodass die Verwechslung der Bedeutung verständlich ist. Ein anderer Fall kann das oben erwähnte *morreteth* sein, das im Verzeichnis für *conari* statt eines anzunehmenden *curare* steht. Das schwedische Bedeutungsäquivalent von *mårretet* ist bei Fjellström *försörja* (= ‚(ver)sorgen‘). Lateinisch *conari* (= ‚versuchen‘) heißt auf Schwedisch *försöka*, veraltet auch *försökja*. Diese Wörter sind zum Verwechseln ähnlich, ein falsch gedeuteter Buchstabe hätte die Bedeutung leicht ändern können.

Es sieht so aus, dass sich in das Verzeichnis in verschiedenen Arbeitsphasen Fehler eingeschlichen haben, sodass auch die weiter oben besprochenen nicht alle dem Hersteller der uns erhalten gebliebenen Kopie zuzuschreiben sind. Trotz seiner Mängel und Ungenauigkeiten zeugt das Verzeichnis von guter Kenntnis des Saamischen und Lateinischen beim Verfasser (und/oder seinen Informanten). Plantinus' Arbeit blieb seinerzeit ungedruckt, und über ihre Bedeutung für die spätere saamische Lexikografie wissen wir nichts Genaueres. Es ist möglich, dass das Manuskript noch im nächsten Jahrhundert von Fjellström benutzt wurde. Die von mir durchgeführte Analyse hat jedoch keine direkte Indizien dafür vorlegen können. Sicher ist hingegen, dass die von Schefferus angeführten saamischen Wörter nicht Plantinus entstammen. Somit ist auch wahrscheinlich, dass ihm nur die Vorrede zu Plantinus Lexicon zur Verfügung gestanden hat.

## 5 Literatur

Aapala, K., Koponen, E., Ruppel, K. (2010). Überblick über die Geschichte, den gegenwärtigen Stand und die Zukunftsperspektiven der etymologischen Forschung des Saamischen (Lappischen) in Finnland. In *Studia etymologica Cracoviensia*, 15, pp 7-14.

Álgu-Datenbank. Álgu-tietokanta. <kaino.kotus.fi/algú>

Bergsland, K. (1982). Den svensk-samiske ABC fra 1638 som sproghistorisk dokument. In *Språkhistoria och språkkontakt i Finland och Nord-Skandinavien. Studier tillägnade Tryggve Sköld den 2. november 1982*. Kungl. Skytteanska Samfundets handlingar 26. Umeå, pp. 11-20.

Fjellström, P. (1738). *Dictionarium Sueco-Lapponicum*. Stockholm.

Graan, O. S. (1668). *Korta och Enfaldiga Spörsmahl Öfwer Catechismum*. Stockholm.

Graan, O. S. (1669). *Manuale Lapponicum*. Stockholm.

Itkonen, E. (1940). Olaus Sirman kotiseudusta ja kielestä. In *Virittäjä* 44, pp. 334-349.

JSFOu = Journal de la Societé Finno-Ougrienne. Helsinki.

- Koponen, E. (2010). Johannes Tornaueuksen Manuale Lapponicum murrepohjasta ja vaikutuksesta ruotsinlappin kirjakielen myöhempään kehitykseen. In *Sanoista kirjakieliin. Juhlakirja Kaisa Häkkiselle 17. marraskuuta 2010*. MSFOu 259, pp. 43-52.
- Larsson, L.-G. (1997). Prästen och ordet. Ur den samiska lexikografins historia. In *LexicoNordica* 4, pp. 101-117.
- Leem, K. (1768, 1781). *Lexicon Lapponicum bipartitum*. Nidrosiae.
- LL = Lindahl, E., Öhrling, J. (1780). *Lexicon Lapponicum*. Holmiae.
- Löw, B. (1956). Einleitung und Kommentare zum Text in *Johannes Schefferus, Lappland*. Översättning från latinet av Henrik Sundin. Granskad och bearbetad av John Granlund, Bengt Löw och John Bernström. Nordiska museet. Acta Lapponica 8. Uppsala, pp. 9-23, 408-408.
- Magga, O.-H. (2012). Lexicography and indigenous languages. In *Euralex 2012 Proceedings*, pp. 3-18. MSFOu = Mémoires de la Société Finno-Ougrienne. Helsinki.
- Qvigstad, J. (1899). Übersicht der geschichte der lappischen sprachforschung. In *JSFOu* 16:3, pp. 11-29.
- Qvigstad, J. (1947). Sproget i Graans Manuale Lapponicum. In *Studia Septentrionalia* 3. Oslo, pp. 8-39.
- Qvigstad, J., Wiklund, K. B. (1899). *Bibliographie der lappischen Litteratur*. MSFOu 13.
- Schefferus, J. (1673). Lapponia Id est, regionis Lapponum et gentis nova et verissima descriptio. Francofurti.
- Schroderus, E. (1637). *Lexicon Latino-Scondicum*. Holmiae Sueonum.
- Setälä, E. N. (1890). Ein lappisches wörterverzeichnis von Zacharias Plantinus. In *JSFOu* 8, pp. 85-104.
- Sköld, T. (1986). Pehr Fjellström och det svensksamiska skriftspråket. In *Saga och sed* 9, pp. 15-26.
- Stipa, G. J. (1990). *Finnisch-ugrische Sprachforschung von dem Renaissance bis zum Neopositivismus*. Redaktionelle Bearbeitung und Zusammenstellung der Bibliographie Klaas Ph. Ruppel. MSFOu 206.
- Tornaueus, J. (1648). *Manuale Lapponicum*. Stockholm.
- VR = Variarum Rerum Vocabula Latina (1579, 1668).
- Wiklund, K. B. (1922). De första lapska böckerna. In *Nordisk tidskrift för bok- och biblioteksväsen* 9. Uppsala, pp. 13-28.

## Appendix 1

- (i) tempus, aetas, annus, ver, aestas, autumnus, hyems, dies, meridies, vespera, nox, lux, umbra, tenebrae, radius, sol, luna, stella, coelum, sonus, strepitus, ventus, tempestas, nubes, nebula, aqua, glacies, nix, grando, pluvia, gutta, ros, bulla, spuma, foetor, odor, nidor, fumus, fuligo, ignis, flamma, pruna, carbo, pulvis, lutum, coenum, arena, lapis, aes, ferrum, locus, terra, sylva, desertum, mons, clivus, vallis, fossa, littus, insula, lacus, flumen, fons;
- (ii) arbor, arbuscula, abies, pinus, radix, cortex, virga, ramus, gummi, arundo, foenum, muscus, animal, alces, canis, equus, fiber, lupus, mus, mustela, taurus, sciurus, avis, aquila, corvus, piscis, perca, serpens, vermis, pediculus, musca, culex;
- (iii) corpus, cadaver, latus, dorsum, gibbus, sinus, pectus, umbilicus, pudenda, venter, cor, jecur, renes, vesica, \*armus, humerus, brachium, manus, pugnus, digitus, pollex, pes, femur, genu, poples, sura, cervix, collum, \*guttur, caput, vultus, nasus, mala, mentum, auris, oculus, pupilla, os ‚Mund‘, labium, lingua, dens, os ‚Knochen‘, pellis, sangvis, crinis, unguis, penna, cornu;
- (iv) cibus, caro, adeps, sebum, butyrum, caseus, lac, farina, bacca, ovum, fames, sitis, somnus, angor, dolor, scabies, livor, tussis, mucus, sudor, urina;
- (v) homo, vir, senex, foemina, puella, pater, mater, liberi, foetus, filius, filia, frater, soror, procus, sponsus, sponsa, conjugium, uxor, vidua, viduus, affinis, gener, socer, socerus, nuntius, dux, mercator, venator, fautor, famulus, subditus, possessor, debitor, mendicus, latro, rusticus, heros, amicus, adulter, claudus, majores, genus, populus, turba;
- (vi) domus, ostium, columna, porticus, gradus, \*angulus, focus, saeptum, vallum, pons, via, cymba, traha, jugum, habena, lorum, vinculum, funis, nodus, vestis, lacinia, pannus, tunica, balteus, calceus, chirotheca, anulus, \*saccus, rete, insidiae, hamus, arcus, sagitta, hasta, clavis, cochlear, acus, forfex, forceps, malleus, pistillum, cuneus, ligo, ignarium, silex, fomes, cos, gluten, scipio, lectus, patina, globus, liber, icon;



(vii) Deus, oratio, spiritus, superstitio, praerogativa, spectrum, bellum, pax, rixa, auxilium, labor, merces, meritum, tributum, pars, magistratus, licentia, institutum, lex, jussum, poena, poenitentia, mos, consuetudo, gestus, opinio, mendacium, veritas, verbum, sermo, colloquium.

## Appendix 2

- (viii) incipere, fluere, diescere, \*fulgere, splendere, sibilare, fremere, defervere(?), haerere, se incurvare, tumescere, tremere, rotari, cadere, silere, desinere;
- (ix) necare, occidere, frangere, lacerare, rumpere, perfundere, siccare, incendere, concremare, urere, extinguere, inficere, minuere, lustrare, vertere, sufflare;
- (x) vivere, fieri, crescere, mori, perire, respirare, comedere, mordere, mandere, rodere, bibere, lactere, sorbere, deglutire, lingere, gustare, odorari, intueri, videre, audire, spuere, ructare, vomere, cacare, tussire, sternutare, oscitare, ringi, scabere, pati, frigere, esurire, sitire, aegrotare, dormire, vigilare, expergisci, fatigare, quiscere;
- (xi) ire, currere, irruere, migrare, proficisci, pergere, venire, redire, sequi, natare, volare, luctari, titubare, scandere, surgere, stare, iacere, sedere, habitare, manere, exspectare, errare, quaerere, invenire, reperire, occurrere;
- (xii) facere, parere, pascere, cibare, trahere, portare, levare, haurire, tundere, fricare, relinquere;
- (xiii) laborare, conari, probare, tentare, moliri, ponere, suspendere, sepelire, occultare, iacere, iaculari, praecipitare, reicere, verberare, abigere, cogere, \*comitari, concitare, congregare, deducere, pellere, venari, ludere, aedificare, statuere, tendere, erigere, tegere, suere, secare, scribere, pungere, ornare, induere, aperire, confirmare, munire, excavare, extendere, excoiare, ungere, miscere, lavare, coquere, assare, implere, consumere, uti, finire, molestare, dolore afficere, simulare;
- (xiv) accipere, captivare, sumere, dare, tradere, donare, egere, possidere, tenere, retinere, \*furari, decipere, mendicare, liceri, vendere, emere, mercari, solvere, liberare, servare, custodire, curare, observare, parcere, \*compercere, desolare, deserere, auxiliari, parere;
- (xv) noscere, scire, recordari, intelligere, posse, consuescere, discere, opinari, cogitare, conicere, desiderare, optare, velle, poenitere, imaginari, excogitare, assentiri, approbare, despiciere, detestari, suspicari, dubitare, mirari, invidere, amare, metuere, timere, irasci, dicere, narrare, depraedicare, notificare, interrogare, respondere, defendere, invocare, iubere, docere, hortari, suadere, dissuadere, negare, consultare, consulere, promittere, iurare, mentiri, fateri, orare, deprecari, gratias agere, laudare, accusare, lacescere, maledicere, illudere, ineptire, [se] iactare, superbire, garrere, iocari, valedicere, blandiri, placare, solari, minari, terrere, ulisci, allicere, mussitare, queri, gemere, singultire, flere, deflare, clamare, vocare, rixari, furere, cachinnare, ridere, gaudere, laetari, canere, saltare, osculari, amplecti.

## Appendix 3

- (xvi) obligatus, reus, insons, liber, iuratus, dives, egenus, miser, tenax, \*carus, honestus, gratus, benevolus, concors, gratus, promptus, contentus, mansuetus, docilis, prudens, diligens, castus, probus, pius, cautus, pavidus, timidus, verecundus, audax, temerarius, anomosus, furibundus, inimicus, malevolus, contumax, ferox, superbus, imprudens, fallax, negligens, petulans, impudicus, impudens, improbus, insipiens, fatuus, stultus, piger, maestus, laetus, macer, nudus, natus, robustus, fortis, imbecillus, agilis, \*festinus, sanus, somnolentus, lassus, sobrius, ebrius, surdus, caecus, mutus, balbus;
- (xvii) simplex, solus, vulgaris, exiguus, sinister, totus, integer, rotundus, rectus, erectus, curvus, inversus, altus, latus, longus, brevis, crassus, rarus, densus, mollis, durus, gravis, levis, vacuus, plenus, tumidus, acclivis, profundus, pallidus, luridus, obscurus, niger, albus, ruber, viridis, versicolor, novus, recens, crudus, mucidus, insipidus, amarus, acidus, dulcis, humidus, siccus, frigidus, calidus, fervidus, constans, tener, remissus, munitus, rudis, erroneus, deformis, ineptus, nequam, malus, ridiculus, extraneus, mirabilis, bonus, acceptus, pulcher, pretiosus, precipuus, sanctus, verus, selectus, utilis, notus, secretus, securus, tutus.



# Compiling a Basic Vocabulary for German Sign Language (DGS) – lexicographic issues with a focus on word senses

Gabriele Langer, Susanne König, Silke Matthes

Institute of German Sign Language and Communications of the Deaf, University of Hamburg

Gabriele.Langer@sign-lang.uni-hamburg.de,

Susanne.Koenig@sign-lang.uni-hamburg.de,

Silke.Matthes@sign-lang.uni-hamburg.de

## Abstract

Nowadays lexicographic work such as lemma selection, identification of word senses and usage information are usually based on large corpora. Sign language lexicographers face the same difficulties as their colleagues of other less studied and previously not written minority languages. They cannot rely on written texts and large corpora. In addition sign language linguists have to cope with sign-language specific issues due to the visual-gestural modality, namely the iconicity of signs, a broad utilization of simultaneity of linguistic signals and the integration of lexical material of a spoken language by the way of mouthings.

Sign language specific lexicographic issues are discussed and exemplified with regard to the Basic Vocabulary of German Sign Language (DGS) that is compiled within the larger context of the DGS Corpus Project. The Basic Vocabulary is not corpus-based but based on previously published sign collections that are used as a starting point. The data is reviewed, sign senses are finer split and disambiguated and the data then undergo a validation process by the sign language community. This validation process is conducted through a feedback system especially designed for surveys involving sign language.

**Keywords:** lexicography of sign language; German Sign Language (DGS); word senses; feedback system; crowd sourcing; basic vocabulary; iconicity; mouthing

## 1 Background

### 1.1 Sign Language Dictionaries

Sign language lexicography is still in the process of striving for the best lexicographic methods to analyse signs, as well as solutions to document and present them in an adequate and user-friendly way. For meaning analyses and descriptions there are two aspects especially relevant in a dictionary pro-

ject: First, what kind of data is the given information based upon, and, second, how can the identified meanings of signs (senses) be made clear and best presented to the dictionary user.

In the past – due to a lack of an everyday writing system and written sign texts – sign dictionaries have hardly been in the position to base their sign selection and information on sign use on corpus analyses. Up to the present, sign dictionaries and sign collections have mainly been based on information drawn from earlier sign collections, word-to-sign elicitations, the introspection of single signers or groups of signers (and their metalinguistic discussions on signs' meanings and usage), on participant observation, or on a mixture of these methods (compare for example Johnston 1989: 8, Brien 1992: x-xi, Kristoffersen & Troelsgård 2010: 5). To our knowledge the first sign language dictionary largely and systematically based on a collection of filmed and transcribed sign language data was the ASL Dictionary published in 1965 (Stokoe, Casterline & Croneberg 1965). Considering the limited technical possibilities at the time (cf. Stokoe 1993) this work was outstanding in many regards and way ahead of its time.

Technical progress in recording equipment and transcription software now makes the collection of relatively large sign language corpora feasible. Sign corpora exist or are being collected in several countries such as Australia, Germany, Great Britain, Italy, the Netherlands, Poland, France and others. Most of these corpora are still in the stage of data collection or transcription and annotation. With the availability of larger annotated sign corpora the information in sign dictionaries – especially on the meanings and usage of signs – can reach new levels of quality.

Without a fully functioning writing system for signs (see below) some effort has to be put into the presentation of sign forms and how to provide a search by sign form in order to ensure bi-directional accessibility in sign dictionaries. However, how to best identify meaning(s) of signs and to make them accessible has not been discussed in great depth so far in sign language lexicography. Common to all general sign dictionaries that we are aware of is the bilingual solution of expressing the meanings of signs by matching translation equivalents (keywords) of the surrounding spoken language. In addition to that some dictionaries also provide meaning explanations (e.g. Johnston 1989), usage notes (e.g. Stokoe, Casterline & Croneberg 1965) or example-like contexts (e.g. Kennedy 2008, cf. esp. xvii) – all using the surrounding spoken language as meta-language – or signed examples (e.g. *Ordbog over Dansk Tegnsprog*).

In this paper we differentiate between sign dictionaries and more simple sign collections. A sign dictionary regards the signs as the linguistic units to be described with information about their form, grammar, meaning and usage. Sign language dictionaries of this kind have sign entries and – if necessary – include several senses of a given sign. Sign collections, on the other hand, are predominantly mono-directional word lists with the direction from spoken language words to signs. Usually, they contain no real sign entries and do not provide much information about the sign other than its form. For German Sign Language (DGS) there exist quite a number of different sign collections but not yet a corpus-based general dictionary.

## 1.2 The DGS Corpus Project

The DGS Corpus Project<sup>1</sup> is a long-term project of the Academy of Sciences and Humanities in Hamburg. The research is carried out at the Institute for German Sign Language and Communication of the Deaf (University of Hamburg). The project has started in 2009 and will continue until 2023. The project has two major goals: building a large general reference corpus of DGS and producing the first corpus-based general dictionary of DGS – German.

The DGS corpus is intended to be a linguistic and cultural resource for further research as well as the language community and other interested persons. It will also provide data for the first corpus-based dictionary DGS – German. For the corpus 330 signing informants from all over Germany were filmed in pairs at 12 different locations. The filmed material includes narrations of personal experiences, discussions, re-narrations, and other kinds of language use elicited in staged communicative events.<sup>2</sup> The raw data are comprised of 900 hours recording time with estimated 540 hours of signed activity containing estimated 3.5 million tokens of signs. A large portion of this data is being made accessible through segmentation, lemmatization, transcription, annotation and translation to become a searchable corpus. Corpus annotation work is carried out using iLex, an annotation environment that is linked to a lexical database (Hanke & Storz 2008). For lemmatization tokens are matched to types that are represented by unique glosses. The citation form of each sign type is described using the notation system HamNoSys (Hanke 2004). This is very time-consuming and cumbersome work, as none of these steps can be carried out in an automated way yet. Hence, the corpus will not have reached a relevant size for a number of years to go. This leads to the situation that at the present time DGS lexicographers have to resort to other sources and methods for dictionary compilation.

Within the DGS Corpus Project two electronic dictionaries of DGS will be produced in consecutive phases. First – while the corpus is still being built and the elicited data not yet fully accessible – a preliminary Basic Vocabulary with about 1500 entries is being produced. It is based on previously published sign collections and validated through feedback from the signing community. At a later stage the corpus data will be used as basis for the production of a general Dictionary of DGS – German with approximately 6000 sign entries to be published in 2023.

## 1.3 Sign Language as a Multi-channelled and Non-written Language

One challenge each sign dictionary project has to face is, that sign language is a visual-gestural language without written form. A sign stream involves signals produced by different parts of the upper body – two hands, mouth movements, facial expressions, head movements, posture, body shifts and eye-gaze. In the visual-gestural modality these articulatory features are easily produced and perceived simultaneously. Since the sign stream consists of several simultaneous but different signals it is diffi-

---

1 For more information on the project see: [www.dgs-korpus.de](http://www.dgs-korpus.de).

2 For more detail see Nishio et al. (2010).

cult to devise a writing system that on the one hand can capture and represent the form and given complexity and on the other hand is simple enough to be written and read easily. In the research of spoken languages it is possible to adapt an existing writing system to a previously unwritten language – at least for a transition period – because the problem of how to write spoken languages has been solved. In sign language, however, there is no functioning writing system for any language that could simply be adapted to DGS for the purpose of dictionary making.<sup>3</sup> For sign language lexicography this absence of an operating writing system poses several challenges:

First, researchers cannot rely on already existing written texts to build a corpus of or to base their analyses on. Corpora have to work with data consisting of filmed signing. For corpus building such films are often recorded specifically for this purpose. Sign films recorded for other purposes are still limited to specific contexts (types of language use, genres). For most signers they are not part of everyday communication and can therefore cover only a small segment of sign language use. Second, for corpus-building and transcription purposes there is no fast and easy way of writing down signed text in a way that represents the form of the language units. More demanding and time consuming notation and transcription methods or indirect forms of annotations have to be used. This also makes it more difficult to analyse and automatically process corpus data. For instance searches on word forms have to be provided for by other means if there is no written representation of the forms available. Third, there is no direct and stable representation of form that enables the analyser to scan and review the sign and its context on a glance directly through their linguistic forms in the same way as a written text would do for example in a concordance line. Fourth, for dictionary production there is no practical way to use sign language as meta-language extensively. Film clips with signed information do not provide the same usability for scanning, browsing, searching and punctual access as does written information, since the content of a film is not permanently visible or machine-readable. Furthermore, films are very time-consuming in production and editing and therefore also for practical and financial reasons no real alternative to the use of written information.

Fifth, since there is no written form of the lemma sign, the question arises of what element to use as lemma. The lemma usually represents the form of the word, serves as an element for quick identification, sorting, searching and ordering and as the address for cross-references. All elements that could possibly be chosen (film, drawing, notation, gloss, number) do not fulfil all of these functions and have their drawbacks.<sup>4</sup> The best solution so far is to use a combination of one element to represent

---

3 There are some notation systems developed for academic purposes (e.g. HamNoSys, Stokoe-Notation). They are suited for detailed descriptions of sign forms. However, these systems do not handle facial expressions, body posture, eye gaze, mouthings or iconic locations well enough to completely rely on them in research. Also stretches of text are not easily written and read. Other systems, intended for general use such as SignWriting (Sutton et al. 2009), are not widely established and also have other drawbacks like not being machine-readable. For a further discussion of merits and shortcomings of the most common writing systems see König & Schmaling (2012).

4 For a more in-depth discussion of the advantages and disadvantages of these elements representing signs in dictionaries see Zwitserlood 2010: especially pp. 445-454 and Kristoffersen & Troelsgård, 2012a: 295-298.

the form of the lemma sign (e.g. drawing or film) and a second element for unique identification, sorting and address for cross-referencing (e.g. unique gloss or number).

## 2 The Basic Vocabulary of DGS

### 2.1 The Basic Vocabulary within the Context of the DGS Corpus Project

The Basic Vocabulary is intended to be a basic learners' vocabulary and to cover the most central signs of DGS and their core meanings. It is based on information about DGS signs contained in different sign collections and learning materials. The information is lemmatized, reviewed, edited and prepared for the signing community's feedback by which it will be validated and supplemented. The Basic Vocabulary is only a first and preliminary step in the lexicographic work within the project. It serves a number of purposes within the overall project design:

It is a means to round up and review information already published on DGS signs and put some of it to the test. This allows for a critical review and evaluation of previous lexicographic work with the chance of identifying and eliminating artefacts and to move towards a more complete and adequate description of signs and their meanings. With the Basic Vocabulary we can also develop and test lexicographic solutions to sign language specific problems that result from the visual modality as well as from the lack of a written form.

In addition, the publication of the Basic Vocabulary is a way to give something back to the signing community for their previous participation in the project, i.e. as informants in the corpus data collection, much earlier than 2023. The feedback process is also a way to involve the language community and let them further participate in this project on their language. Through the feedback they gain some influence and control over how their language is represented in the resulting dictionaries.

For the compilation of the Basic Vocabulary a set of question types to validate lexical items are developed, tested and used to collect data on signers' intuitions about signs and their meanings. Also a feedback system is specifically designed to conduct this survey via sign language. In the context of the Basic Vocabulary we can develop, test and improve both feedback tool and procedures. The gained experience and know-how can at a later stage be utilized for supplementing corpus data in the production of the Dictionary DGS – German.<sup>5</sup>

---

5 With estimated 2.5 million tokens the DGS corpus will still be relatively small in comparison to corpora of well-studied written languages. Complementary methods to gain additional insights on signs and their meanings and uses will be useful.

## 2.2 Properties of the Basic Vocabulary

The Basic Vocabulary is developed with a primarily monolingual oriented perspective: It focuses on the signs and aims at an adequate description of their variants and meanings. The meta-language of the dictionary is written German. Following this approach the dictionary will contain only one type of entries, that is, sign entries. Nearly all information included in the Basic Vocabulary will concern DGS signs while at the same time there is neither intention nor need to reproduce information on German words that can be obtained easily from existing dictionaries of German. Nevertheless, the actual product will exhibit some features of a bilingual dictionary: The signs' senses will be identified and expressed by German translation equivalents and disambiguating contexts. Furthermore there will be an alphabetical index of the German equivalents, making the sign entries accessible also via German (as source language) and thus providing also for a bilingual access.

The sign entries will be ordered by formational parameters of the signs.<sup>6</sup> However, since the Basic Vocabulary will be produced in an electronic form, the primary ordering of signs (macrostructure) is not crucial. The product will include a search function accessing the entries via sign form. For a search the user can select different combinations of manifestations of the sign form parameters as presented in a series of menus.<sup>7</sup>

## 2.3 Sign Collections: Sources for Lemma Selection and Spoken Language Equivalents

For the compilation of the Basic Vocabulary nine previously produced sign collections of DGS are taken as a basis for the selection and analysis of basic DGS signs.<sup>8</sup> The sign collections are used in two ways: for lemma selection and as a starting point for word sense disambiguation.<sup>9</sup>

---

6 Formational parameters are e.g. handshape, location, number of hands, movement types. There is no agreed upon order available and existing sign dictionaries have arrived at different solutions. For the current project the final order has not been decided yet.

7 This search function will be similar to the one implemented for the *Fachgebärdenlexikon Gesundheit und Pflege* ([5e]). To go to the sign form search function click on "Suche über Gebärdenform" at <http://www.sign-lang.uni-hamburg.de/glex/>.

8 Some of these collections consist of several publications that have been produced by the same team or institution and are follow ups of each other or a series of complementary works. For an overview see separate list in the reference section. The list of sign collections also include some teaching materials with vocabulary lists.

9 *Word sense* here in the case of sign language means "sign sense". In linguistics many technical terms exhibit a spoken language bias. Examples are: speaker, word, phoneme, phonology and oral. However, in a more general sense they can be used to describe the same abstract or corresponding phenomenon, category or role in both language modalities. Other technical terms with a spoken language bias are applied to sign language accordingly in this paper.



### 2.3.1 Lemma Selection

The Basic Vocabulary aims at including signs that are most basic with regard to everyday communication needs and that are frequently used and widely known. With frequency lists still unavailable for DGS the second best way to approximate this aim was to look for signs that had been listed in several previously published sign collections. Following this, we based the lemma selection process primarily on a comparison of these signs.

Without a written representation of the sign forms in the products the only way to identify and compare the signs was to look at each visual representation (photos with arrows or films) of the given sign, sign combination or sign sentence, break sign combinations and sentences down into single signs (tokenisation) and lemmatize them consistently. This was done in our working environment iLex following the annotation rules of the DGS Corpus Project. The nine sign collections contained about 40.800 tokens (single signs) that have been processed. Not counting uses of the manual alphabet, pointing signs, number signs and productive signs about 9350 different sign forms of 5440 basic conventional signs have been documented. All basic signs – about 1050 – that were found in at least five of the nine sign collections have been selected for further analysis and inclusion into the Basic Vocabulary.<sup>10</sup>

### 2.3.2 Spoken Language Equivalents

All of the nine sign collections contain to some extent bilingual information matching signs and words as equivalents. They either have a primary make-up using German as source language showing signed equivalents (e.g. [6]) or they have sign entries and use German equivalents and contexts to list the signs' meaning(s) (e.g. [5c-5f]). The vocabulary lists of learning materials (e.g. [9a, 9b]) and the collection of phrases ([2c]) use glosses as written labels for signs where the gloss word is to be taken as an indication of the meaning of the sign.<sup>11</sup> In the children's visual dictionary ([8a, 8b, 8c]) pictures of objects and sign films are presented without any written words. However, the sign film contains an audio track where the corresponding German word is spoken aloud.<sup>12</sup> All available equivalents and meaning indications assigned to one sign form in different sign collections are taken as a starting point for identifying word senses for signs that are to be included in the Basic Dictionary.

In the Basic Vocabulary we aim at the most central meanings of selected signs. Not all of the equivalents listed in the sign collections are equally basic or central, and it is sensible and necessary to make a selection. This selection is operationalized in the following way: German equivalents are compared

---

10 Additional signs will be included in order to make sure all concepts of a specific learner's basic word list of German (Glaboniat et al. 2005) and some Deaf-specific concepts are covered.

11 In corpus annotation and lexicography glosses are used as a unique identifiers for signs: one gloss always stands for the same form and different forms receive different glosses. In learning materials glosses are often used as a means to represent the sign order of example sentences and at the same time to indicate the signs' meanings in this particular context. Depending on the context a sign might therefore be represented by different glosses and the same gloss might be used for different signs.

12 In the print version of the product German equivalents of signs are printed below the signs' drawings (comp. Kestner 2002).

to a combined list of German words that are considered relevant for a basic (German) vocabulary. This combined list is basically a merge of a foreign language learner's basic vocabulary list of 1071 words (adapted from Glaboniat et al. 2005, levels A1 and A2 (active vocabulary)) and a frequency list of the about 4000 most frequent German words (Jones & Tschirner 2006, 2011). Words that are both listed as equivalents in the sign collections and found in the combined German word list are selected for further processing.

### **2.3.3 Reversal of bilingual information**

After the basic German equivalents of a sign have been determined, the next step is to identify individual senses that the words and the sign potentially share. This step can be understood in terms of the reversal of a mono-directional bilingual dictionary. It goes beyond the reversal on the form level and is aimed at the level of meaning, i.e. the level of lexical units (as outlined by Martin 2013: 1447-1448). However, in our case the equivalents do not stem from one uniform source but from different sources and are of different status, quality and granularity as far as meaning specification is concerned.

It is reasonable to assume that any German word listed as translational equivalent must share at least one sense with the corresponding sign. If the German word is polysemous each of its senses has to be considered separately, whether it could also be a potential sense of the sign. At this stage it is inevitable to do a considerable amount of sense splitting. For this we use German dictionaries. The approach we take is to identify relevant senses of the German word according to two criteria: First, the senses chosen are to be the most basic and central ones with regard to everyday communication. More marginal senses are omitted from further processing. Second, the senses selected for the feedback have to be certain or likely candidates for being senses of the sign as well. Whenever the context information provided in a sign collection clearly identifies a particular sense this information is taken into account and the identified sense goes into the feedback process for verification. The same is true for senses that are attested clearly in the data from previous projects or already transcribed corpus data in our database. In some cases it is very straightforward that a particular word sense is clearly not a sign sense and can be omitted, e.g. because it contradicts with the iconic properties of the sign. In many other cases a decision cannot be reached without further information of the sign's use in context. Uncertain basic senses are prepared for the feedback process to be verified or rejected by the language community (see 4.).

## **3 Sign Language Specific Issues concerning Word Sense Disambiguation**

In the process of identifying and disambiguating word senses of signs two sign language specific properties have to be taken into consideration: iconicity and mouthings.

### 3.1 Iconicity

Many signs are iconic. The form of an iconic sign resembles something – the underlying image – which is directly or indirectly associated with one or several of the sign’s meanings. The underlying image can influence the ways a sign’s meaning may be restricted or open for extension. Therefore, when checking possible senses of a sign special attention has to be given to its underlying image.

#### 3.1.1 Underlying Image and Meaning

The underlying image of a sign may restrict its use to certain specific senses and prohibit its use for related senses. Due to iconic aspects the granularity of meaning differentiation of a sign may be much finer than that of the corresponding word of the surrounding spoken language. When the underlying image of an iconic sign is very specific and the sign’s meaning is restricted accordingly, related senses that are not compatible with the iconic properties of the sign in question are often covered by other signs.

Consider for example different DGS signs for *Pfeife*. The German word *Pfeife* can mean either “whistle (object)” or “smoking pipe (object)” while DGS has different signs for these concepts PIPE<sup>13</sup>, WHISTLE1 and WHISTLE2 (see figure 1). The sign form of PIPE resembles a person holding a pipe at its bowl some small distance away from the mouth so that the end of its imagined shank touches the lips for smoking. In WHISTLE1 the fingers indicate a whistle (as used by referees) located with its mouth-piece at the mouth. WHISTLE2 shows how a person is holding a whistle to the mouth. The signs’ meanings are restricted to the different objects in accordance with their underlying image.

---

13 Since there is no conventional writing system for sign languages it is a convention in sign language linguistics to use capital letter glosses with numberings to identify and represent different signs. The spoken language word used as gloss simply is an easily written and readable label or name for the a sign, but neither is the gloss’ part of speech nor are its meanings to be interpreted as indicating the sign’s part of speech or meanings.

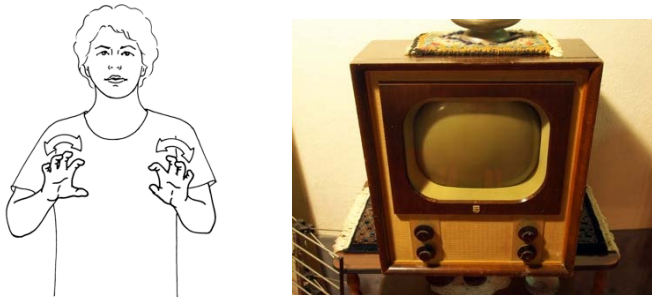


**Figure 1: DGS signs PIPE, WHISTLE1 (Source: [5e]), WHISTLE2 (Source: [6]), AND WHISTLE3 (Source: [5e]).**

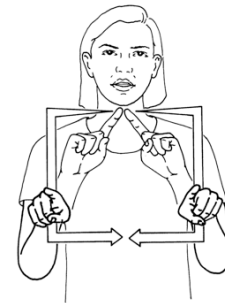
WHISTLE1 and WHISTLE2 can also mean “producing a whistling sound, whistling”. Another DGS sign listed for this meaning is WHISTLE3 (see figure1). On the iconic level the three signs for *whistling* are specific as to how the whistling sound is produced: WHISTLE3 shows how thumb and index finger are rounded to an O and put to the mouth producing the sound without the help of a whistle, while for WHISTLE1 and WHISTLE2 a whistle is involved. The question arises whether these iconic differences are reflected in sign usage or not. In a specific context the sign used will usually be the one that iconically best matches the named or described incident.<sup>14</sup> Thus the “whistling” sense of the signs could be further specified as “producing a whistling sound by blowing into a whistle” for WHISTLE1 and WHISTLE2 and “producing a whistling sound by using your fingers in the indicated way” for WHISTLE3. Sign language users may find this trivial or logical or even unnecessary information since the form of the sign shows these aspects of meaning, but this only proves the point that iconicity plays a role in sign language use. Whenever the selection and use of a particular iconic sign takes its underlying image into account this kind of information should be made explicit in a dictionary and thus be accessible to the language learner. This can be done by listing the specific meaning as determined by the iconic value of the sign as one of the sign’s senses.

However, iconic signs do not always exploit their more specific underlying images for subtle meaning differences. Often sign forms are conventionalized for a general meaning without paying attention to the more specific images their forms are derived from. One example is the sign TELEVISION (see figure 2). Its underlying image shows a person turning two knobs (the kind old television sets had for adjusting the frequency). This sign is used for all kind of television sets regardless of their technical type or outward appearance.

<sup>14</sup> This is a hypothesis based on informal observations and signers’ intuition. To what extend this can be confirmed by corpus data remains an open question.



**Figure 2: TELEVISION and old television set with turning knobs.**



**Figure 3: RECTANGLE.**

Signs – just like words of spoken languages – can extend their meaning from the specific to the more general. With this we come back to the whistling examples: When the particular circumstances of sound production are not known to the signer or are not relevant or if one talks about whistling in general, is it then possible that any of the three signs may be used for the more general concept of “whistling” independent of how the sound is produced? And what about the whistling of a boiling water kettle or of an old steam engine train? Can the three signs for “whistling” also be used in a transferred sense for these kinds of whistling or are there other signs to be used instead? Do WHISTLE1, WHISTLE2 and WHISTLE3 additionally to their very specific sense matching their underlying image have a more general meaning “producing a whistling sound” (no matter how it is produced)? These are questions to be asked and hopefully answered in the feedback and – at a later stage of the project – by analysis of corpus data.

Whenever a sign’s form resembles one particular example of a more general category its meaning can either be restricted to that specific kind resembled, or the sign can be used for the more general concept. There is no general rule as to what applies to a particular sign. For each sign this has to be determined separately. Thus, for iconic signs special attention has to be given underlying image when analysing their meanings.

### **3.1.2 Underlying Image as a Source for Polysemy**

The iconicity of signs is an important source for polysemy and meaning extensions in DGS. This is especially true when the sign’s form represents a very unspecified image that may stand for a number of objects, situations or actions. The underlying image can serve as a common core for rather diverse meanings which either share a common visual trait in their real-world manifestations or which utilize the visual metaphor provided. For example, among the translation equivalents listed in various sign collections as meanings for the sign RECTANGLE (see figure 3) quite a number refer to objects or convey meanings that either consist of or are associated with a piece of paper, such as: “paper”, “piece of paper”, “slip (paper)”, “page”, “form (to be filled in)”, “map”, “recipe (cooking)”, “prescription”, “sick note”, “official notification”, “certificate”, “report card”. Other meanings refer to rectangular and predominantly two-dimensional objects that are solid such as “window”, “signboard”, “mirror”, “screen (computer, TV)”, or that are soft such as “pillow”, “towel”, and “(cleaning) rag”. On the semantic level a

“pillow” does not have much in common with a “computer screen” neither does an “official notification” have anything to do with a “cleaning rag” – the characteristic all these share is their rectangular shape or their association with something of a rectangular shape. In cases like this the iconic potential depicted in the sign’s form serves as the basis for meaning extension of the sign.<sup>15</sup>

This kind of extensive polysemy in DGS is facilitated by an interaction of iconicity and mouthings. The mouthing helps to contextualize and disambiguate the polysemous sign (see below 3.2.).<sup>16</sup>

In DGS there are many ways of how the meaning of a sign can be extended. Often the meaning is extended along the lines of regular polysemy. Bentele, Konrad & Langer (2000: 621) list some conventional and productive uses of the sign STIRRING that exemplify some of these patterns (see Table 1). The form of STIRRING (see figure 4) depicts someone holding a tool, such as a cooking spoon, and stirring something with it.

It is very common in DGS that in addition to a sign’s core meaning (an action represented by the sign form as in STIRRING) it is also used in many other related senses – for example – for a profession typically associated with the activity, for a place or institution where the activity takes place, and for an object produced or manipulated by the activity.

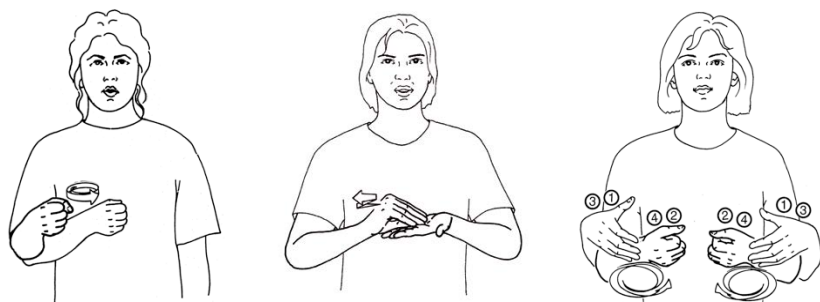


Figure 4: STIRRING, WRITING, and KNEADING.



Figure 5: BEAK.

15 The meanings listed here are conventional uses of the sign RECTANGLE as they appear in sign collections. The list here is not complete. These conventional meanings are strongly and regularly associated with this sign. However, in addition the sign may also spontaneously be used to denote other rectangular two-dimensional objects in a specific context, often in combination with a corresponding mouthing. This is a common strategy to be used, especially when there is no other conventional sign or when the signer doesn’t know such a sign. Such more spontaneous uses are backed by the sign’s iconic potential and will be easily understood in context. In contrast to the more *conventional uses* these have been called *productive uses* of a sign (cf. König, Konrad & Langer 2008).

16 Specific mouth gestures (mouth movements not derived from spoken language words) and facial expressions can also contribute to the disambiguation of meanings. These are not discussed here.

Other examples of this type are WRITING (“writing”, “author, writer”, “school” and “literature”), and KNEADING (“kneading”, “baker”, “bakery”, “dough”). Our impression is that in addition to well-known patterns of regular polysemy found in spoken languages (e.g. Atkins & Rundell 2008: 140) some DGS patterns may be specific to a visual language.<sup>17</sup>

German equivalents	English equivalents	meaning extension pattern
rühren, umrühren	to stir	core meaning: action directly represented by sign’s form
kochen	cook (prepare food)	more comprehensive activity typically involving the represented action
Küche	kitchen	place or area where activity usually takes place
Koch	cook, chef	person who usually or professionally carries out the activity
(Koch-)Löffel	(cooking) spoon	tool used in the represented action (productive use)
Eintopf	stew	object or substance manipulated by action (productive use)
Hauswirtschaft	home economics	superordinate domain the activity belongs to (productive use)

**Table 1: Some meaning extensions for STIRRING (adapted from Bentele, Konrad & Langer 2000: 621).**

The conclusion that we draw from these observations is that the underlying image and image producing technique of an iconic sign can be and should be considered for decisions on lemmatisation and in the analysis of word senses. For us a common underlying image connecting diverse meanings is a valid reason to treat these meanings as different senses of a polysemous sign in one entry rather than treating them as homonyms in different entries (cf. also König, Konrad & Langer 2008).

### 3.2 Mouthings

Many manual signs are accompanied more or less regularly by mouthings. Mouthings are mouth movements that remind of the articulation of spoken language words. Some of these mouthings are quite conventionally attached to certain signs, while others appear to be more occasional, dynamic or even spontaneously combined (cf. Ebbinghaus & Heßmann 1996, 2001). While the linguistic status of these mouthings is not yet finally agreed upon in sign language linguistics, we consider mouthings to be some kind of lexical material from German integrated into DGS signing.

Most signers are to some degree bilingual and have at least some basic knowledge of German words and their core meanings. Mouthings in DGS originate from German words and bear the potential of

<sup>17</sup> Further research might reveal that certain kinds of regular polysemy may be typical for or restricted to certain iconic sign types – whether the underlying image is a representation of an action (manipulative technique as in STIRRING), the outline of the form of an object (sketching technique as in VIERECK) or representing an object and its movement (substitutive technique as in BEAK see figure 5 above). For a description of image producing techniques in DGS cf. Langer (2005) and König, Konrad & Langer (2008: 388-389).



carrying and utilizing the semantic load of the mouthed word. This both supports and at the same time complicates the process of determining the various senses of a sign.

Mouthings can help to disambiguate polysemous signs. Often the mouthing is an important factor to determine the meaning of a polysemous sign in a specific context. It is not unusual to have one sign (for example BEAK see figure 5) to cover a basic category concept (such as “bird”) and a set of subordinate concepts (such as “sparrow”, “blackbird”, “pigeon”, “chicken” and a number of other birds with small narrow beaks) in everyday non-expert signing. One sign form can also cover complementary concepts such as “brother” and “sister” or antonymic concepts such as “dry” and “moist”. The occurrences of such polysemous signs are usually accompanied by mouthings that have a disambiguating function and correspond to the meaning.

When a sign is conventionally associated with a particular mouthed word not all of the word’s meanings automatically are also conventional meanings of that particular sign. Some of a word’s meanings might not match the iconic value of the sign. However, in a specific context the sign form might just be used to contextualize that word, occasionally even for a meaning that is not lexicalized for the sign. Also, a polysemous German word might contribute some of its other or more marginal senses into sign utterances by way of mouthing and finally extend the sign’s meaning in a way that might not be expected or not be in accordance with the sign’s iconic core but parallels the meanings of the German word. One example for this is the sign CRAB. The underlying image of the sign shows the opening and closing of a pair of crab pincers. The German name for this animal is *Krebs*. But *Krebs* has also the meanings “cancer (disease)” and “Cancer (zodiac sign)”. This is paralleled in DGS in that the DGS sign CRAB is also used in all three senses of the German word.

The interplay of a sign’s iconic potential and its productive exploitation, the sign’s lexicalized (conventional) meanings and the meanings of associated mouthed words make it difficult to determine the degree of conventionalization for a particular meaning. One major issue here is that a frequently used mouthing has the potential to bring additional meanings from another language (German) into the equation. This meaning potential of the word has to be considered especially when dealing with rather isolated signs as listed in sign collections and dictionaries that use spoken language equivalents to indicate the sign’s meaning(s). Analyses of usage data from a large corpus can provide evidence for which senses are common usage and which are to be considered marginal or even occasional or creative occurrences. Until corpus data analyses of this kind will be possible we resort to the data as presented in the sign collections, documented in our internal database from previous projects and signer’s intuition to identify and disambiguate possible senses to be validated in the feedback process.



## 4 Feedback from the Signing Community

In order to verify and complement our data, e.g. for regional variation and passive vocabulary, we rely on feedback from the signing community. This holds especially for the signs to be included in the Basic Vocabulary.

### 4.1 Preparation of Content

Each potential combination of sign form and meaning identified in the process described above has to be verified by the community. The crucial point is how to convey the intended meanings. Ideally context information in DGS would be given for each instance, however such an approach is cumbersome and not feasible within the project. Furthermore, in several cases senses seem clear enough from a short disambiguating written context such as “mouse (computer)” vs. “mouse (animal)”. In cases like this we can spare the participants the effort of having to watch an additional film. For the feedback we chose a rather pragmatic approach: In order to make it as easy for the participants as possible, we present German equivalents of the signs, often along with German disambiguating contexts in brackets, and add signed contexts wherever we feel it is needed. German contexts can be a superordinate term, a domain, a relevant dimension (e.g. space, time), further disambiguating synonyms, a short explanation, collocations or other suitable context information. For DGS contexts we mainly use examples, but we also try short explanations, collocations and sign synonyms. For this we rely largely on the linguistic knowledge and intuition of a Deaf colleague. To a smaller extent, as far as it has been already made accessible through transcription, we also use data from our corpus as basis for the construction of examples.

### 4.2 The Feedback System

For the realisation of the feedback we decided for a crowd sourcing approach using an online feedback platform. Our aim is to get as many members of the signing community involved as possible, i.e. not only Deaf people but everybody using DGS.<sup>18</sup> Dealing with a relatively small language community spread across the country, the crucial point is not only how to attract enough participants, but also how to make them check back regularly. The newly developed online platform, running on desktop computers as well as mobile devices, allows for as much freedom for the participants as possible.<sup>19</sup> While the target community’s pride of their own language and interest in supporting the project can be expected to be the main motivation for taking part in the feedback, computer game elements such

---

18 Metadata information will be used to weigh the participants’ answers depending on their proficiency in DGS.

19 A more detailed description of the Feedback System in DGS and German can be found at <http://feedback.dgs-korpus.de>.

as levels and high score are used as extrinsic motivation. An important point for developing the system was to make it accessible by presenting all information in DGS and also to allow for answers and comments to be given in DGS (video upload via webcam). The system is optimised for our use, but the software is open source and can be adjusted for other purposes.

Participants create individual accounts where their answers and metadata are stored. Within the feedback system tasks are organised in categories and work packages. Due to the complexity of the tasks, users are asked to deal with only one type of task at a time. Categories therefore contain work packages of generally one type of task, and a new category is only released after a certain amount of work packages of the previous category have been completed. A so-called “golden trail” leads through the work packages. A help page contains explanations dealing with different aspects of the feedback system and with problems and questions that might arise. A “comment”-button on each page allows the participants to leave comments (in DGS or written German) that provide us with extra information or help us to spot problems regarding the tasks and the system.

### 4.3 Presentation of Content in the Feedback System

Work packages are organised in pages and rows. Task explanations and questions are all presented in DGS and written German (a button allows to switch between both languages), other content is provided as film, text or pictures depending on the purpose.



Figure 6: Example page of a work package “Form and Meaning”<sup>20</sup>.

20 In this example the sign is made by two hands with “V” handshape moving up and apart. The meanings listed for verification are: “famous”, “celebrity”, and “public” (rows are moving upwards when answer button has been clicked).

Our first question type deals with sign forms and meanings as described above. Each sign form is first shown without mouth pattern in order to have the participants concentrate on the hands.<sup>21</sup> Participants are asked whether they know the sign form presented. Clicking “no” leads directly to the next page with a different sign form. When pressing “yes” further questions are revealed concerning the meanings of the respective sign form. In each row one meaning of the sign is presented by a film clip of the sign with a common mouth pattern and one or more German equivalents (often including sense indicating contexts in brackets). If needed an additional film clip provides context information in DGS also. Participants are asked whether they use the sign with the respective meaning themselves, don’t use it but have others seen using it, or don’t know this sign with this meaning at all. At the end of each page the system allows to add further meanings in written German or DGS (via webcam).

The completion time for each work package is planned to be approximately 20 minutes. On average a work package contains about eight different sign forms depending on the number of meanings to be verified (so far up to a maximum of 12 meanings with one sign form). When a work package is submitted the participant receives the number of points as defined for this package (i.e. independent of the number of “positive” or “negative” answers given) and is listed in the high score table. While this game-like approach primarily aims at motivating participants and boosting competition between them it also serves as a tool to stir which tasks are assigned to a participant (i.e. a certain number of points is needed to access the next category with different types of questions).

## 5 Conclusion

While the corpus collected within the project is still in the process of transcription and annotation we rely on other sources for our lexicographic work, especially for the production of our Basic Vocabulary. We have compared the signs of nine other previously produced sources (sign collections, learning materials) and based our lemma selection on the highest overlap in these products. The equivalents listed are taken as a starting point for a process of identifying different senses of the signs in a process of reversal of mono-directional bilingual information. In this process the iconicity of signs is also considered. The given meanings of signs are split into a finer and more sufficient granularity of senses, disambiguated through suitable contexts and presented via an online feedback system to the language community for verification or rejection.

The described procedure is not ideal to identify word senses of signs. It does things a little backwards, starting with the meanings of the German words given as equivalents instead of looking at signs and their contexts of actual use. However, this procedure takes the already published material containing

---

21 Considering mouthings as an essential part of signing raises the issue of how to present sign forms in the feedback as signers should not rate sign-mouthing but sign-meaning combinations. Our solution so far is to omit any mouth pattern for the more general question on sign form and show a separate film of the sign with the presumable most commonly used mouth pattern for each meaning. An accompanying explanation is included in the tasks explanation and pre-test have revealed no major problems.

information on signs' meanings seriously. Also, involving the language community in the feedback process has some advantages: While the corpus contains signing from only 330 informants we expect that through the feedback we can obtain data from even considerably more signers. This way we can obtain a more detailed account on regional variation concerning signs and sign senses. Direct questions also allow us to elicit data on passive as well as active vocabulary.

Nevertheless the result of such a feedback process can only be a first incomplete and approximated account of a sign's senses. This is the reason why we regard the Basic Vocabulary as preliminary. Once the corpus is available it will be the basis for further lexicographic work. Results from the feedback can be confirmed and complemented by corpus data, while the feedback process (crowd sourcing) may prove a valuable complementary method to obtain data for signs and questions the corpus does not cover.

## 6 References

- Atkins, B.T.S, Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. New York: Oxford Univ. Pr.
- Bentele, S., Konrad, R. & Langer, G. (2000). Transkription und Analyse. In R. Konrad, T. Hanke, A. Schwarz, S. Prillwitz & S. Bentele. *Fachgebärdenlexikon Hauswirtschaft*. Vol. 2. Hamburg: Signum, pp. 617-632.
- Brien, D. (ed.) (1992). *Dictionary of British Sign Language/English*. London: Faber and Faber.
- Ebbinghaus, H., Heßmann, J. (1996). Signs and words. Accounting for spoken language elements in German Sign Language. In W. H. Edmondson, R.B. Wilbur (eds.) *International Review of Sign Linguistics* 1. Mahwah, N.J.: Erlbaum, pp. 23-56.
- Ebbinghaus, H., Heßmann, J. (2001). Sign language as multidimensional communication: Why manual signs, mouthings, and mouth gestures are three different things. In P. Boyes Braem, R. Sutton-Spence (eds.) *The hands are the head of the mouth: The mouth as articulator in sign language*. Hamburg: Signum, pp. 133-151.
- Glaboniat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile deutsch. Gemeinsamer europäischer Referenzrahmen*. Berlin, München: Langenscheidt. (Includes CD-ROM).
- Hanke, T. (2004). HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts. In O. Streiter, C. Vettori (eds). *LREC 2004, Workshop proceedings: Representation and processing of sign languages*. Paris: ELRA, 2004, pp. 1-6.
- Hanke, T., Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd (eds.) *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. Paris: ELRA, 2008, pp. 64-67.
- Johnston, T. (1989). *AUSLAN Dictionary. A dictionary of the sign language of the Australian deaf community*. Victoria: Aust. Print Group.
- Jones, R.L., Tschirner, E. (2006). *Frequency Dictionary of German. Core Vocabulary for Learners*. London: Routledge.
- Jones, R.L., Tschirner, E. (2011). *Frequency Dictionary of German. Core Vocabulary for Learners*. (CD-ROM). London: Routledge.
- Kennedy, G. (ed.) (2008). *A Concise Dictionary of New Zealand Sign Language*. Reprint. Wellington: Bridget Williams Books.

- Kestner, K. (2002): Tommys Gebärdenwelt. Das Gebärdensprachbuch zur CD-ROM. Guxhagen: Manual Audio Devices.
- König, S., Konrad, R., Langer, G. (2008). What's in a sign? Theoretical lessons from practical sign language lexicography. In J. Quer (ed.) *Signs of the Time. Selected Papers from TISLR 2004*. Hamburg: Signum, pp. 379-404.
- König, S., Schmaling, C. (2012). Gebärdenschriften: Flüchtlings fixieren. In H. Eichmann, M. Hansen, J. Heßmann (eds.) *Handbuch Deutsche Gebärdensprache. Sprachwissenschaftliche und anwendungsbezogene Perspektiven*. Hamburg: Signum, pp. 341-356.
- Kristoffersen, J.H., Troelsgård, T. (2010). Compiling a sign language dictionary. Some of the problems faced by the sign language lexicographer. In M. Mertzani (ed.) *Sign Language. Teaching and Learning. Papers from the 1<sup>st</sup> Symposium in Applied Sign Linguistics, Centre for Deaf Studies, University of Bristol, 24-26 September 2009*. Bristol: Centre for Deaf Studies, pp. 1-10.
- Kristoffersen, J.H., Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In S. Granger, M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford Univ. Pr., pp. 293-315.
- Langer, G. (2005). Bilderzeugungstechniken in der Deutschen Gebärdensprache. In *Das Zeichen* 70, pp. 254-270.
- Martin, W. (2013). Reversal of Bilingual Dictionaries. In R.H. Gouws, U. Heid, W. Schweickard & H.E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*. (Vol. 5,4 of *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)*). Berlin: De Gruyter, pp. 1445-1455.
- Nishio, R., Hong, S., König, S., Konrad, R., Langer, G., Hanke, T., Rathmann, C. (2010). Elicitation methods in the DGS (German Sign Language) Corpus Project. In P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, G. Martínez Ruiz, A. Schembri (eds.) *Corpora and Sign Language Technologies. 4th Workshop on the Representation and Processing of Sign Languages*. Paris: ELRA, 2010, pp. 178-185.
- Ordbog over Dansk Tegnsprog*. Accessed at: <http://www.tegnsprog.dk/> [20/03/2014].
- Stokoe, W.C. (1993). Dictionary Making, Then and Now. In *Sign Language Studies*, 79 (22), pp. 127-146.
- Stokoe, W.C., Casterline, D.C. & Croneberg, C.G. (1965). *A dictionary of American Sign Language on linguistic principles*. Washington, DC: Gallaudet College Press.
- Sutton, V.J., Paul, F.A., Candelaria, I., Gunderson, J. (2009). *SignWriting Basics Instruction Manual*. SignWriting Pr.
- Zwitzerlood, I. (2010). Sign Language Lexicography in the Early 21<sup>st</sup> Century and a Recently Published Dictionary of Sign Language of the Netherlands. In *International Journal of Lexicography*, 23 (4), pp. 443-276.

## Sign Collections for DGS

Sign collections for the Basic Vocabulary. Related or complementary works are grouped together.

- [1] *777 Gebärden 1-3. Alle 3 Folgen auf einer DVD*. (2002). Version 2.0. Guxhagen: Manual Audio Devices. (DVD-ROM).
- [2a] *DGS-Aufbau-Lexikon*. (1998). Aachen: Microbooks/Desire. (CD-ROM).
- [2b] *DGS-Basis-Lexikon*. (1998). Aachen: Microbooks/Desire. (CD-ROM).
- [2c] *DGS-Phrasensammlung*. (1998). Aachen: Microbooks/Desire. (CD-ROM).
- [3a] Maisch, G., Wisch, F.-H. (1987). *Gebärdenlexikon. Band 1. Grundgebärden*. Hamburg: Verlag hörgeschädigte kinder.
- [3b] Maisch, G., Wisch, F.-H. (1988). *Gebärdenlexikon. Band 2. Mensch*. Hamburg: Verlag hörgeschädigte kinder.
- [3c] *Grundgebärden 1. Für Einsteiger*. (1999). Hamburg: Verlag hörgeschädigte kinder. (CD-ROM).
- [3d] *Grundgebärden 2*. (2000). Hamburg: Verlag hörgeschädigte kinder. (CD-ROM).

- [4a] Metzger, C., Schulmeister, R. & Zienert, H. (2000). *Die Firma. Deutsche Gebärdensprache do it yourself*. Hamburg: Signum. (CD-ROM).
- [4b] Metzger, C., Schulmeister, R. & Zienert, H. (2003). *Die Firma 2. Deutsche Gebärdensprache interaktiv. Aufbaukurs in Deutscher Gebärdensprache - Schwerpunkt Raumnutzung*. Hamburg: Signum. (CD-ROM).
- [5a] Arbeitsgruppe Fachgebärdenlexika (ed.). (1994). *Fachgebärdenlexikon Computer*. 2 Vol. Hamburg: Signum.
- [5b] Arbeitsgruppe Fachgebärdenlexika (ed.). (1996). *Fachgebärdenlexikon Psychologie*. 2 Vol. Hamburg: Signum. Online-Version can be accessed at: <http://www.sign-lang.uni-hamburg.de/plex/> [20/03/2014].
- [5c] Konrad, R., Hanke, T., Schwarz, A., Prillwitz, S. & Bentele, S. (2000). *Fachgebärdenlexikon Hauswirtschaft*. 2 Vol. Hamburg: Signum. Online-Version can be accessed at: <http://www.sign-lang.uni-hamburg.de/hlex/> [20/03/2014].
- [5d] Konrad, R., Schwarz, A., König, S., Langer, G., Hanke, T. & Prillwitz, S. (2003). *Fachgebärdenlexikon Sozialarbeit/Sozialpädagogik*. Hamburg: Signum. Online-Version can be accessed at: <http://www.sign-lang.uni-hamburg.de/slex/> [20/03/2014].
- [5e] Konrad, R., Langer, G., König, S., Schwarz, A., Hanke, T. & Prillwitz, S. (eds.) (2007). *Fachgebärdenlexikon Gesundheit und Pflege*. 2 Vol. Seedorf: Signum. Online-Version can be accessed at: <http://www.sign-lang.uni-hamburg.de/glex/> [20/03/2014].
- [5f] Konrad, R., Langer, G., König, S., Hanke, T. & Rathmann, C. (eds.). (2010). *Fachgebärdenlexikon Gärtnerei und Landschaftsbau*. 2 Vol. Seedorf: Signum. Online-Version can be accessed at: <http://www.sign-lang.uni-hamburg.de/galex/> [20/03/2014].
- [6] Kestner, K., Hollmann, T. (2009). *Das große Wörterbuch der deutschen Gebärdensprache. Der erste umfassende Gebrauchswortschatz in DGS als elektronisches Wörterbuch. 18000 Begriffe in Schrift und Video von A wie Aachen bis Z wie Zypresse. Deutsch – DGS, DGS – Deutsch. Bundeselternverband Gehörloser Kinder (ed.). Guxhagen: Kestner. [DVD-ROM with booklet.]*
- [7] Arbeitsgruppe ProViL. (2006). *DGS 3 und DGS 4. Aufbaukurse in Deutscher Gebärdensprache*. [Located at the eLearning platform WebCT/Blackboard; for members of the University of Hamburg accessible via OLAT - Online Learning And Training. Accessed at: [www.olat.uni-hamburg.de](http://www.olat.uni-hamburg.de) [20/03/2014].
- [8a] *Tommys Gebärdenswelt*. Version 3.0. (2007). Guxhagen: Kestner. CD-ROM.
- [8b] *Tommys Gebärdenswelt*. 2. Tommy und Tina. Version 3.0. (2008). Guxhagen: Kestner. CD-ROM.
- [8c] *Tommys Gebärdenswelt*. 3. Mit Tommys erstem Lexikon. Version 3.0. (2009). Guxhagen: Kestner. DVD-ROM.
- [9a] Keller, J., Zienert, H. (2000). *Grundkurs Deutsche Gebärdensprache. Stufe I. Vokabel CD-ROM*. Hamburg: Signum. (CD-ROM).
- [9b] Keller, J., Zienert, H. (2002). *Grundkurs Deutsche Gebärdensprache. Stufe II. Vokabel-Video*. Hamburg: Signum. (VHS-Video).

## Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.



# Dearcadh na nDéise – Representations of Gaeltacht na nDéise in Dineen’s Bilingual Irish-English Dictionary (1927)

Chris Mulhall, Seamus ó’Diollúin  
Waterford Institute of Technology  
cmulhall@wit.ie, sodiolluin@wit.ie

## Abstract

Dictionaries are a written manifestation of the world, but the selection of words is a reflection of society at a particular time from one or more viewpoints. Irish dictionaries are largely uncharted territories in terms of lexicographic research and contain important regional and national narratives that reflect the ever-changing Irish cultural landscape in the late nineteenth and early twentieth century. The creation of the independent Irish state in the early part of the twentieth century and increasing demand for bilingual Irish-English dictionaries at this time heralded a beginning of a new area of activity in European lexicography. One of landmark editions from this period is *An Foclóir Gaedhilge agus Béarla* by Patrick S. Dineen. The aim of this paper is to explore a regional variety of Irish, Gaeltacht na nDéise, found in the South-East of Ireland and examine some of the socio-cultural narratives behind Dineen’s choice of entries to depict this regional dialect. Based on a recategorisation of the empirical work of Nyhan (2006), a brief discussion of three prevalent socio-cultural themes, namely; anger, anxiety and poverty, recurring in some Gaeltacht na nDéise-specific entries found in Dineen’s work is presented. These offer an interesting window into how the lexis of this particular region was represented and potentially contributed to its regional identity in a national context.

**Keywords:** Bilingual Irish-English Dictionaries; Socio-cultural narrative; Regional Dialect

## 1 An Foclóir Gaedhilge agus Béarla and Gaeltacht na nDéise

*An Foclóir Gaedhilge agus Béarla* by Reverend Patrick S. Dineen stands as the most iconic works in the comparatively short history of bilingual Irish-English lexicography. The early editions of this pioneering work came at a time of upheaval as Dineen sought to encapsulate the regional and national linguistic identities of the newly-liberated Irish State. A substantive part of Dineen’s narrative of the Irish language engages in what Hartmann (2001) refers to as a ‘personal-biographical’ account of the

language<sup>1</sup>, but this was tempered with a widely representative account of regional dialects and a conservative acceptance of borrowings. In the preface, Dineen stresses his efforts to flavour his dictionary with as many dialectal words as possible, most notably from Connaught and Aran, North and Mid-Ulster, West and North Kerry, South and South-West Cork, Meath, Omeath, Clare and West Limerick and the Comeragh District of Waterford. The regional specific lexis offers an interesting panorama into the linguistic sub-cultures in Ireland, but is also productive in highlighting particular values, identities and struggles pertinent to those areas at this point in Irish history. In the preface of *An Foclóir Gaedhilge agus Béarla* Dineen lists Fr. Mícheál McGrath and Riobaird Bheldon (1838-1914) as one of his lexicographical sources<sup>2</sup>. While not mentioned as a source in the first edition (1904) of the *Foclóir Gaedhilge agus Béarla* the 1927 edition contain 96 entries from this ‘Cm.’ source. The absence in the 1904 edition may be explained by the fact that Bheldon’s poetry was only first published in December, 1903 - just months before the first edition of the dictionary was published. Dineen himself admits that the 1927 edition was “practically a new work” and “an effort has been made to secure much representative provincial Irish in word and phrase as possible” (Dineen, 1927: vii, xi). Bheldon and McGrath as immersed as they were in the language, literature and culture of the Déise Gaeltacht region of Waterford, and its surrounding hinterland, proved suitable sources of the living Irish language in south east Munster. The words in *An Foclóir Gaedhilge agus Béarla* denoted by “Cm.” are examples of the living language as spoken in rural county Waterford at the turn of the twentieth century. While the actual provenance of this collection of words might not be certain we are left in no doubt as to Dineen’s reasons for including them in his *Foclóir Gaedhilge agus Béarla*:

an effort has been made to net the chief living elements of the language while there was still time. The materials for the work have been drawn from the living language of Irish-speaking Ireland as well as from the written remains of the modern literature (Dineen, 1927: vii)

## 2 The Work of Riobaird Bheldon

Bheldon was a relatively little known poet until the publication of *Riobaird Bheldon, Amhráin agus Dánta* (1995) by Pádraig Ó Macháin<sup>3</sup>. Ó Macháin (1995:125) mentions that Bheldon was an acquaintance of Fr.

- 
- 1 The first edition was, to a considerable extent, compiled from memory. A large percentage of the illustrative phrases were taken from living expressions, conversations, etc, stored up in my childhood’s memory; so that many of them have the vividness and directness characteristic of the spoken word. In this edition still further use has been made of this source; and the meanings and applications of key words, somatic terms and other important expressions have been considerably expanded (*An Foclóir Gaedhilge agus Béarla* [Second Edition], 1927:vii).
  - 2 Cm.- a list of words from the Comeragh district of Waterford compiled by the late Fr. McGrath (P.P. of Ring) and the late Riobárd Ueldon, the poet (*An Foclóir Gaedhilge agus Béarla* [Second Edition], 1927:xxiv).
  - 3 Pádraig Ó Dálaigh published *Riobaird Bheldon: File an Chomaraigh* in 1925. Bheldon was also the subject of (or merely mentioned) in a small number of articles in newspapers and journals.



Michael McGrath and that, “le cabhair Riobaird dhein an sagart cnuasach de chaint an Chomaraigh a sholáthar don Athair Pádraig Ó Duinnín” (With the help of Riobard the priest, i.e. Fr. McGrath, supplied a collection of words from the spoken Irish of the Comeragh region to Fr. Dineen). Fr. McGrath was an Irish-speaking priest in west county Waterford at the time. It has been suggested that Bheldon may not have been able to read or write in the Irish language (Ó Macháin, 1995: 8-9). One could infer, then, that the task of integrating Bheldon’s input to Dineen’s dictionary may have fell on McGrath. While it cannot be determined with any level of certainty what was the actual origin of the list one can surmise that it not only contained a collection of words and phrases from the spoken language of the region, but also examples from Bheldon’s poetry. It is interesting to note, however, that in his *Sean-chaint na nDéise*, Bishop Michael Sheehan seems to suggest that the list was primarily of McGrath’s making:

A valuable list of Déise words and phrases was supplied to Dr. Dineen by Fr. Michael McGrath which he took down while curate in Kilrossanty (Comeragh) parish. They are indicated in the dictionary by “Cm.” Fr. McGrath later became P.P. of Ring, and died in 1919. Many of his notes, not used by Dr Dineen, are in my possession, and may appear, I hope, in a future edition of this book (Sheehan, 1944: 216).

### 3 A Recategorisation of Gaeltacht na nDéise Entries in Dineen’s Dictionary

The following sections present a narrative account of socio-cultural aspects of the Comeragh dialect contain in Dineen’s *Foclóir Gaedhilge agus Béarla*. This analysis is founded upon the work of Nyhan (2006), who identifies 355 entries in Dineen’s work that have provenance in this particular region. A small number of these entries have been selected, which depict emotional states (anger and anxiety) and social class (poverty).

#### 3.1 Anger

Dineen’s 1927 dictionary includes 15 entries from the Comeragh specific dialect that have connotations of acts of anger, dispute or violence (see Table 1). There are no evident regional factors explaining the choice of these particular words, apart from the backdrop of a national unrest in Ireland against foreign occupation. However, a closer inspection of Dineen’s entries in this particular lexical set communicate certain feelings of regional tension, for example, *reaping with a swing round killed the Munsterman* found in the entry **baic** or in the agricultural endeavours strongly associated with the region, *the dog made a drive at me* in **ablach**. Most interesting is the finality or permanency associated with the acts of violence that are narrated through certain listings, such as *I will give you a lasting wound* (**cio-na-sheicean**), *He injured him for life/He is permanently marked or injured* (**faic**) and *you have finished him*

with that blow (**mart**). The explicit description of physical aggression, particularly referencing males, appears to be a strong theme in anger-specific lexis and could be connected to the powerful autocratic role of the Catholic Church in Ireland during this period, particularly given the religious background of Dineen and his named contributors.

Headword	Source Language Entry	Target Language Translation
<i>Baic</i>	<i>buaint tar b. do mhairbh an Muimhneach</i>	Reaping around with the sword killed the Munsterman
<i>Ablach</i> [2]	<i>Thug an mada a. orm</i>	The dog made a 'drive' at me
<i>Ciona-sheicean</i>	<i>In phr. cuirfead c. ort</i>	I will give you a lasting wound
<i>Cofach</i>	<i>c. chum troda atá ort</i>	You are spoiling for a fight
<i>Donagar</i>	<i>A thuilleadh donagair chughat</i>	May more misfortune be thine
<i>Failc</i>	<i>Chuir se f. ann Ta. f. ann</i>	He injured him for life He is permanently marked or injured
<i>Faimin</i>	----	A blow
<i>Fuarpadh</i>	---	One in a rigid or unconscious state after a blow
<i>Giolcadh</i>	<i>Fuair sé g. maith ón máighistir</i>	The teacher gave him a good beating
<i>Liath-shúil</i>	<i>Thug sé l.orm</i>	He eyed me bitterly
<i>Máiglid</i>	<i>Ag m. Le n-a chéile</i>	Act of wrangling, disputing
<i>Mart</i>	<i>Tá sé 'na mb. agat leis an mbuille sin</i>	You have finished him with that blow
<i>Pliastrail</i>	<i>Ag p. ar fuaid an tigh</i>	Knocking things about the house
<i>Rámhghail</i>	----	Ranting, raving, medley
<i>Sméideadh</i>	<i>ní leomhfá s. air</i>	You dare not wink at him

**Table 1: Entries denoting Anger, Dispute or Violence.**

Fractured personal relationships are also communicated through Dineen's content with some of these relating to formal social structures, such as school, *the teacher gave him a good beating* (**giolcadh**) or towards individuals in the wider regional community, for example, *he eyed me bitterly* (**liath-shúil**) or *you dare not wink at him* (**sméideadh**). Collectively, these entries express a certain malaise experienced by Dineen either in childhood or on the behalf of those who contributed examples of the Comeragh dialect to his dictionary.

### 3.2 Anxiety

Another subset of Comeragh dialect words recorded by Dineen convey a measurable sense of anxiety (see Table 2), which appears to be connected to personal, relating to geographical surroundings or associated with a recent personal loss. A particularly strong emotive theme is the portrayal of being un-

settled in a particular place, which makes possible reference to the region. This can be found in the entry **connuighim**, which includes the listing translating as *he was sad enough until he became familiar with the place*.

Headword	Source Language Entry	Target Language Translation
<b>Cásnach</b>	----	Full of concern
<b>Connuighim</b>	<i>Bhí sé diachrach go leor gur chonnuigh sé léis an áit</i>	He was sad enough until he became familiar with the place
<b>Deighreán</b>	<i>Ag déanamh deighreain dó</i>	Giving him anxiety
<b>Diadhánach</b>	----	Lonesome, as a cow bereft of her calf
<b>Diúdaireacht</b>	<i>chuirfeadh sé d. ar mo chroidhe</i>	It would rejoice my heart
<b>Dorn</b>	<i>dhein an ghaoth d. dubh orn</i>	The wind did me a bad turn
<b>Fógla</b>	<i>f. chum imthighthe</i>	Anxiety to depart
<b>Iolchaing</b>	<i>In phr. Tàim ar i. chum</i>	I am anxious to get at or be at
<b>Ionnas</b>	----	Expectation
<b>Sonntaighe</b>	<i>thuig mé go raibh mo dheirbhshiúr atá curtha le bliadhain lem ais agus tháinig s. orm</i>	I understood my sister, a year buried, was beside me and I became unnerved
<b>Stiugaighil</b>	<i>Ag s. les an mbás</i>	In the throes of death

Table 2: Entries denoting Anxiety.

### 3.3 Poverty

Certain entries chosen by Dineen in his account of the Comeragh dialect make a discernible reference to poverty and lower social class as being associated with its regional identity (see Table 3). These typically centre on the quality of food available and the culinary expectations of Comeragh people. In the case of the former, the entry **gannaire** clearly communicates the state of the social class stratum of this region by their choice of food. Another example of a clear statement of poverty can be found in the entry **beaindin** containing the citation *the tap of the cream vessel through which impoverish milk is withdrawn*.

Headword	Source Language Entry	Target Language Translation
<b>Beaindín</b>	----	The tap of a cream vessel, etc, through which the impoverished milk is withdrawn
<b>Cearbh [1]</b>	<i>Cuirim c. mo shúl ann</i>	I regard it with a covetous eye
<b>Cuthaigh</b>	<i>biadh c.</i>	Stimulative food
<b>Forthain</b>	<i>cead f. den bhiadh a dh'ithe</i>	Permission to eat enough of the food
<b>Gannaire</b>	<i>iosfaimid an g.</i>	We will eat the poor food
<b>Lóta</b>	<i>lótaí na phócaí</i>	Scraps in his pockets
<b>Molaim-mo-lámhadas</b>	<i>Is dóigh le muinntir an ch. gur m. arán agus gruth</i>	The Comeragh people think that bread and curds make the finest of food

**Table 3: Entries denoting Poverty.**

## 4 Conclusion

The socio-cultural narrative of the Comeragh region in Dineen's *Foclóir Gaedhilge-Béarla* (1927) depicts a regional existence characterised by feelings of anger, anxiety and an acute sense of poverty. Although the thematic areas under analysis represent only a small proportion of Dineen's account of this particular Irish dialect, they allow an exploratory insight into the lexicographical portrayal of this geographical area, contextualising its emotional disposition and social stratum. Together, these contributed to shaping its regional identity of this early twentieth century period within a larger national context.

## 5 References

### 5.1 Dictionaries

Dineen, P.S. (1927) *Foclóir Gaedhilge agus Béarla*. Second Edition. Dublin: The Educational Company of Ireland Ltd.

### 5.2 Other Publications

Hartmann, R. R. K. (2001). *Teaching and Researching Lexicography*. London: Pearson Education.

Nyhan, J. (2006). *Findfhocla an Chomaraigh: Cnúsach Riobaird Bheldon*. An Linn Bhuí (10). Kilkenny: Leabhair na Linne.

Ó'Macháin, P. (1995). *Riobard Bheldon: amhráin agus dánta*. Dublin: Puddle Press.

# The *eLexicon Mediae et Infimae Latinitatis Polonorum*. The Electronic Dictionary of Polish Medieval Latin

Krzysztof Nowak  
Institute of the Polish Language, Polish Academy of Sciences  
krzysztof.n@ijp-pan.krakow.pl

## Abstract

The paper presents goals, methods, and results of the project of the Electronic Dictionary of Polish Medieval Latin. First, a brief history of the paper dictionary, as well as an account of its main features are presented. Second, the main problems of the metalexigraphic analysis and the subsequent XML encoding of the lexicographic content are discussed. The main purpose of both being a fine-grained description of linguistic resource, it was necessary to make explicit a fair amount of data which are coded only by means of convention. Third, the web interface of the dictionary is treated in more detail. Its most important of its design principles include separation of the expert and novice user perspective, system of aids and suggestions, integration with external sources.

**Keywords:** electronic lexicography; Medieval Latin; dictionary interface; TEI XML encoding; implicit information; lexicographic convention

## 1 Introduction

The *eLexicon Mediae et Infimae Latinitatis Polonorum* (henceforth referred as the *eLexicon*) is an electronic dictionary based on the first 7 volumes<sup>1</sup> of the paper *Lexicon Mediae et Infimae Polonorum* (henceforth the *Lexicon*) which has been published since 1953 under the auspices and with the financial support of the Polish Academy of Sciences (Plezia, Weyssenhoff-Brożkova, Rzepiela 1953). The *Lexicon* was conceived by its first editor, the Polish eminent philologist Prof. Marian Plezia, as a work which would fully document the use of the Latin language on the Polish territory between the X<sup>th</sup> and the mid-XVI<sup>th</sup> century (Plezia 1958). As such, it was meant to form a part of the European network of the national dictionaries of Medieval Latin which started to emerge at the same time in response to an appeal of the *Union Académique Internationale* (Bautier 1981: 433–436). Users to which the print *Lexicon* has been addressed are in particular members of a research community, which is the reason why so much emphasis has been put, among others, on the completeness of the source material included. The print dictionary provides, then, in-depth etymological, morphosyntactic and semantic description of each

---

1 They include entries from A to Q, which is ca. 6000 pages printed in two columns.

word attested in the Polish Latin during the Middle Ages. Sense definitions are formulated both in Polish and – with foreign readers in mind – in Latin, and are illustrated with appropriate source quotations, if the meaning was not known in the Antiquity. The audience of the *Lexicon* being scholarly community, it becomes partially clear why the *Lexicon* does not make much concessions as far as user friendliness is concerned, with its heavy use of the typographic conventions, tightly printed columns etc. The paper dictionary suffers also from the drawbacks symptomatic for every long-term academic publishing enterprise, and in particular from several inconsistencies of the editorial policy, which affect especially usage labelling system, semantic change description or sense nesting practice, to name only few.

The project of the electronic dictionary which would be based on the *Lexicon* was conceived by the author of the paper and has been carried out between mid-2010 and mid-2014 by the team of the Department of the Medieval Latin of the Institute of the Polish Language (Polish Academy of Sciences) in Kraków.<sup>2</sup> Regardless of its roots, from the beginning the *eLexicon* was expected to become a research tool on its own and not merely a digitized version of the paper work. Firstly, its content was to differ to various extent from what can be found in the print volumes. One source of substantial modifications was the incorporation into the main text of the *addenda et corrigenda*, ‘supplements and corrections’, printed at the end of each of the 7 volumes of paper dictionary. Another one was both manual and automatic update of the lexicographic content. The members of the team (and, in the same time, current authors of the paper dictionary) had to eliminate most obvious errors and, where only it was necessary, to adjust the text to the modern editorial rules.

Secondly, the *eLexicon* was expected to provide research community with capabilities that the print dictionary could not offer. Apart from the simple search and browse features, the on-line dictionary was meant to offer access to the wealth of information encapsulated either explicitly or implicitly in the dictionary entries. At the same time, the *eLexicon* was conceived as a constituent of a larger text analysis framework. It had not only to be integrated with the digital library of the scanned images of paper slips, but also to be actively linked to the bibliography list of the medieval sources and to constitute a *sui generis* wrapper around the Medieval Latin corpus.<sup>3</sup> Moreover, the on-line dictionary was to incorporate a fair amount of external resources, whether it be through locally triggered queries or by means of outward linking.

Finally, the *eLexicon* has been planned as an open-access and open-source project. From the beginning the access to the web service was meant to be free and unlimited, as was also the case of the XML annotated dictionary files, which are to be distributed under liberal licenses. Although there were many reasons to do so, the main of them was assuring the longevity of the project, a major challenge in academic projects with time-limited funding. The other factor expected to contribute to project’s longev-

---

2 Its funding was provided by a grant of the Polish National Science Centre awarded to the chief-editor of the paper dictionary, Prof. Michał Rzepiela.

3 A 5 million words, balanced and representative corpus of the Polish Medieval Latin is now being developed by the same team and is due to be delivered by the end of the 2016.

ity was compliance with standards. Developed firstly as a set of the TEI-conformant files (TEI Consortium 2013), the electronic dictionary allows platform-independent implementations, the fact which implies two major consequences. On the one hand, one can benefit from the open-source technologies and existing text or data retrieval frameworks. On the other, one may hope that the available lexicographic data will be incorporated in other research contexts, integrated with NLP infrastructures, and, consequently, they will be steadily ameliorated and refined, even when the project itself comes to an end.

## 2 Methods

No phase of the e-dictionary creating was outsourced. After the volumes of the print dictionary had been scanned, the image pre-processing and OCR process began as a result of which machine-readable text was, firstly, obtained and, then, carefully proofread. After that metalexigraphic analysis followed, its aim being twofold. First of all, it was expected to reveal the features of the print dictionary macro- and micro-structure to be retained in the *eLexicon*, but also to select lexicographic information worth retrieving by means of the on-line search interface. Contrary to what one might believe, first part of the analysis was far from trivial, since it was often equivalent to questioning the very foundation of the paper dictionary methodology and, at the same time, to designing principles of the future on-line dictionary. The main issues addressed included internal reference system, approach to the entries with deeply nested structure, status of idioms and multi-word expressions as lexical units etc. In what concerns lexicographic data, the guiding principle was to retrieve and make explicit as much linguistic and non-linguistic information as possible, since from the very beginning it was clear that the on-line dictionary should serve researchers of various expertise in medieval studies, from the historians working on the Medieval Latin sources, to the Latin and Polish linguists, to the historians of literature, art, philosophy and science. What is more, one of the goals of the *eLexicon* was also expanding the audience of its paper predecessor beyond the scholarly world to embrace students and teachers of Latin.<sup>4</sup>In order, then, to satisfy needs of the academic users<sup>5</sup>, on the one hand, and to effectively distinguish between expert and lay users on the level of the web interface, on the other hand, a highly structured resource had to be created.

Secondly, the lexicographic analysis served also two other purposes, the first of them being to estimate the feasibility of the data annotation within project's time limits, that is without resorting to advanced NLP methods, the second – to conceptualize the dictionary macro- and microstructure by means of the TEI XML tagset. Although encoding standards in linguistic annotation constitute nowa-

---

4 Not only Medieval, but also Classical Latin, since there does not exist as yet any on-line Polish dictionary of Classical Latin, at least academic one.

5 Or what was believed to be their needs, since to my knowledge there do not exist any empirical studies of the needs of the users of (academic) Latin dictionaries.

days a topic on their own (Garside, Leech, McEnergy 1997; Pustejovsky, Stubbs 2013), I will limit myself to indicating three main reasons why the TEI XML has been chosen as an output format of the dictionary files. First of them has been already mentioned: storing lexicographic data in text files (contrary to binary ones) makes them at least partially immune to platform or software-related issues. XML encoded resources are human-readable, so they may be easily subject to modifications, adaptations and further refinement even by less technical-oriented users. Secondly, the TEI XML encoding serves well the purpose of storing highly structured, paper-born documents. Since the print dictionary being a starting point of the *eLexicon* is a result of 60 years' work, not only is it far from unified, but it also makes heavy use of sense nesting, *ad hoc* usage hints etc., all of which makes putting it into database format a non-trivial task. Thirdly, the use of widely supported formats and standards becomes essential, if one wishes to benefit from the already existing software solutions, on the one hand, and, on the other, to make one's data useful in yet unpredicted research environments. As to the former, there was no intention to create from scratch a proprietary interface to serve the dictionary content. In fact, the *eLexicon*, rather than being a closed interface solution, attempts to initiate discussion about what tool do the medievalists need and to dynamically change as the community will express its expectations. In that it differs significantly from now outdated in their design, closed-source and paid resources, such as the *Database of Latin Dictionaries* published by Brepols<sup>6</sup>. The availability of the XML annotated files, in turn, should encourage dictionary content reuse, whereas applying the TEI recommendations should facilitate data exchange, as, despite their drawbacks, they were generally adapted in other open Medieval Latin dictionary projects, such as precursory digitisations of Lewis and Short's *A Latin-English Lexicon* (1879)<sup>7</sup> by the Perseus Project team (Crane, Seales, Terras 2009; Baman, Crane 2009), DuCange's *Glossarium* by Ecole Nationale des Chartes (Glorieux, Thuillier 2010),<sup>8</sup> *Novum Glossarium* in frame of the project Omnia (Bon 2009; Bon 2010; Bon 2011).<sup>9</sup>

Once the analysis had come to an end, the annotation guide was created and the annotation itself started. After the XML files had been generated through PERL and XSLT processing of the OCR output, they were next distributed among project's team members who diligently proofread them, modified when necessary the dictionary content and adjusted automatic encoding. Verified for their well-formedness, the files were next validated with a previously generated *Document Type Definition* (DTD). The web interface of the *eLexicon*, which will be treated in more detail below, was built around the eXist-db, a free and open-source, XML native, no-SQL database running in the back-end. Programmed as a set of XQuery scripts, it produces on the front-end a light-weight HTML5+CSS web ap-

---

6 <http://www.brepolis.net/>.

7 <http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a1999.04.0059>, now consultable also within Perseus under Philologic Project (<http://perseus.uchicago.edu/Reference/lewisandshort.html>) or, in a more convenient way, within Logeion project (<http://logeion.uchicago.edu/>).

8 <http://ducange.enc.sorbonne.fr/>.

9 The TEI was also employed in other historical lexicography projects, such as the Anglo-Norman Hub (<http://www.anglo-norman.net/>). For details, see (Trotter 2011).



plication. In order to introduce some elements of interactivity, as well as to provide users with instant headword suggestions and similar features, a moderate amount of jQuery scripting has been added.

## 3 Results

### 3.1 The XML Annotation

The principal rule applied in the process of the XML encoding was, as was already mentioned, to make as much lexicographic information explicit as possible without caring so much about the typographic peculiarities of the print dictionary. Formatting information was generally “translated” into appropriate “semantic” annotation and retained only if it could add to the on-line dictionary features. No special effort, then, was made to save indentation or font, since they should be otherwise easily deducible from the semantic encoding. Page and line numbers have been preserved in order to ensure correct resolution of the cross-references and, thus, successful intra-linking<sup>10</sup>.

Each of the output files corresponds to one print volume and is preceded by a metadata header in which basic bibliographic information was recorded. There is, however, nothing that could prevent prospective users from decomposing original files according to their specific needs. As far as the annotation design is concerned, dictionary entries have been first separated from each other<sup>11</sup>. For the purpose of subsequent processing, unambiguous identification of the entries had to be assured by means of the automatically generated identifiers, however, the lemma-based identification has been also retained<sup>12</sup>. The entries were next classified<sup>13</sup>, so as to distinguish between standard and reference entries of the type:

- “LETANIA *cf.* LITANIA”, where one of the orthographic variants points to the canonical word form;
- “LETARG ... *cf.* LETHARG ...”, where a word fragment (most frequently word prefix) points to the position in the dictionary text rather than to a precise headword.

Since the entry access in the *eLexicon* was meant to be subject to major redefinition, further refinements needed to be applied to the selected “secondary headwords” (Atkins, Rundell 2008: 235-236), with the most significant example being derived forms. Although such forms as n. *laureus* ‘laurel’ which is to be found in the paper dictionary as a sub-entry of the adj. *laureus* ‘laurel’, remain embedded in their respective superordinate entries, they will also function as separate lexical units during

---

10 In the *Lexicon* cross-references come generally in two forms and may point either 1) to a precise entry or one of senses (e.g. „*Cf.* LATIO II” under LEGISLATIO), or 2) to a volume, page and line(s) of the dictionary (e.g. „*cf. supra* I 1076,49 *sqq.*” under LEX which may be rendered as ‘cp. above, [volume] I, [page] 1076, [line] 49 and foll[owing]’).

11 For this purpose <entryFree> tag („unstructured entry”) was used which allows for a more liberal encoding of the paper-born and, thus, text-oriented dictionaries.

12 Here, the attributes @xml:id (“identifier”) and @n (“number”) were used.

13 By means of the @type attribute, with e.g. homonyms labelled as @type=“hom”.

alphabetical browsing or when listed in results lists.<sup>14</sup> The same can be said about other instances of secondary entries, such as multi-word or idiomatic expressions.

Orthographic, etymological and morphosyntactic information was subject to diligent, fine-grained encoding. In spite of the privileged position that a headword occupies in traditional lexicography, careful annotation of the variant orthographic forms is essential for a Medieval Latin dictionary to be a serious tool of research, and that from many reasons. With orthography changing often within one manuscript from scribe to scribe, a lexicographer can never know which word form dictionary user may be looking after, which makes selecting unique, “canonical” form somewhat anachronistic, if one takes into account the medieval sense of language correctness.<sup>15</sup> What is more, in the print *Lexicon* one can find headwords which serve only a purpose of identifying entries and they have been never attested in medieval texts. This is the case of such entries as “[LAVANDA] s. LAVENDA” ‘lavender’, where the square bracket is employed to indicate that, unlike *lavenda*, the form *lavanda* does not occur in sources, but, in turn, was used by the lexicographer only as a conventional representation of the entry as an ideal, in his or her opinion, reconstruction of the Italian *lavanda*.<sup>16</sup> In a case like the one just mentioned, separation of the genuine linguistic material from what is only a pure convention becomes crucial.

Without going into unnecessary detail, let it suffice to say that the other elements of the grammatical description of the headwords were subject to likely minute encoding, which, apart from its direct and obvious goal, namely that of description of linguistic resource, had two secondary objectives. First of them was making implicitly coded lexicographic data fully explicit. The variety of the information which is tacitly conveyed in the *Lexicon* is striking, however, what may be only challenging for a human reader, if she is not accustomed enough to dictionary convention, makes a good deal of data inaccessible for automatic processing. Among those pieces of information which could be lost, if they were not scrupulously deduced from sometimes cryptic metalanguage and, then, redundantly added to the original files, one can mention part-of-speech labelling, which is explicit (that is, expressed with appropriate labels) for adverbs or sub-headwords,<sup>17</sup> but for verbs, nouns and adjectives is to be inferred from the inflectional information.<sup>18</sup> The same is true about the language from which the headword originated, since appropriate labels are in the print dictionary employed uniquely for languages other than Latin, so, for instance, while the entry LEXICON ‘a dictionary’ includes a self-explanatory etymology “*Gr. λεξικόν*” (where *Gr.* stands for “Greek”), in the entry LICENTIO ‘to give a license’ one

14 Here <re>, ie. „related entry”, tag was used.

15 The problem of the abundance of word forms is even more striking for Medieval Latin as was used in France or Spain, where it experienced substantial assimilation to a vernacular language.

16 *S.* is here abbreviation for the Lat. *siue* ‘or’. Such notation can be found on a regular basis when hypothetical Classical Latin form of the word is reconstructed, see, for example, „[RHINOCERON] s. RINOCERON” ‘rhinoceros’.

17 See, for example, LICENTIOSE *adv.* ‘violently’.

18 See, for example, entries for a verb LICENTIO, -are, -avi, -atum ‘to give a licence’, a noun LICENTIA, -ae *f.* ‘licence’ or an adjective LICENTIOSUS, -a, -um ‘licentious’, for which PoS information should be determined from inflectional description.

finds notation “licentia”, from which one should infer that the word was coined during the Middle Ages from a Classical or Medieval Latin term. To make things even less transparent, the entries like LEX have no etymology at all, which, in turn, means they were inherited from the Classical Latin.

The list of information types which are encoded only indirectly is, naturally, far from complete and should also include such important features of historical lexicography as time and place of word’s attestation. While geographical information is never explicitly given in the paper *Lexicon*, chronological data are provided for the sake of precision, that is only if the source quotation comes from a work which includes multifarious, chronologically diverse material.<sup>19</sup> Otherwise, spatio-temporal characteristics of the quotation should be deduced from the alphabetical list of the dictionary sources. Yet, there are many reasons why information of this sort should be accessible in the on-line dictionary and, thus, why it should be also explicitly declared in the XML files. The reason that comes first to mind is, obviously, more efficient and straightforward retrieval of these data within the search interface of the on-line dictionary. The other reason why aforementioned, but also e.g. genological properties of source quotations should be explicitly encoded is that it could greatly facilitate its interactive representation in form of maps, timelines or charts, as the example of the Wiki Lexicographica (Bon, Nowak 2013) demonstrated.

The other of the secondary goals of dictionary encoding was standardisation of the lexicographic description. This included, for instance, eliminating domain or usage labels which are now obsolete or were coined *ad hoc* at some point of the dictionary writing process and shortly after fell from use.<sup>20</sup> On the contrary, some of the subtle or nowadays less useful distinctions were subsumed on the encoding level under general or more frequently used ones. This was the case of the labels indicating direction of the semantic change. Thanks to their unification, one will get an access to words which experienced metaphorical extension, although they were originally marked in the paper dictionary either with the standard label *metaph.* standing for *metaphorice* ‘metaphorically’, or with a more verbose label, *in imagine* ‘in the image of’.

It should be also added that the XML encoding, apart from its obvious, data-oriented objectives, has many practical, user-oriented ramifications. In the print dictionary, for instance, sense definitions are given, as was already noticed, both in Polish and Latin. Clear separation of the definition strings not only allows their subsequent retrieval and reuse, but also, on more practical level, allows Polish users to consult the on-line dictionary in their mother tongue, while serving foreign researcher with a Latin version of the entry. Fine-grained linguistic data encoding, in turn, facilitates differentiating basic

---

19 For example, the only quotation which can be found under the headword LEXICON is labelled as „AKap p. 61 (a. 1540)”, where „AKap” is a source identifier (pointing to a multifarious collection of the chapter tribunal), „p.” stands for *pagina* ‘page’ and „a. 1540” is a chronological hint which should be resolved as „anno 1540”, ie. ‘in the year 1540’.

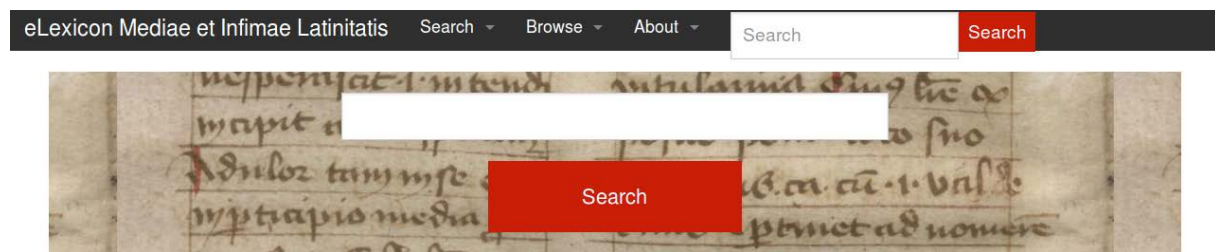
20 Naturally, XML encoding allowed for more obvious ameliorations as well. It was possible, for example, to introduce explicit distinction between domain (e.g. *astr.* for ‘astrology’, *eccl.* ‘ecclesiastical term’) and attitude (eg. *in malam partem* ‘pejorative’) labels on the one hand and the syntax markers (such as *intrans.* for ‘intransitive’ or *refl.* for ‘reflexive’) on the other.

and advanced user scenarios, and enables adapting lexicographic content perspective to the varying user needs.

### 3.2 The Web Interface

Apart from the obvious goal of overcoming the well-known drawbacks of paper dictionaries, the web interface of the *eLexicon* was created in order to facilitate advanced retrieval of the data obtained in the process described above. Thus, expected to constitute the main entry point to the electronic dictionary and other tools of textual studies, it was meant to provide professional users with a fully-fledged research platform. At the same time, however paradoxical it may appear, it had to satisfy the needs of less-advanced users, students and language teachers, by clearly separating basic and advanced perspective on lexicographic content. In order to serve well both groups, namely that of expert, as well as that of novice users, the guiding principle of the web interface creation became to help users better understand what they are looking for and to produce meaningful output, even if the phrase the user looked for, was not found in the dictionary.

When visiting the *eLexicon* page for the first time, users are proposed a quick tour of the search and browse features the dictionary offers. The main page is not meant, however, to overwhelm a visitor with a plethora of options (Figure 1). Rather the contrary is true, since apart from the simple menu, which gives direct access to the search and browse interface, it does not display anything but a simple search form which is, though, an actual entry point to the dictionary content.



**Figure 1: The on-line dictionary: the main page.**

Its underlying logic is to support two expected use scenarios:

- a user is looking for a lemmatised word form, for which there exists a corresponding main or secondary headword;
- a user is looking either a) for an inflected form, a Polish or other non-Latin term, a Latin word which is not attested in Polish sources, or b) for an incorrect word form.

The first scenario, i.e. successful lookup of a headword included in the *eLexicon*, is promoted by means of the Ajax-based suggestion list which appears once the user types in three first letters of the phrase

she is looking for.<sup>21</sup> The list of the words suggested consists of all the orthographic variants of the headwords included in the dictionary, as well as of the multi-word expressions and idioms, which are, however, still distinguishable from the former thanks to their different formatting. The suggestion list not only should significantly speed up the lookup process, but also may handle potential typing errors and, since Latin is an inflected language, point user to a correct lemma of the word she is querying. Once suggested option is selected, the user is redirected to the appropriate entry.

Here, two perspectives on the dictionary content are provided as separate tabs which, once clicked, reveal, respectively, a basic and a full content view. First of them (Figure 2), under clearly separated headings presents selected morpho-syntactical properties of the word, a brief overview of its meaning, as well as various summaries of its use.<sup>22</sup> As such, it should aid the novice, as well as expert users to get the general idea of the word they are looking for, without necessarily overwhelming them with the full apparatus of the academic lexicography.

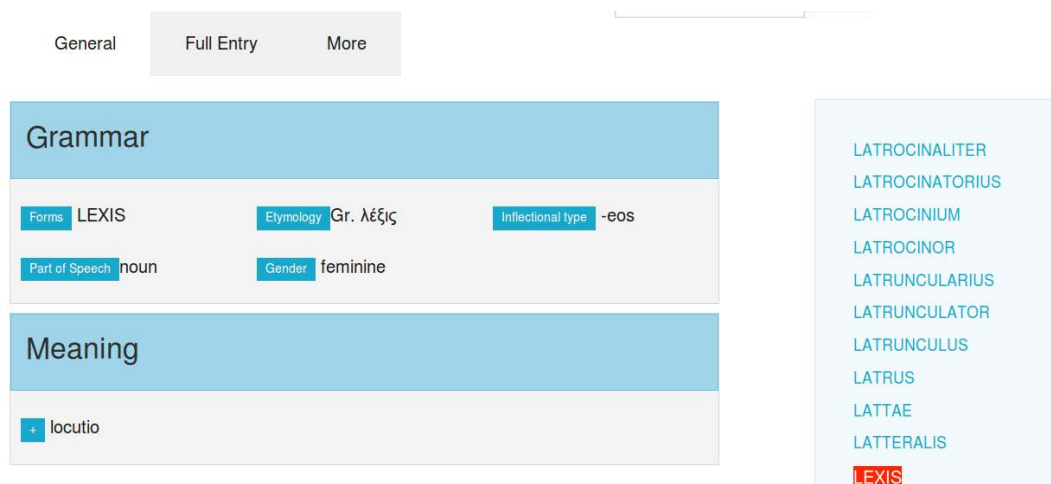


Figure 2: The on-line dictionary: a single entry (“Basic View” tab).

For both user groups the basic view may also be a convenient access point to the full version of the entry. The latter, in turn, makes heavy use of the CSS and JavaScript styling in order to improve the readability of the paper-born entry and facilitate information retrieval.

If the user does not decide to follow the suggestions and her query does not correspond directly to one of the entries, she is taken to the disambiguation page, where the second of the aforementioned scenarios is handled. User’s input is being processed and searched for in the sections of the *eLexicon* different than headwords.

21 The threshold was selected as a compromise between acceptable server load and usefulness. It is certain that it will be adjusted, once the user search logs are collected and analysed.

22 It is strongly inspired by the basic view previously implemented in the WikiLexicographica (Bon, Nowak 2013).

**Figure 3: The on-line dictionary: disambiguation page.**

The user is next presented with a result list (Figure 3) which, depending on the case, includes all or only some of the following parts:

- lemmatised form, in case if the user's input was an inflected form of a word;
- results of the word lookup in source quotations and/or definitions, if the user was using the dictionary as a source of attestations or synonyms, if she was after translation of a Polish term or if she was mimicking onomasiological search;<sup>23</sup>
- suggestions of similar words, if the input does not yield any meaningful result, so instead a correct word form should be suggested.<sup>24</sup>

The *eLexicon* content, however, may be accessed not only from the simplified main page search form, but also from the browse and expert search pages. The former functions as an equivalent of turning pages of the paper dictionary. Entries may be, then, selected by specifying the respective volume, page and line of the print edition. What is, however, more important, the browse interface offers dynamic (changing as the user is typing), synchronous lookup of the user's input at the beginning, in the middle and at the end of the dictionary headwords. This seemingly trivial feature was included to serve especially the Medieval Latin paleographers and manuscript readers in general who used to consult dictionaries looking for a reading suggestion of a hardly legible characters rather than for precise sense explanation.

<sup>23</sup> By looking, for example, for all words of which the Latin definitions employ word *color* 'colour'.

<sup>24</sup> Since the eXist-db makes use of the Lucene engine for text search, this feature is implemented as a fuzzy search of the user's phrase in the dictionary headwords. As the default Levenshtein distance value (Jurafsky, Martin 2009: 152) looks to be too liberal to produce helpful output, it will be certainly adjusted, once the data about the actual queries are collected.

The third access point is constituted by the advanced search facility, which takes full advantage of the scrupulous XML encoding of the lexicographic information (Figure 4)

The image shows a web interface for an advanced search. At the top left is a text input field. Below it is a 'Matching:' section with radio buttons for 'full' (selected) and 'partial'. To the right of the input field are three checkboxes: 'in headwords' (checked), 'within definitions' (unchecked), and 'within quotations' (unchecked). Below these is a blue 'Search' button. Underneath the search button are five blue buttons arranged in two rows: 'Part of Speech', 'Etymology', and 'Syntax' in the first row; 'Meaning' and 'Quotations' in the second row. At the bottom, there is a section titled 'Part of Speech, Inflectional Type, Gender' containing three dropdown menus. The first dropdown is set to 'substantivum', the second to 'Paradigma', and the third to 'Genus'.

**Figure 4: The on-line dictionary: an advanced search page.**

The user is here provided with a search form consisting of two main parts:

- the text input field, in which query string should be typed. The user may further specify scope of her search,<sup>25</sup> as well as matching strategy of her choice.<sup>26</sup>
- the list of additional restrictions to apply to search results. The user is free to refine her query and limit results by means of the morphosyntactic, etymological, semantic and chronological criteria.<sup>27</sup> In case the query string is not specified, selected criteria are applied to all dictionary entries and the interface functions as a tool of the exploratory analysis of the Medieval Latin lexicon. The results page allows for further refinement, since it contains a filtering list of linguistic properties of the previously queried headwords, which behaves in a manner similar to faceted browsing widgets.

As was already mentioned above, the *eLexicon* is expected to become a centre of a fully-fledged research platform. For that purpose, it is meant to be closely bound with the corpus of the Polish Medieval Latin. Although there still remains much work to be done, even now, whenever possible, use is made of the already existing external resources that are expected to be of help for the *eLexicon's* expert users. Since, as for now, resources in question are stored externally, only appropriate links to the freely available corpora, dictionaries or on-line encyclopedias may be provided to the users. In not so distant future, however, external resources are planned to be exploited locally and the content of at least some

25 That is, specific section of the dictionary entry within which the phrase should be looked for (currently options are limited to headwords, quotations and definitions).

26 That is, whether exact or approximate matching search should be applied.

27 Therefore, each query may be restricted, for instance, only to entries belonging to a specified inflectional class, originating from certain language, labelled as technical terms of a given domain or attested only in certain period.



of them will be directly embedded in the *eLexicon* search results.<sup>28</sup> In the current state of the interface, external resources are displayed:

- as a sidebar on the disambiguation page, as a way to suggest to the user other localisations in which she may find word absent from the *eLexicon* (see Figure 3 above);
- as a separate tab under a single entry view, so as to extend lexicographic perspective with corpus and knowledge base extracted data (Figure 5).<sup>29</sup>

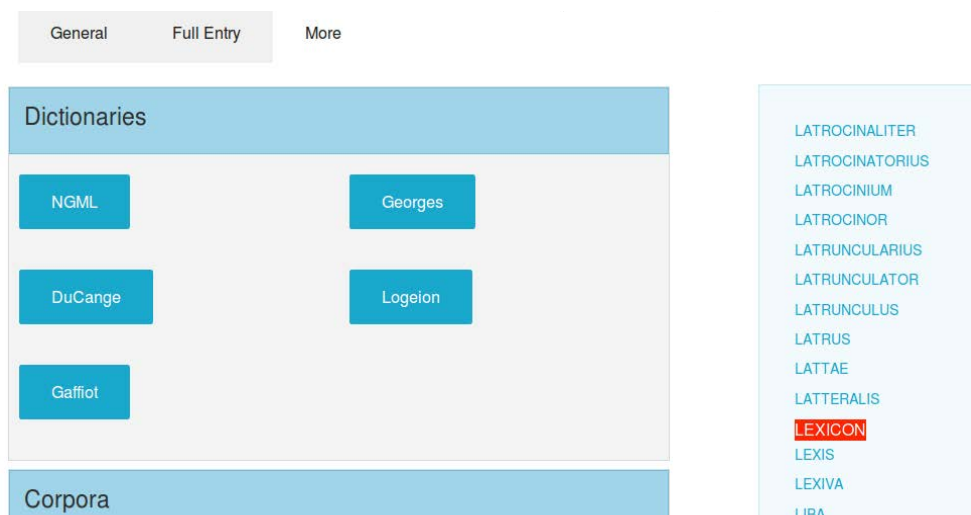


Figure 5: The on-line dictionary: a single entry (“More” tab).

## 4 Conclusions

The electronic dictionary of Polish Medieval Latin may become an important tool of the medieval studies, and this for many reasons. It will be freely accessible both in form of diligently encoded XML files and as a research-driven web application. To provide user with better insight into the medieval lexicon, the internet dictionary employs external sources, either by means of dynamically linking or direct embedding. Advanced users not only should find its single entry more readable, but they also will benefit from a configurable expert search and browsing interface, which provides the *a tergo*-like lookup. In turn, clear separation of a basic and advanced perspective on lexicographic content, as well as the use of suggestion lists and disambiguation pages may contribute to its becoming an effective tool for the Latin language students and teachers.

At the same time, there, naturally, still remains much room for improvement. As far as data presentation layer is concerned, maps, timelines, charts and other alternative displays need to be implement-

28 This is the case of the freely available volumes of the *Novum Glossarium Mediae Latinitatis* or the *Glossarium* of DuCange, but also of the texts collected in the Perseus Library. The similar approach has been already applied in such inspiring tools as *Logeion* (<http://logeion.uchicago.edu>) or *Le Dictionnaire vivant de la langue française* (<http://dvlf.uchicago.edu/>).

29 In its current form, it is clearly still very far from being fully implemented.



ed. There is also a serious NLP work which has to be done, since the *eLexicon* is expected to provide conceptual search interface and to better integrate with knowledge bases.

## 5 References

- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bamman, D., Crane, G. (2009). Computational Linguistics and Classical Lexicography. In *Digital Humanities Quarterly*, 3(1). Accessed at: <http://www.digitalhumanities.org/dhq/vol/3/1/000033/000033.html> [10/04/2014].
- Bautier, A.-M. (1981). La lexicographie du latin médiéval. Bilan international des travaux. In *La lexicographie du latin médiéval et ses rapports avec les recherches actuelles sur la civilisation du Moyen Age: Paris, 18-21 octobre 1978*, Paris: CNRS, pp. 433–53.
- Bon, B. (2009). OMNIA – Outils et Méthodes Numériques pour l’Interrogation et l’Analyse des textes médiolatins. In *BUCEMA. Bulletin du centre d’études médiévales d’Auxerre*, 13, pp. 291–92. Accessed at: <http://cem.revues.org/11086> [10/04/2014].
- Bon, B. (2010). OMNIA: outils et méthodes numériques pour l’interrogation et l’analyse des textes médiolatins (2). In *BUCEMA. Bulletin du centre d’études médiévales d’Auxerre*, 14, pp. 251–52. Accessed at: <http://cem.revues.org/11566> [10/04/2014].
- Bon, B. (2011). OMNIA : outils et méthodes numériques pour l’interrogation et l’analyse des textes médiolatins (3). In *BUCEMA. Bulletin du centre d’études médiévales d’Auxerre*, 15. Accessed at: <http://cem.revues.org/12015> [10/04/2014].
- Bon, B., Nowak, K. (2013). Wiki Lexicographica. Linking Medieval Latin Dictionaries with Semantic Media-Wiki. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langements, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013*. Tallinn - Ljubljana: Trojina, Institute for Applied Slovene Studies, Eesti Keele Instituut, pp. 407–420. Accessed at: [http://eki.ee/elex2013/proceedings/eLex2013\\_28\\_Bon+Nowak.pdf](http://eki.ee/elex2013/proceedings/eLex2013_28_Bon+Nowak.pdf) [10/04/2014].
- Crane, G., Seales, B., Terras M. (2009). Cyberinfrastructure for Classical Philology. In *Digital Humanities Quarterly*, 3 (1). Accessed at: <http://www.digitalhumanities.org/dhq/vol/003/1/000023/000023.html> [10/04/2014].
- Garside, R., Leech, G. N., McEnery, T. eds. (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London-New York: Longman.
- Glorieux, F., Thuillier, S. (2010). Grec ancien, latin médiéval, balisage comparé de deux dictionnaires, vers des ressources linguistiques. In *ALMA. Archivum Latinitatis Medii Aevi*, 68, pp. 161–81.
- Jurafsky, D., Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River: Pearson Prentice Hall.
- Lewis, Ch. T., Short, Ch. (1879). *A Latin Dictionary*. Oxford: Clarendon Press.
- Plezia, M. (1958). Lexicon mediae et infimae latinitatis Polonorum. In *ALMA. Archivum Latinitatis Medii Aevi*, 28, pp. 271–84.
- Plezia, M., Weyssenhoff-Brożkova, K., Rzepiela, M. eds. (1953). *Słownik łaciny średniowiecznej w Polsce. Lexicon mediae et infimae Latinitatis Polonorum*. Vols. 1–8. Kraków: Wydawnictwo IJP PAN.
- Pustejovsky, J., Stubbs, A. (2013). *Natural language annotation for machine learning*. Sebastopol, CA: O’Reilly Media.
- TEI Consortium. (2013). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version: 2.5.0. Accessed at: <http://www.tei-c.org/Guidelines/P5/> [10/08/2013].

Trotter, D.A. (2011). Bytes, Words, Texts: The Anglo-Norman Dictionary and Its Text-Base. In *Digital Medievalist*, 7. Accessed at: <http://www.digitalmedievalist.org/journal/7/trotter/> [10/04/2014].

### **Acknowledgments**

The *eLexicon Mediae et Infimae Latinitatis Polonorum* as well as the work on the present paper were supported by a research grant of the Polish National Science Centre (*eLexicon Mediae et Infimae Latinitatis Polonorum* (A-Q), nr 3736/B/H03/2011/40).

# From DANTE to Dictionary: The New English-Irish Dictionary

Pádraig Ó Mianáin, Cathal Convery  
Foras na Gaeilge, Dublin  
pomianain@forasnagaeilge.ie, cconvery@forasnagaeilge.ie

## Abstract

Most major bilingual dictionary projects tend to largely involve adapting an existing bilingual dictionary, by adopting or amending the existing source language material as required and then supplying the necessary target language material. This paper describes the innovative approach followed in the production of the New English-Irish Dictionary (NEID) project. The online version of NEID was launched in 2013 ([www.focloir.ie](http://www.focloir.ie)) and the project is to be completed when a printed version is produced in 2016. The English language content of NEID is based on the Database of ANalysed Texts of English (the DANTE database), a ground-breaking corpus-based lexical database developed specifically for the project. Attention is drawn to how the DANTE entry frameworks evolved through the various translation and editing processes to the final entries now available in the online. The paper also discusses the development of the Irish-language material and details some of the challenges, practical and editorial, encountered in the course of the production of the dictionary, and the working solutions that were developed.

**Keywords:** DANTE; bilingual; Irish-language

## 1 The New English-Irish Dictionary

The New English-Irish Dictionary (NEID) is being produced and funded by Foras na Gaeilge, the inter-governmental body with responsibility for the promotion of the Irish language, with a project budget of €6m. It is the first major English-Irish dictionary since de Bhaldraithe's English-Irish Dictionary (1959) and the first major bilingual dictionary in Ireland since Ó Dónaill's *Foclóir Gaeilge-Béarla* [Irish-English Dictionary] (1977). The first version of NEID was launched online in January 2013 with new material and revisions being uploaded at regular intervals until the envisaged completion of the online edition in 2015. The NEID will then be published in printed format in 2016.

## 2 Project Timeline and Phases

The NEID project started in 2000 and is on target to be completed at the end of 2016. The project is supported by IDM's DPS platform along with the Entry Editor interface. The online version of NEID

will eventually contain c. 130,000 sense units (c. 40,000 headwords) by December 2015, with a print version to follow in 2016. The project is divided into three major phases:

- **Phase 1: Planning and design** (started 2000, completed 2006). The planning and design phase was carried out by Lexicography Masterclass and delivered key elements such as an overall project plan, the English-language and Irish-language corpora which underpin the entire project, as well as sample entries, headword lists, draft style guides for each phase etc.
- **Phase 2: Compilation, Writing and Editing** (started 2008, to be completed 2016). This phase concerns the actual writing of the dictionary, and is discussed in detail below.
- **Phase 3: Publication** (started 2012, to be completed 2016). The online and mobile platforms were launched in January 2013 ([www.focloir.ie](http://www.focloir.ie)) with about 30% of the final content of the dictionary; the online content is being added to on an incremental basis. The dictionary will also be made available as an app in 2014.

A number of separate databases were created to facilitate the progress of entries from the DANTE database through the various translation and editing stages to the online entry. The main databases are shown in Figure 1 below along with their purpose and some of the more significant changes between databases:

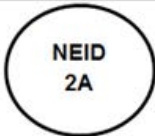
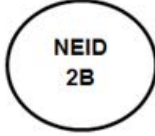
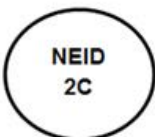
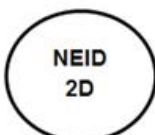
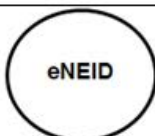
DATABASE	DESCRIPTION	PROCESSED BY
 NEID 2A	The source-language phase	English-language lexicographers
 NEID 2B	The target-language phase Changes automatically implemented: <ul style="list-style-type: none"> <li>• Compounds extracted as full HWDs</li> <li>• FREQUENCY field inserted</li> <li>• TRID (translator ID) field inserted</li> <li>• TrCnt (translation container) inserted at set positions within entries</li> </ul>	Translators
 NEID 2C	The editing phase (at sense unit level) <ul style="list-style-type: none"> <li>• Initial editing</li> <li>• Senior editing</li> </ul> Changes automatically implemented: <ul style="list-style-type: none"> <li>• EDID and SENED (editor &amp; senior editor ID) fields inserted</li> <li>• reporting of style guide non-compliance</li> </ul>	Editors
 NEID 2D	The final entry phase <ul style="list-style-type: none"> <li>• Sense units merged where possible</li> <li>• Disambiguators inserted</li> <li>• Streamlining of DOMAINS etc</li> </ul> Changes automatically implemented: <ul style="list-style-type: none"> <li>• Sense units grouped in blocks by POS</li> <li>• Links to grammar and sound files automatically inserted</li> </ul>	Senior Editors
 eNEID	Content uploaded to website Changes automatically implemented: <ul style="list-style-type: none"> <li>• All fields not to be displayed are deleted (MEANING, FREQ, EDID etc)</li> </ul>	IDM

Figure 1: The main databases involved in the phased transition between the DANTE database (designated NEID 2A within the context of the NEID project) and the final online entries.

### **3 The Source-Language phase: the DANTE database (began 2008, completed 2010)**

Lexicography Masterclass also supplied the English-language lexical database on which the source language content of the dictionary is based. It has subsequently been made available as a resource for various other lexicographical and linguistic projects as the Database of Analysed Texts of English or the DANTE database. The entire database can be viewed and explored at [www.webdante.com](http://www.webdante.com).

DANTE is based on systematic analysis of a 1.7-billion-word corpus of English supported by the Sketch Engine corpus query package and the GDEX feature for optimising the corpus example selection. The final package is a highly-structured 16-million-word lexical database of English that can be used not only by lexicographers but also by linguists, researchers and teachers. From a purely lexicographical point of view, DANTE is target-language-neutral in the sense that, despite its origins as the foundation for a bilingual dictionary, its sole focus is on describing and evidencing the English language without regard for any potential companion language (the Irish language specifically in the NEID project). As such, DANTE is equally a primary resource for monolingual and bilingual English dictionaries.

The headline statistics of the DANTE database give an indication of its coverage: 42,000 headwords (62,000 if compounds are counted as separate headwords), 12,000 phrases and 3,000 phrasal verbs, 149,000 sense units in total, over 600,000 relevant corpus examples and 16 million words in all. However the real value of the database is seen in the fine detail of its coverage within the entries, for instance:

- 42 different grammatical structures for verbs, 16 for nouns and 15 for adjectives (see Table 1 below);
- all headwords classified by 12 levels of complexity;
- entry templates;
- over 150 domain categories;
- regular proformas for closed set entries; etc.

From a purely NEID perspective, the DANTE database more than adequately served its purpose of providing the translation team with a comprehensive and detailed lexical profile of each English-language headword which was a candidate for inclusion in the final dictionary.

### **4 The target-language phase: adding the Irish-language content (began 2009, completed 2012)**

The aim of the translation phase of the NEID project was to maintain the structure and detail of the English-language entry frameworks while providing the editing team in the next phase with a comprehensive and detailed translation database, by adding as many relevant Irish-language equivalents

as possible to the English-language frameworks. It was at this stage that Foras na Gaeilge's own lexicography team took over the writing of the dictionary, as the translation and editing phases would be entirely driven by the Irish-language requirements.

The most efficient way of implementing this phase was to clone the DANTE database as a separate translation database and, as part of the cloning process, to automatically insert translation-specific fields in specified positions within the entry frameworks. Also, all compound entries were extracted from their mother entries (as they are in DANTE) and were promoted to full headword status, not least to facilitate the distribution of work batches in a more efficient manner. The end product of the translation phase is a rich database of bilingual entry frameworks consisting of the unabridged DANTE, plus 4 million words in Irish in over 600,000 translation fields.

Given the size and nature of the DANTE database, the translation task posed significant challenges at a number of levels, particularly given the lack of experienced Irish-language lexicographers available to the project. There were three main challenges:

- **The size of the translation task.** It would have been impractical and quite pointless to translate the entire English-language database, particularly as the specific remit of the translators was to provide as many relevant translations as possible in Irish. The translators were directed to concentrate on the key element or node of the word or phrase in question, and not to translate any of the surrounding text in the supporting examples unless it impacted on the translation; this was facilitated by automatically highlighting the node as part of the cloning procedure. Also, as each lexicographically significant fact in DANTE is contained in a specific structure container with corpus examples to match, there was no need to add translations to each individual example; to minimise such duplication, translation fields were automatically inserted and highlighted at the head of each structure container as a guide to the translators:



**Figure 2: An example of how translation fields (highlighted in green) were automatically inserted at specified positions within the DANTE entries for the translation phase: the first sense unit of *expect*.**

- The nature of the translation task.** Given the fifty-year gap since the previous English-Irish dictionary, a significant proportion of the English content of the NEID (approximately 30%) had to be translated into Irish from scratch, and a lot of the content that could be sourced in pre-existing dictionaries was dated or even obsolete in both English and Irish in terms of language, register etc. It was consequently decided from the outset to instruct the translation team to work *ex tempore* and NOT to consult published sources, except for technical terms. This self-reliance allied to the requirement to record as many Irish-language equivalents as they could think of came as a culture shock to translators who normally work from reference sources using one Irish-language word or phrase to translate a given English-language word or phrase. In addition to the translation challenges, the most frequent 1,000 headwords in English (20% of the sense units in DANTE) were each translated by three translators, one from each of the main dialects, who also recorded the relevance of each translation to their own dialect. Only technical terminology was translated from existing sources, with the translators recording those sources to facilitate the decision-making process at the editing stage.
- The layout and structure of DANTE.** To the uninitiated – and bearing in mind that the translators were primarily Irish-language specialists with no experience of lexical databases – DANTE is a

daunting beast. The hierarchical structure, consisting as it does of headwords, multiword entries, senses, structures and patterns, took some getting used to, and some of the grammar labels (or underlying structures) are less than transparent to all but the linguistically trained. The early solution here was to instruct the translators to ignore the grammar or structural information and concentrate on the corpus examples as a prompt for their translations in those cases.

## 5 The editing phase (began 2011, ongoing)

The initial step in the editing phase is to clone the completed translation frameworks to provide the editors with the full English-language profile plus the accompanying translations. This also ensures that the macrostructure of the edited content remains secure while the editors are drafting the final entry. Working on a sense-unit by sense-unit basis, the editors first decide on the English-language content of the entry, then the matching Irish-language content. Examples in the edited entries are included according to strict guidelines depending on the type and level of entry or sense involved, and the editors can either amend one of the corpus examples from DANTE or compose a new example entirely. Finally, the remainder of the framework is discarded once the English and Irish content of the entry has been decided.

It is also during the editing phase that new sense units are added to the dictionary database for senses not present in the original DANTE database. This may occur

- if a single sense unit in DANTE requires significantly different translation solutions in Irish (for instance *bassist* as a single sense in DANTE covers the ‘double base’ and ‘bass guitar’ but each instrument requires a different translation in Irish):



Figure 3: The entry *bassist* as one sense unit in DANTE (left) and how it was split in the editing database (right).



- if a word sense in usage in Ireland is not covered in DANTE (for instance *hurl* as verb = ‘to play hurling’ and as noun = ‘hurling stick’);
- if senses have come to the fore in the intervening years (for instance *to friend sb* in a social network context).

Also as part of the final streamlining of metalanguage etc, most of the grammar and structure labels are either removed entirely or converted to the smaller subset in use in the published dictionary; for instance, the current dictionary entries show only three verb structure labels (intransitive, transitive and modal) where the underlying DANTE database has forty-two:

Field	Examples	Range of fields present in English frameworks (DANTE)	Range of fields present in final entry
DOMAIN labels	<i>agri, food, ornith</i>	156	98
GRAM (grammar labels)	<i>abbrev, c_u, proper</i>	28	3
STRADJ (adjective structures)	<i>AVP_premod, that_0</i>	17	0
STRN (noun structures)	<i>AJ_pert, PP_X</i>	18	1
STRV (verb structures)	<i>AJP, Part, that_0_cond</i>	42	3
STYLE labels	<i>child, euph, pc, tech</i>	21	9

**Table 1: Examples of the reduction in the number of fields between DANTE and the final entry.**

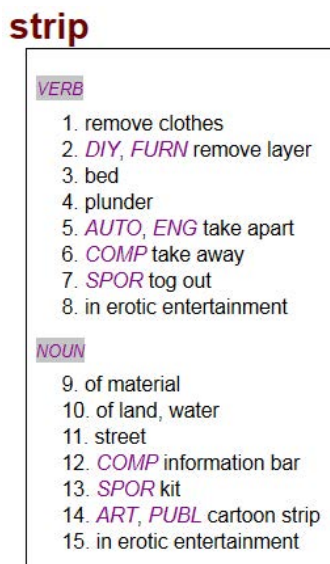
The editors can also recommend and mark entire entries or individual sense units for omission from the final dictionary, but as this is the prerogative of the senior editing team at the next phase no entries or sense units are discarded at this stage.

Though still not at a point where it could be published, the edited database is now a much leaner version of the translation database (for instance, the word count in the 19,000 frameworks edited by April 2014 was 2.2 million compared to 8.5 million in the same translation frameworks).

## 6 Publication phase (began 2012, ongoing)

In this phase the final dictionary entries are arranged and prepared for uploading to the website (and in 2016 for the print version). The most obvious change is the re-ordering of all sense units by part of speech: in the DANTE database, associated senses are clustered together regardless of part of speech, but NEID follows the traditional POS-based order. This re-ordering is done automatically, reflecting the order of precedence of the sense units in DANTE, and occasionally requires minor adjustment where the order of precedence of senses under one POS may not mirror another POS. In the case of *strip*, for instance, the first sense as a verb is ‘to remove clothes’, but as a noun the sense ‘narrow band

of sth’ would be much more common than ‘an instance of removing clothes’, and thus was brought to the beginning of the noun section:



**Figure 4: The manually adjusted sense order under *strip* where the noun sense associated with the first verb sense was manually transferred to a lower position in the noun order.**

Another instance where the automatic re-ordering needed to be manually tweaked was *turf* as a noun, where the sense ‘peat’ is much more common in Ireland than the sense ‘sod’ and was brought to the top of the order. It is also at the publication phase that entries or senses are marked for exclusion from the published dictionary though they are not deleted from the dictionary database.

Sense units are then merged where possible in order to avoid the user having to scroll unnecessarily through numerous senses with similar translations. For instance, the Irish equivalent remains the same for seven DANTE senses of the word *studio* as ‘work area’, so those seven senses were amalgamated into one single sense in NEID with examples added to indicate the breadth of coverage. Similarly four sense units under *coffee* in DANTE became one in NEID with pertinent examples:

 <p><b>NOUNBLK</b></p> <ul style="list-style-type: none"> <li><b>FWKSENCNT [1]</b> <ul style="list-style-type: none"> <li><b>POS [N]</b></li> <li><b>LABELGP</b> <ul style="list-style-type: none"> <li><b>DOMAIN [ART]</b></li> <li><b>DOMAIN [PHOT]</b></li> <li><b>DOMAIN [TV-RAD]</b></li> <li><b>DOMAIN [ETC]</b></li> </ul> </li> <li><b>MEANING</b> a room where an artist, sculptor or photographer works</li> <li><b>MEANING</b> a room or building where television and radio programmes are recorded or broadcast</li> <li><b>MEANING</b> a place where cinema films are made or produced</li> <li><b>MEANING</b> a room where sound or music recordings are made</li> <li><b>MEANING</b> a room where dancers, actors, people doing exercise etc can practise</li> <li><b>MEANING</b> a film or television production company</li> <li><b>MEANING</b> a place in a company where new products are designed</li> </ul> </li> </ul>	<p><b>studio</b></p> <p>1 <i>NOUN ART, PHOT, TV-RAD, ETC</i>  <i>stiúideo masc4</i> 🗣️ <b>C M U</b>  <b>aerobics studio</b> <i>stiúideo aeróbaice</i>  <b>art studio</b> <i>stiúideo ealaíne</i>  <b>film studio</b> <i>stiúideo scannán</i>  <b>photographic studio</b> <i>stiúideo grianghrafadóireachta</i>  <b>recording studio</b> <i>stiúideo taifeadta</i>  <b>television studio</b> <i>stiúideo teilifíse</i></p> <p>2 <i>NOUN</i> studio flat  <i>árasán stiúideo</i> 🗣️ <b>C M U</b></p>
 <p><b>NOUNBLK</b></p> <ul style="list-style-type: none"> <li><b>FWKSENCNT</b> <ul style="list-style-type: none"> <li><b>POS [N]</b></li> <li><b>LABELGP</b> <ul style="list-style-type: none"> <li><b>DOMAIN [DRINK]</b></li> <li><b>DOMAIN [FOOD]</b></li> </ul> </li> <li><b>MEANING</b> the liquid as a drink</li> <li><b>MEANING</b> in its dried form, for making into a drink</li> <li><b>MEANING</b> in the form in which it grows on plants</li> <li><b>MEANING</b> a unit or containerful of the drink</li> </ul> </li> </ul>	<p><b>coffee</b></p> <p>1 <i>NOUN DRINK, FOOD</i>  <i>caife masc4</i> 🗣️ <b>C M U</b>  <b>a cup of coffee</b> <i>cupán caife</i>  <b>I like strong coffee</b> <i>is maith liom caife láidir</i>  <b>he takes two spoonfuls of coffee</b> <i>glacann sé dhá spúnóg chaife</i>  <b>she ordered two coffees</b> <i>d'ordaigh sí dhá chaife</i>  <b>black coffee</b> <i>caife dubh, caife gan bhainne</i>  <b>decaffeinated coffee</b> <i>caife gan chaiféin</i>  <b>instant coffee</b> <i>caife ar an toirt</i>  <b>white coffee</b> <i>caife bán</i>  <b>coffee jar</b> <i>próca caife</i></p> <p>2 <i>ADJECTIVE COL</i>  <i>ar dhath an chaife</i></p>

Figure 5: Multiple senses of *studio* and *coffee* merged into a single unit within the 2D database (left) and how the final entry appears in the online dictionary (right).

This approach, which would not be possible in a decoding dictionary, is facilitated by the fact that virtually all users of NEID are fully fluent in English and have an intuitive understanding of the English content.

To complete the editing process, sense disambiguators and domain labels are added or removed as required, and finally, when the text content of the entry is finalised, grammar and sound files are attached to the translation fields in preparation for publishing online.

## 7 Practical challenges

The project to produce a modern bilingual English-Irish dictionary faced a number of significant practical challenges. Some of these challenges may apply to similar projects in any language, some may apply in particular to other lesser-used languages, while others stemmed from the previous gap in bilingual lexicography in Irish and the consequent problems of finding suitably qualified and experienced staff at editorial and managerial level. The main such challenges were:

- **An innovative approach to dictionary compilation.** The approach followed in the NEID project is ground-breaking in that the final content of the dictionary is entirely derived from a lexical database based on systematic corpus analysis. This required all translators and editors to exercise a higher level of judgement throughout.
- **Technical challenges.** The technical working environment of modern lexicography posed a significant challenge, both at the organisational and at the individual level, where continuous training and monitoring was required.
- **Management of staff and processes.** The varied nature of the project and the overlapping of the various phases and sub-phases required careful organisation and management, particularly as a lot of the processes in the translation and editing stages were being developed from scratch.
- **Training and upskilling of staff.** The project required a significant and continuous programme of training and monitoring as the processes for each phase of the project were being tested and implemented, and this burden was exacerbated by the shortage of experienced people at all levels in the project.

## 8 Conclusion

Notwithstanding the challenges posed by undertaking such an innovative model for this dictionary project and the particular challenges arising in relation to Irish lexicography, the fundamental benefit to NEID is that DANTE enabled the project team to produce a dictionary which they can claim is uniquely Irish, in as much as every word of both English and Irish in the dictionary is there by the editors' choice. The customisations and additions to the DANTE database at the subsequent three stages of the compilation process of NEID show that DANTE is a very flexible resource which can be adapted to the requirements or wishes of any English-language dictionary project.

## 9 Further information

The Database of Analysed Text of English (DANTE): [www.webdante.com](http://www.webdante.com)

The New English-Irish Dictionary: [www.focloir.ie](http://www.focloir.ie)

## 10 References

- Atkins, B. T. S. (2010). The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data. In G.-M. de Schryver (ed.) *A Way with Words: Recent Advances in Lexical Theory and Analysis: A Festschrift for Patrick Hanks*. Kampala: Menha Publishers ([www.menhapublishers.com](http://www.menhapublishers.com)).
- Atkins, B. T. S., Kilgarriff, A. & Rundell, M. (2010). Database of ANalysed Texts of English (DANTE): the NEID database project. In *Proceedings of the Fourteenth EURALEX International Congress*. Leeuwarden: Fryske Akademy, pp. 549-556.
- Atkins, B. T. S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Convery, C., Atkins, B. T. S., Kilgarriff, A., Rundell, M., Ó Mianáin, P., & Ó Raghallaigh, M. (2010). The DANTE Database (Database of ANalysed Texts of English). In *Proceedings of the Fourteenth EURALEX International Congress*. Leeuwarden: Fryske Akademy, pp. 293-295.
- Convery, C., Ó Mianáin, P., & Ó Raghallaigh, M. (2010). Covering All Bases: Regional Marking of Material in the New English-Irish Dictionary. In *Proceedings of the Fourteenth EURALEX International Congress*. Leeuwarden, pp. 609-619.
- Kilgarriff, A. (2010). DANTE: A Detailed, Accurate, Extensive, Available English Lexical Database. In *Proceedings of a meeting of the North American Association for Computational Linguistics (NAACL-HLT)*, Los Angeles, June.
- McCarthy, D. (2010). DANTE: a new resource for research at the syntax-semantics interface. In *Proceedings of Interdisciplinary Workshop on Verbs, Pisa*.
- Ó Mianáin, P. (2013). The New English-Irish Dictionary. In Stickel, G. & Varadi, T. (eds.) *Lexical Challenges in a Multilingual Europe: Contributions to the Annual Conference 2012 of EFNIL in Budapest*. Frankfurt um Main: Peter Lang, pp. 111-114.
- Rundell, M. & Atkins, B. T. S. (2011). The DANTE database: a User Guide. In *Proceedings of eLex 2011*. Trojina: Institute for Applied Slovene Studies, pp. 233-246.
- Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end?. In Meunier, F., De Cock S., Gilquin G. & Paquot M. (eds) *A Taste for Corpora: A tribute to Professor Sylviane Granger*. Benjamins, pp. 257-281.



# User Support in e-Dictionaries for Complex Grammatical Structures in the Bantu Languages

Danie J. Prinsloo\*, Theo J.D. Bothma\*, Ulrich Heid\*\*

\*University of Pretoria, \*\*University of Hildesheim

danie.prinsloo@up.ac.za, theo.bothma@up.ac.za, heid@uni-hildesheim.de

## Abstract

This paper discusses direct user guidance as a mechanism in an e-dictionary to provide user support for complex grammatical structures in the Bantu languages. We present a design study to show that user support through direct user guidance can provide solutions in the case of complex concordial relationships between nouns and pronouns. The compilation of the complex relative construction is taken as a case in point. The concept of user support appropriately puts the user in focus. Our approach to user support also caters for the casual, on-the-fly user, who is not interested or in a position to devote time to the learning of a foreign language.

**Keywords:** e-Dictionaries; Bantu languages; complex grammatical structures; user support; the relative construction; Northern Sotho

## 1 Introduction

In the Bantu languages, there are many grammatical constructions that are insufficiently treated in current dictionaries because of the complexity of the constructions. Other solutions need to be designed as an integral part of a dictionary and additional levels of user support are required within the dictionary. Such support should be available to the dictionary user “on demand”, and different options can be available for a specific information need. The nature of the support could typically also link to a user’s level of knowledge of the grammatical system of the language. A user with a very limited knowledge of the language or a casual user, for example, may prefer a machine translation option in the dictionary, with links to the grammar rules which may be consulted on demand. On the other hand, a user who has a fair knowledge of the language may require a different type of support, e.g. through *inter alia* decision trees, structured paths or direct user guidance. Such technologies, integrated in the dictionary, may enable the user to find the correct information at exactly the right level of detail and complexity (s)he requires to solve his/her information need (cf. Bothma 2011).

The innovative use of decision trees and structured paths as tools to support users have been dealt with in some detail in Prinsloo et al. (2011, 2012) and it has been shown that these solutions can provide significant decision support to users for complex text production situations such as copulative constructions and kinship terminology in Northern Sotho. The purpose of such tools is to guide users

to the information they are looking for, i.e. without having to first study complicated grammatical structures in order to find the required information. This guidance process is done through decision trees (i.e. a series of basic choices made by the user) or through structured paths (e.g. visually linking kinship terms in a schematic illustration of a family tree) as discussed in Prinsloo et al. (2011, 2012) or through direct user guidance, as discussed in this paper.

Direct user guidance as an additional technique in the dictionary to provide user support for complex grammatical structures in the Bantu languages is not a solution for all user support. We regard it as a complementary technology that may be used in conjunction with other user support technologies for specific grammatical constructions, available to the user on demand, depending on the user's level of language knowledge, the nature of the information need and choice of support tool. We present a design study to show that user support through direct user guidance can provide solutions in the case of complex concordial relationships between nouns and pronouns. In terms of the Function Theory of Lexicography (Tarp 2008, Bothma and Tarp 2012) the design provides for text production, text reception and cognitive information needs. No case studies or user evaluation of these techniques has been done to date, as we feel that it is important to first define a range of techniques and the range of complex grammatical structures where such techniques could offer relevant user support before any serious implementation in actual real world scenarios would be warranted. This does not mean that small scale prototypes of individual techniques should not be developed to establish the technological feasibility of such techniques. However, to do proper usability studies on such prototypes that are not fully integrated into a full dictionary will have only limited value, as it will not be possible to determine whether (or to what extent) users would use such techniques in real world situations as an integral part of dictionary use. As will be clear from the discussion below and from Prinsloo et al. (2011, 2012), such techniques are made available "on demand", i.e., users are not forced to use them if they feel that their information needs have been solved by the "standard" dictionary article. In every case, the use of such a technique is therefore a conscious choice of the user to find more information or information that is easier to use / digest / apply than the information available in the dictionary, the outer text of the dictionary or other reference tools such as grammar books that the user may have available.

The importance of the user perspective as the main thrust in the compilation of modern dictionaries has been emphasized in numerous publications, e.g., Gouws and Prinsloo (2005), Tarp (2008, 2011, 2012). The concept of user support appropriately puts the user in focus. Compare Tarp's (2012:253) idea of individualization when he refers to "quicker, more accurate and personalized satisfaction of the corresponding user needs". Our approach to user support furthermore does not necessarily put the user into a specific category (e.g., as a learner of the language): it is not profile-based and does not assume that the user will be interested to study a complete grammatical paradigm before being able to produce (or understand) text and speech. We therefore also cater for the casual, on-the-fly user, who is not interested or in a position to devote time to the learning of a foreign language.



## 2 Phenomena and Data: Grammatical Distinctions as a Problem for Bantu Lexicographers

### 2.1 The Notion of Grammatical Distinctions

Due to the richness of grammatical distinctions, a given grammatical property may be expressed in many different forms. For example, there are different equivalents for a pronoun such as he in Bantu, determined by the grammatical class of the noun. Nouns in Bantu languages are subdivided into different noun classes and these classes have their own sets of, e.g., subject concords and object concords, as well as different sets of pronouns such as demonstrative, possessive, emphatic and quantitative. This means that in Northern Sotho a basic English pronoun such as he can be expressed by up to ten different subject concords, a form like him by ten object concords and more than 20 pronominal forms. Consider table 1 which distinguishes 15 different noun classes each having their own subject concords (Sc.); object concords (Oc.); demonstratives (Dem.); possessive concords (Poss.); emphatic pronouns (Ep.) and quantitative pronouns (Qp.).

Person or noun class	Example	Sc.	Oc.	Dem.	Poss.	Ep.	Qp.
1st Person singular	nna 'I'	ke	n-				
1st Person plural	rena 'we'	re	re				
2nd Person sing.	wena 'you' (singular)	o	go				
2nd Person plural	lena 'you' (plural)	le	le				
Class 1	monna 'man'	o/a	mo	yo	wa	yena	yohle
Class 2	banna 'men'	ba	ba	ba	ba	bona	bohle
Class 3	molato 'trouble, problem'	o	o	wo	wa	wona	wohle
Class 4	melato 'problems'	e	e	ye	ya	yona	yohle
Class 5	lesogana 'young man'	le	le	le	la	lona	lohle
Class 6	masogana 'young men'	a	a	a	a	ona	ohle
Class 7	selo 'object, thing'	se	se	se	sa	sona	sohle
Class 8	dilo 'objects, things'	di	di	tše	tša	tšona	tšohle
Class 9	ntlo 'hut'	e	e	ye	ya	yona	yohle
Class 10	dintlo 'huts'	di	di	tše	tša	tšona	tšohle
Class 14	bogobe 'porridge'	bo	bo	bjo	bja	bjona	bjohle
Class 15	go reka 'to buy'	go	go		ga		
Class 16	fase 'below'			fa			
Class 17	godimo 'above'	go	go		ga	gona	gohle
Class 18	morago 'behind'			mo			

**Table 1: The noun class system of Northern Sotho with a few sets of concords and pronouns.**

In table 1 the demonstrative ‘this’ varies depending on the class of the noun, e.g., class 1: *monna yo* ‘this man’ but class 14: *bogobe bjo* ‘this porridge’. Likewise, the possessive ‘of’ differs for each class, e.g., class 1: *mosadi wa monna* ‘wife of the man’ but class 2: *basadi ba monna* ‘wives of the man’. ConCORDs and pronouns representing subjects and objects also vary according to the nominal class, e.g.:

(1) O e bone ‘He saw it’

<i>o</i> (e.g. <i>monna</i> class 1)	<i>e</i> (e.g. <i>tau</i> class 9)	<i>bone</i>
<i>he</i> (the man)	<i>it</i> (the lion)	saw

## 2.2 Grammatical Distinctions in the Sentence Context

In table 1 the grammatical distinctions paradigm is mono-dimensional in the sense that it is always given for a single source language item which diverges into a single set of equivalents. More than one instance of grammatical distinction can, however, co-occur in a single construction or phrase:

### A single occurrence

Example 1: *he*, as the subject of a sentence (subject conCORDs):

(2) *O/a/le/se/e*    *thušitše mosadi.*  
*He*                helped the woman

Example 2: how to express *all* (quantitative pronouns):

(3) *Go bolaya*    *bohle/yohle/ohle/tšohle*  
 To kill                *all*

### Two occurrences: he as a subject and them as an object:

What is at stake here is direct guidance in terms of the simultaneous handling of subject and object conCORDs:

(4) *O/a/le/se/e*    *tlo*    *ba/e/a/di*    *thuša*  
*He*                will    them                help  
*He* will help *them*.

### Three occurrences (the verbal relative): he as a subject, and as a demonstrative and them as an object:

(5) *Yo/wo/le/se/ye*    *a/wo/le/se/e*    *ba/e/a/di*    *thušitšego*  
*He*                *he*                them                helped  
*He who* helped *them*.

### 3 Direct Guidance for Concords and Pronouns

Guidance is given by means of three possible access points depending on the user's need in terms of text production (access points 1 and 2) or text reception (access point 3), and his/her knowledge of the language:

- **Access point 1:** Step by step guidance: build your own Northern Sotho sentence/construction.
- **Access point 2:** The user enters an English phrase and the system then assists him/her in a step-by-step build-up process of the Northern Sotho construction.
- **Access point 3:** The user enters a Northern Sotho phrase and the software analyses it.

Utilising *Access point 1* simply requires the user to enter the Northern Sotho noun and the software will suggest the correct pronoun and subject/object concord from table 1. Where more than one option is applicable, the user has to select the correct one or utilise clickable help functions to guide him/her to the correct option or (s)he can revert to the *Access point 2* option.

Taking *Access point 2* as departure, the user can type in “the man is walking” and the software will return the noun (*monna*) + the subject concord (*o*) + the present tense marker (*a*) + the verb *sepela*, guiding the user to build *monna o a sepela*. Clickable help functions and ‘more information’ boxes are also provided.

Entering a Northern Sotho sentence from Access point 3 will result in the reverse process, e.g. *monna* ‘man’(noun) + *o* ‘he’ (subject concord) + *a* (present tense marker) + *sepela* ‘walk’ (verb).

The full set of necessary data, for these cases, is thus as follows: In the case of single and two occurrences given in 2.2, subject and object pronouns/concords are independent from each other, and their choice is only conditioned by the noun class of their antecedents.

To be able to provide the above mentioned kinds of guidance for single and two-occurrence of grammatical distinctions a word (token) list tagged for part of speech for nouns, verbs, subject concords, object concords and pronouns, and a basic bilingual dictionary for word forms are required.

To be able to provide guidance for cases like example (5) more than just word lists are needed as an internal knowledge source for the guidance tool: the agreement between subject and relative demonstrative must be encoded as well.

To guide users in the creation of the relative construction, also the morpho-syntactic structure of this construction must be explained.

The verbal relative case (three occurrences) is different in so far as the relative demonstrative (“who”) is in grammatical agreement with the subject. In addition to the requirements for single and two-occurrences cases, a basic five-element formation rule for the verbal relative: noun + demonstrative + subject concord + verb + relative suffix (-go) is required.

## 4 Example of Direct Guidance for the Verbal Relative

### 4.1 User Support for Text Production

**Access point 1:** This provides step by step guidance on how to build your own relative construction. The user with a basic knowledge of the grammatical system would like to express “the man who loves her”. (S)he knows the different nouns and verbs in Northern Sotho but needs guidance in terms of the concordial system. In this case the user consults the article for “who” in an English to Northern Sotho dictionary, selects the button “Build your own relative construction” and types the Northern Sotho word for “man” = *monna*.

The system subsequently suggests the relevant concords from table 1. In the consultation, the relevant section for the appropriate noun class is highlighted while being placed in context within the grammar table. For cognitive use, clickable options to see larger portions of the table are provided; cf. table 2.

Person or noun class	Example	Dem.	Sc.		Oc.
		This	He/she		Him/her
Class 1	monna ‘man’	yo	o	a	mo
Class 2	banna ‘men’	ba	ba	ba	ba
Class 3	molato ‘trouble’	wo	o	o	o
Class 4	melato ‘problems’	ye	e	e	e
Class 5	lesogana ‘young man’	le	le	le	le
Class ...					
		Click to add	Click to add	Click to add	Click to add
		More information	More information	More information	More information
Click for full list	Click for full list	Click for full list	Click for full list		Click for full list

**Table 2: The noun class 1 of Northern Sotho highlighted for selection of the correct concords.**

Based on his/her knowledge of the relative construction (noun + demonstrative + subject concord + verb + relative suffix (-go)), the user can now build the full relative construction, to arrive at the full phrase *monna yo a mo ratago*. If the user is knowledgeable about the subject concords of class 1 in the relative, (s)he will click the subject concord *a* directly. If the user nevertheless needs more support, e.g., choosing between the subject concords *o* or *a*, the cursor could be momentarily rested on any of the “more information” boxes in the table triggering a pop-up box to support him or her in the selection task.

The user who needs full support can type a complete relative phrase in either English or Northern Sotho as portrayed in access points 2 and 3.

**Access point 2:** The user enters an English phrase: “The man who loves her”, similarly to access point 1 in the dictionary. The system then assists the user in a step-by-step build-up process of the relative construction:

- (i) *the man*: the tool provides the correct equivalent from the dictionary, i.e. *monna* tagged for part of speech as *NO1* (noun of class 1, cf. table 1);
- (ii) *who*: keeping the agreement constraint from the sentence formation rule (noun + demonstrative + subject concord, + verb + relative suffix (-go) in 4.1), the tool extracts the demonstrative for class 1 from the closed-class list of demonstratives, i.e. *yo*;
- (iii) (subject concord): The insertion of the SC is coded in the rule for relatives: it requires, in addition to the demonstrative in (ii), the subject concord for the noun in (i). As in (ii), the tool proposes *a*, i.e., the subject concord for class 1 which is appropriate, among others, for the relative *as* opposed to *o*.
- (iv) *her*: unspecified, as there is no unique referent e.g., *the woman*. On the basis of corpus frequency the software suggests the top five ranked possible options for the object concord, i.e. classes 1, 9, 3, 7, 5 (calculated from the Pretoria Sepedi Corpus, PSC). For human nouns, class 1 stands out in terms of frequency, and the selection is therefore for class 1 = *mo*. If, however, the relevant word is, e.g., from class 7, the user can type the word or click on the full list of object concords and selects the concord *se* for class 7.
- (v) *loves*: as for (i), the task is only to find the correct Northern Sotho equivalent: *rata*, plus adding the relative suffix *-go* which is, as for (iii), built into the relative construction rule.

In the consultation, the relevant section for the appropriate noun class is highlighted while being placed in context within the grammar table. For cognitive use clickable options to see larger portions of the table are provided, as shown in table 2. The construction rules ensure contextually appropriate highlighting, e.g., only of the subject concord, in step (iii) or of the object concord in step (iv).

## 4.2 User Support for Text Reception

**Access point 3:** A Northern Sotho phrase: *monna yo a mo ratago* (or part thereof, e.g., *ratago*, *mo ratago*, *yo a mo*, etc.). The software analyses the phrase in terms of the formation rule for the relative (automatically or user selected), in fact reversing the strategy explained with respect to access point 2 for “the man who loves her”.

## 5 Conclusion and Future Work

User support through direct guidance (and other support mechanisms) for complex grammatical structures allows the user to navigate via the shortest route to the information (s)he is looking for in an dictionary without having to work through long and often complicated grammar-type representations of complex grammatical structures. Such guidance is always available on demand, i.e., the user is not forced to work through any such support mechanisms if (s)he finds that the “standard” data in the dictionary are sufficient to solve his/her information need in a given situation. However, if more information is needed or if the standard presentation of the information (be this in the dictionary, in outer texts or in reference tools) is too difficult or complex to be easily understood, the user would have an alternative mechanism (or alternative mechanisms) to obtain the relevant information. It also successfully combats information overload and fulfils the needs of not only the learner of the language but also of the casual on-the-fly-user of the language; its flexibility is intended to provide a step towards individualization.

Different access points are available to the user depending on his/her pre-existing knowledge. It is not a profile-based dictionary. We envisage that such mechanisms be implemented as “plug-in modules” in entries of specific lemmas of the dictionary, i.e., an additional link is shown to the user on screen which (s)he can follow on demand. Since such modules can exist independently from the dictionary database, it would be feasible to reuse them in other environments as well. It would therefore be feasible to use such tools as writing tools integrated in a word processor, again activated by the user on demand, if (s)he requires to check the correct formulation of a complex grammatical construction, similar to spelling and grammar checkers that currently occur in popular word processing software.

Future work includes the development of a working prototype and possibly the full-scale implementation of user support for complex structures proposed in this paper as a module of electronic dictionaries. Identifying and categorising additional support techniques and developing prototypes and the full-scale implementation of such additional support techniques are also envisaged, as well as identifying further complex grammatical structures for which additional user support techniques may need to be developed. We will also investigate the possibility of the reuse of all such modules in writing tools for user support.

## 6 References

- Bothma, T.J.D. (2011). Filtering and adapting data and information in the online environment in response to user needs. In: Fuertes-Olivera, PA and Bergenholtz, H (Eds.). 2011. *e- Lexicography: The Internet, Digital Initiatives and Lexicography*. London & New York: Continuum. Pp. 71-102.
- Bothma, T.J.D. & Tarp, S. (2012). Lexicography and the relevance criterion. *Lexikos* 22: 86-108.

- Gouws, R.H. and Prinsloo, D.J. (2005). *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PRESS.
- Prinsloo, D.J., Heid, U., Bothma, T.J.D. & Faaß, G. (2011). Interactive, dynamic electronic dictionaries for text production. In Kosem, I. and Kosem, K. 2011. *Electronic Lexicography in the 21st Century. New Applications for New Users. Proceedings of eLex 2011*. Bled, 10-12 November 2011. Bled: Trojina. Pp. 215-220. [http://www.trojina.si/elex2011/elex2011\\_proceedings.pdf](http://www.trojina.si/elex2011/elex2011_proceedings.pdf).
- Prinsloo, D.J., Heid, U., Bothma, T.J.D. & Faaß, G. (2012). Interactive, dynamic electronic dictionaries for text production. *Lexikos* 22: 290-320. <http://lexikos.journals.ac.za/pub/article/view/1009/526>. PSC, Pretoria Sepedi Corpus. University of Pretoria.
- Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge. General lexicographical theory with particular focus on learner's lexicography*. (Lexicographica Series Maior 134). Tübingen: Niemeyer.
- Tarp, S. (2011). Lexicographical and other e-tools for consultation purposes: Towards the individualization of needs satisfaction. In: Fuertes-Olivera, P.A. & Bergenholtz, H. (Eds.) 2011. *e-Lxicography: The Internet, Digital Initiatives and Lexicography*. London & New York: Continuum. Pp. 54-70.
- Tarp, S. (2012). Online dictionaries: today and tomorrow. In: *Lexicographica* 28. 253-267.

### **Acknowledgement**

This research is conducted within the SeLA project (Scientific e-Lxicography for Africa), supported by a grant from the German Ministry for Education and Research, BMBF, administered by the DAAD.





# Concerning the Treatment of Co-existent Synonyms in Estonian Monolingual and Bilingual Dictionaries

Enn Veldi  
University of Tartu, Estonia  
Enn.Veldi@ut.ee

## Abstract

Estonian is a lesser used language of Europe (with about one million users) that belongs to the Finnic group of Finno-Ugric language. It has a large proportion of international words, which has resulted in extensive co-existent synonymy between native and international words. However, because of linguistic purism Estonian monolingual dictionaries do not treat the members of such synonym pairs on an equal footing. Usually, they give preference to native words, which in practice implies that international words are provided together with their native-language equivalents, but native words are without any reference to international words. Unfortunately, this kind of asymmetrical treatment is not helpful for language users and compilers of bilingual dictionaries; they need to develop synonym competence, which covers both international and native synonyms. Probably the best solution is symmetrical treatment, that is, both synonyms should be provided side by side on a regular basis, which contributes to their better knowledge. It is argued that the quality of bilingual dictionaries could be significantly improved by providing the synonym pairs together on a systematic basis.

**Keywords:** synonyms; monolingual dictionaries; bilingual dictionaries; language planning; English; Estonian

## 1 Introduction

Synonymy can be regarded as “a relation of similarity / identity of meaning between senses associated with two (or more) lexical forms” (Cruse 2002: 486). For practical purposes, it is useful to follow the “synonymy rule of thumb: X is Y and Y is X” (Atkins and Rundell 2008: 135). It is also important to bear in mind that synonymy is a language-specific phenomenon (Gouws 2013: 349).

The purpose of the present study is to examine the lexicographic practice of treating co-existent intralingual synonyms in Estonian monolingual dictionaries from the perspective of its usefulness for lexicographers who compile English-Estonian and Estonian-English bilingual dictionaries. Estonian is a lesser used language of Europe (with about one million users) that belongs to the Finnic group of Finno-Ugric languages. Estonian has a large proportion of international words; therefore, it is not surprising that the size of the latest edition of the dictionary of foreign words (VSL 2012) almost equals the size of the single-volume dictionary of Standard Estonian (ÕS 2013). In many cases synonymy

concerns pairs of synonyms between native (Estonian) and foreign (international) words. However, sometimes both synonyms can be foreign words (e.g. *kabatšohk*, *tsukiini* ‘courgette Br, zucchini Am). Extensive co-existent intralingual synonymy in Estonian is the result of conscious efforts of language planners to increase the proportion of native vocabulary. It is typical of Estonian specialized vocabulary but can be found in general language too. The following examples come from linguistic terminology: *nimetav*, *nominatiiv* ‘nominative (case)’, *omastav*, *genitiiv* ‘genitive case’; *osastav*, *partitiiv* ‘partitive (case)’; *täishäälik*, *vokaal* ‘vowel’; *kaashäälik*, *konsonant* ‘consonant’; *vormiõpetus*, *morfoloogia* ‘morphology’; *lauseõpetus*, *süntaks* ‘syntax’.

Stephen Ullmann discussed similar synonymy in German:

The German linguist, for instance, can choose between *Lautlehre* and *Phonetik*, *Formenlehre* and *Morphologie*, *Bedeutungslehre* and *Semantik* (or *Semasiologie*), and as these synonyms are used in the same contexts, and sometimes even in the title of the same book, one can hardly speak even of stylistic differences between them. (Ullmann 1962: 142)

It could well be that Estonian may have followed the German example. Historically one could explain the development of extensive native-language terminology with national pride and identity building. In the case of Dutch, for example, Roel Vismans claims that “in the early Republic pride in the Dutch language also led to the introduction of many newly coined words for concepts expressed by Latinisms in other languages” (Vismans 1998: 133).

Estonian language planners have regarded the introduction of co-existent synonymy between international and native words as an important method of establishing native-language identity and increasing the proportion of native vocabulary. In 1980 Henn Saari, an Estonian linguist, published a seminal paper where he discussed at length the importance of such word pairs in Estonian terminology (Saari 1980)

Uno Mereste, another Estonian academic whose views have been influential in language planning, explained the need for intralingual synonymy as follows:

There is no such thing as positive polysemy. However, in addition to undesirable synonymy, there is also positive sameness of meaning or synonymy, which one should try to preserve and develop further – it is synonymy between pairs of native and international terms, for example, *nimetav* or *nominative* ‘nominative (case)’, *ainsus* or *singular* ‘the singular’, *tasuvus* or *rentaablus* ‘profitability’, *tehis-* or *süntetiline aine* ‘synthetic substance’, etc.. (Mereste 2000: 87)

In Estonia, one of the purposes of national word coining competitions is to seek native synonyms for the existing international words. A recent example is the winning entry *taristu* of the 2010 competition, which means the same as *infrastruktuur* ‘infrastructure’. The word was immediately adopted by Estonians and is now widely used.

On the one hand, there is good reason to regard the co-existent intralingual synonymy in Estonian as a luxury because it contradicts the principle economy in language. D. Alan Cruse has claimed that

“there is very little semiotic motivation for such a state of affairs: the only possible utility for absolute synonyms is aesthetic, to avoid repetition of forms (2002: 488).

## 2 Treatment of Intralingual Synonyms in Monolingual Dictionaries

In order to understand the treatment of intralingual synonyms in Estonian bilingual dictionaries, one has to examine the practice of presenting intralingual synonyms in monolingual dictionaries. It appears that intralingual synonyms do not receive equal treatment in monolingual dictionaries. They tend to prefer native-language synonyms and provide the native-language synonym in the entry for the international word but not the other way round (see the treatment of the synonyms *regionaalne*, *piirkondlik* ‘regional’ below; EKSS stands for *Eesti keele seletav sõnaraamat* ‘Explanatory Dictionary of Estonian’).

- (1) **regionaalne** adj mingit regiooni hõlmav, selles esinev v sellele isloomulik, piirkondlik (EKSS)
- (2) **piirkondlik** adj piirkonda hõlmav, selles toimuv v sellele iseloomulik, piirkonna- (EKSS)

This practice of asymmetrical treatment is intended to discourage the use of international words. It is difficult to say whether it serves its purpose, but one could likewise regard it as not user-friendly. Ordinary dictionary users, as well as compilers of bilingual dictionaries, need to develop competence with regard to the knowledge of both synonyms. In some cases, a prescriptive monolingual dictionary, such as *ÕS 2013 (Õigekeelsussõnaraamat 2013* ‘Dictionary of Correct Usage 2013’), may even indicate that the use of a foreign synonym is undesirable. Such is the case with *multikultuurne* and *mitmekultuuriline* ‘multicultural’.

- (3) [**multikultuurne**] → *mitmekultuuriline* (ÕS 2013)
- (4) **mitmekultuuriline**. Kanada on *mitmekultuuriline* maa (ÕS 2013)

This dictionary uses brackets for undesirable lexemes; in this case language planners regard the use of the combining form *multi-* undesirable and recommend its native equivalent *mitme-* ‘several’ instead. On the other hand, the following example shows that there can be inconsistencies even in dictionaries published by the same institutions. The new edition of the dictionary of foreign words VSL (*Võõrsõnade leksikon* ‘Dictionary of Foreign Words’) lists two forms of the ‘undesirable’ item, fails to provide the native synonym (*mitmekultuuriline*) recommended by the dictionary of correct usage, and suggests a different synonym (*paljukultuuriline*).

- (5) **multikultuurne**, **multikultuuriline** mitut kultuuri sisaldav, paljukultuuriline (VSL)

The next example shows that a dictionary user needs to develop a critical attitude towards the data provided in monolingual dictionaries and should not take everything at face value. The example concerns the notion of *kabatšokk*, *tsukiini* ‘courgette Br, zucchini Am’. A similar example of *aubergine* Br, *eggplant* Am and its Polish equivalents was discussed by Adamska-Sałaciak (2013: 338–339). Such lexical differences between British and American English are regarded as cross-varietal synonyms. The present co-existence of *kabatšokk* and *tsukiini* in Estonian can be explained by the fact that before 1991 the vegetable was known in Estonia by its Russian name (which ultimately has a Tatar origin). In recent years, however, *tsukiini* has been adopted as well, and at present both terms are used. It remains to be seen whether the younger generation will continue to use *kabatšokk*, or it will gradually be replaced by *tsukiini*.

However, the analysis of the treatment of *kabatšokk* and *tsukiini* in dictionaries revealed several problems.

- (6) **kabatšokk** s AIAND kõrvitsa pikerguste viljadega teisend, suvikõrvits, tsukiini (EKSS)
- (7) **tsukiini** s kabatšokk, suvikõrvits (EKSS)
- (8) **suvikõrvits** suvel v varasügisel kasutatavate kõrvitsate üldnimetus. *Kabatšokk, taldrikkõrvits ja spagetikõrvits on suvikõrvitsad* || kabatšokk, tsukiini. *Täidetud suvikõrvitsad* (EKSS)

The problem with the treatment of *kabatšokk*, *tsukiini*, and *suvikõrvits* ‘lit. summer pumpkin’ in EKSS is that *suvikõrvits* is a broader term and should be regarded as a superordinate of the first two terms. However, it is now time to turn to examine how the data provided in monolingual dictionaries tends to influence the treatment of co-existent synonyms in bilingual dictionaries.

The treatment of the synonyms *kabatšokk* and *tsukiini* in the dictionary of foreign words showed remarkable differences.

- (9) **kabatšokk** s BOT suvikõrvits, melonkõrvits, puhmikkõrvits, hariliku kõrvitsa piklike viljadega teisend (VSL)
- (10) **tsukiini** s BOT kabatšokk, hariliku kõrvitsa teisend (*Cucurbita pepo*) (VSL)

While the entry *tsukiini* claims that *tsukiini* is *kabatšokk*, the same is not true of the entry *kabatšokk*, which provided a number of other intralingual equivalents. It seems that different lexicographers were responsible for contributing material to different letters and no systematic harmonization was carried out.

### 3 Treatment of intralingual Synonyms in Bilingual Dictionaries

Analysis of the material of the Estonian bilingual dictionaries shows that more accuracy and systematicity is needed in the treatment of co-existent intralingual synonyms. At first the coverage of *courgette* and *zucchini* will be examined in three dictionaries. The English-Estonian Dictionary by Johannes Silvet (Silvet 4) is the best-known dictionary of this category; its first edition was published in the

late 1930s; at present its fourth enlarged edition of 2002 is widely used. The Estonian-English Dictionary by the TEA publishers (TEA 2005) is a recent Estonian-English dictionary. The Contemporary Estonian-English Dictionary by Andres Aule (see Aule 2003) is a large-scale Estonian-English dictionary; so far two volumes have been published (the letters A–J and K). So far the Estonian-English dictionary by Paul Saagpakk is the largest as to its coverage; unfortunately, it includes a considerable amount of outdated material while many more recent words are absent.

(11) **courgette** (no entry in Silvet 4)

(12) **zucchini** *s bot* kabatšokk, suvikõrvits, tsukiini (Silvet 4)

(13) **tsukiini** (no entry in TEA 2005)

(14) **kabatšokk** zucchini <pl zucchini, zucchinis>, (it), squash <pl squash, squashes>, (vegetable) marrow (TEA 2005)

(15) **kabatšokk** [melonkõrvits] vegetable marrow, marrow (*üldk*) (Aule)

(16) **kabatšokk** (no entry in Saagpakk)

(17) **tsukiini** (no entry in Saagpakk)

The presented examples reveal a considerable degree of inconsistency in the treatment of co-existent synonyms; nor are the suggested translation equivalents always accurate. In fact, the analysis shows that a bilingual lexicographer needs to critically study the material provided in monolingual dictionaries. However, in addition, primary corpus data is badly needed during the preparation of bilingual dictionaries.

Another area of intralingual synonymy concerns the situation where the international word belongs to a specific domain and the native synonym can be regarded as belonging to general language. In such cases the synonym that belongs to a specific domain should be labelled accordingly (e.g. *kardioloog med* ‘cardiologist’ and *südamearst* ‘heart doctor’).

(18) **kardioloog** *med* kardioloogia eriteadlane, südamearst (VSL)

However, analysis of a number of monolingual and bilingual dictionaries showed surprisingly that VSL was the only dictionary that provided the native synonym *südamearst*, which, in fact, is an everyday word in Estonian. It was also surprising that none of the other monolingual dictionaries (EKSS and ÕS 2013) listed the entry *südamearst*. Nor could it be found in the studied bilingual dictionaries. This finding suggests once again that a bilingual lexicographer should not rely too much on the existing monolingual dictionaries and has to use primary data as well.

What could be a solution to this problem? The answer is consistent symmetrical treatment of co-existent synonyms. Are there any Estonian dictionaries that treat the co-existent native and foreign synonyms on an equal footing? One such dictionary is the Estonian-English Dictionary of Linguistics by Erelt et al. (2012) (*Eesti-inglise keeleteaduse sõnastik*, EIKS). In this dictionary, synonyms are provided after the sense, separated by a comma. The treatment is consistent in that the same

information is provided in two places. This approach is also helpful for students who can see co-existent native and international terms side by side, which contributes to their synonym competence.

(19) **kaasa võnkuma, resoneerima** resonate (EIKS)

(20) **resoneerima, kaasa võnkuma** resonate (EIKS)

(21) **parasiitsõna, nugisõna** parasitic word, overused word (EIKS)

(22) **nugisõna, parasiitsõna** parasitic word, overused word (EIKS)

## 4 Conclusion

The study of co-existent intralingual synonyms in Estonian showed that monolingual dictionaries tend to promote native synonyms and do not cross-reference international words. While the rationale of the language planners is understandable, the disadvantage of this approach is asymmetry in the treatment of native and international synonyms. However, language users, as well as compilers of bilingual dictionaries, need to be familiar with the entire repertoire of such synonyms. For this reason, the symmetrical approach where the co-existent synonyms are provided side by side on an equal footing is fully justified. It is argued that the quality of bilingual dictionaries could be significantly improved by providing synonym pairs together on a systematic basis.

## 5 References

- Adamska-Salaciak, A. (2013). Equivalence, synonymy, and sameness of meaning in a bilingual dictionary. In *International Journal of Lexicography*, 26(3), pp. 329–345.
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Aule = Aule, A. (2003). *The Contemporary Estonian-English Dictionary*. Vol. 2, Tallinn: Estonian Language Foundation.
- Cruse, D.A. (2002). Synonymy. In D.A. Cruse et al. (eds.) *Lexikologie. Lexicology*. Berlin: Walter de Gruyter, pp. 485–497.
- EIKS = Ereht, M., Ereht, T., Veldi, E. (2012). *Eesti-inglise keeleteaduse sõnastik. Teine, täiendatud ja parandatud trükk*. Tallinn: Eesti Keele Sihtasutus [Estonian-English Dictionary of Linguistics, second enlarged and revised edition].
- EKSS = *Eesti keele seletav sõnaraamat*. Accessed at: <http://www.eki.ee/dict/ekss> [10/11/2013] [Explanatory Dictionary of Estonian].
- Gouws, R.H. (2013). Contextual and co-textual guidance regarding synonyms in general bilingual dictionaries. In *International Journal of Lexicography*, 26(3), pp. 346–361.
- Mereste, U. (2000). Oskuskeele korraldamise mõningaid põhimõtteid (Teesid). In U. Mereste. *Oskuskeel ja seaduse keeleline rüü*. Tallinn: Eesti Keele Sihtasutus, pp. 84–88 [Some principles of LSP planning, Abstract].
- Saagpakk = Saagpakk, P. (2000). *Estonian-English Dictionary*. Tallinn: Koolibri.
- Saari, H. (1980). Omasõna ja võõrsõna paarid eesti oskussõnavaras. In *Keel ja Kirjandus*, 11, pp. 654–666, 12, pp. 737–743 [Pairs of native and foreign words in Estonian specialized vocabulary].

- Silvet 4 = Silvet, J. (2002). *English-Estonian Dictionary*. 4<sup>th</sup> enlarged and revised edition. Tallinn: TEA.
- TEA 2005 = Mägi, R. (ed.) (2005). *Estonian-English Dictionary*. First edition. Tallinn: TEA.
- Ullmann, S. (1962). *Semantics. An Introduction to the Science of Meaning*. New York: Barnes & Noble.
- Vismans, R. (1998). Dutch. In G.. Price (ed.) *Encyclopædia of the Languages of Europe*. Oxford: Blackwell, pp. 129-136.
- VSL = Paet, T. (ed.) (2012). *Võõrsõnade leksikon*. 8., põhjalikult umber töötatud trükk. Tallinn: Eesti Keele Instituut & Valgus [Dictionary of Foreign Words, 8<sup>th</sup> thoroughly revised edition].
- ÕS 2013 = Raadik, M. (ed.) (2013). *Eesti õigekeelsussõnaraamat*. ÕS 2013. Tallinn: Eesti Keele Sihtasutus [Dictionary of Correct Usage of Estonian].





# **Phraseology and Collocation**



# Unusual Phrases in English MLDs: Increasing User Friendliness

Stephen Coffey  
Università di Pisa  
coffey@cli.unipi.it

## Abstract

This paper investigates the presentation of compositionally anomalous phrases in English monolingual learners' dictionaries (MLDs). In particular, it argues that it would be pedagogically useful to explain to the dictionary user, where possible, the reason why certain types of anomaly exist. Two types of phrase are discussed: firstly, idiomatic expressions in which the relationship between phrasal meaning and original meaning may not be clear to the learner (e.g. *run the gauntlet*); secondly, phrases which include particularly unusual word forms or word senses. These include lexical fossils (as in *the whys and wherefores*) and phrases partially motivated by phonological characteristics (as in *bits and bobs*). In order to form an impression of how anomalous phrases are currently treated in MLDs, samples of items were looked for in both print and online editions. It was found that, overall, little attention is paid to the motivation of phrasal composition, and it is suggested that more should be done in this direction. This would involve integrating current description, almost entirely synchronic in nature, with historical data, at least in the case of some types of phrasal unit.

**Keywords:** phraseology; learners' dictionaries; idiomatic expressions; lexical fossils; alliteration; rhyme; etymology

## 1 Introduction

The description of phraseological units has always been an important feature of the English monolingual learner's dictionary (MLD), ever since the publication of Hornby et al's ground-breaking *Idiomatic and Syntactic English Dictionary* in 1942. In the last few decades, lexicographical description of phraseology has reaped enormous benefits from advances made in the field of information technology: specifically, 1) the availability of large language corpora has allowed the description of phraseology to become more complete and more precise, and 2) the arrival of MLDs in digital form, first as CD-ROMS and later through the internet, has meant that learner's dictionaries are now in a much better position to deal with the composite, and sometimes complex, nature of phraseology.<sup>1</sup>

---

1 For an overview of the treatment of phraseology in successive editions of MLDs, from the beginnings till the late 1990s, see Cowie 1999 (52-81, and *passim*).

However, although lexicographical description of phraseological phenomena is now generally of a high standard, there are still aspects which could be improved. In this paper I focus specifically on one, pedagogically defined, sub-category of phrases, those which might appear to the learner to be unusual with respect to their lexico-semantic composition.

## 2 Perceived usualness and unusualness in phrasal composition

### 2.1 Usualness

There are a vast number of phrases in the English lexicon, and in many cases there is no noticeable clash, for the foreign language learner, between form and meaning. Even if the learner has never before come across (or never noticed) a particular item, the meaning may still be relatively clear (and, indeed, the learner may be unaware of the phrase's status as a lexical unit). Examples of such phrases are *bank manager*, *blood pressure*, *reading glasses*, and *road sign*. Some phrases may seem a little more unusual from a lexico-semantic perspective (e.g. *plastic money*, *money laundering*, *bottle bank*, *blood orange*, *full house*, *front office*, *fuel rod*), but the relationship between form and meaning should be relatively clear once a dictionary has been consulted. The phrase *blood orange*, for example, is defined in OALD 8 as “a type of orange with red flesh”, and *bottle bank* (British English) is defined in LDOCE 5 as: “a container in the street that you put empty bottles into, so that the glass can be used again”.<sup>2</sup>

Even phrases consisting wholly of a figurative use of the component words will in many cases cause no problems, at least once a dictionary definition has been read. Cross-cultural metaphor may be involved, or else the relationship between physical and figurative meanings may be very evident. Consider, for example, the following phrases and their dictionary explanations:

- (1) *playing with fire* MEDAL 2: doing something dangerous or risky that could cause lots of problems for you – “He knew he was playing with fire by encouraging her attentions.”
- (2) *Out of the frying pan into the fire* OALD 8: from a bad situation to one that is worse.
- (3) *water under the bridge* CALD 4: problems that someone has had in the past that they do not worry about because they happened a long time ago and cannot now be changed – “Yes, we did have our disagreements but that's water under the bridge now.”
- (4) *to have a frog in your throat* CALD 4: to have difficulty in speaking because your throat feels dry and you want to cough.
- (5) *not have a leg to stand on* OALD 8: to be in a position where you are unable to prove sth or explain why something is reasonable – “Without written evidence, we don't have a leg to stand on”.

---

2 Most of the dictionaries cited in the present study are referred to in initialized form (e.g. OALD 8); full bibliographic details may be found in the References.

In cases such as these, it should not be difficult for most learners to connect the literal and lexicalized metaphorical meaning (though with some figurative expressions, recognizing this connection may depend on the linguistic and cultural background of the individual learner).

## 2.2 Unusualness

Alongside phrases such as those mentioned in Section 2.1, there are also phrases which may give rise, *even after* dictionary consultation, to a perception of discrepancy between form and meaning. Some lexicalized figurative phrases are of this sort, for example:

(6) *throw in the towel* COB 5: If you **throw in the towel**, you stop trying to do something because you realize that you cannot succeed – “It seemed as if the police had thrown in the towel and were abandoning the investigation”.

The learner who can appreciate why we say *play with fire* and *water under the bridge*, may well be confused by the phrase *throw in the towel*.<sup>3</sup>

The majority of phrasal verbs could also be described as being compositionally anomalous, especially those in which the verb itself is highly delexicalized, for example *put up with*, *get round sb*, and *take off* (in the sense of “imitate”). A further set of phrases which may appear to the learner as unusual in their form are those which have been called “cranberry collocations” (Moon 1998: 21). This set of phrases is very disparate in nature as regards both form and meaning, but may be grouped together by virtue of the fact that they all include “items that are unique to the string and not found in other collocations” (*ibid*). Many of these items, as Moon points out, “are rare fossil words, or have been borrowed from other languages or varieties” (*ibid*: 78); some of the author’s examples are *run AMOK*, *to and FRO*, and *SLEIGHT of hand*. Another set of anomalous items described in Moon’s study are those which, for one reason or another, are *grammatically* ill-formed. Examples are *be seeing you*, *by and large*, *in brief*, and *put pen to paper*. Moon also mentions phrases which are highly anomalous from a collocational point of view (e.g. *look daggers at sb*).

## 2.3 Which ‘unusual’ phrases to annotate in the MLD

It would be counter-productive to comment on every phrase in which the form-meaning relationship of its component words was notably distant from what would be expected to be a normal juxtaposition of single-word lexical items in modern English. There would be a very large number of items involved, and each explanation would take up space, on the page or on the screen, and perhaps distract from more important information. Furthermore, what is anomalous may go largely or wholly un-

---

3 In the case of figurative phrases such as *throw in the towel*, learners will have an advantage if there is an analogous phrase in their own mother tongue. For examples and discussion of such cross-language phrasal pairs and sets, see Piirainen (2012) and, from the perspective of pedagogical lexicography, Coffey (2002).

noticed by the language learner, especially where relatively frequent words are concerned and intuitable meanings. There would be little point in drawing attention, for example, to the fact that, phrases such as *in brief*, *in general* and *at last* are, in effect, PREPOSITION + ADJECTIVE sequences, or that in the phrase *in question* there is no article before the noun. Nor would it be necessary to comment on phrases such as the above-mentioned *blood orange* and *playing with fire*.

Exactly which phrases (or which types of phrase) it would be useful to comment on from the point of view of their composition may be best ascertained through specifically designed dictionary-user studies. However, it is perhaps not unreasonable to suppose that the following two general types of phrase would be strong candidates. Firstly, idiomatic expressions, of one sort or another, in which the motivation behind the idiom's form is wholly or partially hidden from the learner. Examples are the already mentioned *throw in the towel* and the phrase (*as*) *mad as a hatter*. Secondly, phrases which include a word not normally used as a single-word lexical item in present-day English, for example the word form "sleight" in the phrase *sleight of hand*. In actual fact, some phraseological items would fit perfectly well into both categories, since idiomatic phrases sometimes include fossilized word forms or word meanings. An example is the phrase (*to buy*) *a pig in a poke* which is both a semantically obscure idiomatic phrase and includes the lexical fossil *poke*.<sup>4</sup>

With both types of phrase, some sort of explanation of unusual form will address the learner's curiosity (even if it will not always be possible to fully satisfy that curiosity). In addition, in the case of idiomatic phrases, an explanation of origin (and therefore of phrasal composition) may help the learner to remember a given expression. In the case of lexical fossils, explicit comments on unusual words will help ensure that learners realize that the words in question are (virtually) phrase-bound and cannot normally be used in other ways.

In order to obtain an overall picture of current practice in MLDS, the following dictionaries were examined: a) print dictionaries: CALD 4, COB 5, LDOCE 5, MEDAL 2, MW, OALD 8; b) online dictionaries: e-CALD, e-LDOCE, e-MEDAL, e-MW, e-OALD. In actual fact, virtually no differences in content were found between the print and online versions of the dictionaries examined, and below I will often cite from, or refer to, the print dictionaries.

### 3 Idiomatic expressions

The phraseology of English, as is well known, is very complex to describe, and specific categories that we identify (or create) can themselves be very varied in nature. The examples of "idioms" described in this section are no exception to this; there are differences in grammatical role, type of meaning, and the relationship between phrasal meaning and the meaning of individual parts.

---

4 For the sake of precision, it should be noted that the word *poke*, which here has the sense of "bag", is still used in some regional varieties.

In order to see whether dictionaries offer the learner any extra guidance with regard to the composition of partially or wholly idiomatic phrases, a total of 37 items were looked for, consisting of two different (pedagogically speaking) groups. The first group was composed of phrases which, it was considered, would definitely benefit from some explicit comment. The second group consisted of items which would not necessarily need any explanation, assuming they were located at the right headword or the appropriate sense of a given headword. The actual composition of these two groups was adjusted slightly once dictionaries had been consulted and current lexicographical data observed.

With regard to the inclusion of the phrases in the dictionaries examined in the present study, 25 out of the 37 phrases were present in all MLDS, and 34 in all dictionaries but one. One dictionary (COB 5) had significantly fewer phrases than the others, with eleven of the phrases being absent.<sup>5</sup>

### 3.1 Opaque idiomatic phrases

The following are the phrases in the first group (27 items):

*a red herring; a feather in your cap; an ivory tower; a gravy train; a fifth column; Bob's your uncle; Coals to Newcastle; the penny dropped; the gloves are off; be the bee's knees; send sb to Coventry; kick the bucket; be grist to/for the mill; play gooseberry; pass the buck; face the music; run the gauntlet; throw down the gauntlet; give sb a wide berth; throw in the towel; pull someone's leg; know etc the ropes; draw/get the short straw; live the life of Riley; lock, stock and barrel; hook, line and sinker; as mad as a hatter.*

As regards the form-meaning relationship of these phrases, it was found that, overall, there was very little comment in the dictionaries examined, and that some MLDS had no comment at all. This comes as no surprise, since explanation of this sort would involve introducing the historical dimension to language and the latter has never been a priority in the MLD; indeed, it has usually been completely absent.

As far as I am aware, few writers dealing with pedagogical lexicography have discussed or commented on the absence of historical in data in MLDS. One exception is Ilson (1983), who points out the potential usefulness of providing at least some etymological information, and includes mention of its relevance to phraseology. Much more recently, Boers (2007) points out the usefulness, for comprehension, of being made aware of the origin of idioms. Exemplifying with the expression *show sb the ropes*, he points out that, "It would help if you knew that the expression was originally used in the context of sailing, where an experienced sailor had to show a novice how to handle the ropes on a boat". In the context of the present article, this quotation has added significance since it comes from a short artic-

---

5 It might also be mentioned here that all items but two are recorded in the *Collins COBUILD Dictionary of Idioms* (1995) – the exceptions are *fifth column* and *Beauty is in the eye of the beholder*. The latter is not actually an archetypal "idiom", since it is fairly transparent in nature. However, I have included it because the word *beholder* will render the phrase partially opaque to many learners and also because of the relatively unusual phrase "in the eye of".

le entitled *Understanding Idioms*, which is part of the Language Awareness section of MEDAL 2 (pp. LA2-3).<sup>6</sup>

Of the dictionaries investigated in the present study, the only one which has at least a few explanations of idioms is OALD 8. Two examples are:<sup>7</sup>

- (7) OALD 8 **COAL** [...] **carry, take, etc. coals to Newcastle** [UK] to take goods to a place where there are already plenty of them; to supply sth where it is not needed. ORIGIN: Newcastle-upon-Tyne, in the north of England, was once an important coal-mining centre.
- (8) OALD 8 **RED HERRING** an unimportant fact, idea, event, etc, that takes people's attention away from the important ones. ORIGIN: From the custom of using the smell of a smoked, dried herring (which was red) to train dogs to hunt.

Examples of definitions with no explanation of phrasal composition are the following:

- (9) CALD 4 **MAD** [...] (as) *mad as a hatter / March hare* extremely silly or stupid.
- (10) COB 5 **GAUNTLET** [...] PHRASE If you **run the gauntlet**, you go through an unpleasant experience in which a lot of people criticize or attack you – “The trucks tried to drive to the British base, running the gauntlet of marauding bands of gunmen.”

There are differing degrees of difficulty in understanding the connection between form and meaning. For example, whereas *run the gauntlet* will be highly obscure to the uninitiated learner, form and meaning should be much more connectable in the case of the following phrase description, even though there is no explicit explanation of the type seen previously in example (7):

- (11) CALD 4 **COAL** [...] **carry / take coals to Newcastle** [UK] to supply something to a place or person that already has a lot of that particular thing – “Exporting pine to Scandinavia seems a bit like carrying coals to Newcastle.”

The phrase *ivory tower* may also be relatively clear after reading definition and example, and may not attract too much curiosity on the part of the dictionary user. However, since the phrase has a well documented origin, it might be useful to include it.

Another situation worth commenting on is that wherein a phrase is located at the entry for a single-word headword, but the headword itself has no description. This happens in all dictionaries, for example, with the word *grist*, found in the phrase *be grist to/for the mill*. The following is one such entry:

---

6 Data regarding phrase origin is found, by contrast, in some dictionaries devoted to idioms, notably ODCIE (1983) and LDEI (1979), (whereas the previously mentioned *Collins COBUILD Dictionary of Idioms* does not include explanatory data of this sort). Mainstream language teaching publications have also shown little interest in the historical dimension of phraseology, though there are some applied linguists who recognize the potential of etymological explanation; see, for example, Boers et al (2004) and Boers et al (2007).

7 From this point of the text onwards, in numbered examples I will use **bold SMALL CAPS** to indicate the headwords. The square brackets which sometimes follow this ([...]) indicate that there is other lexical description before that of the phrase I am discussing.



(12) COB **GRIST** PHRASE If you say that something is **grist to the mill**, you mean that it is useful for a particular purpose or helps support someone's point of view.

The “problem” with *grist* is that it is not commonly found in modern English outside of this phrase. This does not mean, however, that it has become a lexical fossil (except, perhaps, from a language teaching point of view). A short explanation of its meaning (in relation to this phrase) would be useful, together with an indication that it is usually only found in this phrase and variants thereof. The same happens with *hook, line and sinker* in COB 5, which is listed at the headword *sinker*, which, however, has no single-word explanation.

The phrase *Bob's your uncle*, present in five MLDs, is a similar case, with the expression being recorded in all dictionaries at the unexplained headword *Bob* (with a capital letter, and thus distinguished from entries with the headword *bob*). Actually, since *Bob's your uncle* is the only phrase at the headword *Bob* in the various dictionaries, it might be simpler and neater to have the saying itself as the headword, in the same way as many noun phrases and other multiword items regularly constitute headwords.

Another phrase worth commenting on is the *bee's knees*. The five dictionaries which record this phrase give no explanation for its form. If they did, regardless of whether or not they were in a position to explain its meaning from the point of view of its composition, it would be useful to underline the rhyme in the phrase, which is almost certainly at least part of the motivation behind its form. It is worth noting in this respect that sound repetition might well help memorization, and the very act of pointing out the (in this case) rhyme, may be of added benefit to learners (For discussion of this topic, see Boers & Lindstromberg 2005 and Lindstromberg & Boers 2008).

If a dictionary does adopt the policy of commenting on phrases of the type being considered here, it will sometimes be necessary to say that the phrase is of “unknown” or “uncertain” origin. In the current state of our knowledge of phrasal origins, this would happen, for example, with *kick the bucket*, *Bob's your uncle*, *face the music*, and *pull someone's leg*. Where there are several contesting theories as to the origin of a phrase, it would probably make little sense to go into details, and be best to label the phrase as being “of uncertain origin”. It could be argued that it is of little help to the learner to read that the origin of a phrase is “unknown”, but I think that this is more user-friendly than keeping silent.

### 3.2 Potentially comprehensible idiomatic phrases

The following are the phrases in the second group (10 items):

*Beauty is in the eye of the beholder; go against the grain; jump the gun; bury the hatchet; play second fiddle; pull out the stops; make a mountain out of a molehill; green about the gills; cheek by jowl; like a bolt from the blue.*<sup>8</sup>

In each of these phrases, there is a word which may present problems for an understanding of the motivation behind the form of the phrase; of course, whether or not this is actually a problem will depend on the individual learner. The words in question are: *beholder, grain, jump/gun, hatchet, fiddle, stops, molehill, gills, jowl, and bolt*. In the case of *beholder, hatchet, molehill* and *jowl*, dictionaries only record one sense for each of the words, and if the phrase was explained at the same point of the dictionary at which the single word is explained, then the learner should be in a position to appreciate the motivation of phrasal form. This is what happens, for example, in the case of the following definition:

(13) COB5 **MOLEHILL** (1) A **molehill** is a small pile of earth made by a mole digging a tunnel; (2) If you say that someone is **making a mountain out of a molehill**, you are critical of them for making an unimportant fact or difficulty seem like a serious one.

It is to be noted also that the entry for the word *mole* itself is very close by in the text. So, we have *making a molehill* defined very close to *molehill*, which is itself defined very close to *mole*. And the relationship between the physical and figurative meanings should also be fairly clear to learners. Whereas this may seem a fairly easy case, and the COBUILD treatment of the phrase a useful one for learners, only one other dictionary locates this phrase at the headword **molehill**, the other four placing it at **mountain**. A similar situation is found with the phrase *cheek by jowl*, present in two dictionaries at **jowl**, and in four at **cheek**. We may contrast the following two descriptions:

(14) MEDAL 2 **JOWL** The lower part of your cheek, especially if the skin hangs down and covers your jaw. PHRASE **cheek by jowl** If two or more people or things are cheek by jowl, they are very close to each other.

(15) LDOCE 5 **CHEEK** [...] **cheek by jowl (with sb/sth)** very close to someone or something else – “An expensive French restaurant cheek by jowl with a cheap clothes shop.”

One of the reasons that we find dictionary explanations such as that in (15), is the fact that in some MLDS there is a standard rule whereby a phrase should be placed at the headword for the first content word in a phrase. I believe that there are pros and cons to having rules of this type. One important factor which is sometimes overlooked, is that while lexicographers and language teachers have a clear idea (most of the time) about what is, and what is not, a lexical phrase, the average language learner is not as linguistically sophisticated. A learner who reads, for example, of “An expensive French restaurant cheek by jowl with a cheap clothes shop” may view any eventual comprehension problem

<sup>8</sup> The verbal idiom *pull out (all) the stops* is also an example given by van der Meer (1996) in an article dealing with the MLD treatment of figurative meaning more generally (i.e. not just with reference to phraseology).

only in terms of not knowing the word *jowl* – everything else has a familiar look to it, so the problem must be with this word in particular.

Let us turn now to cases where the problem lies not with a possibly unknown word form, but with knowing which meaning of a word is involved (which may or may not be one that is known to the learner). Consider, for example, the phrase *go against the grain*. The relevant sense of the word *grain* is explained in CALD 4 in the following way:

(16) CALD 4 **GRAIN** [...] **the grain** the natural patterns of lines in the surface of wood or cloth – “to cut something along/against the grain”.

It is interesting to note that the example phrase includes the phrase “against the grain”. However, the meaning is the physical one, not that of the figurative idiom. From the point of view of understanding the motivation behind the idiomatic phrase, this would have been the ideal place to record the figurative expression *go against the grain*; however, it is not recorded at this point in the dictionary. In CALD 4, as in all the other dictionaries, the phrase is presented at the entry for *grain*, but as a phrase with no direct connection to any of the single-word senses of *grain*. A similar situation was found for the phrase *pull out all the stops*, (which comes from the idea of an organist pulling out all the organ stops in order to increase the volume). This use of the word *stop* is present in all the dictionaries, but none of them makes a direct association between word meaning and phrasal meaning. Also, three dictionaries place the phrase at the headword *stop* and three at *pull*.

An example of good dictionary treatment is the explanation of *like a bolt from the blue* in MW:

(17) MW **BOLT** a bright line of light that appears in the sky during a storm; a flash of lightning *a bolt of lightning = a lightning bolt* — often used figuratively in the phrases **a bolt from the blue** and **a bolt out of the blue** – “The news of his firing came as/like a bolt from the blue.” [= like a bolt of lightning from the sky; it was surprising and unexpected]

The phrase *green about the gills* creates particular problems, since, in order to appreciate the form of the phrase, it may be necessary to see both a definition of *gills* and to understand which sense of *green* is involved. MW satisfies the second need very well:

(18) MW **GREEN** 5. [informal] having a pale or sick appearance – “Our flight hit some turbulence, and half the passengers started turning green.” — often used in the phrase **green around/about the gills** – “The passengers were looking green around the gills.”

Here, we not only find the phrase at the right sense of *green*, but also see the specification “often used in the phrase ...”. The problem still remains, however, of the basic meaning of *gills*, which the learner would have to look up separately.

As has been argued, from the point of view of appreciating the original logic of the phrases, it would make sense (wherever phrases are explained at the entry for one of the component words, as opposed to having an entry of their own), for the phrase to be explained close by the less commonly known

word. However, overall it was found that there were few dictionary entries in which the phrases examined were placed at an entry or subsense which would allow motivation of phrasal form to be understood (without explicit commentary). In all, 56 out of the 60 possible phrasal entries were present in the MLDS (10 entries x 6 dictionaries), but in only 12 cases were phrases explained at the appropriate point in the text. In the case of the e-dictionaries, and where a phrase is explained at the “wrong” entry, there is slightly less of a problem, since the reader can go quickly from one entry to another. But it is still a problem, in that the two definitions (of the single word and the phrase) do not appear on the screen together.

## 4 Other phrases which include unusual word forms or word senses

The second general phrase type for which I suggest it would be useful to have comment on phrasal composition are phrases which include a word not normally used on its own in modern English. In this case, there are two main types of information which the dictionary could provide. The first is, quite simply, the fact that the word in question (or that particular meaning, where homonymy is involved) is normally found just in the phrase indicated. The second data type is the explanation of the unusual word (what sort of word it is, and why it isn’t used elsewhere).

There are a number of different reasons for the presence of phraseologically-bound word forms and meanings, and the specific reason will at least in part determine what the dictionary should say about the phrase. From an investigation of many different dictionary entries for lexical phrases, it would appear that the majority of such words are lexical fossils of one sort or another, and it is these that I will look at first.<sup>9</sup>

### 4.1 Lexical fossils

There are word forms which used to be freer lexical items but which are now found above all “fossilized” in lexical phrases. Some are found in the types of phrase discussed in Section 3. The word *poke* as in (*to buy*) *pig in a poke* has already been mentioned in this respect; other examples are the words *lurch* and *truck* found in, respectively, *leave sb in the lurch*, and *have no truck with sb/sth*. Examples in phrases which are less likely to be referred to as “idioms” are the already mentioned “fro” (*to and fro*) and “sleight” (*sleight of hand*), and further examples can be seen in the following phrases: *take UMBRAGE*, *the whys and WHEREFORES*, *a DAB hand*, *by DINT of*, and *in fine FETTL*.

---

9 The term “fossil”, in a linguistic sense, is defined in the OED (3rd edn) as “A word or other linguistic form which has become obsolete except in isolated regions or in set phrases, idioms, or collocations”. For a discussion of the notion of “lexical fossil”, see Coffey 2013.

Some fossils are close in form to related words in modern English; an example is *afield*, used in phrases such as *far afield* and *farther afield*. Fossils may also have exactly the same form as a modern word, and thus be less noticeable. This applies to the already mentioned *poke* and *truck*. Other examples are the word forms “let” and “hue”, as found in *without let or hindrance* and *hue and cry*. Grammatical word category may also be of relevance: the word *pale* is not usually found as a noun in modern English, but this usage (and relevant meaning) can still be seen in the phrase *beyond the pale*.

The fact that the words or word forms are not normally used in modern English as single-word lexical items has a number of consequences for language learners. Firstly, the learner may be puzzled as to the presence of a word in a phrase – why do we say *a pig in a poke*: what is a “poke” in this case? Secondly, the learner may feel a sense of frustration at not knowing anything about a word. The phrase *in high dudgeon*, for example, appears most frequently in MLDS under the headword **dudgeon**, implying, therefore, that the latter exists as a free-standing word, which, however, it doesn't. Thirdly, the learner may remember the unusual word (precisely because it is unusual), and later use it in an inappropriate way, for example by taking the word “umbrage” out of its usual phrasal environment (*take umbrage*).

Taken as a whole, the MLDS examined do not have much to say about the composition of items such as the above. The following are some examples of presentation:

- (19) MEDAL 2 **POKE** a quick push with your finger or a pointed object. [...] **a pig in a poke** something that you have bought without seeing it first.
- (20) OALD 4 **WHY** [...] *noun* **the whys and (the) wherefores** the reasons for sth – “I had no intention of going into the whys and the wherefores of the situation.”
- (21) E-LDOCE **FETTLE** *noun* **in fine/good fettle** [old-fashioned] healthy or working properly.
- (22) CALD 4 **DINT** *noun* **by dint of sth** [formal] as a result of sth – “She got what she wanted by dint of pleading and threatening.”
- (23) COB 5 **WEND** PHRASE If you **wend** your **way** in a particular direction, you walk, especially slowly, casually, or carefully, in that direction [LITERARY] – “Sleepy-eyed commuters were wending their way to work.”
- (24) MW **PALE** *noun* **beyond the pale** offensive or unacceptable – “conduct that was beyond the pale”.

With regard to the lexical fossils within phrases, the most useful feature I have found in the dictionaries examined is wording which is used sometimes in MW, and an example of which involves the word *umbrage*:

- (25) MW **UMBRAGE** a feeling of being offended by what someone has said or done — usually used in the phrase **take umbrage** – “I imagine some people will take umbrage [= will be offended] when they hear the quote.”

The important words here are “usually used in the phrase ...”. This type of comment is not found for all fossils in MW (nor is it found only for fossils), but it is a step in the right direction.<sup>10</sup>

How much information should be given for a fossilized phrase will depend on the word in question. Often, it will be enough to indicate that it is a fossil and therefore probably only found in that particular phrase, but sometimes other information could be useful. For example, in the case of *WEND one's way*, it may be of interest to the learner to know that the verb *wend* is historically related to the word form *went*, now considered to be part of the verb *go*; and in the case of *without further/more ADO*, reference could be made to the play title *Much ado about nothing*.

## 4.2 Other unusual words in phrases

I will now briefly mention other types of phrase which may strike the learner because of one or more unusual word forms. First, I list a number of example descriptions in MLDS, and thereafter I comment on the features I wish to point out.

(26) COB 5 **BOB** [...] PHRASE **Bits and bobs** are small objects or parts of something [mainly British, informal] - “The microscope contains a few hundred dollars-worth of electronic bits and bobs.”

(27) MW **ODDS AND SODS** [UK, informal] = **odds and ends** - “The store sells art supplies and other odds and sods.”

(28) e-LDOCE **KITH AND KIN** *noun* [old-fashioned] family and friends.

(29) e-LDOCE **LOVEY-DOVEY** *adj* [informal] behaviour that is lovey-dovey is too romantic - “a lovey-dovey phone call”.

(30) MEDAL 2 **CHIT-CHAT** *noun* [informal] friendly conversation about things that are not very important.

(31) CALD 4 **BUTCHER** *noun* [...] **have a butcher's** [UK, old-fashioned, slang] to look at something - “Let's have a butcher's at your present then.”

In (26) can be seen an example of a word (*bob*, or rather *bobs*) which does not exist on its own with this type of meaning. Nor is it known to be a lexical fossil. The phrase is recorded in the oed (2nd edn) at the entry for *bit*, and there is no meaning of *bob* which might relate to this phrase. It may be presumed, therefore, that the phrase was coined for its alliterative effect. Example (27) is similar but involves rhyme rather than alliteration. The second phrase indicated, *odds and ends*, also involves sound repetition, since the two parts of the binomial are both monosyllabic, begin with a vowel, and end with the spelling/sound *-ds*. In (28), the phrase *kith and kin* combines alliteration with the presence of a lexical fossil (*kith*). Examples (29) and (30) also exhibit, respectively, rhyme and alliteration. However,

---

10 Information about some lexical fossils may also be found in the brief etymological notes on individual words (“word origin”) in e-LDOCE and the CD-ROM version of OALD 8; however, there is no explicit statement about the fact that the words in question are fossils nor that they are only usually found in the phrases indicated.

they are formally different from the preceding examples, in that they are reduplicatives, and are written, usually, as single, hyphenated words. In the case of *lovey-dovey*, both morphological parts are easily associated with other words (*love* and *dove*), though the word *lovey* also exists. The word *chit-chat* is different, in that only *chat* exists with a relevant meaning.

Examples (26) to (30), then, all involve sound repetition of some sort, in addition to the presence of unusual words or morphemes. Example (31) is a little different, in that there is no apparent sound repetition, just a word sense which seems unconnected to any of the various meanings of the headword *butcher*. However, rhyme is involved, indirectly, as can be seen in the explanation of the same phrase in MEDAL:

(32) MEDAL 2 **BUTCHER** *noun* [...] **have/take a butcher's** [UK, informal] to have a look at something -  
 From *butcher's hook*, rhyming slang for 'look'.

Generally speaking, there is little comment in current MLDS on the types of phrase mentioned in this section. And whereas there are no more than a handful of word usages dependent on rhyming slang, there are many more items in which sound repetition has a fundamental role to play. It would not be difficult to point this out in the dictionary, and may help to satisfy the reader's curiosity as to the origin of the phrase.

### 4.3 Grouping phrases which share certain features

As part of the process of improving learners' general knowledge of the English lexicon, it would be useful if phrases which share a certain feature or features were brought together. This would be much easier in the e-dictionary, where space is not a problem and where the user could click on a link to find other examples of the phenomenon being looked at in a particular entry. Some of the phrasal types mentioned in 4.1 and 4.2 could be brought together in this way. In the case of fossils, since there are quite a large number involved, those with a further common characteristic could be brought together, for example those in coordinated phrases involving sound repetition (e.g. *kith and kin*, *the whys and wherefores*). Some idiomatic phrases could also be interlinked, for example with reference to the area of "original" meaning of phrases (e.g. "SHIPS AND THE SEA") or through the form of the phrases (e.g. *as [mad] as a [hatter]*).

Using the terminology employed by Apresjan (1993: 80), interconnections of this type would allow the dictionary to enhance its information on "lexicographic types", while at the same time, the improvement of data regarding phrasal composition would also allow each individual "lexicographic portrait" (*ibid*: 86) to be enriched.



## 5 Conclusions

English lexical phrases come in many shapes and sizes, and some of these shapes and sizes are dependent on factors which are not obvious to many present-day native speakers, let alone learners of English as a foreign language. However, whereas lexico-phraseological oddities cause no problems for native speakers and also usually go largely unnoticed, the same cannot be said for language learners, who tend to be more aware of form and have to reconcile it with meaning. Whereas I believe it is generally a positive asset that learners' dictionaries do not dwell too much on the compositional nature of lexical phrases, I think that they should do so when appropriate. In the types of phrase described in the present paper, this also involves bringing in the historical dimension of language, which has perhaps been left out in the cold too long. Given, especially, the enormous potential of web-based dictionaries, this should be perfectly possible.

## 6 References

### Print dictionaries

- CALD 4. *Cambridge Advanced Learner's Dictionary*, 4th edn (2013). Cambridge: Cambridge University Press.
- COB 5. *Collins Cobuild Advanced Dictionary of English*, 5th edn (2006). Glasgow: HarperCollins.
- Collins Cobuild Dictionary of Idioms* (1995). London: HarperCollins.
- Hornby, A.S., Gatenby, E.V. & Wakefield, H. (1942). *Idiomatic and Syntactic English Dictionary*. Tokyo: Kaitakusha. [later, 1948, published by Oxford University Press as *A Learner's Dictionary of Current English*, and subsequently, 1952, retitled *The Advanced Learner's Dictionary of Current English*].
- LDOCE 5. *Longman Dictionary of Contemporary English*, 5<sup>th</sup> edn (2009). Harlow: Longman.
- LDEI. *Longman Dictionary of English Idioms* (1979). Harlow: Longman.
- MEDAL2. *The Macmillan English Dictionary for Advanced Learners*, 2nd edn (2007). Oxford: Macmillan Education.
- MW. *Merriam-Webster's Advanced Learner's Dictionary* (2008). Springfield, Massachusetts: Merriam-Webster Inc.
- OALD 8. *Oxford Advanced Learner's Dictionary*, 8<sup>th</sup> edn (2010). Oxford: Oxford University Press.
- ODCIE. Cowie, A.P., Mackin, R. & McCaig, I. R. (1983). *Oxford Dictionary of Current Idiomatic English, Volume 2: Phrase, Clause & Sentence Idioms*. Oxford: Oxford University Press.

### On-line dictionaries

- e-CALD. Accessed at: <http://dictionary.cambridge.org/dictionary>
- e-LDOCE. Accessed at: <http://ldoce.longmandictionariesonline.com>
- e-MEDAL. Accessed at: <http://www.macmillandictionary.com>
- e-MW. Accessed at: <http://www.learnersdictionary.com>
- e-OALD. Accessed at: <http://oald8.oxfordlearnersdictionaries.com/dictionary>, and at: <http://www.oxfordlearnersdictionaries.com/>
- OED. *The Oxford English Dictionary*. Oxford: Oxford University Press. Accessed at: <http://www.oed.com>
- All online dictionaries were accessed at various times during the period September 2013 – April 2014



### Other literature

- Apresjan, J. D. (1993). Systematic Lexicography as a Basis of Dictionary-making. In *Dictionaries*, 14 (1992/93), pp. 79-87.
- Boers, F. (2007). Understanding Idioms. In the *Macmillan English Dictionary for Advanced Learners*, 2nd edn, pp. LA2-3. This article is also available in the *MED Magazine*, Issue 49, February 2008. Accessed at <http://www.macmillandictionaries.com/MED-Magazine/February2008/49-LA-Idioms-Print.htm> [02/04/2014]
- Boers, F., Demecheleer, M. & Eyckmans, J. (2004). Etymological elaboration as a strategy for learning idioms. In P. Bogaards, B. Laufer (eds.) *Vocabulary in a Second Language: Selection, acquisition, and testing*. Amsterdam / Philadelphia: John Benjamins, pp. 54-78.
- Boers, F. Eyckmans, J. & Stengers, H. (2007). Presenting figurative idioms with a touch of etymology: more than mere mnemonics? In *Language Teaching Research*, 11 (1), pp. 43-62.
- Boers, F. & Lindstromberg, S. (2005). Finding ways to make phrase-learning feasible: The mnemonic effect of alliteration. In *System*, 33, pp. 225-238.
- Coffey, S. (2002). Interlingual Phrasal Friends as a Resource for Second Language Learning: Outline of a lexicographical project. In A. Braasch, C. Povlsen (eds.) *Proceedings of the Tenth EURALEX International Congress*. Copenhagen: Center for Sprogteknologi, pp. 315-323.
- Coffey, S. (2013). Lexical Fossils in Present-Day English: Describing and Delimiting the Phenomenon. In R. W. McConchie, T. Juvonen, M. Kaunisto, M. Nevala & J. Tyrkkö (eds.) *Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis (HEL-LEX 3)*. Somerville MA: Cascadilla Proceedings Project, pp. 47-53.
- Cowie, A. P. (1999). *English Dictionaries for Foreign Learners: A History*. Oxford: Oxford University Press.
- Ilson, R. (1983). Etymological information: can it help our students? In *ELT Journal*, 37(1), pp. 76-82.
- Lindstromberg, S. & Boers, F. (2008). Phonemic repetition and the learning of lexical chunks: The power of assonance. In *System*, 36, pp. 423-436.
- Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.
- Piirainen, E. (2012). *Widespread Idioms in Europe and Beyond: Toward a Lexicon of Common Figurative Units*. New York: Peter Lang.
- van der Meer, G. (1996). The Treatment of Figurative Meanings in the English Learner's Dictionary (OALD, LDOCE, CC and CIDE). In M. Gellerstam, J. Järborg, S-G. Malmgren, K. Norén, L. Rogström, C. Røjder Pappmehl (eds.) *Euralex '96 Proceedings*. Göteborg: Department of Swedish, Göteborg University, pp. 423-429.



# Harvesting from One's Own Field: A Study in Collocational Resonance

Janet DeCesaris, Geoffrey Williams  
Universitat Pompeu Fabra, Université de Bretagne-Sud  
janet.decesaris@upf.edu, geoffrey.williams@univ-ubs.fr

## Abstract

This paper presents an initial study in the collocational resonance of three words: *field*, *champ*, and *campo*. *Field* and *champ/campo* do not share the same etymology, yet *field* displays sense extension that in some cases is parallel to that displayed by *champ* and *campo*. Collocational resonance posits that meaning associated with one context of use may be activated by speakers in another context, even though the original meaning context may fall into disuse in the language. To determine collocational resonance, the study considers both the historical development of senses as represented in dictionaries and the current behaviour of these words as represented in dictionaries and corpora. It is suggested that an approach to word meaning including prototypes and resonance can improve the representation of polysemy in dictionaries.

**Keywords:** collocational resonance; polysemy, metaphor

## 1 Introduction

Both homonymy and polysemy have long posed problems for lexicography. Whilst the former can be relatively easily handled from a synchronic perspective, this is not so in diachrony. Thus, a historical dictionary may well have to lump senses together in a single entry when a dictionary of the contemporary language would split them into different entries, as historically, what is now justifiably analysed as a homonym, might well be a polyseme. Even in dictionaries of contemporary language, however, the representation of polysemy is problematic, as the requirement for discrete senses denies the obvious continuum between senses. Obviously, these issues are rendered even more complex as language moves from a so-called literal sense to a figurative one through metaphor. In such movements, to adopt the terminology of Hanks (2013), the exploitation of the norm becomes the norm itself, which will, in turn, inevitably be exploited. Four issues thus require a tool for their joint management: etymology, homonymy, polysemy and metaphor.

Hanks (2000) has already proposed a solution to polysemy by proposing lexicographical prototypes, a series of simple propositions that are activated in each individual sense. Hanks (2005) also proposed a means of handling metaphor by proposing collocational resonance. At the same event, Williams (2008) independently also proposed collocational resonance, albeit approaching the data from a more

inter-textual angle. Both researchers are heavily influenced by John Sinclair, so it is not altogether surprising that, working independently, they should reach very similar conclusions and a selfsame term. Since these landmark presentations, work has gone on to combine lexicographical prototypes and collocational resonance in order to handle the variation of meaning potentials (Hanks 2013) across time and between languages (Williams 2012).

In this paper, we shall look at the word *field* and its equivalents in French, *champ*, and Spanish, *campo* as the beginnings of a longer study into the norms and exploitations of the agricultural metaphor in contemporary use and at the treatment of these words in dictionaries. In this study, we shall concentrate on the dictionary definitions that provide the initial prototype and the current usages of the word as found in corpora.

## 2 Collocational Resonance

The idea behind collocational resonance is that over time words have been attributed different meanings in different contexts. We define meaning as referring to a particular usage that can be defined through a series of propositions that in a stable generalised form provide what can be recognized as a dictionary sense. Meanings are created within a given textual environment within a given context of culture. Thus, meaning elements are shared within a society, with general agreement as to broad senses. As society and contexts of culture evolve, so do the meaning potentials of words. There can be a slight contextual variation, or a more radical one as an exploitation becomes a norm.

Collocational resonance posits that although the earlier senses attributed to a word may be lost, we live in a world of cumulated knowledge so that some meaning attributes may subsist, consciously or unconsciously, thereby colouring a user's use of a word. When there is a deliberate exploitation, for example through active metaphor, the user is quite conscious of the exploitation being made of meaning attributes, and the reader or hearer is expected to share this explicit knowledge. Collocational resonance eschews any so-called cognitive knowledge and posits that while a dead metaphor is simply dead, the knowledge of earlier usage may be found in the unconscious as this knowledge comes from an encounter with the word in a different meaning context, either in a text, a dictionary or through the educational process. The unconscious aspect of meaning attribute carry-over has been demonstrated by Williams (2012) in the case of biologists using rather Lamarckian terms when referring to neo-Darwinian concepts. This variation need not be diachronic; resonance can equally well show variations between general and specialised usage of language.

The aim of studies in collocational resonance, then, is to make meaning variation explicit over time or area of expertise and to show what meaning attributes may remain active. This has been shown for certain verbs used in the sciences, such as *probe*, and also for more general words such as *culture*. The word *culture* is an interesting case from a cross-linguistic perspective, as it has a common root in all Romance languages and has developed through metaphor in a similar way in all of them. Neverthe-

less, the earlier, literal meaning of *culture* ('the cultivation of soil; tillage'<sup>1</sup>) has remained active in some languages, but has virtually disappeared from the use of *culture* in English, although it is clearly present in the use of the derived word *agriculture*. Collocational resonance can be traced using lexicographical prototypes built using a mixture of sources such as historical dictionaries as the *Oxford English Dictionary*, earlier dictionaries as the *Vocabolario* of the Accademia della Crusca, 17<sup>th</sup> century French dictionaries, 17<sup>th</sup> century Spanish dictionaries, the Spanish Royal Academy's many editions of the *Diccionario de la lengua castellana* (which would become the *Diccionario de la lengua española* starting with the 1925 edition) and diachronic corpora, when available. Collocational resonance may be demonstrated by showing variations in the collocational networks of lexical items: as the meanings of a word change, so do the collocational patterns associated with the word in question.

When prototypes are used to show variation of meaning potentials over time, contexts or languages, there is no privilege interlanguage acting as a translation hub. The prototypes are there to show meaning potentials and can be started in any language, the aim being to find translatable comparable units in other languages to see what potentials are activated. If here we are starting with the English, it is only because of the excellence of the *Oxford English Dictionary* as a starting point for looking at earlier usage. For French, we have used the *Trésor de la Langue Française informatisé* and the *Dictionnaire Historique de la Langue Française* from Le Robert and for Spanish the Spanish Royal Academy's *Nuevo tesoro lexicográfico de la lengua española*. In this study, we look at words with no etymological relation at all, *field* and *campo/champ*, but which are generally accepted as translation equivalents. In all cases, current usage is analysed making use of the Sketch Engine<sup>®</sup> *tenten* series of WaC corpora.

### 3 *Field*: From ploughing the soil to ploughing through data

Whether it be through culture or through units of land, agricultural metaphors are rife in that our modern urbanised societies could not exist without the organised production of food. The interest of the words under study lies in their totally different etymological origins, which signifies a potentially major difference in resonance. The etymology of *field* is unclear, but by the early Middle Ages it has taken on the notion of open land rather than woodland. In the Romance languages, it is easier to trace back to a Latin source that distinguishes plain and mountain. By the time of the *Vocabolario*, *campo* as being an agricultural area in which seeds are sown, but also a wide variety of other senses, including that of battlefield. Moving forward to the *Dictionnaire Universel* of Antoine Furetière (1690), *champ* has as its first sense an area of ploughed land. Thus, both Italian and French sources confirm an agricultural meaning of cultivated land as opposed to pasture land. Neither has any clear boundaries. Furetière also gives a large number of non-agricultural terminological uses, from heraldry to comb making.

---

1 Paraphrase of sense I.1.a. of culture in the Oxford English Dictionary and also sense 4 of culture in The American Heritage Dictionary of the English Language.

In Spanish, the situation is somewhat different. Covarrubias (1611) lists the idea of flat land capable of being cultivated or an enclosed area used for farm animals as the first sense of *campo*. The entry for *campo* in the *Diccionario de autoridades* (1729, for letter C) is quite long, comprising several subsenses for the word in addition to several lexicalised phrases (such as *campo de batalla* ‘battlefield’ or *hombre de campo* ‘man who works in fields’). The first sense defines *campo* as a wide, open plain that is outside a populated area; the notion of lying outside a populated area is quite prevalent throughout the twenty-two editions of the Spanish Royal Academy’s dictionary and is still present in the first sense given for the word in the current dictionary (“*Terreno extenso fuera de poblado*” ‘Large piece of land outside a populated area’). The second sense of *campo* in the *Diccionario de autoridades* is described as metaphorical and defines *campo* as the space or period including the whole of something, and the third sense of *campo* includes the idea of cultivated area (thus, the examples *están buenos los campos* ‘the fields are fine’, *los campos están perdidos* ‘the fields are lost’, and *buen año para los campos* ‘good year for the fields’). Those three senses would be maintained with little variation as the first three senses listed until the Academy’s dictionary of 1837, in which the idea of plain as opposed to mountain is listed as the second sense. Other subsenses listed in the *Diccionario de autoridades* for *campo* refer to *campo* as an army, *campo* used in textiles and in heraldry, and *campo* as the location chosen for a duel.

What emerges from this brief historical examination of data from three languages is that a wide variety of senses from two root etymologies have arisen and have merged in certain areas over time. Thus, *field*, *champ* and *campo* as agricultural units of land provide a common starting place for an exploration that will lead to areas of scientific endeavour, fields of study.

An initial prototype is drawn starting from dictionaries, notably the Oxford English Dictionary, Trésor de la Langue Française informatisé and early editions of the *Diccionario de la lengua española* (see Table 1) and consists of extracts from the initial entries for *field*, *champ*, and *campo*. As Table 1 shows, there is no direct correspondence across entries: in English and French, the idea of a piece of land that is delimited arises early on, whereas in Spanish the idea that the land lies outside of, and thus contrasts with, a populated settlement is important.

English	French	Spanish
a piece of ground	espace d’une certaine étendue	terreno extenso fuera de poblado
open land as opposed to woodland		en contraposición a sierra o monte, <i>campiña</i>
land or a piece of land appropriated to pasture or tillage	étendue plate de terre arable	tierra laborable
usually parted off by hedges, fences, boundary stones, etc.	plus ou moins nettement délimité	
	étendue plate de terre arable caractérisée par l’absence de clôture	
		sembrados, árboles y demás cultivos.

Table 1. Some prototypes for *field*, *champ*, and *campo*.

### 3.1 Field

Collocational networks have been drawn up for the first 5 occurrences in each category of four areas of Sketch Engine® output: ‘object\_of,’ ‘subject\_of,’ ‘modifier’ and ‘modified.’ An initial sweep of the object collocates brings forth four large meaning classes: [agriculture], [disciplines], [opportunities], and [sport]. Widening to the three other areas provided by the Sketch Engine® data, we can add [computing and mathematics], [physics], [in vivo situations] and [rural].<sup>2</sup> This done, it is now possible to see what aspects of the prototypes are activated in each case. In the text that follows, the concept areas are square bracketed and the collocates are in italics.

Unsurprisingly, the agricultural and rural senses are closely linked as these are the oldest recorded senses. Both are areas: more specifically, enclosed areas of unwooded ground. The delimitation of the agricultural *field* may be explicit or implicit, as we are hindered by our limited view of clearances, commons and enclosures as acts of creating and appropriating open spaces. The important aspects underlined are tillage, *ploughing* and *sowing*, and food production, which includes *grazing*, although the latter does not require enclosure. This area is relatively level so as to permit tillage. [Rural] is similarly areas of land, and it is opposed to urban areas, which they *surround*. It is possible that woodlands are included, as this is simply an area found in proximity to another area, *villages*, or *towns*. What we term [in vivo situations] are partially related to these, such as the notion of *field trial*, as opposed to laboratory testing, which implies getting out of an enclosed environment into an open space, which when linked to farming is a field as a place for food production. The notion of getting out into the open also comes with *field trip* and *field recording*. The exploitation of the prototype is thus an area, which is delimited, used for agricultural production, is not woodland, lies outside of an urban area, and is a closed space. [Sport] makes use of this as well, but further limits the area. Note, however, the salience of the parameters of outdoors, level and treeless in [sport].

Agricultural metaphors as *cultivate* [the mind] and *culture* [the arts] are frequent. It could be considered surprising thus that *fields of study* largely pre-date the metaphor of *culture*. What is carried over is the notion of a defined area, but a more dynamic one than in agriculture as new *fields* can *emerge*, and as enclosed spaces, they can be *entered*. What is emerging implies new [opportunities] that can be *wide* or *narrow*. Taking the metaphor of delimited area further, we can find the notions defined in [physics] such as electromagnetism, which also hark back to the more or less delimited area as there are no clear barriers here, unlike in computing, where a *field* in a database is limited and needs to be filled in, and is possibly like a *playing field* in that it is rectangular.

---

2 Although the Sketch Engine® is an extremely powerful and useful tool, the data provided in a Word Sketch must be carefully analysed and, in some cases, checked against the examples in because sometimes the results can be misleading. Errors in tagging can occur; for example, in the Word Sketch for campo under the category ‘subject of,’ the proper noun mauthausen is given as the verb with the highest MI index of 6.72. Other proper nouns, such as Valderrama, Covadonga, and Huelva, also erroneously appear in the ‘subject of’ column, so presumably the grammar used for the analysis needs to be revised. Past participles are considered verb forms, but in many contexts are used as adjectives and as such the results in the column ‘sujet de’ for the French corpus includes constructions in which the word champ is not a subject (for example, expressions like un champ cultivé or les champs fleuris). Nevertheless, we have found this tool to be useful for providing a quick, overall picture of a word’s behaviour.

What the above shows is the subtle linkage of concept areas that in dictionary terms would be called senses. Rather than simply describing them, the prototype approach can be used to show linkage between senses through what is being activated. It also allows us to map change across time and between languages.

### 3.2 A *champ* is and is not a *field*

As was done for *field* and *champ*, collocational networks have been drawn up for the first 5 occurrences in each category of four areas of Sketch Engine® output: ‘object\_of,’ ‘subject\_of,’ ‘modifier’ and ‘modified.’ Although [agriculture] and [disciplines] are common object collocates with *champ*, the notion of extending something (which inherently must have a limit) is very salient in the ‘object of’ output (the verb *élargir* ‘broaden’, for example, shows a Mutual Information (MI) index of 9.3 and *étendre* ‘extend’, an MI index of 6.87, in the French tenten corpus). The salience of the notions of ‘used for agricultural production,’ especially of crops, and ‘treeless’ is clear, as *champ* often occurs in a context in which *fôret* and *pâturage* are also listed and contrast with the idea evoked by *champ*.

The concept area with the highest MI index for ‘*champ* + modifier’ is [physics], with *magnétique* and *électromagnétique* displaying both high frequency and high MI indices. Perhaps surprisingly, the concept area of [language] is salient in the French corpus data: we have *champ lexicque* (MI index, 9.07) and *champ sémantique* (MI index, 7.7). The extension of the agricultural metaphor in present in French but has taken a somewhat different direction from that in English; for example, the verb *cultiver*, which is a very strong collocate for *champ*, prefers crops and plants as a direct object, although one can also *cultiver le paradoxe*, *cultiver l’ambiguïté*, and *cultiver la nostalgie* (none of which are typically *cultivated* in English, according to the corpus data).

### 3.3 A *campo* is and is not a *field*

As was done for *field* and *champ*, collocational networks have been drawn up for the first 5 occurrences in each category of four areas of Sketch Engine® output: ‘object\_of,’ ‘subject\_of,’ ‘modifier’ and ‘modified.’ Although [agriculture] and [disciplines] are common object collocates with *campo*, as in English, [computing] is very salient in the ‘object of’ output (the verb *rellenar* ‘fill in’, for example, shows an MI index of 9.27 in the European Spanish tenten corpus). Interestingly, the notion of [enclosure], which is not included in the dictionary definitions for the several senses of *campo*, does seem to underlie some usage in contemporary Spanish, as the verb *delimitar* ‘delimit’ shows a reasonably high MI index for ‘object of’ (6.66). The area of [confinement of people], to which the underlying notion of enclosure is inherent, is also very salient for *campo*: we find *deportar* ‘deport’ (‘subject of’); *concentración* ‘concentration,’ *exterminio* ‘extermination,’ and *refugiado* ‘refugee’ (n\_modifier). Of course, that subject area in English is not associated with *field* but rather with the etymological cognate of *campo*, *camp*.



The notion of getting out into the open, which gives rise to much phraseology and several lexicalized expressions in English (e.g. *field trip*, *field work*), is important to the collocational network for *campo* mainly in conjunction with the nouns *trabajo* ‘work’, *experimento* ‘experiment’ and *estudio* ‘study’; this notion is much less salient in the network for *campo* than it is in the network for *field*.

An important difference between the behaviour of *field* and *campo* is related to the area [military]: one can *invadir* ‘invade’ *campos* and be the *mariscal de campo* (‘field marshall’) on a *campo de batalla* (‘battlefield’; MI index of 9.54). The fact that English *battlefield* is a morphological compound surely explains the fact that this subject area is not as salient for *field* as it is for *campo*.

In English, the verb *cultivate* is a common collocate of *field*, and similarly in Spanish, *cultivar* is a common collocate of *campo*. The agricultural metaphor appears to have been extended further in English, however, as *cultivate* often takes abstract nouns as a direct object (*mindfulness*, *friendship*, *relationship*, *virtue*), whereas the corpus data show that Spanish *cultivar* overwhelmingly takes crops and plants as its direct object. In fact, the Word Sketch only lists one direct object in 25 that is not literally related to agriculture and that word is *amistad* ‘friendship.’

In Spanish, the concept area [sport] is present mainly because of *campo* is used to refer to golf courses (Spanish, *campo de golf*) and football fields (*campo de fútbol*), which can be stepped on (*pisar el campo*) once it has been opened and inaugurated (*abrir/inaugurar el campo*). This concept area, however, appears to play a smaller role in Spanish than it does in English.

Interestingly, the prototype of *campo* as being located away from a populated centre is still prominent in Spanish. Under the category of ‘and\_or’, and discarding proper nouns which produce errors, the two nouns that are in some sense complementary to *campo* are *bosque* ‘forest’ (MI index of 6.09) and *montaña* ‘mountain’ (5.85). Using the Sketch Engine to consult a different corpus (the Spanish web corpus), the noun that appears with the highest MI index in this category is *ciudad* ‘city,’ which clearly evokes the contrast with *campo*.

## 4 Dictionary representation of collocational resonance

Collocational resonance, as stated earlier, claims that meaning attributes can be carried over from one context to another. As such, it can, and we believe, should, be taken into account in dictionary representation because it can help to show that there is a relationship between senses that are depicted as discrete items on a list. Such listing practice, of course, may be unavoidable in dictionary representation, but as Fillmore (1975) and Hanks (2000; 2013) have argued, is not necessarily a proper approach to word meaning. Monolingual dictionaries could attempt to highlight the relationship between senses, perhaps by ordering senses to show derived meanings and or by stating that a meaning has developed as an extended sense and fits into a metaphor that is operative in the language. To date, few monolingual dictionaries have attempted to represent metaphor in their entries, although the *Macmillan English Dictionary* (both in the printed and online versions), with ‘Metaphor Boxes’ and a section on Metaphor in the body of the dictionary, stands out in this respect. For the purposes of this

study, we shall consider the entries for the noun *field* in the *Macmillan English Dictionary Online* and in *The American Heritage Dictionary of the English Language*, shown in Figures 1 and 2, respectively.

1 [COUNTABLE] an area of land used for keeping animals or growing food

*There were horses grazing in the next field.*

*a corn/wheat field*

**field of:**

*We drove past huge fields of barley and hay.*

a. an area of land covered in grass and used for sport

*The England striker left the field with a knee injury.*

*a sports/football field*

**take the field (=walk onto it in order to start playing):**

*The crowd gave Ripken a standing ovation when he took the field.*

**on/off the field:**

*He behaves badly both on and off the football field.*

b. a large area of land or water where something is found

*a gas field*

c. a large area of land or water covered in a particular substance

*an ice field*

d. MAINLY LITERARY an area of land where people fight a battle

2 [COUNTABLE] a subject that you study, or a type of work that you do

**field of:**

*a chemist working in the field of polymer research*

**a field of study/endeavour/enquiry:**

*She has the ability to succeed in any field of endeavour.*

**a specialist/expert in a field:**

*Professor Edwards is one of the main experts in his field.*

3 [SINGULAR] all the people or animals taking part in a race or competition: can be followed by a singular or plural verb

*Henderson will be competing against a very strong field today.*

4 [COUNTABLE] COMPUTING a part of a database that contains information of a particular type  
*Type your name in the User field.*

5 [COUNTABLE] PHYSICS an area where a particular force has an effect  
*a magnetic field*

6 [COUNTABLE] an area that a person or piece of equipment can see at one time

7 **the field** the team in baseball, cricket etc that is throwing the ball and trying to catch it when the other team hits it: can be followed by a singular or plural verb

**Figure 1: Macmillan English Dictionary Online.**

1. a. A broad, level, open expanse of land.  
b. A meadow: *cows grazing in a field*.  
c. A cultivated expanse of land, especially one devoted to a particular crop: *a field of corn*.  
d. A portion of land or a geologic formation containing a specified natural resource: *a copper field*.  
e. A wide unbroken expanse, as of ice.
2. a. A battleground.  
b. *Archaic* A battle.  
c. The scene or an area of military operations or maneuvers: *officers in the field*.
3. a. A background area, as on a flag, painting, or coin: *a blue insignia on a field of red*.  
b. *Heraldry* The background of a shield or one of the divisions of the background.
4. a. An area or setting of practical activity or application outside an office, school, factory, or laboratory: *biologists working in the field; a product tested in the field*.  
b. An area or region where business activities are conducted: *sales representatives in the field*.
5. *Sports*
  - a. An area in which an athletic event takes place, especially the area inside or near to a running track, where field events are held.
  - b. In baseball, the positions on defense or the ability to play defense: *She excels in the field*.
  - c. In baseball, one of the three sections of the outfield: *He can hit to any field*.
6. A range, area, or subject of human activity, interest, or knowledge: *several fields of endeavor*.
7. a. The contestants or participants in a competition or athletic event, especially those other than the favorite or winner.  
b. The body of riders following a pack of hounds in hunting.  
c. The people running in an election for a political office: *The field has been reduced to three candidates*.
8. *Mathematics* A set of elements having two operations, designated addition and multiplication, satisfying the conditions that multiplication is distributive over addition, that the set is a group under addition, and that the elements with the exception of the additive identity form a group under multiplication.
9. *Physics* A region of space characterized by a physical property, such as gravitational or electromagnetic force or fluid pressure, having a determinable value at every point in the region.
10. The usually circular area in which the image is rendered by the lens system of an optical instrument. Also called *field of view*.
11. *Computers*
  - a. An element of a database record in which one piece of information is stored.
  - b. A space, as on an online form or request for information, that accepts the input of text: *an address field*.

**Figure 2: field in The American Heritage Dictionary of the English Language.**

In the Macmillan entry, the fact that sense (1a), that of a *field* used in the concept area [sports], is separated from sense (3), the people taking place in a sporting competition, makes it difficult to see the relationship between these two senses. Notice that the same problem occurs in the American Heritage Dictionary, in which changing the order of senses (5) and (6) might make things clearer. Although the notion of boundary is latent in the wording of the definitions (notice the frequent occurrence of the preposition *in*, as in ‘An area *in which* an athletic event takes place’, the idea of *field* as an enclosure is not really explicit in either entry.

Entries for *field* in even very good, large bilingual dictionaries do not address the subtle differences an analysis of collocational resonance can reveal. Let us look at the entry for *field* in the well-regarded *Collins Spanish Dictionary*.<sup>3</sup>

1. a. (*agriculture*) campo *m*  
(= *meadow*) prado *m*
- b. (*geology*) yacimiento *m*
2. (*sport*) campo *m*, terreno *m* de juego, cancha *f* (*LAm*)  
(= *participants*) participantes *mpl*  
(*for post*) opositores *mpl*, candidatos *mpl*  
⇒ is there a strong field? ¿se ha presentado gente buena? ⇒ to lead the field (*sport, business*) llevar la delantera ⇒ to take the field (*sport*) salir al campo, saltar al terreno de juego
- IDIOM: to play the field (*informal*) alternar con cualquiera
3. (= *sphere of activity*) campo *m*, esfera *f* ⇒ field of activity esfera *f* de actividades, campo *m* de acción  
⇒ my particular field mi especialidad ⇒ it’s not my field no es mi campo *or* especialidad, no es lo mío ⇒ what’s your field? ¿qué especialidad tiene Vd? ⇒ in the field of painting en el campo *or* mundo de la pintura ⇒ to be the first in the field ser líder en su campo
4. (= *real environment*) ⇒ a year’s trial in the field un año de prueba en el mercado ⇒ to study sth in the field estudiar algo sobre el terreno
5. (*computing*) campo *m*
6. (*military*) campo *m* ⇒ field of battle campo *m* de batalla ⇒ to die in the field morir en combate
7. (*electricity and electronics*) campo *m* ⇒ field of vision campo *m* visual
8. (*heraldry*) campo *m*

**Figure 3: The noun field in the Collins Spanish Dictionary.**

Although the dictionary does an admirable job of grouping some related senses together, notice that there is no indication that the word *campo*, which is clearly the main equivalent for *field* as it appears in all but one of the eight identified senses, evokes land outside where people live, and like *field*, is a clearing that contrasts with mountains. That is why *field* is used in expressions like copper field. Fa-

3 Boldface and color typesetting have been removed from the original in this figure.

ced with this entry, which is typical of bilingual dictionaries in that there is little attempt to link sense developments to one another, the speaker of Spanish may be bewildered by a word that can mean a cultivated area where human intervention is required (*campo*, with the subject label ‘agriculture’), as well as an open space *prado* ‘meadow’ and area where a mineral is found naturally underground (*yacimiento*), all of which are grouped together in sense (1).

## 5 Conclusion

There is much to do in the analysis of resonance and of agricultural metaphor. This text aims only to show the potential of collocational resonance as a means of showing language change by mapping variations by use of mono- and multilingual lexicographical prototypes and of collocational networks. How dictionaries should incorporate collocational resonance into their descriptions is an open question at this point. For monolingual dictionaries, information about resonance is essential to show linkage across senses and accounts for collocational patterns, yet most contemporary dictionaries do not provide enough information from resonance for users to grasp the linguistic consequences of the metaphor. From a multilingual perspective, usually uncontroversial equivalents such as *field*, *champ*, and *campo*, are shown to develop different patterns of resonance, which—in our view—should have consequences for their representation in bilingual dictionaries.

## 6 References

- Collins Spanish Dictionary Online*. Accessed at: <http://www.collinsdictionary.com/dictionary/english-spanish> [06/03/2014]
- Dictionnaire Historique de la Langue Française. (2006). Paris: Dictionnaires Le Robert.
- Fillmore, C.J. (1975) An alternative to checklist theories of meaning. In *Papers from the First Annual Meeting of the Berkeley Linguistics Society*, pp.123–132.
- Furetière, Antoine. (1690). *Dictionnaire Universel*. La Haye. Accessed at :<http://gallica.bnf.fr/ark:/12148/bpt-6k50614b> [09/03/2014]
- Hanks, P. 2000. Do Word Meanings Exist. In *Computers and the Humanities* 34, pp.205-215.
- Hanks, P. (unpublished 2005). Resonance and the Phraseology of Metaphors. Paper presented at Phraseology 2005: The Many faces of Phraseology Conference. Louvain-la-Neuve.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: The MIT Press.
- Macmillan English Dictionary Online*. Accessed at: <http://www.macmillandictionary.com/dictionary> [02/04/2014]
- Oxford English Dictionary Online*. Accessed at: <http://www.oed.com> [03/03/2014]
- Sketch Engine*® tenten corpora. Accessed at: <https://www.sketchengine.co.uk> [10/02/2014]
- The American Heritage® Dictionary of the English Language. (2011). Boston: Houghton Mifflin Harcourt.
- Tresor de la Langue Française informatisé. <http://atilf.atilf.fr> [10/03/2014]

Williams, G. (2008b). The Good Lord and his works: A corpus-based study of collocational resonance. In S. Granger, F. Meunier (eds.) *Phraseology: an interdisciplinary perspective*. Amsterdam: John Benjamins, pp. 159-174.

Williams, G. (2012). Bringing Data *and* Dictionary Together: Real Science in Real Dictionaries. In A. Bolton, S. Thomas, E. Rowley-Jolivet (eds.) *Corpus-Informed Research and Learning in ESP: Issues and Applications*. Amsterdam: John Benjamins, pp. 219-240.

### **Acknowledgements**

We gratefully acknowledge that work on this paper was supported by research project FFI2012-38847 *Análisis léxico basado en corpus y su incidencia en los diccionarios* funded by the Spanish Ministry of Economy and Competitiveness (DeCesaris, principal researcher and Williams, team project member).

# The Use of Corpora in Bilingual Phraseography

Dmitrij Dobrovol'skij  
Russian Academy of Sciences, Russian Language Institute  
Austrian Academy of Sciences, AAC-Austrian Academy Corpus  
dobrovol'skij@gmail.com

## Abstract

The present paper discusses issues in the compilation of bilingual dictionaries of idioms based on an analysis of corpus data. The advantages of using corpora consist not only in more detailed and well thought-out illustrations of the expressions being described, but also in the additional possibilities that the corpus materials provide for compiling the idiom list and structuring entries. Thus the corpus allows us to determine the degree of frequency of an expression (at least in the written language). The relevant principles are illustrated by data taken from a new German-Russian dictionary of idioms that is being constructed by an international team of linguists and lexicographers. Fragments of this dictionary are available on the website of the German Language Institute in Mannheim: "Deutsch-russische Idiome online" <[http://wvonline.ids-mannheim.de/idiome\\_russ/index.htm](http://wvonline.ids-mannheim.de/idiome_russ/index.htm)>. Relevant information is also made available via the Europhras homepage on the website <<http://www.europhras.org>>. All examples of idiom usage in this dictionary are taken from the text corpora DeReKo and DWDS, and in individual cases from the German-language Internet. Parallel German-Russian texts from the Russian National Corpus (RNC) are also used.

**Keywords:** corpus; bilingual lexicography; phraseology; idiom; German; Russian

## 1 Preliminary Remarks

Bilingual lexicography widely acknowledges the role of phraseology; for a discussion of relevant theoretical issues see (Lubensky & McShane 2007). Considerable work has been done recently on the compilation of bilingual phraseological dictionaries in languages such as English, German, Russian, Czech, Spanish, French, Italian and Portuguese; cf., for instance, (Heřman et al. 2010), (Kraus & Baumgartner 2011) and a series of German bilingual idiom dictionaries initiated and co-compiled by Hans Schemann. The dictionary in this field that is especially remarkable and meets the highest lexicographic standards is (Lubensky 2013). This most complete Russian-English dictionary of idioms first came out in 1995 in New York. It was subsequently published twice in Moscow (in 1996 and in 2004), and now it has appeared in an enlarged and revised version that includes about 550 new entries. (Lubensky 2013) offers virtually the only lexicographic description of Russian phrasemes with their Eng-

lish counterparts that is based on contemporary notions of linguistically significant features of idioms.

Against this background, it seems especially surprising that modern bilingual phraseography scarcely makes use of text corpora. Though Lubensky (2013: vii) points out that the “availability of language corpora made it possible to check the idioms’ register and usage in multiple contexts”, none of the aforementioned dictionaries is really corpus-based. This fact makes it necessary to address the question as to how corpora can be used as a primary source for compiling a bilingual dictionary of idioms. Today, as lexicography is experiencing “the corpus revolution” (Hanks 2012), this is a question of vital importance. The various uses of corpora in bilingual phraseography will be discussed here on the basis of data taken from a new German-Russian dictionary of idioms that is now under construction.<sup>1</sup>

## 2 German-Russian Phraseography: State of the Art

The need for a new German-Russian phraseological dictionary is motivated by the fact that existing such dictionaries do not meet present requirements. Both the vocabulary and the examples in Binovič and Grišin’s German-Russian phraseological dictionary (Бинович, Гришин 1975) are out of date, and the work fails to satisfy current needs with respect to a number of other parameters as well. Although Dobrovol’skij’s *Немецко-русский словарь живых идиом* “German-Russian Dictionary of Current Idioms” (Добровольский 1997) is on the whole more up to date, it also has certain shortcomings. Its idiom list is rather limited, and illustrative examples are often arbitrary and unpersuasive, which may be because it was written back in the “pre-corpus era”. Actually, one of the basic goals of our new lexicographical project is to eliminate all the shortcomings of this dictionary and to significantly expand its idiom list.

Yet another dictionary of this type has appeared recently: *Новый немецко-русский фразеологический словарь* “The New German-Russian Phraseological Dictionary” (Шекасюк 2010). Its phraseme list is fairly large and up to date, but the work is difficult to use, primarily because the illustrative examples are not translated into Russian, and the division of entries into meanings and selected equivalents often appears hasty and arbitrary.

Thus there is an unquestionable need for a new dictionary containing the most widely used contemporary German idioms together with carefully selected Russian equivalents, explanations facilitating the correct use of these idioms, and good, authentic examples translated into Russian. It is also important that such a dictionary exist not only in print, but also (at least in part) in an online version,

---

1 „Moderne deutsch-russische Idiomatik: Ein Korpus-Wörterbuch“, unter der Leitung von Dmitrij Dobrovol’skij. Wissenschaftliche Redaktion: Dmitrij Dobrovol’skij, Artem Šarandin, Irina Parina und Tat’jana Filipenko; erarbeitet von Elena Krotova, Dmitrij Dobrovol’skij, Tat’jana Filipenko, Artem Šarandin, Viktorija Kosteva, Irina Parina und Denis Zaxarov. Russische Akademie der Wissenschaften, Moskau / Österreichische Akademie der Wissenschaften, Wien.



which will not only provide easier access to the information but will also ensure continuous revision and improvement.<sup>2</sup>

### 3 Corpus-based Bilingual Phraseography and Cross-linguistic Equivalence

The lexicographic treatment of the notion of equivalent in dictionaries based on corpus data encounters certain problems. Not infrequently, the generally accepted equivalent of an idiom cannot always be used to translate authentic texts.

Let us take an example. The German idiom *sich (D) die Beine in den Bauch stehen* (literally ~ “to stand one’s legs into the stomach”) has a “standard” equivalent in Russian, namely the expression *отстоять себе все ноги* (literally ~ “to stand on one’s feet as long as they fall off”), both meaning something like ‘to stand out’ or ‘to stand through’. It would be somewhat odd to doubt that these expressions are basically equivalent, since they are identical with respect to both their lexicalized meaning and have similar image components. Nevertheless, it turns out that it is far from always possible to translate the idiom *sich (D) die Beine in den Bauch stehen* with the Russian expression *отстоять себе все ноги*. Numerous contexts with the idiom *sich (D) die Beine in den Bauch stehen* can be found in text corpora in which this idiom has to be translated into Russian either by the verb *простоять/простаивать* ‘to stand for some time’ or by the collocations *стоять в очереди* ‘to queue up’ and *выстраиваться в длинную очередь* ‘to stand in a long queue’.

(1) Schon am Nachmittag *standen sich* die Fans *die Beine in den Bauch*, um ein Autogramm Ullrichs zu bekommen. Fast 200 Meter lang war die Schlange bis zum Tisch, an dem der Radstar [...] Autogramme schrieb. (Mannheimer Morgen, 28.08.2004)

*Уже во второй половине дня фанаты выстроились в длинную очередь, чтобы взять у Ульриха автограф. Очередь к столу, за которым звезда велогонок раздавал автографы, была почти 200 метров.*

(2) Endlose Warteschlangen winden sich um das Moskauer Puschkin-Museum. Biedere russische Hausfrauen, Veteranen mit Orden am Sonntagsanzug [...], elegante Moskauerinnen – sie *stehen sich stundenlang die Beine in den Bauch* für ein paar Blicke auf den Schatz. (Zürcher Tagesanzeiger, 23.04.1996)

*Бесконечная очередь вьётся вокруг московского Пушкинского музея. Простые российские домохозяйки, ветераны с орденами на груди, элегантные москвички – все они часами стоят в очереди, чтобы взглянуть на сокровища.*

---

2 It goes without saying that putting a dictionary online does not automatically mean an easy access to data for its permanent revision. However, an online dictionary provides a better opportunity to improve the entries.

Consequently, despite the intuitively felt equivalence of the expressions *sich (D) die Beine in den Bauch stehen* and *отстоять себе все ноги*, this equivalence cannot be considered complete. For the lexicographer interested in a maximally precise description of the material, such instances are problematical. Either we acknowledge that *sich (D) die Beine in den Bauch stehen* and *отстоять себе все ноги* are equivalent, in which case it is necessary to explain why the “standard” equivalent is unacceptable in a number of contexts, or we deny that a relationship of bilingual equivalence obtains between *sich (D) die Beine in den Bauch stehen* and *отстоять себе все ноги*, and focus exclusively on translating specific contexts. Such a solution, however, is counterintuitive.

There are at least two possibilities to solve this problem. Either we refrain from giving equivalents and replace them with an explanation, or we provide the given equivalents with a commentary indicating relevant limitations.

In our dictionary we have followed the second path. Thus for the German idiom *sich (D) die Beine in den Bauch stehen* we give the Russian equivalent *отстоять себе все ноги* and explain divergences in the use of the idioms in the commentary, where we point to the fact that the Russian idiom *отстоять себе все ноги* is a perfectiva tantum, i.e. it cannot normally be used in the imperfective aspect.

Another example. The German idiom *jmdn. an der Nase herumführen* (cf. English *to lead s.o. (around) by the nose*) is not fully equivalent to its seemingly ideal Russian counterpart *водить за нос кого-л.* because this Russian idiom is an imperfectiva tantum and can be used in the perfective aspect only in non-veridical contexts such as *а народ не дурак, за нос его так просто не проведешь* or *за нос такого провести нетрудно*, which are encountered quite rarely. For more detail see (Dobrovolskij 2013). Normally, when used in contexts focusing the result, the German idiom *jmdn. an der Nase herumführen* has to be translated into Russian either by the verbs *надуть* and *одурачить* or by the idiom *обвести вокруг пальца*.

(3) Die Aktionäre fühlen sich vom größten deutschen Industriekonzern *an der Nase herumgeführt*. (Mannheimer Morgen, 08.08.1995)

*У акционеров такое чувство, что самый большой промышленный концерн Германии обвел их вокруг пальца.*

(4) In Wahrheit hatte er [Wolfgang Schäuble] aber 100.000 Mark [...] bekommen [...]. Und das hat er im Deutschen Bundestag [...] verschwiegen und hat das erst später, vier Wochen später in einem Fernsehinterview aufgedeckt und da haben viele gesagt, [...] der hat den Deutschen Bundestag *an der Nase herumgeführt*. ([www.stroebele-online.de/themen/spendenaffaere/29273.html](http://www.stroebele-online.de/themen/spendenaffaere/29273.html))

*На самом деле он [Вольфганг Шойбле] получил 100.000 марок. Причем он скрыл это от бундестага и только позднее, спустя четыре недели, признался в этом во время телеинтервью. И многие сказали тогда: он просто одурачил немецкий парламент.*

A question that arises from the perspective of phraseological theory (especially its contrastive aspects) concerns the essence of cross-linguistic equivalence of idioms. It seems expedient to distinguish two different aspects of equivalence: (a) equivalence in translation; that is, the relationship between an idiom of language L1 and its translation into language L2 in a particular text, and (b)

equivalence in the language system; that is, the relationship between the compared idioms of L1 and L2 on the systemic level.<sup>3</sup>

One of the most important differences between translational and systemic equivalence (besides the fact that the former has to do with a concrete text and the latter with the lexical system) consists in the circumstance that equivalence in translation is a unilateral relationship, whereas equivalence in the language system is defined as bilateral. In other words, if a phraseme of language L1 is equivalent to a phraseme in language L2 (in terms of (b)-equivalence), this means that the L2 phraseme is also equivalent to the corresponding L1 expression. With respect to equivalence in translation, all that is being said is that an expression in language L2 is being used in the translation of some specific text in language L1 in such a way that between the L1 phraseme from this particular text and the L2 expression there is a relationship of semantic correspondence. The fact that the translation of some L1 phraseme into language L2 is its equivalent (at least with respect to this particular context) does not, of course, mean that the relationship can be reversed. That is, the L1 phraseme should not be regarded as an equivalent of the expression used in the translation of this phraseme into language L2 (even if this expression is a phraseme, which is not at all obligatory). Obviously, the study of equivalence in translation broadens our notions about the possibilities of cross-linguistic paraphrasing and about the role of contextual conditions in the selection of adequate correspondences, and it contributes to the development of both translation theory and contrastive phraseology.

As for equivalence in the language system, its study has both theoretical and practical significance for phraseology. Deserving of special attention from the theoretical point of view is the question of why one and the same concept is expressed by means of an idiom in one language but not in another. Another (no less important) problem concerns the fact that between basically similar idioms in language L1 and language L2, there are practically always certain semantic, pragmatic, and collocational differences that must be discovered and described. This is especially important in cases where a traditional description postulates a relationship of “full equivalence” but ignores the absence of functional interchangeability between the idioms. The practical aspect of systemic equivalence is what is reflected in bilingual dictionaries, where the entry consists of a phraseme of language L1 (in the lemma) and its idiomatic (to the extent this is possible) correlates in L2. Can these correlates be regarded as equivalents of the L1 phraseme? Yes and no. On the one hand, they must be at least “partial equivalents”, for otherwise they could not be placed in the corresponding dictionary entry. On the other, often they cannot be used in the translation of specific texts. The reason, as a rule, is that the phrasemes of L1 and L2 display certain differences in their semantic, pragmatic, and collocational features. They can be considered cross-linguistic equivalents only in a rather approximate comparison of the idioms of the given languages, and are the starting point of a thorough contrastive analysis that attempts to

---

3 That (a) and (b) represent different aspects of the equivalence phenomenon has been noted in various theoretical contexts. For example, Zgusta distinguishes between *explanatory* or *descriptive* and *translational* or *insertable* equivalents, Hausmann between *prototypical* and *textual* equivalents, and Gouws between *semantic* and *communicative* ones. For more detail see (Adamska-Sałaciak 2010: 392-397).

discover the unique properties of each idiom and thereby improve the lexicological and lexicographical description of phraseology.

Obviously, aspects (a) and (b) are, as it were, two sides of the same phenomenon or two approaches to studying it. We assume that one of the principal goals of contrastive phraseology is to discover genuine equivalents – that is, those that are as close as possible with respect to their actual meanings and – ideally – with respect to the image basis of the expressions, and that function equally well in analogous types of situations, which does not at all imply an obligatory “phraseme – phraseme” relationship. What is important for cross-linguistic correspondence, after all, is not “phraseologicalness,” but functional equivalence.<sup>4</sup> It is this type of equivalence that is most interesting from the perspective of bilingual lexicography. To find out functional equivalents we have to simultaneously go two ways: from text to language system and from language system to text. On the one hand, not all systemic equivalents can function as counterparts in authentic texts and, on the other, not all translational equivalents can be included in the dictionary as typical parallels suitable for using in neutral contexts.

In contrast to a conception that is wide-spread within traditional phraseology, I claim that lexical units of any kind (i.e., not only idioms) in L2 which have the identical meaning and, in the ideal case, near-identical metaphorical basis as the L1-idioms from the source text are excellent functional equivalents, so they have to be considered not only more or less appropriate translational solutions, but also real functional equivalents, i.e. parallels in the lexicons of L1 and L2, which have to be fixed lexicographically.

## 4 Parameters of the new Corpus-based Dictionary

The basic parameters upon which dictionaries can be described and compared are (i) the word list (in our case, the idiom list), (ii) the corpus of illustrative examples, (iii) the macrostructure, and (iv) the microstructure, that is, the structure of the entries. Each of these parameters is briefly described below.

---

4 I consider functional equivalents to be a kind of compromise between the translational and systemic approaches, i.e. functional equivalents are lexical items that on the one hand, semantically resemble each other as closely as possible, i.e. are intuitively felt similar in a contextless, isolated presentation, and, on the other, mostly can be used in similar situations. Thus, my interpretation of functional equivalence differs from, for example, Zgusta's approach. Zgusta (1984: 151) points out that “a translation should convey to its reader the same message with the same aesthetic and other values which are conveyed by the original text. Since languages differ in all imaginable respects, the translator-lexicographer must sometimes use means quite different from those used in the original in order to obtain the same results. If the different means do produce the same effect, the texts are considered functionally equivalent”.

## 4.1 The Idiom List

The idiom list of our new dictionary is based primarily on that of Dobrovol'skij's *Немецко-русский словарь живых идиом* "German-Russian Dictionary of Current Idioms" (Добровольский 1997), which contains in all about 1000 items. While working on the monograph (Dobrovol'skij 1997), I conducted a detailed survey in which informants were asked to take into account not only the units that they felt were widely used in contemporary speech, but also those that were judged to be generally known although not necessarily used. In other words, a distinction was drawn between passive and active command of the phraseology. Combining these two idiom lists resulted in a new, expanded idiom list that was supplemented in the course of working with the corpora. At present our idiom list contains some 2000 idioms with variants. There is reason to believe that it covers a majority of commonly used and most familiar idioms of the contemporary German literary language.

Vulgar expressions were deliberately excluded, since such idioms are ill suited for active use by non-native speakers of German. Since the dictionary aspires to a certain extent to be active, its idiom list focuses not so much on understanding as on use.

## 4.2 The Body of Illustrative Examples

The basic difference between the present dictionary and traditional ones is that all examples of idiom usage in it are taken from the text corpora DeReKo and DWDS, and in individual cases from the German-language Internet. Parallel texts from the Russian National Corpus (RNC) are also used. These examples are especially valuable because they have been translated by professional translators rather than by the authors and editors of the dictionary. Since this part of the parallel corpus of the RNC is still rather modest in size, however, examples needed for the dictionary were rarely encountered.

The use of authentic examples based on text corpora is a new approach in bilingual lexicography. Traditional dictionaries were based on a limited body of generally randomly selected examples, and the use of the idioms was often not even exemplified. The advantages of using corpora consist not only in more detailed and well thought-out illustrations of the expressions being described, but also in the additional possibilities that the corpus materials provide for compiling the idiom list and structuring entries. Thus the corpus allows us to determine the degree of frequency of an expression (at least in the written language). For example, the expression *ich fresse einen Besen* occurred in DeReKo 60 times, *Blech reden* 128 times, *bei Adam und Eva anfangen [beginnen]* 236 times, *jmdm. um den Bart gehen* 41 times, *Gift und Galle spucken [speien...]* 312 times, and *bittere Pille* 2804 times. The lower occurrence threshold for an expression to be included in the idiom list can be set differently for different dictionaries. The important point is that together with surveys of informants, the lexicographer now has a supplemental resource for determining the frequency of each individual idiom.

Yet another advantage of using corpora is that it increases our ability to determine the peculiarities of the formal and semantic structure of idioms, particularly in the description of the ambiguity and

variation of a form. Although an analysis of examples of use clearly indicates that polysemy in phraseology is an extremely widespread phenomenon (for further detail see Dobrovol'skij & Filipenko 2009), traditional dictionaries rarely distinguish the different meanings of idioms, and seldom reflect the full diversity of variants actually represented in texts. Dictionaries often register only a single "canonized" form of an idiom that in many cases proves to be not the most frequent one.

In a number of instances text corpora allow us not only to determine the form of a lemma and a selection of its most frequent variants, but also to establish whether a given expression belongs to the sphere of phraseology. For example, Duden 11 (2002) cites two synonymous idioms with the verb *abberufen* in the passive: *abberufen werden: in die Ewigkeit abberufen werden* and *aus dem Leben abberufen werden*. The following synonymous expressions with this verb form are given in DeReKo: *aus dem Leben abberufen werden, zur großen Armee abberufen werden, in die Ewigkeit abberufen werden, ins Jenseits abberufen werden, in die ewigen Jagdgründe abberufen werden, in die ewige Heimat abberufen werden, von/aus dieser Welt abberufen werden, aus diesem irdischen Leben abberufen werden, aus unseren Reihen [aus unserer Mitte] abberufen werden, zu den Scharen der Engel abberufen werden, in eine andere Welt abberufen werden, in den ewigen Frieden abberufen werden, in ein besseres Jenseits abberufen werden, für uns alle viel zu früh abberufen werden, vom Schöpfer abberufen werden, von Gott (dem Herrn) abberufen werden, vom Tod (ins Jenseits) abberufen werden, von einem gnädigen Tod abberufen werden*. There are also expressions close in meaning in which the verb *abberufen* is used in the active voice: *jmdn. will Gott abberufen, jmdn. hat der Tod abberufen*. These findings suggest that the sense of "calling/summoning s.o. from life" is simply a metaphorical meaning of the verb *abberufen*. Consequently, what we have to do with here is not an idiom but a series of relatively free collocations based on a metaphor.

Another example. Duden 11 (2002) cites four idioms with the noun *Mundwerk*: *jmds. Mundwerk steht nicht still* (ugs.) 'jmd. redet ununterbrochen'; *ein böses/lockeres/loses/frechtes* o.ä. *Mundwerk haben* (ugs.) 'gehässig/vorlaut/frech o.ä. reden'; *ein gutes/flinkes Mundwerk haben* (ugs.) 'sehr gewandt reden'; *ein großes Mundwerk haben* (ugs.) 'großsprecherisch reden'. Corpus analysis has shown that the noun *Mundwerk* has a much broader combinatorial profile. Compare, e.g., *flottes, vorlautes, geschliffenes Mundwerk*. This noun can also be used without any adjectives, combining with verbs of various meanings. Cf.

(5) Manchmal wäre es vielleicht sinnvoller, mein *Mundwerk* etwas zu zügeln, nach dem Motto „Reden ist Silber, Schweigen ist Gold“. (St. Galler Tagblatt, 08.04.1999)

Hence, we are dealing here not with four idioms but with a free noun. It seems that the combinatorial profile of this noun is relatively restricted. The only way to describe the collocational constraints in question is the consistent analysis of corpus data.

### 4.3 Dictionary Macrostructure

The dictionary has two parts: the *body*, consisting of entries listed alphabetically by headword, and the *index*, which makes it possible to find an idiom from any of its constituents.

The idioms are arranged alphabetically by headword, selected according to the following hierarchy:

- nouns
- adjectives (including adjectivized participles)
- adverbs (including adjectives in adverbial position and adverbialized participles)
- numerals
- verbs
- particles (with the exception of the negative particle *nicht*)
- pronouns (with the exception of the reflexive pronoun *sich*)
- prepositions
- conjunctions
- interjections

The order of this hierarchy is motivated by the variation features of the lexical structure of the idiom. Thus the verb can often be replaced by a synonym (or more rarely by an antonym), whereas adjectives and adverbs are more stable elements of the structure, and it is this that accounts for their higher position in the hierarchy. Adjectives and adjectivized participles, in turn, are more stable than adverbs. For example, cf. the structurally and semantically similar idioms *es ist (nicht) gut bestellt (um jmdn., etw. A) = дела обстоят (не очень) хорошо (с чем-л. / у кого-л.)* и *es ist (nicht so) schlecht bestellt (um A) = дела обстоят (не так) плохо (с чем-л. / у кого-л.)*. Alphabetizing them according to the adverbial constituent would necessitate entering them in different parts of the dictionary (under GUT and under SCHLECHT, respectively), which is counterintuitive. Alphabetization according to the constituent BESTELLT, which is an adjectivized participle, is much more convenient for the user.

An example of a group of idioms based on the headword under which they are arranged is given in the Appendix.

### 4.4 Dictionary Entry Structure

Dictionary entries open with the *headword*, i.e. the word on which alphabetization is based. This (word, if it is a noun) is always given in the nominative singular (e.g. КОПФ), even if the idioms following the headword contain forms such as *Kopfes, Köpfe, Köpfen* etc. The headword is followed by the *lemma* – the idiom in traditional dictionary form (nominative for nominal expressions, the infinitive with valencies for verbal ones).

Idioms in *propositional* (or *personal*) form are indicated in cases where the subjective valency is filled in a non-trivial way or when the infinitive of the idiom translates poorly into Russian. Compare, e.g., **ei-**



**nen dicken Kopf haben** (*von etw. D*): (*jmd.*) hat (*von etw. D*) einen dicken Kopf = (*у кого-л.*) голова болит (*из-за чего-л.*), (*у кого-л.*) голова раскалывается (*от чего-л., после чего-л. – особенно с похмелья*). The propositional form often helps to discriminate the senses of a polysemous idiom; cf. **jenseits von Gut und Böse sein**: **1.** (*jmd.*) ist jenseits von Gut und Böse = (*кто-л.*) чужд плотским удовольствиям (*часто о пожилых людях*) **2.** (*jmd.*) ist jenseits von Gut und Böse = (*кто-л.*) не от мира сего, (*кто-л.*) потерял связь с реальностью, (*кто-л.*) неадекватен (*часто о людях, находящихся в состоянии сильного алкогольного опьянения*) **3.** (*jmd.*) ist jenseits von Gut und Böse = (*кто-л.*) по ту сторону добра и зла **4.** (*etw.*) ist jenseits von Gut und Böse = (*что-л.*) выходит за привычные рамки, (*что-л.*) невероятно (*что-л. очень хорошо либо очень плохо; часто о слишком высоких либо слишком низких ценах*).

The lemma or propositional form (if there is one) is followed by stylistic *labels*. The use of which follows the principles set forth in (Баранов, Добровольский 2008). Thus the label *разг.* (colloquial) is not used at all, since most idioms belong to the colloquial register. In other words, this label “works” by remaining silent. The following labels are used: *высок.* (high style) – for high style expressions, *книжн.* (literary) – for literary and bookish expressions, *офиц.* (formal) – for expressions in official language and business communication, *нейтр.* (neutral) – for expressions in the neutral register (that is, for idioms higher than colloquial expressions on the scale of stylistic registers), and *снижен.* ( $\approx$  very informal) – for idioms felt to be not entirely acceptable in the standard colloquial style (i.e. lower than *разг.*).

The *translation* of the idiom into Russian is generally oriented toward the system of the language, i.e. toward (b)-equivalents, rather than toward contextual conditions. Relevant functional and context-sensitive properties are additionally explained and illustrated in other parts of the entry, mostly in the commentary and illustrative field. That is, if in the examples of usage an idiom is translated in a non-standard manner, this does not mean that these – often unique – ways of translating it must be registered in the translation field. There it is often expedient to indicate several equivalents, first of all, those translations that with respect to their actual meaning and image basis maximally approximate the German idiom being described. The syntactic parallelism of suggested equivalents is also taken into account as far as possible. If an equivalent parallel to the lemma cannot be found or if it sounds strange, what is recorded in the field of the propositional form is the syntactic version of the German expression that would best correspond to the suggested Russian translation. The translation field can also contain explanatory commentaries that further indicate in which of the possible meanings the suggested Russian translation is equivalent to the German expression.

The *variant field* follows the translation field. Describing variants in a separate field makes it possible not only to reflect more completely the actual variation of the structure of the idiom, but also to avoid having to burden the notation of the lemma with a series of parentheses.

As for selecting *illustrative examples*, preference is given to modern examples, that is, to contexts with idioms dating from the past fifteen years. The basic source for illustrative examples is the corpus of the Mannheim Institute of the German Language (DeReKo). For more detail see section 4.2. In selec-



ting illustrative examples, we have tried not to include examples peculiar to Austrian or Swiss usage, since these deviate from standard literary German (and due to their regional and cultural distinctiveness) do not fully satisfy the needs of a bilingual dictionary with educational goals.<sup>5</sup>

The search for contexts relies not only on the standard options but also on so called “co-occurrence analysis” (Kookkurrenzanalyse). This program helps to determine the lexical contexts in which a given idiom occurs especially often.

All contexts are given in the current (i.e. the “new”) orthography. The peculiarities of Swiss orthography (for example, *ss* instead of the normative *ß*) are not preserved. Such deviations from prevailing standards are given in conformity with the spelling norms of the common German language. For the sake of convenience in using the dictionary, the authors have simplified extremely complex and verbose contexts. Deletions in abbreviated contexts are marked by [...]; cf. examples (1), (2), and (4) above. This indication is not repeated in the Russian translations. Contexts that are overloaded with specific information that is not relevant to conditions for using a given idiom are slightly modified. For example, unfamiliar proper names are replaced with neutral designations of the participants of a situation. In such cases the source is indicated (in parentheses immediately following the context) by *Nach.*. Compare examples in the Appendix.

The *commentary field* contains information significant for the correct use of the expressions if such information cannot be derived from the valency model and/or the semantic and syntactic features of the Russian equivalents. The commentary field indicates, for example, the syntactic and combinatorial properties of the idiom. Also reflected in the commentaries are features relating to the polarization (especially the negative polarity) of expressions, their aspectological peculiarities, possibilities of nominalization, characteristic metonymical shifts, etc., as well as any significant transformational properties of the idiom, especially if they do not coincide with the syntax of the Russian equivalent. Thus the idiom *Blech reden* (unlike its Russian equivalents *пустословить*, *нести чепуху*, *болтать языком*) can be passivized. The commentary field has no fixed position in the structure of the dictionary entry. Accordingly, it can be located in any part of the dictionary entry (depending on the nature of the information being provided).

---

5 This does not mean that Austrian and Swiss sources of empirical data were excluded.

## 5 Conclusion

The use of corpora clearly expands the resources available to the lexicographer for creating the illustrative component of the dictionary entry, but it also offers a number of additional possibilities. Let us attempt to list the most obvious such advantages. Working with corpora makes it possible

- to determine the frequency of each idiom included in the dictionary;
- to determine whether a particular word group is an idiom;
- to determine the standard form of a lemma from the point of view of modern usage;
- to clarify the government models of relevant idioms;
- to determine the most significant variants of each idiom;
- to determine the polysemy structure of each idiom and refine the description of its concrete meanings;
- on the basis of corpus materials, to select the most adequate correspondences, including translations of concrete examples, for each meaning of an idiom;
- to describe the typical modifications of the structure of each idiom;
- to determine the typical environment of the idioms being described and the types of contexts in which they are perceived to be most natural.

Literature on the subject distinguishes two approaches to the use of corpora in lexicographical research: corpus-based and corpus-driven.<sup>6</sup> In the first approach, corpus data are used to confirm already existing hypotheses, while in the second it is the corpus itself that constitutes the data about linguistic structures, and it is only later that these data are interpreted by the linguist. It is clear that on the whole, lexicographers use corpora as the source of additional information about already given linguistic forms (that is, the corpus-based approach). As the material discussed in this paper shows, however, lexicographical work also presumes elements of the corpus-driven paradigm. In other words, in a number of cases corpus data provide lexicographers with knowledge about the structure and semantics of idioms to which they would not have had access even on the hypothetical level prior to consulting the corpus.

## 6 References

- Adamska-Sałaciak, A. (2010). Examining equivalence. In *International Journal of Lexicography*, 21(4), pp. 387-409.
- Dobrovol'skij, D. (1997). *Idiome im mentalen Lexikon: Ziele und Methoden der kognitivbasierten Phraseologieforschung*. Trier: WVT Wissenschaftlicher Verlag Trier.

---

6 Cf., first of all (Tognini-Bonelli 2001), where this distinction is discussed. However, “good corpus research almost always uses both” (Kilgarriff 2013: 96).

- Dobrovol'skij, D. (2013). German-Russian phraseography: On a new dictionary of modern idiomatics. In I. Gonzáles Rey (ed.) *Phraseodidactic studies on German as a foreign language*. Hamburg: Verlag Dr. Kovač, pp. 121-138.
- Dobrovol'skij, D.O. & Filipenko, T.V. (2009). Polysemie in der Idiomatik. In C. Földes (ed.) *Phraseologie disziplinär und interdisziplinär*. Tübingen: Gunter Narr, pp. 109-115.
- Duden 11 (2002) = *Duden – Redewendungen. Wörterbuch der deutschen Idiomatik. 2.*, neu bearb. und aktualisierte Auflage. (=Der Duden, Band 11). Mannheim etc.
- Hanks, P. (2012). The corpus revolution in lexicography. In *International Journal of Lexicography*, 25(4), pp. 398-436.
- Heřman, K. et al. (2010). *Deutsch-tschechisches Wörterbuch der Phraseologismen und festgeprägten Wendungen*. Prag: C. H. Beck.
- Kilgarriff, A. (2013). Review of Tony McEnery & Andrew Hardie. *Corpus linguistics: Method, theory and practice*. In *International Journal of Lexicography*, 22(1), pp. 95-97.
- Kraus, R. & Baumgartner, P. (eds.) (2011). *Phraseological Dictionary English-German: General Vocabulary in Technical and Scientific Texts*. Berlin & Heidelberg: Springer.
- Lubensky, S. (2013). *Russian-English dictionary of idioms. Revised Edition*. New Haven & London: Yale University Press.
- Lubensky, S. & McShane, M. (2007). Bilingual phraseological dictionaries. In H. Burger, D. Dobrovol'skij, P. Kühn and N.R. Norrick (eds.) *Phraseology: An international handbook of contemporary research*. Vol. 2. Berlin & New York: Walter de Gruyter, pp. 919-928.
- Schemann, H et al. (2013). *Idiomatik Deutsch-Spanisch*. Hamburg: Buske.
- Schemann, H et al. (2012). *Idiomatik Deutsch-Portugiesisch. 2.*, durchgesehene Auflage. Hamburg: Buske.
- Schemann, H & Dias, I. (2013). *Idiomatik Portugiesisch-Deutsch. 2.*, durchgesehene Auflage. Hamburg: Buske.
- Schemann, H., Fenati, B. & Rovere, G. (2011). *Idiomatik Deutsch-Italienisch. 2.*, durchgesehene Auflage. Hamburg: Buske.
- Schemann, H. & Knight, P. (2011). *Idiomatik Deutsch-Englisch. 2.*, durchgesehene Auflage. Hamburg: Buske.
- Schemann, H. & Raymond, A. (2011). *Idiomatik Deutsch-Französisch. 2.*, durchgesehene Auflage. Hamburg: Buske.
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Zgusta, L. (1984). Translational equivalence and the bilingual dictionary. In R.R.K. Hartmann (ed.) *LEXeter'83 Proceedings*. Tübingen: Max Niemeyer, pp. 147-154.
- Баранов, А.Н. & Добровольский, Д.О. [Baranov, A.N. & Dobrovol'skij, D.O.] (2008). *Аспекты теории фразеологии* [Aspects of the Theory of Phraseology]. Москва: Знак.
- Бинович, Л.Э. & Гришин, Н.Н. [Binovič, L.E. & Grišin, N.N.] (1975). *Немецко-русский фразеологический словарь* [German-Russian Phraseological Dictionary]. Москва: Русский язык.
- Добровольский, Д.О. [Dobrovol'skij, D.O.] (1997). *Немецко-русский словарь живых идиом* [German-Russian Dictionary of Current Idioms]. Москва: Метатекст.
- Шекасюк, В.П. [Šekasjuk, V.P.] (2010). *Новый немецко-русский фразеологический словарь* [The New German-Russian Phraseological Dictionary]. Изд. 2-е, перераб. и доп. Москва: Либроком.

## 7 Digital Resources

- DeReKo – Das Deutsche Referenzkorpus des IDS Mannheim im Portal COSMAS II (Corpus Search, Management and Analysis System) <<https://cosmas2.ids-mannheim.de/cosmas2-web>>
- DWDS – Corpora des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts <<http://www.dwds.de>>

Online dictionary “Deutsch-russische Idiome online“ <[http://wvonline.ids-mannheim.de/idiome\\_russ/index.htm](http://wvonline.ids-mannheim.de/idiome_russ/index.htm)>

RNC (НКРЯ) – Russian National Corpus (Национальный корпус русского языка) <http://www.ruscorpora.ru>

## Acknowledgements

This paper is based on work supported by the RGNF under Grant 12-04-12041, by the RFFI under Grant 13-06-00403 and by the Basic Research Program “Corpus Linguistics” of the Presidium of the Russian Academy of Sciences.

## Appendix

### ЛИХТ

#### Licht bringen (in etw. A)

нейтр.

внести ясность (во что-л.); прояснить (что-л.); пролить свет (на что-л.)

☞ Archäologen *haben* endlich *Licht* in einen Abschnitt der Geschichte Londons *gebracht*, der vom Abzug der Römer aus Britannien 410 bis ins Mittelalter reicht. (Berliner Morgenpost, 09.09.1999)

Археологам наконец-то удалось *пролить свет* на период истории Лондона с момента ухода римлян из Британии в 410 г. и до Средневековья.

☞ Die Aussage eines 37 Jahre alten Beifahrers in einem anderen Auto *hatte Licht* in die zunächst rätselhafte Kollision *gebracht*. (Frankfurter Rundschau, 26.03.1999)

Показания 37-летнего пассажира, сидевшего рядом с водителем в другом автомобиле, *прояснили* это сперва казавшееся загадочным столкновение.

☞ Auf jeden Fall sind die Funde aus Engers sehr wichtig, um *Licht* in das für Historiker „dunkle fünfte Jahrhundert“ zu *bringen*. (Rhein-Zeitung, 25.05.2007)

В любом случае, находки в Энгерсе очень важны для историков, поскольку могут *пролить свет* на «тёмный пятый век».

#### das Licht der Welt erblicken: (jmd., etw.) erblickt das Licht der Welt

высок.

(кто-л.) появился на свет, (кто-л.) родился; (что-л.) родилось (напр. об изобретениях); (что-л.) увидело свет

☞ In Deutschland *erblicken* pro Jahr rund 60 000 Säuglinge vor der 37. Schwangerschaftswoche *das Licht der Welt* – Tendenz steigend. (Rhein-Zeitung, 09.11.2011)

В Германии за год более 60 000 детей *появляются на свет* до 37-ой недели беременности, и эта тенденция растёт.

☞ Die Anzahl seiner Geburtspartys hält sich in überschaubaren Grenzen, denn er *erblickte* an einem 29. Februar *das Licht der Welt*. (Hamburger Morgenpost, 08.02.2009)

Ему не так часто доводилось устраивать вечеринки в честь своего дня рождения, потому что он *появился на свет* 29-го февраля.

☞ 1995 war auch das Geburtsjahr des Internet-Dienstes Yahoo und Ende 1998 *erblickte* die Suchmaschine Google in einer Garage *das Licht der Welt*. (Hamburger Morgenpost, 15.03.2009)

В 1995 году появился ещё и Интернет-сервис «Yahoo», а в конце 1998-го в гараже *была рождена* поисковая система «Гугл».

### **ein Licht geht auf** (*jmdm.*)

глаза открылись (*у кого-л.*); (*кто-л.*) догадался

- 📖 Идиома часто употребляется с наречиями *plötzlich*, *endlich*, *langsam* – перевод соответственно модифицируется.
- 📖 Die italienische Schauspielerin Gina Lollobrigida kommt nach Nürnberg! *Langsam ging* den Verehrern des Filmstars *ein Licht auf*: April! April! (Nach: Nürnberger Zeitung, 29.04.2010)  
Итальянская актриса Джина Лоллобриджида приезжает в Нюрнберг! *Постепенно* до поклонников кинозвезды дошло, что это была первоапрельская шутка.
- 📖 Aber nachdem ich Hunderte von Interviews mit Trainern gelesen und gehört habe, *ist mir endlich ein Licht aufgegangen*: Ich bin eigentlich gar kein Arbeitnehmer oder Kunde, ich bin ein Helfer. (Nach: Rhein-Zeitung, 28.02.2002)  
Но после того как я прочитал и прослушал сотни интервью с тренерами, *я наконец-то понял*: я никакой не наёмный рабочий и не клиент, я помощник.
- 📖 Идиома может употребляться с валентностью *über etw.* А в значении '(кто-л.) осознал (что-л.)'.
- 📖 Und *geht* der Mannschaft endlich *ein Licht auf* über den Ernst der Lage? (Hannoversche Allgemeine, 12.03.2009)  
И *осознает* ли команда наконец всю серьёзность положения?

### **grünes Licht geben** (*für etw. A*)

нейтр.

дать зелёный свет (*для чего-л.*), дать разрешение (*на что-л.*)

- 📖 Das Bauamt *hat grünes Licht* für einen Anbau ans Museum *gegeben*. Dort soll das Archiv untergebracht werden. (Rhein-Zeitung, 08.09.2011)  
Строительное ведомство *дало разрешение* на постройку флигеля музея. Там будет располагаться архив.
- 📖 Das Arbeitsgericht Frankfurt *hat* im Tarifstreit beim Flugzeugbauer Airbus *grünes Licht* für Warnstreiks *gegeben*. (Rhein-Zeitung, 01.10.2011)  
Суд по трудовым спорам Франкфурта *дал зелёный свет* на предупредительную забастовку рабочих самолётостроительного концерна «Эйрбас» из-за тарифного конфликта.
- 📖 Die Suche nach der seit Jahrzehnten verschwundenen Leiche von Lolita Brieger auf einer ehemaligen Mülldeponie in der Eifel kann beginnen. Experten *gaben grünes Licht*. (Rhein-Zeitung, 05.10.2011)  
Поиски трупа Лолиты Бригер, исчезнувшей уже десятки лет назад, на бывшей мусорной свалке в Эйфеле, могут начаться. Эксперты *дали зелёный свет*.
- 📖 Ben Bernanke (56) kann aufatmen. Er bekommt eine zweite Amtszeit als US-Notenbankchef. Der mächtige Bankenausschuss des Senats *gab* dafür *grünes Licht*. (Hamburger Morgenpost, 18.12.2009)  
56-летний Бен Бернанке может вздохнуть с облегчением. Он во второй раз получил должность директора эмиссионного банка США. Вчера влиятельная банковская комиссия сената дала соответствующее *разрешение*.
- 📖 Bis 2012 hat der Iran die Atombombe, glauben viele Beobachter. „Dann ist es zu spät“, lautet das Credo des israelischen Premiers Netanjahu. Er drängt auf einen Angriff. Von George Bush *gab* es dafür *grünes Licht*, von Obama nicht. (Hamburger Morgenpost, 14.04.2010)  
К 2012 году у Ирана появится атомная бомба, полагают многие наблюдатели. «Тогда будет слишком поздно», – уверен премьер министр Израиля Нетаньяху. Он настаивает на нападении. Джордж Буш *давал* ему на это *зелёный свет*, а Обама – нет.
- 📖 Warschau. Ein Gericht in Polen *gab grünes Licht* für die Auslieferung eines mutmaßlichen Agenten des israelischen Geheimdienstes Mossad an Deutschland. (Hamburger Morgenpost, 08.07.2010)  
Польский суд *дал зелёный свет* на экстрадицию в Германию предполагаемого агента израильской разведки «Моссад».
- 📖 Возможна атрибутивная модификация.

- ☞ Für entsprechende Projekte in Hammelburg, Freising, Kempten und Passau *gab* Wirtschaftsminister Otto Wiesheu jetzt *das lang ersehnte grüne Licht*. (Nürnberger Nachrichten, 10.08.1994)  
Министр экономики Отто Визхой *наконец-то дал разрешение* на соответствующие проекты в Хаммелбурге, Фрайзинге, Кэмптене и Пассау.

### in rosigem Licht

*нейтр.*

в розовом [радужном] свете

☞ редко **in rosa(rotem) Licht**

- ☞ Frankfurt sieht nach zehn trüben Jahren plötzlich die eigene Zukunft *in rosigem Licht*. (Nach: Nürnberger Nachrichten, 20.02.2001)

После десяти мрачных лет будущее Франкфурта неожиданно предстаёт в *розовом свете*.

- ☞ Прилагательное *rosig* может употребляться с неопределённым артиклем (*in einem rosigen Licht*). Возможны также сравнительная или превосходная степени прилагательного (*in einem rosigeren Licht*, *im rosigsten Licht*).

- ☞ Ich war 13 oder 14 Jahre alt, als wir in der Schule einen Aufsatz über unsere Zukunftspläne und Lebensziele schreiben mussten. Aufsätze schrieb ich gern, und trotz aller Nachkriegseinschränkungen sah ich die Zukunft *in einem rosigen Licht*. (Nach: Mannheimer Morgen, 27.12.1997)

Мне было лет 13-14, когда нам в школе задали написать сочинение о планах на будущее и целях в жизни. Я любил писать сочинения, а будущее, несмотря на все тяготы послевоенного времени, видел в *розовом свете*.

- ☞ Die Pannenserie der vergangenen Monate an Bord der Mir hat paradoxerweise dazu beigetragen, dass die Zukunft der russischen Raumfahrt heute wieder *in einem rosigeren Licht* erscheint. (Nürnberger Nachrichten, 25.08.1997)

Как ни парадоксально, но серия поломок на орбитальной станции «Мир» за последние месяцы способствовала тому, что сегодня будущее российской космонавтики снова кажется *более радужным*.

- ☞ Прилагательное в составе идиомы может модифицироваться с помощью указательных местоимений, адвербиалов и пр.

- ☞ Warum ist er nicht Lehrer geworden, wenn er die materiellen Vorteile des Lehrerberufes *in solch herrlich rosarotem Licht* sieht? (Frankfurter Rundschau, 23.07.1997)

Почему же он сам не стал учителем, если материальные выгоды этой профессии ему видятся *в столь прекрасном розовом свете*?

- ☞ В контекстах с отрицанием идиома употребляется в форме *in keinem rosigen Licht* или *nicht in rosigem Licht*.

- ☞ Beschwerden von Nachbarn über zu laute Musik und die anhaltende Ebbe im Stadtsäckel lassen die Zukunft des Jugendkulturhauses „Schillers“ *in keinem rosigen Licht* erscheinen. (Nach: Mannheimer Morgen, 19.03.2010)

Из-за жалоб соседей на слишком громкую музыку и постоянного недостатка в городской казне будущее молодёжного дома культуры «Шиллерс» видится отнюдь *не в розовом свете*.

- ☞ Der scheidende Privatsekretär der Queen wollte zu Charles Entsetzen ein weiteres Diana-Buch genehmigen, das die Verstorbene *nicht gerade in rosigem Licht darstellt*. (Berliner Morgenpost, 02.11.1998)

Увольняющийся личный секретарь королевы хотел к ужасу Чарльза дать согласие на публикацию ещё одной книги о Диане, в которой покойная предстаёт отнюдь *не в розовом свете*.

- ☞ Идиома может также употребляться в форме (*etw. A*) *in (ein) rosiges Licht tauchen [stellen, rücken ...]*.



☞ Bei den Neujahrsansprachen haben Politiker und Unternehmen die Wirtschaftswirklichkeit *in rosigen Licht* getaucht. Von drei Prozent Wachstum und der Wende am Arbeitsmarkt war da die Rede. (Rhein-Zeitung, 07.01.1998)

В своих новогодних поздравлениях политики и бизнесмены *представили* реальную экономическую ситуацию страны *в розовом свете*. Речь шла тогда о трёх процентах роста и переменах на рынке труда.

## **TUCH**

### **ein rotes Tuch sein** (*für jmdn.*)

действовать как красная тряпка на быка (*на кого-л.*)

☞ **wie ein rotes Tuch wirken** (*auf jmdn.*)

☞ Die Steuererklärung *ist für fast jeden ein rotes Tuch*. Viele sind schon in Hektik, denn der Termin rückt jetzt langsam näher: Spätestens am 31. Mai muss die Steuererklärung abgegeben werden. (Hamburger Morgenpost, 26.03.2010)

Налоговая декларация *действует* почти на каждого *как красная тряпка на быка*. Многие уже в панике, так как срок медленно, но верно приближается: декларация должна быть сдана не позднее 31 мая.

☞ Als Außenminister Guido Westerwelle seinen Antrittsbesuch in Warschau machte, wusste er sehr genau, dass Frau Steinbach *ein rotes Tuch* in Polen *ist*. Die CDU-Abgeordnete hatte 1991 im Bundestag nicht für die Oder-Neiße-Grenze gestimmt. (Nach: Nürnberger Nachrichten, 12.02.2010)

Когда министр иностранных дел Гидо Вестервелле Германии совершил свой первый визит в этой должности в Варшаву, он знал очень хорошо, что госпожа Штайнбах *действует* на поляков *как красная тряпка на быка*. Будучи депутатом от партии ХДС в бундестаге, она не проголосовала в 1991 году за границу по Одере-Нейсе.

☞ Die Sicherheitskonferenz bleibt eine kitschige Angelegenheit für die bayerische Polizei, zumal unter den Gästen auch einige sind, die auf die Gegner *wie ein rotes Tuch wirken*. (Nürnberger Zeitung, 11.02.2005)

Конференция по безопасности остается весьма щекотливым мероприятием для баварской полиции, тем более что среди гостей есть такие, *которые* действуют на своих оппонентов *как красная тряпка на быка*.

### **in trockene Tücher bringen** (*etw. A*)

☞ **in trockene Tücher bekommen** (*etw. A*)

окончательно оговорить [согласовать] (*что-л.*)

☞ Der Wechsel des Holländers von Real Madrid nach München soll kurz vor dem Abschluss stehen! Die Bayern und Real wollen den Transfer heute *in trockene Tücher bringen*. Nur noch Details seien zu klären. (Hamburger Morgenpost, 27.08.2009)

Переход голландского игрока из мадридского «Реала» в мюнхенский клуб уже практически завершён! «Бавария» и «Реал» хотят *окончательно оговорить* трансфер сегодня. Осталось только согласовать детали.

☞ Seehofer warnte vor einem Scheitern der Verhandlungen. Zur Not müsse die Ministerrunde die ganze Nacht zum Mittwoch verhandeln, um die Reform *in trockene Tücher* zu *bekommen*. Würde das Problem in das nächste Jahr verschoben, wäre eine Lösung noch schwieriger als jetzt. (dpa, 18.12.2007)

Господин Зеehoфер предостерег о возможном провале переговоров. По его словам, коллегии министров в крайнем случае придётся заседать всю ночь до утра среды, чтобы *окончательно согласовать* реформу. Если же отложить этот проблемный вопрос на следующий год, решить его потом будет ещё сложнее.

☞ Ср. также идиомы *in trockenen Tüchern sein*, *in trockene Tücher kommen* и *in trockenen Tüchern haben* (*etw. A*).

### **in trockenen Tüchern haben** (etw. A)

довести до конца (что-л.), (окончательно) уладить (что-л.)

☞ Wir *haben* jetzt schon alles *in trockenen Tüchern* und wollten den neuen Trainer auch so schnell wie möglich der Mannschaft vorstellen. (Braunschweiger Zeitung, 26.01.2012)

Мы уже всё *уладили* и хотим как можно скорее представить нового тренера команде.

☞ Vorrang vor allem anderen hat für Obama nach wie vor die Gesundheitsreform, die er möglichst noch in diesem Jahr *in trockenen Tüchern haben* will. (Mannheimer Morgen, 10.12.2009)

Предпочтение Обама по-прежнему отдаёт реформе здравоохранения, которую он, по возможности, планирует *довести до конца* в этом году.

📖 Ср. также идиомы *in trockenen Tüchern sein*, *in trockene Tücher bringen* (etw. A) и *in trockene Tücher kommen*.

### **in trockene Tücher kommen**

реализоваться, быть принятым

☞ Während die Finanzmärkte in Asien und Europa zunächst positiv auf die Einigung in Washington reagierten, fiel der Enthusiasmus an der Wall Street gedämpft aus. Zudem warten die Märkte ab, ob die Vereinbarung zwischen Obama und den Kongressführern tatsächlich *in trockene Tücher kommt*. (Nach: St. Galler Tagblatt, 02.08.2011)

В то время как финансовые рынки Азии и Европы положительно отреагировали на достигнутую в Вашингтоне договорённость, на Уолл Стрит она была встречена с вялым энтузиазмом. Кроме того, рынки выжидают, будет ли *принято* соглашение между Обамой и конгрессменами.

☞ Die Nervosität im Regierungslager steigt. Schließlich sollen innerhalb der nächsten fünf Wochen die ehrgeizigsten Reformen *in trockene Tücher kommen*, die sich die Koalition vorgenommen hat. (Nach: dpa, 02.06.2006)

В правительственном лагере растёт нервное напряжение. Ведь в течение следующих пяти недель предстоит *реализовать* самые смелые реформы, о которых заявила правящая коалиция.

📖 Ср. также идиомы *in trockenen Tüchern sein*, *in trockene Tücher bringen* (etw. A) и *in trockenen Tüchern haben* (etw. A).

### **in trockenen Tüchern sein:** (etw.) ist in trockenen Tüchern

(что-л.) (окончательно) решено, (что-л.) доведено до конца

☞ Neben der Arbeit an Großveranstaltungen trifft der Coach auch Personalentscheidungen für seine Mannschaft. „Aber zu meiner Philosophie gehört es, erst über Namen zu sprechen, wenn alles *in trockenen Tüchern ist*“, sagt er. (Nach: Braunschweiger Zeitung, 11.01.2012)

Тренер не только организует крупные мероприятия, но и принимает решения, касающиеся состава команды. «Таков мой принцип – называть имена, только когда всё *решено окончательно*», – замечает он.

☞ Allerdings *ist* die Finanzierung des Großprojekts noch nicht *in trockenen Tüchern*. Es steht noch nicht ausreichend Geld zur Verfügung. (Nach: Mannheimer Morgen, 13.01.2012)

Однако вопрос о финансировании этого крупномасштабного проекта ещё не *решён окончательно*. Пока не было выделено достаточно средств.

📖 Ср. также идиомы *in trockene Tücher bringen* (etw. A), *in trockene Tücher kommen* и *in trockenen Tüchern haben* (etw. A).



# Comparing Phraseologisms: Building a Corpus-Based Lexicographic Resource for Translators

Laura Giacomini  
University of Heidelberg  
laura.giacomini@iued.uni-heidelberg.de

## Abstract

Today there is still a significant need for specific lexicographic resources in digital form, as they can remarkably improve access to data and actively assist the process of text production. This paper describes the way in which corpus data can be explored to retrieve suitable material for the representation of a particular class of culture-specific phraseologisms and similes in a bilingual dictionary for translators.

**Keywords:** phraseologisms; corpora; translation

## 1 Introduction

The characteristic formal stability and contentual figurativeness of phraseological expressions as a result of cultural encoding best reflect a society's deeply rooted patterns of world interpretation. Given the strong presence of phraseologisms in the lexicon of a language and the translatability issues they inevitably raise, it is necessary to support the practice of translation by means of specific and up-to-date lexicographic resources. This paper describes the way in which corpus data can be explored to retrieve suitable material for the representation of a particular class of culture-specific phraseologisms, similes, in a bilingual lexicographic resource for translators. It is based on a larger study carried out in 2013 at the Australian National University (Canberra) on the language pair Australian English (AuE) - Italian and recently published (Giacomini 2014). Translatability issues arise from semantic obscurity linked to the presence of culture-specific words and concepts.

General language dictionaries usually define similes and other multiword expressions through a paraphrase, without providing synonymic idioms even if these are available. This prevents translators from finding functionally equivalent phraseological data, which would be particularly useful for reproducing semantic and pragmatic features as well as the overall familiarity of the multiword expression in the target language. In the case of language pairs with phraseology reflecting distinctive cultural marks (Wierzbicka 2010), the exploration of corpus data can efficiently support the selection of adequate phraseological equivalents through reliable quantitative measures, thus forming a useful dictionary basis.

## 2 Object and sources

Similes are based on an explicit comparison between entities and are semantically related to metaphors, in which resemblance becomes implicit and one thing is understood and experienced in terms of another (Lakoff/Johnson 2003: 21 ff., Wikberg 2008: 128). In the usually regular syntactic structure of similes, the resemblance relation between the two compared entities is expressed by a connective, mostly *like/as* in English and (*così*) *come* in Italian. In addition, similes can reveal a different phraseological status: they can be defined either as collocations or as semi-idioms, according to the transparency of the comparison.

Two comparison patterns in AuE have been considered, the first being similes containing the phrase *full as* (e.g. *full as a tick*) meaning a) “having no empty space”, b) “having eaten to one’s limits or satisfaction”, or c) “drunk”, and the second involving a single culture-specific lexeme as the second compared entity, mostly a native animal. The semantic pivot is the contextual reading of the word that designates the shared property (*tertium comparationis*) and that determines the referential object of the multiword expression as a whole, both on the denotative and the connotative level. However, whereas similes belonging to the first pattern are made up of elements that are semantically transparent in their literal and figurative meanings, both for the English and the Italian native speaker, the others confront the translator with the presence of *realia* (cultural keywords) involving a culture- or environment-specific referent (Peters 2007: 249-251).

AuE similes were chosen on the basis of their relevance in a large-scale digital corpus of full-text Australian general news sources<sup>1</sup>, major general language dictionaries of AuE, and selected dictionaries of idioms or colloquialisms. Italian monolingual general language dictionaries and a comparable newspaper corpus (articles published between 2000 and 2013 in major Italian newspapers, totalling around 980 million words) were also employed for this purpose. Due to their stable structure, similes can be split into bigrams. The closest Italian equivalents of the semantic bases (e.g. *full* ≈ *pieno*, *sazio*, *ubriaco*) can be used to query the corpus for their collocators. The absolute frequency of the extracted bigrams can be compared with their log likelihood value, which provides reliable information on the association strength of a certain bigram and thus on its suitability as a phraseological equivalent (cf. Dobrovolskij 2009<sup>2</sup>). The results of data analysis are displayed in Table 1 according to an onomasiological procedure, which assigns Italian phraseological units to the concepts expressed by the AuE similes. Up to five equivalents for each concept are shown and arranged according to their absolute frequency *F* in the corpus and the log likelihood ratio *LL*.

---

1 National and regional newspaper texts covering the period 1985 to 2012, the Australian Corpus of English (ACE), and the Trove database (National Library of Australia).

2 Log likelihood has been chosen because of its reliability with sparse data, which is the case of the chosen words in the AuE corpus (for the topic of *LL* and normal distribution cf. McEnery et al. 2006, 53).

<b>to be full as a goog/ state school (hat rack)/ catholic school/ fat lady's sock</b>	
(a) "full/ overcrowded"	essere pieno zeppo (1799/>100), essere pieno come un uovo (194/>100), essere pieno da scoppiare (10/>100), essere pieno come una botte (1/9)
<b>to be full as a goog/ tick/ boot/ fairy's phone book/ fat lady's sock</b>	
(b) "full up/ satiated"	essere pieno come un uovo (194/>100), essere pieno da scoppiare (10/>100), essere pieno come un otre (1/17)
(c) "full/ drunk/ intoxicated"	essere ubriaco fradicio (567/>100); avere bevuto come una spugna (5/21); essere pieno come un otre (1/17), essere pieno come una botte (1/9)
<b>to be mad as a cut snake, to be pissed as a parrot</b>	
(d) "angry/ nervous"	essere (incavolato/...) nero (245/>100); essere arrabbiato/ incavolato/... come una bestia (27/>100)/ una belva (7/>100)/ una iena (6/>100)/ una biscia (5/>100)
<b>to be pissed as a parrot: cf. meaning (c)</b>	
<b>to be mad as a cut snake, to be mad as a gum tree full of galahs</b>	
(e) "crazy/ eccentric"	essere fuori di testa (1548/>100), essere tutto matto (343/>100), essere pazzo/ matto da legare (96/>100), essere pazzo/ matto come un cavallo (27/>100), essere fuori come un balcone (27/>100)
<b>to be (as) game as Ned Kelly</b>	
(h) "game/ brave/ bold"	avere coraggio da vendere (154/>100), avere un coraggio da leone/leoni (130/>100); avere il coraggio di un leone (7/59); essere coraggioso come un leone (5/83)
<b>to be flat out like a lizard drinking</b>	
(j) "fully extended"	essere/ stare lungo disteso (60/39)
(k) "with the utmost effort"	col massimo impegno (251/>100); impegnarsi al massimo (139/>100); col massimo sforzo (9/42); sforzarsi al massimo (2/7)
(l) "very busy"	essere pieno/ oberato di lavoro (404/>100), essere pieno di impegni (376/>100)
(m) "at full speed"	cf. (i)
<b>to be miserable as a bandicoot</b>	
(n) "wretchedly unhappy"	essere un povero diavolo (243/>100), essere un povero Cristo/cristo (238/>100)
(o) "contemptible"	-
(p) "needy"	essere povero in canna (139/>100), essere povero come Giobbe (2/31)

**Table 1: Corpus data in the target language with F and LL values.**

### 3 Translatability

AuE similes turn out to have close counterparts among phraseological Europeanisms, even though this may happen to varying degrees of equivalence. *Full* usually retains in the simile both its literal and figurative meanings, thus determining the total or partial compositionality of the multiword expression and contributing to its transparency.

Compositionality can be stated on the denotative (either literal or figurative) and connotative semantic level, but not always on the pragmatic level. The observations concerning compositionality of the multiword expression are also true of the second comparison pattern, in which a variable adjective (e.g. *mad, full, miserable...*) or a verb (e.g. *to shoot through*) designating shared property is combined with a culture-specific entity, used with a predicative or an adverbial function. In the case of similes belonging to this pattern, *realia* inevitably produce a lexical gap. Compatible data in the target language and culture cannot be sought for in terms of denotatively equivalent phraseological expressions, which is possible in the case of the *full*-pattern, but, at the most, in similes sharing the semantic pivot (*matto, veloce, coraggioso*, etc.) and with an equal degree of compositionality.

### 4 Lexicographic representation

The extracted clusters of equivalence candidates disclose the presence of alternative comparative patterns in Italian. For instance, we have prepositional phrases headed by *da* (*avere un coraggio da leone*) or *di* (*avere il coraggio di un leone*). In the Italian language, a significant part of the extracted lexical components in similes and other idioms stereotypically refer to animal behaviour and belong to a common European cultural heritage.

Dictionary data have been a useful resource to identify initial information on some widespread phraseologisms, but they fail to cover context-dependent phraseological variation and variability. The comparable newspaper corpus, instead, has revealed that, in concrete language usage, non-lexicalized and not conventionalised collocative or semi-idiomatic variants performing more specific message functions are constantly created on the basis of already existing patterns. The evaluation of corpus data can also disclose differences in phraseological distribution among languages (e.g. a strong tendency of the Italian languages towards metaphorical comparisons for purely descriptive purposes) and point out recent lexical formations which have not yet been recorded in dictionaries (cf. *essere fuori come un balcone*) but are already perceived by native speakers as familiar.

Corpus analysis in the target language supports the creation of a lexicographic basis for a bidirectional dictionary that is suitable for both translation directions. On the one hand, it activates passive knowledge in the native speaker of Italian by providing him/her with pragmatic tags in the source language and a wide choice of equivalents in the target language. On the other hand, it supports the AuE native speaker who is performing an active translation task by 1) allowing for a statistic evalua-

tion of the word combinations, 2) tagging equivalents with pragmatic marks and, most of all, 3) categorizing phraseologisms with varying idiomatic range (*pieno zeppo*, *pieno da scoppiare*) and distinguishing them from non-phraseological material (cf. Wiegand 2002: 52-53). Among non-phraseological equivalents are often single lexical items, usually an emphasised adjective (*strapieno*, *affollatissimo*; *sbronzo*) that can be specifically sought for in syntagmatic or paradigmatic dictionaries and further tested for phraseological strength. Every time a semantically and pragmatically equivalent phraseologism is missing, dictionary users are provided with an open set of non-phraseologisms, which function as reproducible syntactic models (e.g. superlative adjectives) and are particularly helpful for non-native speakers of the target language.

In order to take full advantage of the rich corpus materials and its bifunctionality, the dictionary should be designed as a digital resource, which should allow the translator to access lexicographic data along different combinable criteria, grasp the semantic connections existing between phraseological expressions, and retrieve unabridged corpus examples for each of them, in both the source and the target language.

The first two goals, *data accessibility* and *the disclosure of semantic connections*, can be primarily achieved through a coherent onomasiological macrostructure, which should group phraseologisms together along a common denotative/connotative meaning, and a systematic mediostructure, the aim of which should be to link each meaning to the correspondent phraseologisms and vice versa. The entry examples below show how lexicographic data in the section AuE-Italian can be displayed in a functional microstructural frame, and arranged based on a specific kind of search input (Table 2 according to a specific concept, Table 3 according to a specific phraseologism)<sup>3</sup>.

PHRASEOLOGISMS MATCHING THE CONCEPT IN THE SOURCE LANGUAGE	EQUIVALENTS IN THE TARGET LANGUAGE
<p><b>to be full as a goog</b> <i>coll.</i>                      = <b>state school (hat rack)</b> <i>coll.</i>                      = <b>catholic school</b> <i>coll.</i>                      = <b>fat lady's sock</b> <i>coll.</i></p>	<p>PHRAS: essere (pieno) zeppo, pieno come un uovo <i>coll.</i>, pieno da scoppiare <i>coll.</i>, pieno come una botte <i>coll.</i>                      ◆                      essere pienissimo, affollatissimo, strapieno <i>coll.</i>, stracolmo</p>

**Table 2: Search input: the concept FULL/OVERCROWDED.**

3 ◆ marks the division between phraseological and non-phraseological equivalents, = indicates similes referring to the same concepts.

CONCEPTS MATCHING THE PHRASEOLOGISM IN THE SOURCE LANGUAGE	EQUIVALENT PHRASEOLOGISMS IN THE SOURCE LANGUAGE	EQUIVALENTS IN THE TARGET LANGUAGE
FULL/OVERCROWDED	≈ <b>state school (hat rack) coll.</b> ≈ <b>catholic school coll.</b> ≈ <b>fat lady's sock coll.</b>	PHRAS: essere (pieno) zeppo, pieno come un uovo <i>coll.</i> , pieno da scoppiare <i>coll.</i> , pieno come una botte <i>coll.</i> ◆ essere pienissimo, affollatissimo, strapieno <i>coll.</i> , stracolmo
FULL UP/SATIATED	≈ <b>tick coll.</b> ≈ <b>boot coll.</b> ≈ <b>fairy's phone book coll.</b> ≈ <b>fat lady's sock coll.</b>	PHRAS: essere pieno come un uovo <i>coll.</i> , pieno da scoppiare <i>coll.</i> , pieno come un otre <i>coll.</i> ◆ essere pienissimo <i>coll.</i> , strapieno <i>coll.</i> , stracolmo <i>coll.</i>
FULL/DRUNK	≈ <b>tick coll.</b> ≈ <b>boot coll.</b> ≈ <b>fairy's phone book coll.</b> ≈ <b>fat lady's sock coll.</b>	PHRAS: essere ubriaco fradicio; avere bevuto come una spugna <i>coll.</i> ; essere pieno come un otre <i>coll.</i> , una botte <i>coll.</i>

**Table 3: Search input: the phraseologism to be full as a goog.**

The modular microstructure includes the following items: concept, phraseologism in the source language, phraseologisms in the source language referred to the same concept, and equivalents in the target language (subdivided into phraseological and non-phraseological equivalents). Pragmatic tags are added to a phraseologism or equivalent whenever required.

According to the lexicographic corpus, these Italian similes do not have a marked level of usage. However, a glance at their concordances in the newspaper corpus reveals a frequent tendency towards a colloquial register. In comparison with the source text similes, a generally more neutral level of usage has to be clearly stressed. The equivalents are selected on the grounds of their statistical relevance in the corpus and are not meant to cover the whole spectrum of equivalence in the target language. For the professional translator, they constitute the starting point from which further translation proposals can be generated.

CONCEPT	PHRASEOLOGISM	CONCORDANCES
FULL/ OVERCROWDED	<b>PHRAS: to be full as a goog coll.</b>	<p>The carpark in a certain flat pack emporium, starting with I and ending with A, middle letters K and E, was <b>as full as a goog</b>.</p> <p>We drove about 4000km with five adults and a lot of luggage and although the car was <b>as full as a goog</b> it was a good performer.</p> <p>She took a chance and opened up Swansea cafe. <b>Full as a goog</b>. And she must be doing something right because her business is a finalist in the Cafe category.</p> <p>How weird, though, that Old Trafford can hold only - even when it's <b>as full as a goog</b> - 23,000? They reckon they could have sold 70,000 tickets for the last day.</p>
	<b>PHRAS: essere pieno come un uovo coll.</b>	<p>Non possiamo farvi entrare – gridano gli organizzatori ai tornelli - dentro è <b>pieno come un uovo</b> e non si respire.</p> <p>La platea è quella di lavoratori provenienti da tutta la regione, ieri mattina al PalaDozza (<b>pieno come un uovo</b>)</p> <p>E ci piacerebbe che il palasport fosse <b>pieno come un uovo</b> (i biglietti numerati sono già stati esauriti ieri in prevendita</p> <p>acclamato come una star nella sua tappa aretina del tour in camper. <b>Pieno come un uovo</b> l'auditorium del palaffari</p>

**Table 4: Links to corpus concordances.**

In order to account for context-dependent variation in meaning and register, each equivalent needs to be hyperlinked to the correspondent corpus concordances both in the source and in the target language, which provide the translator with a large-scale database of real language examples (cf. Table 4 for corpus concordances of phraseologisms related to the concept FULL/OVERCROWDED).

This concept-based macrostructure could also constitute the architecture of a multilingual resource aimed at the representation of a core of cultural scripts shared by different languages (for recent research on Europeanisms cf. Piirainen 2012, Reichmann 2001).

## 5 Conclusions

Today there is still a significant need for specific lexicographic resources in digital form for translators, as they can remarkably improve access to data and actively assist the process of text production. In the best-case scenario, such resources could be integrated, together with other dictionaries and language tools, in multi-layer databases, allowing for advanced and customised search options.

This study shows that syntactic and semantic patterns can be effectively extracted from corpora and serve as lexicographic data in a digital resource which is specifically designed for supporting translation of culture-specific word combinations thanks to an onomasiological/conceptual macrostructure and a systematic mediostructure. Moreover, the study demonstrates that a corpus-based procedure is able to adequately account for phraseological variation and variability.

## 6 References

- Dobrovolskij, D. (2009), Zur lexikografischen Repräsentation der Phraseme (mit Schwerpunkt auf zweisprachigen Wörterbüchern). In Mellado Blanco, C. (ed.), *Theorie und Praxis der idiomatischen Wörterbücher*, Lexicographica Series Maior, pp. 149-168
- Giacomini, L. (2014), Languages in Comparison(s): Using Corpora to Translate Culture-Specific Similes. In: SILTA Studi Italiani di Linguistica teorica e Applicata, Pacini Editore 3/2013.
- Lakoff, G./Johnson, M. (2003), *Metaphors We Live By*, Chicago/London, University of Chicago Press.
- McEnery, T. et al. (2006), *Corpus-Based Language Studies: An Advanced Resource Book*, Milton Park/Abingdon/Oxon, Routledge.
- Peters, P. (2007), Similes and other evaluative idioms in Australian English". In Skandera P. (ed.), *Phraseology and Culture in English*, Berlin, Mouton de Gruyter, pp. 235-256.
- Piirainen, E. (2012), *Widespread Idioms in Europe and Beyond: Towards a Lexicon of Common Figurative Units*, Frankfurt am Main, Peter Lang.
- Reichmann, O. (2001), *Das nationale und das europäische Modell in der Sprachgeschichtsschreibung des Deutschen*, Freiburg (Schweiz), Universitätsverlag.
- San Vicente, F. (ed.), *Lessicografia bilingue e traduzione*, Milano, Polimetrica
- Wiegand, H.E. (2002), Äquivalenz, Äquivalentendifferenzierung und Äquivalentpräsentation in zweisprachigen Wörterbüchern: Eine neue einheitliche Konzeption. In *Symposium on Lexicography XI: Proceedings of the Eleventh International Symposium on Lexicography*, Copenhagen, pp. 17-57.
- Wierzbicka, A. (2010), *Experience, Evidence, and Sense: The Hidden Cultural Legacy of English*, OUP.
- Wikberg, K. (2008), Phrasal similes in the BNC. In Granger S., Meunier F. (eds.), *Phraseology: An Interdisciplinary Perspective*, Amsterdam/Philadelphia, John Benjamins, pp. 127-142.



# Lexical Variation within Phraseological Units

Tarja Riitta Heinonen  
Institute for the Languages of Finland  
tarja.heinonen@kotus.fi

## Abstract

This paper discusses lexical variation in phraseological units from theoretical and lexicographical perspectives. The starting point is the observation that the existence of lexical variation is sometimes disputed in principle. It has been argued that a change in a single word is sufficient to change the meaning of the whole, thus creating a new expression. Another argument is that after allowing variation in one word one cannot but allow it in multiple words, which quickly turns the original expression unrecognizable. In contrast to these views, this paper argues that lexical variation is not arbitrary but follows certain principles. When all contributing factors are taken into account, the variation in phraseological units is often item-specific, and yet it conforms to general patterns. Towards the end of the paper, Petrova's (2011) multi-level model will be introduced, offering a promising view for theoretical analysis. However, in dictionary work it is reasonable to adhere to generally accepted conventions and not to complicate the structure of the entries too much. An ideal entry gives a sound corpus-based description with representative examples of usage.

**Keywords:** phraseology; lexical variation; usage-based; idiomatic meaning; lexicographical practices

## 1 Introduction

Lexical variation within phraseological units raises theoretical and practical problems. One of the major questions is how phraseological units are learned and recognized if not with the help of the lexical items that constitute them. This paper starts with the question of whether lexical variation is acceptable in the first place (either in lexicography or in theory). Most dictionaries I have consulted allow variation but there are differences in “how much” and “what kind”.

I will show that lexical variation is not radically different from other types of variation, grammatical and structural variation, and that their workings can be described in a common model.

## 2 Two Opposing Views

It is common knowledge today (Moon 1998a: 92; Atkins & Rundell 2008: 168 among others) that there is a considerable amount of variation within multiword expressions, or phraseological units as I call

them in this paper. However, there are two opposing views on how to deal with expressions that are very similar to each other, except for single lexical choices as in, for instance, (1):

(1) rats desert/leave/quit/forsake a (sinking) ship (ODEI)

Some researchers exclude lexical variation from phraseological units in principle: substituting a word for another would automatically mean that the result no longer represents the same item as the original one (Wulff 2008: 76). In contrast, other researchers consider lexical substitutability as one type of variation alongside morphosyntactic variation (e.g. Sköldbberg 2004).

The opposite views may be due to differences in theoretical frameworks, but they may also be related to the amount of real life data researchers are familiar with. In my data set, collected from newspapers and the Internet, there is a considerable amount of variation in which two or more (near)synonyms occur in one and the same context without difference in meaning, as in the case *rats desert/leave a sinking ship*. The sheer number of such cases requires attention.

If lexical substitutability is allowed, it leads to a question of how phraseological units are defined – and originally recognized – if not with the help of the words that constitute them. On the other hand, if all expressions that are not lexically identical represent different phraseological units, we are left with plenty of units that highly resemble each other and without any formal means to record this in the lexicon.

### 3 Variation Patterns

In order to get a more detailed idea of what kind of alternation patterns there are in phraseological units, I will briefly consider a few examples from earlier studies on variation. Here I will mostly rely on Moon's systematic work on fixed expressions in English (1998b) and my own studies on Finnish verb phrase idioms (Heinonen 2013, 2007), but I will cite examples from various sources.

#### 3.1 Grammatical, Constructional and Lexical Variation

I will start by dividing the area of variation into three subfields: grammatical, constructional and lexical. These three phenomena are conceptually separable, but they co-occur in actual utterances. For instance, a constructional variant often occurs with specific lexical choices. By grammatical variation I mean variation in morphosyntactic features such as number, definiteness, voice and tense. Typically, the predicate verb inflects, albeit not quite freely (Moon 1998b: 94), but its nominal complements tend to be fixed in specific forms (2). However, in some cases also features in noun phrases may vary, especially if the idiomatic expression is metaphorically analyzable (3).

(2) she *gets* ~ *got* cold feet (! a cold foot)<sup>1</sup>

(3) (Swedish:) *dra sitt strå* (~ *sina strån*) till stacken

literally: drag one's straw (~ straws) to the stack

'contribute one's share to a common purpose' (Sköldberg 2004: 203, 311-312)

Inflectional restrictions are idiom-specific: for instance, some phraseological units passivize, some do not. Uttermost, the list of restrictions covers almost all the features: Čermák (2001: 15) cites a dictionary entry for the Czech idiom *tahat někoho za nohu* 'to pull someone's leg' which states that the predicate verb does not normally occur in the interrogative, negative, passive, conditional, imperative, future, or in the 1st person. It looks like these restrictions cannot be purely grammatical, but they describe the way the idiom is normally used.

For (2), the idiom dictionaries CCID and ODEI give also constructional variants with different predicate verbs (4a-b). This alternation pattern with verbs 'get', 'have' and 'give' is typical of possessive idioms.

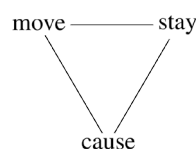
(4a) *get* cold feet or *have* cold feet about something (CCID)

(4b) *get*, (begin to) *have*, or *give* somebody cold feet (ODEI)

The constructional variants also include the alternative expressions of causation, states and processes. An example of causative variant is given in (5). – Moon (1998b: 139ff) lists these and some more patterns under the label "systematic variation".

(5) *go* through the wringer – *put* someone through the wringer (Moon 1998b: 141)

Generally speaking, the variation in the above patterns forms a simple grid with three points: causation, change and state, see figure 1. All these points can have a possessive interpretation as well: one could represent that as parallel to the basic grid. Verbs like *give*, *put* and *throw* are among the verbs that occur in the causative pattern, verbs like *get* and *go* express change or movement, and the verbs *be* and *have* are the primary state verbs.



**Figure 1: A basic constructional grid for causation, change or movement, and state.**

---

1 ! before the variant means that the idiomatic meaning gets lost.

Constructional and lexical variations also meet in extra modification (6).

(6) a *recent* tempest in a *publishing* teapot (Ernst 1981: 54)

Finally, the main focus of this paper, “pure” lexical variation occurs within a single construction. In examples (7) to (9) two or more near-synonymous or otherwise conceptually related words are in paradigmatic alternation. In (7) and (8), both choices are conventionalized, in (9), the first two.

(7) a *chink* ~ *crack* in one’s armour (ODEI)

(8) *flog* ~ *beat* a dead horse (Moon 1998b: 133)

(9) (Finnish:) kahvihammasta *pahottaa* ~ *kolottaa* ~ *särkee* (and several other verbs meaning ‘ache’)

literally: a coffee tooth is aching

‘someone has a craving for coffee’ (Heinonen 2013: 123-124)

It is important to notice that *not* all lexical variation is like this. In example (10b), the substitute for *family*, “the EU”, is not a semantically related word to *family*, but an expression whose referent can be seen as a kind of family. Here we are not operating inside the lexicon, but with categories or sets defined by common attributes (Philip 2008: 105-106).

(10a) black sheep of the *family*

(10b) black sheep of the *EU* (Philip 2008: 106)

Two or more lexical elements of one phraseological unit may also co-vary. Often they vary basically independently of each other (11), even though some combinations are more typical than others.

(11) shake ~ quake ~ quiver in one’s boots ~ shoes (Moon 1998a: 95, Moon 1998b: 161)

As far as I know, covariance is not widely studied in phraseology. Moon (1998b: 161ff) refers to cases like (11) as “idiom schemas”. The variants are listed and given some cover terms like ‘oscillate in one’s footwear’. In Moon (2008), she studies a similar phenomenon in similes of the structure ‘thin’ + *as* + ‘something which is very narrow in comparison to its length’. Here, it is easy to see that the concrete realizations differ from each other in connotations even though the schema represents the variants in very much the same way as in (11). *Skinny as a rat* sounds less attractive than *thin as a whippet* (id. 9-11). There are also collocational preferences.

In some phraseological units, however, one variant is dependent on the other (12):

(12) (Finnish:) lähtee kuin *hauki rannasta* ~ *talonmies peltikatolta* ~ *faarao sarkofagista* (and many others)

literally: leaves like a pike from a shore ~ a janitor from a tin roof ~ a pharaoh from a sarcophagus  
‘leaves quickly’

In this simile schema, someone or something leaves a place where they either belong or that they are at least strongly associated with. An attempt at explaining the mechanisms behind covariance is made by Stefanowitsch and Gries (2005). Their answer is, unsurprisingly, that the varying elements cohere semantically. More exactly, they divide semantic coherence into (at least) three different kinds: coherence based on frame-semantic knowledge, coherence based on prototypes, and image-schematic coherence. To my understanding, the coherence in similes in (12) would represent “frame-semantic” coherence.

### 3.2 Lexical Variation in Dictionary Entries

The presentation of dictionary entries has traditionally been dense. This also shows in how phraseological variants may be placed side by side, for example, in ODEI and in Duden (13).

(13) einen klaren/kühlen Kopf bewahren/behalten (Duden s.v. *Kopf*).

The practices vary, but generally speaking dictionaries have specific means to indicate if a phraseological unit has more than one lexical choice (a slash in 13, the word *or* and parentheses in 14, a comma in 15). It follows that lexical variation in phrases is usually *de facto* accepted in lexicography. How systematically lexical variation is taken into account depends mostly on the type of the dictionary: phraseological dictionaries such as ODEI and CCID are systematic, general dictionaries, such as KS, are less so.

(14) throw (or pour) cold water on (NODE s.v. *cold*)

(15) (Finnish:) erottaa, seuloa jyvät akanoista (KS s.v. *jyvä*)  
'separate, sift the wheat from the chaff'

Dictionary conventions actually work equally well for bundles of idiomatic expressions that have different, perhaps even opposite meanings, as in (16):

(16) say the right/wrong thing (ODEI)

Seen this way, a dictionary entry could also stand for partly formal idioms, following the terminology by Fillmore, Kay & O'Connor (1988). In their parlance, formal idioms are lexically open syntactic patterns or constructions, while what we have traditionally called idioms are substantive or lexically specified idioms. What I find interesting here is the area between these, ie. partially open idioms. An actual example of a partially open, partially specified (sub)entry is given in (17):

(17) play the ---- card e.g., *he saw an opportunity to play the peace card* (NODE s.v. *card*)

Two or more alternating slots predict multiplied combinations. In the case of (13), two adjectives (*klar, kühl*) and two verbs (*bewahren, behalten*) combine in four different ways. In this case, the generalization is valid: all predicted forms actually occur in texts. However, all combinations are not as common: corpus studies via DWDS reveal differences in frequencies. As far as I know, there are no lexicographic conventions that help the user pick the most idiomatic combination(s) in such situations.

Sometimes the variation is not limited to specific words but the alternation set is lexically open. Many Finnish verb phrase idioms allow plenty of variation in the predicate verb. I searched for verb variants in the idiom *heittää hapuloita rattaisiin* ('throw batons to (the wheels of) a carriage [in order to prevent something from succeeding]') in a newspaper corpus FTC and found about 20 different predicate verbs (some of them listed in 18a-e). Of these, six verbs are rather common (16–30 hits), three occur about five times and the rest are mostly hapaxes. In (18a-e), the attested verbs are divided into meaning groups (based on Heinonen 2007: 155 and Heinonen 2013:156-157), and the main six variants are in boldface.

(18a) 'throw': **heittää, heitellä**, viskoa, viskellä

(18b) 'put': **panna, laittaa**, asettaa, asetella

(18c) 'stick': **pistää**, pistellä, tuikkia

(18d) 'push': työntää, pötkkiä

(18e) 'hit': **lyödä**, iskeä

The general dictionary KS mentions the verbs *heitellä* and *panna* in this idiom, and the phraseological dictionary SSIS lists all the six common ones. (SSIS is based on the same corpus as my search.)

One solution to the problem of long lists is to generalize over the choices as in (18a-e). However, it is not always clear how these sets should be labeled and interpreted. Also, some lexical items are preferred over others with similar meanings. Notice that in (17), the idiom contains an open, unspecified slot, but the explanation given – that it should refer to an “issue or idea” that can be exploited especially for “political advantage” – does not really limit the choice of appropriate fillers, and this is possibly the most we can say, besides giving attested examples. Another problem is that the sets (in whatever way they are defined) tend to leak. There are a few Finnish idioms with a lexical item that refers to a human head. Still, the appropriate sets for a ‘head’ are partly lexically specified: in one idiom you can refer to a head metaphorically as a ‘cabbage’, in another as a ‘pin’, but not the other way round (Heinonen 2013: 197-198). Jezek and Hanks (2010) make the same observation, saying that paradigmatic sets of words do not map neatly onto conceptual categories, and neither are there stable generalizations over different contexts.

## 4 One or More Units

There are two further points that speak for lexical variation in phraseological units: interpretation of regular derivational variants (4.1) and language learning as usage-based process (4.2). Issues of variation vs. modification and canonical forms are dealt with briefly in (4.3).

### 4.1 Difference in Meaning as a Criterion

A difference in meaning has been a crucial factor in separating phraseological units from each other. However, it should be kept in mind that a change in one element usually contributes to the full meaning quite predictably. In Finnish, certain derivative affixes can change the aspect of the verb. Substituting a verb with its derived counterpart expressing frequentative aspect counts as lexical substitution (cf. 19a and 19b); still, the resulting difference in meaning is quite straightforward: it is actually comparable to how inflectional affixes are interpreted (cf. 19b and 19c):

(19a) (Finnish:) *heittää* kapuloita rattaisiin

(literally:) throws batons to (the wheels of ) a carriage

'places obstacles in order to prevent something from succeeding'

(19b) *heittelee* kapuloita rattaisiin

keeps throwing batons [...]

(19c) *heitti* kapuloita rattaisiin.

threw batons [...]

I would suggest that the predictability of a change in meaning also applies when a word is substituted with an unrelated word. As long as the phraseological unit is recognized, its meaning can be modified according to the contribution the substitute part carries along.

### 4.2 Where are the Limits of One Unit?

At which point does a phraseological unit change into another one? Čermák (2001: 7) raises this question of an idiom's identity if we allow lexical variation within them. I believe that it is not possible to define from outside how the units are organized in the mental lexicons of speakers. It is likely there is not just one way in which phraseological units and lexical items are connected to each other. The individual lexicons develop hand in hand with language use. When we learn an expression, we also learn, little by little, how it is used: inflection, meaning(s), suitable contexts.

Čermák's example idiom 'to pull someone's leg' referred to earlier is given a voluminous description in his and his colleagues dictionary (Čermák, Hronek & Machač 1994): not only does it cover the idiom's valency, inflectional restrictions, meaning, style and appropriate context of use, but also the rela-

ted expressions and even equivalents in various other languages. Lacking knowledge of Czech, I cannot comment on this particular case but, in general, this sounds like a database and network of idioms as independent, lexically static units. Observations on language use may therefore lead to opposite conceptions on what phraseological units are like.

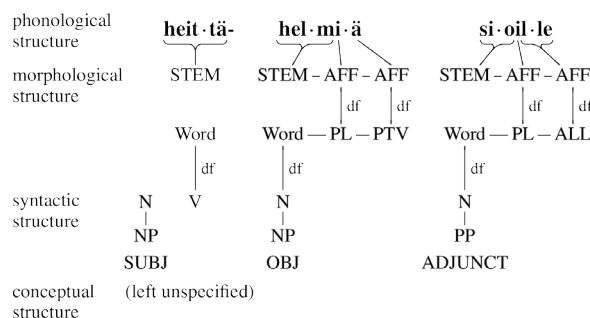
### 4.3 Canonical Forms and Variation

Talk about variation raises the question what the constants are. A canonical form of a phraseological unit appears to be a paradox. Philip (2008: 95, 103) refers to canonical forms as the most typical forms and, at the same time, as something that are generally outnumbered by corresponding non-canonical forms in language corpora. Canonical forms are often identified with dictionary citation forms, as if lexicographers would receive them by some divine announcement.

There is also a lexicographic tradition to keep established variation and temporary modification as separate phenomena (e.g. Burger, Buhofer & Sialm 1982). However, seen through corpora, variation and modification blend. They also derive from the same sources.

### 4.4 A Multi-Level Approach

If we accept lexical substitutability within one phraseological unit, we need to demonstrate how such a phraseological unit can be defined and identified. One applicable idea is the multi-level model by Petrova (2011), which is derived from Jackendoff’s Conceptual Semantics and more directly from Nikanne’s Tier-net model (Nikanne 2005: 191-210; Petrova 2011: 110ff). In Petrova’s model, one unit can vary at several tiers (phonological, morphological, syntactic, conceptual) at the same time (see figure 2). The idea is that the lexical items, as well as some of the inflectional affixes, are chosen by default (this is represented by df-links in the figure), but it is possible to substitute or even leave out one or more of the default items. The predicate verb inflects freely in the example case. The conceptual structure is relevant when substituting lexical items, since any other words following the same syntactic pattern would not do. For instance, *to throw balls to boys* does not represent the same phraseological unit as *to throw pearls to pigs*.



**Figure 2: The structure of the idiom X heittää helmiä sioille ‘X casts pearls before swine’ (literally ‘X throws pearls to pigs’) according to Petrova (2011: 151). The figure is heavily simplified from the original by the present author. PTV (partitive) and ALL (allative) are case markers.**



The model supports the view that lexical variation can be dealt with together with other varying elements and not as a separate issue. Petrova cites a good number of actual uses of the idiom *heittää helmiä sioille*, ‘cast pearls before swine’, which illustrate how the different sources of variation function together (20a-d). One typical pattern is a verbless construction (20a); in fact, it is the most common variant when all default features are taken into account simultaneously (ca. one quarter of all 480 tokens in Petrova’s study) and far more common than a “canonical” citation form (Petrova 2011: 219-220). In quite a few examples the syntactic pattern, inflectional forms or the words are not the default ones (20b-d).

(20a) Ei helmiä sioille, kuten sanonta kuuluu.

‘No pearls to pigs, as the saying goes.’ (Petrova 2011: 278)

(20b) *possulle* heitetty helmi

‘a pearl thrown to a piggy’ (id. 284)

(20c) viisauden helmien *jakamiseen* myös *meille typerämmille yksilöille*

‘distributing pearls of wisdom to us more stupid individuals’ (id. 281)

(20d) *Menikö* taas helmet *sinne kaukaloon* [?]

‘Did pearls go to that trough again [?]’ (id. 279)

The conceptual structure (left unspecified in figure 2 for ease of representation) contains detailed information on verb semantics, argument structure, what sorts of objects nominal complements refer to etc. (id. 137-138). It is straightforward to substitute *pig* for its near-synonym *piggy* (20b), but to connect *a trough* with *pigs* as in (20d) requires a larger situational framework. Since the predicate verb ‘throw’ involves causation, it is predictable that the idiom participates in the causative alternation pattern (20d, cf. figure 1). Petrova points out that the thrown objects, *pearls*, are evaluated by the speaker as good, and the recipients, *pigs*, as inadequate in some way (id. 24, 138). The variant to ‘pigs’ in (20c), ‘more stupid individuals’ reflects this.

It could be argued that this particular idiom is not representative in allowing much more variation than is normally the case. As can be seen, the idiom functions even with just one default lexical choice, *helmi*, ‘pearl’. One could claim accordingly that the noun *helmi* bears a metaphorical meaning that is available in any syntactic context. There are other similar cases of syntactically “free” idiomatic nouns, one often cited example is that of *carrot and stick* (Moon 1998a: 96). This is all true, but in examples (20b-d) the syntax is still pretty regular, and the lexical choices are bound to the default ones. I would claim that all idioms have specific ways of expressing themselves: *heittää helmiä sioille* and *heittää hapuloita rattaisiin* look similar on the front but differ in what options they offer for a language user.

The status of the lexicon in Petrova’s model has probably made the model especially receptive to lexical variation. The lexicon is seen as an interface between phonological, syntactic and conceptual le-

vels. As a drawback, this sometimes complicates discussion, as lexical items are referred to in terms of their phonological structure.

How the model keeps track of encountered usages of phraseological units is not, however, clear to me. All non-default choices are after all not as predictable. Sometimes, it may even be hard to pick the default among many potential ones (cf. 18a-e). Actually, the linking system emerges from usage, and is continually modified usage-based. Instead of one default link, there could be several, stronger or weaker links. In Petrova's model, non-default lexical choices are mostly licensed via conceptual structure (id. 317-344) or referential tier (id. 370-374). For instance, the door to constructional alternation opens through the conceptual structure of the predicate verb, in this case the verb *heittää* expressing caused motion. Cases like *black sheep of the EU* (10b above) fall into the referential tier, the substitution being partly based on extra-linguistic factors.

## 5 Practical Considerations

A phraseological unit is typically limited not only to the core lexical items but to a specific combination of restrictions and preferences with respect to inflectional features, grammatical patterns, contexts etc. All these factors should be taken into account when formulating a dictionary entry. However, the result should illustrate the most common patterns in usage, instead of overwhelming the user with unrelated details. I believe that one of the most successful ways to convey information on usage is to select representative examples (Fox 1987). It is also important to notice that phraseological units are not alike. For instance, the two *heittää* idioms cited in this paper behave differently in some respects even though they are almost identical morphosyntactically. For example, the verb *heittää* is not a clear default choice in the idiom *heittää kapuloita rattaisiin*, but it is in the idiom *heittää helmiä sioille*. The most common lexical variants are given in recent corpus-based phraseological dictionaries (CCID, SSIS).

## 6 Conclusion

Lexical variation within phraseological units raises theoretical and practical problems. The scope of lexical variation is quite wide overall and it is entangled with grammatical and contextual factors. One of the major questions is how phraseological units are learned and recognized in all their varying forms.

The view on phraseology and lexicon taken in this paper is usage-based. As I see it, all items are learned over and over again in contexts, and this leads to memorized items with a collection of information on different aspects, such as pronunciation, style, inflectional forms, conveyed meanings, and situational contexts. The memorized items are not stable, and there are numerous ways in which

they can be modified. However, since these items are usage-based, earlier experiences on their modifiability guides the future variations.

Meanwhile practical considerations guide us on how to write articles in dictionaries. Dictionary users should not be overwhelmed with detailed information on something that does not help them to understand and use the current expression. Besides, language learners are not as dependent on dictionaries as they used to be: it is common practice today to search the net to check if a specific wording is in use or not. Moreover, it is not even possible to list all thinkable options that are available to a language user.

## 7 References

- Atkins, B.T.S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Burger, H., Buhofer, A. & Sialm, A. (1982). *Handbuch der Phraseologie*. Berlin: Walter de Gruyter.
- CCID = *Collins COBUILD Idioms Dictionary*. 2<sup>nd</sup> edn. Glasgow: HarperCollins Publishers, 2002.
- Čermák, F. (2001). Substance of idioms: perennial problems, lack of data or theory? In *International Journal of Lexicography*, 14(1), pp. 1-20.
- Čermák, F., Hronek, J. & Machač, J. *Slovník české frazeologie a idiomatiky. Výrazy slovesné*. [‘Dictionary of phraseology and idiomatics, (part 3) verbal expressions’.] Praha: Academia, 1994.
- Duden = *Duden Deutsches Universal Wörterbuch*. 6<sup>th</sup> edn. Mannheim: Dudenverlag, 2006.
- DWDS = Das Projekt Digitales Wörterbuch der deutschen Sprache: DWDS-Kernkorpus. Accessed at: <http://www.dwds.de> [5/12/2011].
- Ernst, T. (1981). Grist for the linguistic mill: idioms and “extra” adjectives. In *Journal of Linguistic Research*, 1, pp. 51-68.
- Fillmore, C.J., Kay, P. & O’Connor, M.C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. In *Language*, 64, pp. 501-538.
- Fox, G. (1987). The case for examples. In J.M.Sinclair (ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins Publishers, pp. 137-149.
- FTC = Finnish Text Collection. Accessed at <https://sui.csc.fi/group/sui/lemmie> [2.4.2014]. Information on the contents at <http://www.csc.fi/english/research/software/ftc>.
- Heinonen, T.R. (2013). *Idiomien leksikaalinen kuvaus kielenkäytön ja vaihtelun näkökulmasta*. [‘Idioms as lexical constructions: usage and variability’.] PhD thesis. University of Helsinki, Finland. Accessible at: <http://urn.fi/URN:ISBN:978-952-10-8555-0> [17/03/2014].
- Heinonen, T.R. (2007). Variation and flexibility within verb idioms in Finnish. In M. Nenonen, S. Niemi (eds.) *Collocations and Idioms 1, Papers from the First Nordic Conference on Syntactic Freezes*, Joensuu, 19-20 May, 2006. Joensuu: University of Joensuu, pp. 146-159.
- Jezek, E., Hanks, P. (2010). What lexical sets tell us about conceptual categories. *Lexis*, 4, *Corpus Linguistics and the Lexicon*, pp. 7-22. Accessed at: [http://lexis.univ-lyon3.fr/IMG/pdf/Lexis\\_4.pdf](http://lexis.univ-lyon3.fr/IMG/pdf/Lexis_4.pdf) [17/03/2014].
- KS = *Kielitoimiston sanakirja*. [‘Authoritative Dictionary of Contemporary Finnish’.] Helsinki: Institute for the Languages of Finland, 2012.
- Moon, R. (2008). Conventionalized *as*-similes in English: a problem case. In *International Journal of Corpus Linguistics*, 13(1), pp. 3-37.
- Moon, R. (1998a). Frequencies and forms of phrasal lexemes in English. In A.P. Cowie (ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press, pp. 79-100.

- Moon, R. (1998b). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.
- Nikanne, U. (2005). Constructions in Conceptual Semantics. In J.-O. Östman, M. Fried (eds.) *Construction Grammars: Cognitive Grounding and Theoretical Extensions*. Amsterdam: John Benjamins, pp. 191-242.
- NODE = *The New Oxford Dictionary of English*. Oxford: Oxford University Press, 2001.
- ODEI = Cowie, A.P., Mackin, R. & McCaig, I.R. (1993) [1983]. *Oxford Dictionary of English Idioms*. [= *Oxford Dictionary of Current Idiomatic English, Volume 2.*] Oxford: Oxford University Press.
- Petrova, O. (2011). *Of Pearls and Pigs: A Conceptual-Semantic Tiernet Approach to Formal Representation of Structure and Variation of Phraseological Units*. PhD thesis. Åbo: Åbo Akademis Förlag. Accessible also at: <http://urn.fi/URN:ISBN:978-951-765-583-5> [17/03/2014].
- Philip, G. (2008). Reassessing the canon: 'fixed' phrases in general reference corpora. In S. Granger, F. Meunier (eds.) *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins, pp. 95-108.
- Sköldberg, E. (2004). *Korten på bordet: Innehålls- och uttrycksmässig variation hos svenska idiom*. ['Cards on the table: Variations in content and expression in Swedish idioms.'] PhD thesis. Gothenburg: Meijerbergs institut för svensk etymologisk forskning, University of Gothenburg, Sweden.
- SSIS = Muikku-Werner, P., Jantunen, J.H. & Kokko, O. (2008). *Suurella sydämellä ihan sikana. Suomen kielen kuvailuva fraasisanakirja*. ['Descriptive Dictionary of Finnish Phraseology.'] Jyväskylä: Gummerus.
- Stefanowitsch, A., Gries, S.Th. (2005). Covarying collexemes. In *Corpus Linguistics and Linguistic Theory*, 1. pp. 1-43.
- Wulff, S. (2008). *Rethinking Idiomaticity: A Usage-Based Approach*. London: Continuum.

### **Acknowledgements**

I am grateful to Maria Vilkuna and three anonymous referees whose comments and suggestions helped me clarify my presentation on several points.

# ***Prendere il toro per le corna o lasciare una bocca amara?* – The Treatment of Tripartite Italian Idioms in Monolingual Italian and Bilingual Italian-English Dictionaries**

Chris Mulhall  
Waterford Institute of Technology  
cmulhall@wit.ie

## **Abstract**

This paper undertakes an empirical investigation into the treatment of tripartite Italian idioms in selected monolingual Italian and bilingual Italian-English dictionaries. Tripartite idioms are phrasal constructs typically arranged into one of the three following syntactic forms: V+N+N, V+ADJ+N or V+N+ADJ. From an organisational viewpoint, these idioms are somewhat more problematic for lexicographers due to the presence of a third lexical constituent. Current lexicographical practice adopts a largely subjective approach to dealing with idioms, which for the most part are V+N forms, therefore those with a wider syntactic form require more considered decision making when determining their point(s) of entry in a dictionary. Certain tripartite idioms also collapse into binomial (N+N) or nominal-adjectival forms (ADJ+N/N+ADJ), thus placing a question mark over the necessity to record their verb element or not. Together, these issues contribute to making tripartite idioms one of the most acutely difficult phrasal categories for lexicographers. This paper examines the entry points of 100 tripartite Italian idioms in selected monolingual Italian and bilingual Italian-English dictionaries published within the last decade. Firstly, it discusses relevant theoretical viewpoints relating to the lemmatisation of idioms over the last 30 years. Thereafter, it focuses on the notion of an idiom as a lexicographical entry in consideration of its varied intrinsic features and how these are productive in selecting an appropriate entry point. The presentation of the empirical data in the following section is stratified according to the publishing house of the analysed dictionaries and gives a comparative discussion on their respective approaches to dealing with tripartite idioms. Finally, the last section unifies the theoretical arguments and practical approaches and proposes a theoretical framework model for lemmatising tripartite idioms to offer a more coherent and consistent platform for their organisation in monolingual Italian and bilingual Italian-English dictionaries.

**Keywords:** Italian Lexicography; Tripartite Idioms; Lemmatisation

## 1 The Lemmatisation of Idioms: Theoretical Viewpoints

Lexicographical theorising over the last 30 years has regularly put forward proposals to remedy the problematic issue of how best to position idioms within a dictionary text. In their entirety, these arguments aim to achieve a more considered and structured lexicographical coverage of idioms, but fail to reach a consensus on the most suitable approach. Certain theorists advocate embedding idioms within the microstructure but diverge significantly on the number and position of their assigned entry points, for example, there is little theoretical agreement on a single listing strategy: Petermann (1983) (notional point of entry); Burger (1989) (unspecified entry point); Lorentzen (1996) (noun entry point); Mulhall (2010) (lexico-semantic entry point). Contrastingly, Tomaszczyk (1986) suggests entering idioms under each of their constituent's lemmas. The unitary semantic function of idioms equates them to having a word-like function in the lexicon, which Al-Kasimi (1977), Gouws (1991) and Botha (1992) argue is a substantive rationale for their lemmatisation in a dictionary. This, as Gouws (1991:86) states, prevents an 'ambiguous reading' of an idiom's lexical status by dictionary users. Lemmatising idioms accurately portrays their semantic status in the lexicon, but its lexicographical practicality remains questionable and untried. Harras and Proost (2005) advance a bespoke entry model for idioms, configuring their point of entry in accordance with their semantic features; resulting in semantically opaque idioms being lemmatised and semantically interpretable idioms sub-lemmatised. Adopting this particular method would move dictionaries to a more semantic-based lemmatisation model for idioms but is potentially anomalous given its proposal of different organisational principles for the same category of phrases.

An overview of the proposed entry methods reveals that semantics motivates many decisions relating to the most appropriate entry point with Mulhall (2010) taking into account that any lemmatisation model must also incorporate the notion of lexical variability, which, according to Moon (1998) can occur in between 12 to 40 percent of idioms. Another characteristic feature of idioms that is less topical in lexicographical debates is that of syntactic form. Therefore, considering the different semantic, lexical and syntactic properties of idioms may offer a more robust decision-making platform for identifying their most suitable entry point in a dictionary.

The majority of Italian idioms fall into the standard syntactic category of Verb (V) + Noun (N) with potential syntactic expansion to V/V+N or V+N/N if lexical variation is permissible in either the verb or noun component. In such cases, a lexicographer only has two (or possibly three) available point of entry options. A more problematic subset is that of tripartite idioms, some of which may have two constituents of the same word class (V+N+N) or contain three distinct word classes with two different syntactic structures (V+ADJ+N or V+N+ADJ). Tripartite idioms are particularly challenging for lexicographers on a number of levels. Firstly, the subjective identification of the most important or prominent element becomes more complicated due to the presence of a third lexical component. In the case of V+N+N idioms, lexicographers favouring a noun-based listing model must, from the outset, decide whether the first or second noun element is the most appropriate entry point. Secondly, the

V+ADJ+N and V+N+ADJ categories contain a small number of expressions that retain their idiomatic identity in a nominalised form, for example, *a gonfie vele*, *duro d'orecchio*, *il nodo Gordiano*, *la pecora nera*, etc. Important issues arise from these syntactically truncated forms; such as the necessity to record their verb element(s) or not and the importance of the adjectival element in their syntactic binding and lexical identity. The entirety of these issues and the failure of lexicographical practitioners to adopt and integrate recent theoretical suggestions or propose alternative practical solutions contribute to the long-term status of idioms as arguably the most problematic dictionary entry.

## 2 Redefining Idioms for Lexicography

Theoretical linguistics offers a multitude of rich and varied definitions for an idiom, but these typically comprise singular, one dimensional features, such as 'fixed expression' or 'non-compositional' or describe it through vague terms of references, such as 'relatively fixed expression'. Idioms, by their nature, are a linguistic concept; therefore such definitions may not offer the requisite scope to give lexicographers a wider understanding of their form and behaviour to deal with them accurately in the context of a dictionary. Therefore, to ensure a more representative lexicographical treatment of idioms, it is important to factor in their three most salient characteristics; semantics, lexis and syntax. In consideration of these characteristics, Table 1 sets out three feature-specific maxims to reconsider idioms as both a lexicographic entry and linguistic unit.

Feature	Definition
Semantics	Idioms are a semantically complex and compositionally divergent subset of expressions. This often results in a clear semantic disconnect between the lemma as a stand-alone lexical unit and as an idiom constituent.
Lexis	Idioms display different layers of lexical fixity. Their potential variability necessitates a dictionary entry strategy that not only recognises variable expressions but also records them in a consistent and representative way.
Syntax	Idioms are syntactically heterogeneous. Therefore, the number and word class of idiom constituents within any given idiomatic frame may potentially influences the number and position of allocated entry points in a dictionary.

**Table 1: Lexicographical Definitions of Idiom Features.**

Redefining the notion of an idiom as a linguistic unit and a dictionary entry, the following operational definition is proposed to achieve a more holistic and tailored lexicographic treatment of idioms based on their inherent features:

Idioms are a category of multi-lexical, syntactically diverse expressions showing various degrees of semantic compositionality, some of which contain lexical constituents that can substituted for idiomatic equivalents.

This broader definition encapsulates the most salient characteristics of idioms; accurately portraying their status in the lexicon and describing features that are potentially influential in their lexicographical description and organisation.

### 3 Organisational Approaches to Idioms in Monolingual Italian and Bilingual Italian-English Dictionaries

An analysis of monolingual Italian and bilingual Italian-English dictionaries from the eighteenth century onwards shows that idioms, as a lexicographic entry, failed to gain a centrality and an organisational foothold in the design and content of these reference works. This, in part, can be traced to their perceived linguistic impurity in eighteenth and nineteenth century Italian society, which resulted in their limited coverage in mainstream dictionaries. A change in this outlook came in the early twentieth century due to new linguistic models, burgeoning dictionary content and a reformatted microstructure. But in many dictionaries idioms still remained a peripheral entity; a notable exception to this was the *Sansoni-Harrap Standard Italian-English Dictionary* [SHSIED] (1970-1975), which made the singular attempt of this era to systematise the coverage of idioms, but its subjective,<sup>1</sup> rather than substantive, criteria failed to address this problem for dictionary users. Decisions about restructuring the organisation of idioms and providing this information to users remain relevant, but overlooked, in modern day lexicographical practice. This paper aims to provide further evidence identifying the need for a lexicographical reform, at least in an Italian context, in the approach to organising idioms. The research sample includes 100 tripartite Italian idioms; subdivided into the following syntactic categories: 50 V+N+N, 30 V+N+ADJ and 20 V+ADJ+N. Selecting and organising the empirical sample brought to light two recurring trends: the prominence of the V+N+N syntactic structures within the Italian tripartite subset and the presence of a high frequency verb<sup>2</sup> (HFV) element in 39 of 100 expressions. To gain a comparative insight into any converging or diverging entry strategies based on the expectations of Italian speaking users a monolingual Italian and bilingual Italian-English dictionary from three publishing houses formed part of the research corpus. Selecting different dictionaries from the same publishing houses allowed an interesting exploration into ascertaining whether or not certain publishing houses follow any systematic procedure when attempting to record tripartite idioms in their monolingual and bilingual reference works. The dictionaries used are as follows: *Il Sansoni Inglese* [ISI] (2006); *Il Sabatini-Coletti* [ISC] (2007); *Il Ragazzini* [ZIR] (2009); *Lo Zingarelli* [ZLZ] (2009); *Hoepli Dizionario Inglese* [HDIN] (2007) and the *Hoepli Dizionario Italiano* [HDIT] (2008).

1 “The phrases, idiomatic expressions, proverbs, etc., that make up the phrase section are generally found under the first important word in the phrase” (SHSIED 1970: viii).

2 The following Italian verbs occur with a high frequency across a number of different phrasal and idiomatic expressions: *andare, avere, dare, essere, fare, mettere, prendere, stare, tenere, venire*.



Dictionary Publishing House	Sansoni		Zanichelli		Hoepli	
	Il Sansoni Inglese	Il Sabatini-Coletti	Il Ragazzini	Lo Zingarelli	Hoepli Inglese	Hoepli Italiano
V+N+N (N=50)						
Verb Entry	17	2	4	2	5	2
N1 Entry/N2 Entry	13/0	8/2	15/3	9/3	19/8	7/3
Double Noun Entry	0	12	7	11	2	13
Verb/Noun Combination Entries	12	19	14	20	4	20
Listed as a Nominalised Idiom	2	2	3	3	5	2
Showing a different Verb Element	1	3	2	0	2	2
Not Listed	5	2	2	2	5	1
V+ N+ADJ (N=30)						
Verb Entry	2	1	0	0	4	1
Noun Entry	15	1	0	7	11	5
Adjective Entry	0	5	11	0	3	1
Verb/Noun/Adjective Combination Entries	6	16	11	14	5	14
Listed as a Nominalised Idiom	1	3	7	4	2	6
Showing a different Verb Element	4	2	0	0	1	1
Not Listed	2	2	1	5	4	2
V+ADJ+N (N=20)						
Verb Entry	3	0	3	0	2	0
Adjective Entry	9	1	7	1	0	1
Noun Entry	0	4	0	5	9	3
Verb/Noun/Adjective Combination Entries	3	8	6	6	4	9
Listed as a Nominalised Idiom	1	0	1	2	0	1
Showing a different Verb Element	1	1	0	0	1	2
Not Listed	3	6	3	6	4	4

**Table 2: Empirical Data on the Entry Points of Italian Tripartite Idioms in Monolingual Italian and Bilingual Italian-English Dictionaries.**

### 3.1 Il Sansoni Inglese (2006) and Il Sabatini-Coletti (2007)

An often lamented failing of dictionaries is their failure to provide any guidance to users about the exact location of idioms. A notable exception in this regard is ISI (2006), which in the preface clearly informs users where to locate such expressions with accompanying examples. In contrast, its monolingual equivalent, the ISC (2007), takes a different approach, instead exemplifying certain idioms without explicitly indicating their position in the dictionary (see Table 3).

Il Sansoni Inglese (2006)	Il Sabatini-Coletti (2007)
<p>The phrases, idiomatic expressions and proverbs that make up the phraseology section are listed under the first key word contained in the expression (be it verb, noun or adjective). Therefore, for example, the proverb <i>he who pays the piper calls the tune</i> is found under the verb <b>pay</b> and the phrase <i>as hard as iron</i> is listed under the adjective <b>hard</b>. Likewise, the Italian proverb <i>le bugie hanno le gambe corte</i> is given under <b>bugia</b> and the phrase <i>cavalieri della Tavola Rotonda</i> under <b>cavaliere</b>.</p> <p>As an exception to this, certain extremely common verbs (<i>be, can, come, do, get, give, go, have, keep, let, make, must, put, take, will</i> in English and <i>andare, avere, dare, dovere, essere, fare, lasciare, mettere, potere, prendere, stare, tenere, venire, volere</i> in Italian) have been ignored in listing the phrases under the headwords. As a result, the phrase, <i>to get one's cards</i> is given under the headword <b>card</b>, and <i>prendere qcu. in castagna</i> is given under <b>castagna</b>.</p> <p>(<i>Il Sansoni Inglese</i> 2006:14)</p>	<p>All'interno delle parole piene che lo sviluppano, ma con grande evidenza, sono tratte anche <b>le unità polirematiche grammaticali</b>, ossia le locuzioni che hanno valore di preposizione, di congiunzione o di congiunzione testuale (<i>a conti fatti, a costo di, modo che, nella misura in cui...</i>).</p> <p>Ben diverso è il caso delle espressioni idiomatiche, tutte di senso figurato, che appartengono alla lingua comune e sono in genere ben familiari ai parlanti (<i>essere un pozzo di scienza; dare carta bianca; andare per le lunghe; tendere la mano; voltare pagina; cambiare registro</i>). Come appare evidente, queste fanno nesso fisso con un verbo).</p> <p>(<i>Il Sabatini-Coletti</i> 2007:16)</p>

**Table 3: Organisational Criteria for Idioms in Sansoni Publishing House Dictionaries.**

Like the diversity of their organisational approaches to these entries, the empirical analysis also reveals disparities in the treatment of the same syntactic idiomatic categories in the ISI (2006) and the ISC (2007). On a general level, this divergence can be measured through the number of assigned entry points; for example, the ISI (2006) allocates a single entry to the 61/100 expressions in contrast to a multiple listing strategy favoured by the ISC (2007) for 64/100 expressions. A possible explanation for this different approach is the strict adherence by the ISI (2006) compilation team to inserting idioms under the first key word, but the application of this method is not entirely rigid. For example, data from the empirical sample reveal that 21/100 expressions are listed twice or more and 54/100 expressions are recorded directly in line with the information given in the preface. From an Italian speaking user perspective, locating tripartite idioms with a high frequency verb may prove more labourious in the ISI (2006) than in its monolingual equivalent, thus requiring the dictionary user to engage in the subjective assessment of whether the noun or adjectival element can be considered as the first key word.

### 3.2 Il Ragazzini (2009) and Lo Zanichelli (2009)

Listing patterns for tripartite idioms found in the two dictionaries from the Zanichelli publishing house reveal a largely unstructured arrangement, a problem exacerbated by the lack of any information detailing their location. This omission is problematic for users, but is, to a certain degree, offset by the multiplicity of idiom listings in both the ZIR (2009) and the ZLZ (2009). This pattern is apparent across all analysed tripartite syntactic groups in the monolingual version, in particular the V+N+N category with 33/50 expressions recorded twice or more. Similar patterns emerge in the ZIR (2009), but on a lesser scale, with a considerable number of V+N+N (23/50) and V+N+ADJ (15/30) expressions listed twice or more. Another overlapping feature of both empirical samples is the comparably higher number of tripartite idioms recorded in nominalised forms in Zanichelli dictionaries; accounting for 11/100 in the ZIR (2009) and 9/100 in the ZLZ (2009). The removal of the verb element is an inherent feature of certain tripartite Italian idioms, but it negates the objective of a dictionary, which is to record lexical items in their fullest and most descriptive forms. Furthermore, the ZLZ (2009) contains the equal lowest coverage of the analysed expressions with 13/100 not recorded.

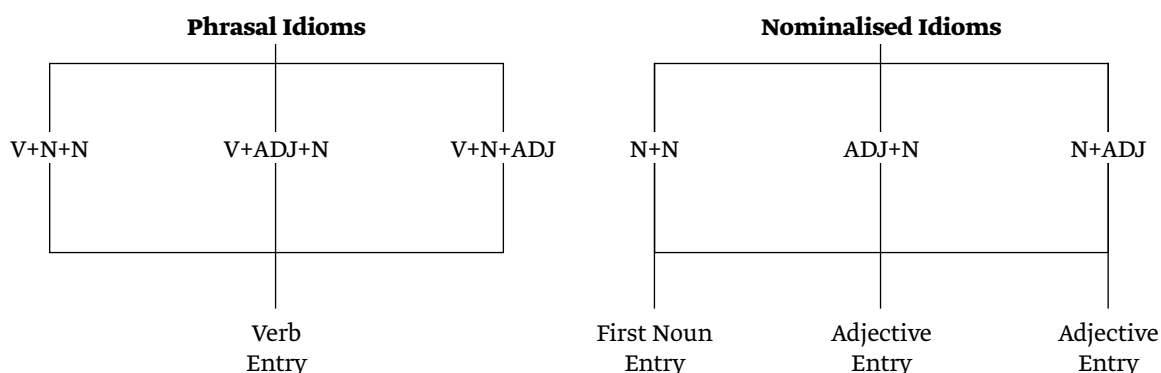
### 3.3 Hoepli Inglese (2007) and Hoepli Italiano (2008)

Both Hoepli-published dictionaries follow individually contrasting, but somewhat consistent, approaches in their treatment of tripartite Italian tripartite idioms. The use of a noun-based entry strategy emerges clearly from the analysis of the three syntactic categories in the HDIN (2007), but its consistency is diluted by scattered recording of similar expressions under alternative entries. A noun entry strategy features most prominently in the V+N+N group (27/50), but the division of this into 19/27 under N1 and 8/27 under N2 is a microcosm of the internally inconsistent approach to their general recording. Keeping a consistency with the other monolingual dictionaries, the HDIT (2008) favours a multi-entry strategy for tripartite idioms, in particular V+N+ADJ (14/20) and V+ADJ+N (9/20) forms, but a similar incongruity to that found in the V+N+N group in the HDIN (2007) resurfaces in the HDIT (2008). In this case, 20/50 binomial idioms are found under a verb and noun element(s) with 13/50 inserted under both noun components. The rationale for recording under both noun elements may be explained by the presence of a HFV element, but this does not extend to the entire subset with *cercare un ago nel pagliaio*, *dire peste e corna di qualcuno*, *finire in una bolla di sapone* and *tirare sassi in piccionaia* not inserted in their verb entries.

## 4 Conclusion

Idioms appear to remain on the periphery of lexicographical importance, at least in an Italian context. Empirical data from monolingual Italian and bilingual Italian-English dictionaries reveal an in-

consistent, unscientific approach to the coverage of tripartite idioms. Generally, idioms tend to be defined by their problematic status rather than their linguistic uniqueness, thus veiling their rich lexical, semantic and syntactic features. This requires a more analytical look at how these characteristics can be influential in systematising the overall accessibility of idioms for dictionary users as well as resolving a perpetual practical difficulty. Figure 1 presents an alternative entry model for tripartite Italian idioms, whether as a verb phrase (VP) or a nominalised form, on the basis of their syntactic composition.



**Figure 1: A Theoretical Framework for the Lexicographical Treatment of Tripartite Italian Idioms.**

A universal verb entry strategy for tripartite idioms with HFV elements contrasts with current lexicographical practice as HFV entries are overpopulated and thus are considered too long for including such expressions. Tripartite idioms are generally VP structures, thus giving the verb element an elevated syntactic importance and identity for dictionary users. Its syntactic position at the head of the expression also increases its probability as a likely point of consultation for users. Nominalised forms of tripartite idioms present a greater organisational challenge, which unlike their VP equivalents, requires a more tailored entry model with due consideration afforded to syntactic order and word class. Therefore, recording N+N structures in their N1 entry and both ADJ+N and N+ADJ types in their adjective entries. These choices are predicated on the following pertinent criteria; the N1 element assumes the role of the syntactic head in binomial idioms, whereas the idiomaticity of those containing adjective and noun element is preserved by the retention of the adjective, compare, for example, the disparate meanings of *giocare a carte* and *giocare a carte scoperte* due to the presence of the idiomatically-inducing adjectival element *scoperte*. In conclusion, the multifaceted nature of idioms is complex, but also provides a substantive platform for choosing their most appropriate point of entry in a dictionary. The current systemic failure of dictionaries to address this issue reinforces the notion of idioms being subservient to words in both the lexicon and lexicography. Therefore, understanding and reprioritising the notion of an idiom and its associated features is an important objective for lexicographical practice in the twenty-first century.

## 5 References

### Monolingual Italian Dictionaries

- [ISC] *Il Sansoni Coletti* Dizionario della Lingua Italiana (2007). Milano: RCS Libri S.p.A.  
[ZLZ] *Lo Zingarelli* (2009). Bologna: Zanichelli.  
[HDIT] *Grande Dizionario Hoepli Italiano* (2008). Milano: Ulrico Hoepli Editore S.p.A.  
Bilingual Italian-English Dictionaries  
[ISI] *Il Sansoni Inglese* (2006). Milano: Edigeo.  
[ZIR] *Il Ragazzini* (2009). Bologna: Zanichelli.  
[HDIN] *Grande Dizionario Hoepli Inglese* (2007). Milano: Ulrico Hoepli Editore S.p.A.  
[SHSIED] *Sansoni-Harrap Standard Italian and English Dictionary* by Vladimiro Macchi, Four Volume, Firenze-Roma: Sansoni Editore, 1970-1975.

### Other Publications

- Al-Kasimi, A. (1977). *Linguistics and Bilingual Dictionaries*. Leiden: E.J. Brill.  
Botha, W. (1992). The Lemmatization of Expressions in Descriptive Dictionaries, in H. Tommola, K. Varantola, T. Salmi-Tolonen and J. Schopp (eds.). *EURALEX '92 Proceedings I-II*. Euralex International Congress, Tampere, August 4-9, 1992. Department of Translation Studies, University of Tampere. pp. 493-502.  
Burger, H. (1989). Phraseologismen im allgemeinen einsprachigen Wörterbuch in F.J. Hausmann, F. Josef et al. (eds.), *Wörterbücher: Ein internationales Handbuch zur Lexikographie*. Volume I, Berlin/New York: De Gruyter. pp. 593-599.  
Gouws, R. H. (1991). Toward a Lexicon-based Lexicography. *Dictionaries: Journal of Dictionary Society of North America*, **13**. pp. 75-90.  
Harras, G. and Proost, K. (2005). The Lemmatization of Idioms in H. Gottlieb, J.E. Mogensen and A. Zettersten (2005) (eds). *Symposium on Lexicography XI, Proceedings of the Eleventh International Symposium on Lexicography*. Copenhagen, May 2-4 2002 Tübingen: Max Niemeyer. pp. 277-291.  
Lorentzen, H. (1996). Lemmatization of Multi-word Lexical Units: In which Entry? in M. Gellerstam, J. Järborg, S.G. Malmgren, K. Norén, L. Rogström and C.R. Pappmehl (eds.), *EURALEX '96 Proceedings I-II*. Seventh Euralex International Conference, Göteborg, August 13-18, 1996, Department of Swedish, Göteborg University. pp. 415-421.  
Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus Based Approach*. Oxford: Clarendon Press.  
Mulhall, C. (2010). A Semantic and Lexical-Based Approach to the Lemmatization of Idioms in Bilingual Italian-English Dictionaries in A. Dykstra and T. Schoonheim (eds.) (2010) *Proceedings of the XIV Euralex International Congress*. pp. 1355-1371.  
Tomaszczyk, J. (1986). The Bilingual Dictionary under Review in M. Snell-Hornby (1986) (ed.) *ZüriLEX 1986 Proceedings, Euralex International Congress*. Zurich, 9-14 September, 1986 University of Zurich. pp. 289-297.



# Especialización y Prototipicidad en Binomios N y N

Ignacio Rodríguez Sánchez  
Universidad Autónoma de Querétaro  
igrodsan@uaq.mx

## Abstract

El trabajo que aquí se presenta es una investigación sobre binomios de estructura *N y N* (dos sustantivos unidos por la conjunción *y*). Esta investigación tiene un carácter exploratorio y se inscribe en la corriente neofirthiana de la lingüística de corpus. A nivel teórico y metodológico partimos de un enfoque guiado por datos, basándonos en los siguientes corpus: Corpus del Español, CORDE, Googlebooks, y EsTenTen11 (Sketchengine).

Aquí se abordan asuntos relacionados con la naturaleza de este tipo de colocaciones: su frecuencia, la estadística de la información mutua, la dispersión y el grado de reversibilidad para identificarlas; las relaciones que se establecen entre ambas partes; la especialización del nodo (como primera o segunda parte de los binomios) y, finalmente, la prototipicidad de algunos elementos. Se concluye que el concepto de binomio tal como se entendía a partir de su definición clásica, se diluye, y se propone una visión integrada en un sistema dinámico de interacciones complejas e impredecibles.

**Keywords:** sustantivo y sustantivo; binomio; colocación; lingüística de corpus

## 1 Antecedentes

La definición de Sinclair del principio de idiomaticidad representa para algunos el momento de un cambio de paradigma en la lingüística: “*The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments*” (Sinclair, 1991: 110). Emparentadas con este trabajo, cristalizan visiones del lenguaje como la gramática de patrones (Hunston & Francis, 2000), la teoría de activación léxica (Hoey, 2005), los trabajos de Biber sobre grupos léxicos (1999; Biber & Barbieri, 2007), el análisis coloconstruccional, y otros trabajos como Wray (2002, 2008) y Corrigan et al. (2009a, 2009b).

La definición clásica de binomio corresponde a Malkiel (1959), que en su estudio comparativo de binomios irreversibles en varias lenguas los define como “[...] *the sequence of two words pertaining to the same form-class, placed on an identical level of syntactic hierarchy, and ordinarily connected by some kind of lexical link*”. Según García-Page (2008: 347), un “binomio fraseológico comprende, básicamente, las construcciones simétricas compuestas por dos sintagmas coordinados y los esquemas prepositivos, y, marginalmente ciertas construcciones asindéticas o yuxtapuestas”. Para muchos autores (Almela Pérez, 2006; García-Page, 1998, 2008; Malkiel, 1959) un binomio digno de estudio es básicamente irreversible, mien-

tras que para otros (Moon, 1998) puede no serlo. Un binomio irreversible sería el que no permitiera revertir el orden de sus dos componentes, como por ejemplo *a tontas y a locas*, *coser y cantar*, *cal y canto*. A diferencia de otras expresiones fraseológicas, el significado del binomio a veces sí se puede deducir de la suma de sus partes.

En este trabajo esperamos mostrar que es productivo estudiar los binomios desde una perspectiva que los incluya como colocaciones y no exclusivamente como expresiones idiomáticas. En la tradición fraseológica, sin embargo, hay posturas encontradas sobre este asunto. García-Page (2008: 12) argumenta que la colocación no es una estructura fija y que por tanto no debe ser objeto de estudio de la fraseología (concepción estrecha de la fraseología). Sin embargo, él mismo reconoce la dificultad de establecer definiciones claras: “¿Es *mesa redonda* o *dinero negro* una locución (...), una colocación o un compuesto?” (p.13). Por otra parte, otros estudiosos coinciden con nuestra visión: Corpas Pastor (1996: 52) propone una división de la fraseología en la que las colocaciones tienen perfecta cabida y que coincide plenamente con la de los fraseólogos anglosajones (concepción ancha de la fraseología).

Evert (2009: 1212) señala que el concepto de colocación es uno de los más controvertidos de la lingüística. Las diferencias entre lo que los neofirthianos y los fraseólogos entienden por ese mismo término ha creado una gran confusión en todos los campos. Desde nuestro punto de vista, coincidimos con Stubbs (1996: 172) sobre el hecho de que las intuiciones de los hablantes nativos sobre las colocaciones son muy imprecisas y no pueden de ninguna manera documentar con detalle dichas colocaciones. A una conclusión similar llega Alderson (2007) en el estudio en que compara datos de corpus y apreciaciones de lingüistas sobre la frecuencia de ciertas palabras.

Desde la psicolingüística, la teoría de activación léxica (*Lexical Priming Theory*) de Hoey (2005) supone un marco teórico adecuado para estudiar el fenómeno que nos proponemos abordar. Esta teoría considera que los hablantes hacemos de modo subconsciente complicadas asociaciones léxicas (semánticas, pragmáticas, de colocaciones y de coligaciones) con un género, estilo y situación social. Finalmente, Hoey sostiene que, también de manera subconsciente, percibimos la posición que ocupa una palabra en un texto, la cohesión que esta produce o deja de producir y las relaciones textuales que contribuye a formar. La teoría de activación léxica se apoya en principios psicolingüísticos como que las palabras de mayor frecuencia se activan antes y más en la mente de los hablantes, lo cual favorece un acceso rápido y fácil al lexicón. Esta activación favorece (y a la vez limita y restringe) la combinación entre las palabras. Si, como dice Giammarresi (2010: 262), almacenamos en el lexicón secuencias formulaicas enteras para ahorrar esfuerzo en el procesamiento, entonces es más lógico esperar que, dada una elección entre dos formas de transmitir un mensaje, una formulaica y otra no formulaica, sea la primera opción la que se genere antes.

En español, además del texto de García Page (2008) hay dos excelentes trabajos sobre binomios: el ya mencionado de Almela Pérez (2006) sobre binomios irreversibles y otro del propio García-Page (1998) sobre binomios antitéticos. Ambos trabajos tienen un carácter descriptivo basado en (no guiado por) corpus. Estos trabajos contienen listados de binomios que, aparentemente, se han ido recogiendo de



manera intuitiva. Para nuestra investigación y por motivos de espacio nos referiremos exclusivamente a algunos datos que nos proporciona Almela Pérez (2006).

Los criterios que usa Almela Pérez (2006, págs. 141-146) para definir qué es un binomio son:

- (1) Los binomios constan de dos lexemas.
- (2) Son una secuencia infratextual.
- (3) Tienen una estructura paralelística.
- (4) Sus formas -léxicas y funcionales- son inmutables.
- (5) Forman una secuencia indescomponible.
- (6) Los miembros son inseparables.
- (7) Tienen un significado composicional o idiomático.

## 2 Preguntas de Investigación

Nuestra intención es señalar que, aparte de los criterios mencionados arriba, hay explicaciones de tipo psicolingüístico (basadas en datos cuantitativos) que contribuyen a explicar las formas que adoptan los binomios:

- (1) ¿La frecuencia con que ocurre el binomio y los índices de relación (en concreto la Información mutua -IM en adelante) pueden servir para identificar los binomios? Es decir, si una colocación binomial tiene 20 casos en un corpus y una información mutua de 8 ¿habremos identificado a un binomio irreversible?
- (2) ¿Qué relaciones se establecen entre  $N_1$  y  $N_2$ ? ¿La frecuencia de uso de cada una de las palabras que constituyen el binomio va asociada a su posición en el binomio? Tomando el ejemplo de *aventuras* y *desventuras*, el hecho de que *aventuras* sea más frecuente que *desventuras* ¿nos está diciendo algo sobre el orden en que ese binomio se lexicalizó?
- (3) ¿Se especializan algunas palabras en ser  $N_1$  o de un binomio (siendo  $N_1$  la primera parte del binomio, y  $N_2$  la segunda)? Es decir, si sabemos que la palabra *señor* siempre aparece como  $N_2$  cuando se combina con ciertos sustantivos de un mismo campo semántico (*amigo y señor, amo y señor, dueño y señor, esposo y señor, marido y señor, padre y señor, primo y señor, tío y señor* que aparecen como  $N_1$ ), ¿no es acaso lógico que el binomio *rey y señor* también haya lexicalizado en el mismo orden que sus cohipónimos?
- (4) Aparte de las relaciones semánticas, ¿qué otro tipo relaciones se establecen entre los diferentes colocativos de un binomio?

### 3 Extracción de datos a partir de un corpus

Los datos para este estudio proceden en primera instancia de Corpus del Español ([Davies, 2002] en adelante, CDE). En este corpus se realizó la búsqueda [NN\*] Y [NN\*] de los siglos XIX y XX, con un límite de 5000 casos. Del resultado obtenido se eliminaron los casos con frecuencia menor a cuatro y quedaron 2482 colocaciones binomiales. Por razones prácticas, se eliminaron posteriormente los casos de binomios con sustantivo repetido (por ejemplo, *años y años*). Nuestra base de datos original tampoco incluyó los binomios en los que se intercalan artículos, posesivos o preposiciones entre los dos sustantivos.

En una investigación previa (Rodríguez, 2013: 291) se mencionó la conveniencia de contrastar esos datos con los de corpus más grandes, que es lo que se empieza a hacer en esta investigación, complementando los datos del CDE con los del CREA, Googlebooks (interface de Mark Davies), y EsTenTen11 (Sketchengine).

## 4 Resultados

### 4.1 Información Mutua y Frecuencia

En respuesta a la primera pregunta de investigación, se calculó un índice de relación (IM) entre los dos términos del binomio para cada caso, siguiendo la fórmula basada en Oakes (1998, págs. 63-65). Según Evert (2009: 1229), la IM se debe combinar siempre con una frecuencia mínima (en nuestro caso 4) para equilibrar el sesgo hacia palabras de alta frecuencia.

Tal y como se esperaba, esta estadística mostró que la inmensa mayoría binomios tenían una IM significativa. Solo 140 binomios de los 2482 tenían una IM inferior a 3, que es, según Hunston (2002: 71), la cifra a partir de la cual se suele considerar que una colocación es significativa. Ejemplos de colocaciones y binomios cuya IM mutua es menor a 3 pero con cierto grado de fijación serían: *tiempo y forma*, *fondo y forma*, *padre y señor*, *tierra y libertad*, *forma y manera*, *cuerpo y sangre*, *vida y muerte* (5% de los casos). Esto apunta la imposibilidad de distinguir exclusivamente por métodos estadísticos una colocación de una no colocación (tal y como mencionan Evert [2009: 1242] y Cantos & Sánchez [2002]).

La frecuencia y la IM, pues, pueden ayudar a descubrir formulaicidad de un binomio pero puede que no sean los únicos criterios que los identifiquen. Junto a estos dos criterios, el de la dispersión es otro elemento útil, que contribuye a eliminar elementos estilísticos propios de los textos incluidos en el corpus. Por ejemplo, “Silvicultura y pesca” es un binomio que se repite 67 veces en el corpus pero solo aparece en un texto, que es una enciclopedia.

Sobre la segunda pregunta de investigación, es decir el orden distributivo de los dos términos de los binomios (las relaciones de  $N_1$  con  $N_2$ ) encontramos dos tipos de explicaciones. Por un lado, García-Paige (2008: 347) señala que las posibles explicaciones del orden de los binomios son tanto de tipo semán-

tico (“los principios del “egocentrismo” (...), de jerarquía social (...), de ordenación cronológica u espacial (...), de disposición de contrarios”) como “fonéticos, morfológicos y léxicos”. Estamos de acuerdo con él cuando comenta que este es un tipo de trabajo pendiente en el español, pero creemos que la semántica cognitiva (la tesis de la cognición corporeizada que desarrollaron casi simultáneamente Talmy [1988] y Johnson [1990] parece ser el referente de “los principios del egocentrismo”) tendría algo que añadir en este sentido, en especial la integración (*blending*) conceptual que sugieren Fauconnier & Turner (2003).

Por otro lado, García-Page (2008: 348) hace referencia a que “el orden distributivo también se ha querido ver a veces en el esquema que dispone el constituyente silábicamente más corto en el primer lugar del binomio y el más largo en el segundo (...) una prueba del valor icónico de los binomios”. De ser cierta esta apreciación -no se aportan datos empíricos en español- nosotros optaríamos por ofrecer para este fenómeno una explicación de carácter estadístico y psicolingüístico: simplemente que la primera parte del binomio es más frecuente que la segunda, como mencionamos antes al hablar de la teoría de activación léxica de Hoey. Este orden basado en la frecuencia permitiría al hablante un acceso más rápido al lexicón y facilitaría, por tanto, la fluidez en la comunicación. Lo corto de las palabras más frecuentes (y en primer término) vendría explicado por el corolario a la ley de Zipf (1949) que se refiere a que existe una relación directa entre la longitud de una palabra y su frecuencia (Davies, 2006: 164).

Así, en el binomio *tiempo y forma*,  $N_1$  (*tiempo*) tiende a ser más frecuente que  $N_2$  (*forma*) y en el binomio *oferta y demanda*,  $N_1$  (*oferta*) tiende a ser más frecuente que  $N_2$  (*demanda*). Esta tendencia viene confirmada por los datos de la Tabla 1.

Casos en que $N_1$ es más frecuente que $N_2$	Casos en que $N_2$ es más frecuente que $N_1$	Casos en que $N_1$ es tan frecuente como $N_2$	<b>Total</b>
784	462	17	<b>1263</b>

**Tabla 1: Comparación de la frecuencia de uso de  $N_1$  y  $N_2$  de los binomios.**

## 4.2 Dispersión

El valor de la estadística de información mutua combinado con una medida simple de dispersión contribuye a una mayor precisión en la identificación de binomios. Del total de 2483 binomios identificados inicialmente, 320 aparecen en el CDE de Mark Davies en solo un texto (independientemente de la frecuencia), lo cual equivale al 12.9% del total de binomios. Asimismo, se procedió a buscar en el CORDE el número de documentos en los que aparecía cada uno de los 2483 binomios. En este caso se obtuvo que 446 (18%) aparecían como máximo en un solo documento (de estos, 277 no aparecen ni una sola vez).

El número de binomios que reunían ambas condiciones (solo un documento del CDE de Mark Davies y uno o ninguno en el CORDE) es de 198 (es decir, el 8%).

### 4.3 Reversibilidad

En cuanto a la irreversibilidad del binomio, ya se ha mencionado que esta es la característica esencial del binomio (Almela Pérez, 2006: 155; Malkiel, 1959: 113; García-Page, 2008: 329). García-Page es el único que reconoce que hay excepciones a la irreversibilidad. Según nuestra investigación, la reversibilidad de los binomios es un fenómeno que ocurre con mayor asiduidad de lo que suele considerarse.

En primer lugar, la en cuanto a la irreversibilidad, hay 346 binomios revertidos (del total de 2483 binomios), lo cual supone un 28% del total. Quince ejemplos y sus frecuencias aparecen en la Tabla 2

En la lista de Almela Pérez (2006) hay aproximadamente unos 90 binomios que coinciden con la estructura de los binomios que analizamos nosotros en nuestra base de datos. De estos 90 analizamos una muestra de 30 binomios que encajan con la estructura que se examina en este trabajo. Así, en la lista de binomios irreversibles que da Almela aparecen binomios que en realidad son reversibles, como *día y noche* y *noche y día* (Almela Pérez, 2006: 148-9); y otros casos como *pan y agua*, *cuerpo y alma*, *besos y abrazos*, *calidad y cantidad*, *cielo y tierra*, *uñas y dientes*, y *pies y manos*). Según nuestras estimaciones, un 30% de los binomios que Almela Pérez lista como irreversibles, no lo son en realidad.

<b>Binomio</b>	<b>FREC.</b>	<b>Binomio revertido</b>	<b>FREC.</b>
hombres y mujeres	426	hombres y mujeres	53
oro y plata	213	plata y oro	54
día y noche	171	noche y día	107
cuerpo y alma	136	alma y cuerpo	22
puertas y ventanas	136	ventanas y puertas	15
blanco y negro	136	negro y blanco	4
mujeres y niños	120	niños y mujeres	17
flora y fauna	116	fauna y flora	20
calles y plazas	100	plazas y calles	24
pies y manos	85	manos y pies	22
usos y costumbres	82	costumbres y usos	9
radio y televisión	80	televisión y radio	6
sangre y fuego	73	fuego y sangre	14
petróleo y gas	65	gas y petróleo	7
flor y nata	46	nata y flor	5

**Tabla 2: Ejemplos de pares de binomios revertidos en el CDE con sus frecuencias.**

La Tabla 2 muestra que en la mayoría de los casos, hay una gran diferencia entre las frecuencias de los dos binomios. Hay varias posibles explicaciones para esta situación. Por un lado podemos estar ante casos en que los procesos de lexicalización de la colocación no han terminado todavía pues el binomio supuestamente irreversible todavía no ha desplazado completamente al binomio revertido. Otra explicación, tal vez complementaria, es que un binomio, puede usarse solamente en determinados contextos.

Para nosotros resulta, pues, importante señalar que más que hablar de binomios irreversibles tal vez sea más fértil hablar de un cierto tipo de colocaciones, en  $N_1$  y  $N_2$ , con un alto grado de lexicalización.

#### 4.4 Especialización como $N_1$ o $N_2$

Con la muestra de 30 binomios  $N_1$  y  $N_2$  de Almela Pérez realizamos un análisis más pormenorizado.

Binomio	$N_1$ y N	Nº de tipos	N y $N_1$	Nº de tipos	$N_2$ y N	Nº de tipos
acoso y derribo	acoso y N	5	N y acoso	2	derribo y N	0
agua y ajo	agua y N	107	N y agua	95	ajo y N	6
ajos y cebollas	ajos y N	7	N y ajos	2	cebollas y N	0
alfa y omega	alfa y N	3	N y alfa	0	omega y N	0
armas y bagajes	armas y N	100	N y armas	70	bagajes y N	4
besos y abrazos	besos y N	19	N y besos	13	abrazos y N	12
bombo y platillo	bombo y N	3	N y bombo	1	platillo y N	0
bromas y veras	bromas y N	16	N y bromas	12	veras y N	3
cal y canto	cal y N	12	N y cal	12	canto y N	0
calidad y cantidad	calidad y N	80	N y calidad	75	cantidad y N	22
capa y espada	capa y N	11	N y capa	12	espada y N	19
cara y cruz	cara y N	33	N y cara	45	cruz y N	18
carne y hueso	carne y N	46	N y carne	41	hueso y N	14
carretera y manta	carretera y N	5	N y carretera	3	manta y N	2
causas y efectos	causas y N	24	N y causas	16	efectos y N	21
cielo y tierra	cielo y N	25	N y cielo	11	tierra y N	90
ciencia y conciencia	ciencia y N	61	N y ciencia	32	conciencia y N	38
cruz y raya	cruz y N	18	N y cruz	16	raya y N	1
cuenta y riesgo	cuenta y N	23	N y cuenta	37	riesgo y N	16
cuerpo y alma	cuerpo y N	59	N y cuerpo	33	alma y N	43
día y noche	día y N	34	N y día	11	noche y N	24
garbo y salero	garbo y N	15	N y garbo	7	salero y N	2
golpe y porrazo	golpe y N	10	N y golpe	4	porrazo y N	0
ida y vuelta	ida y N	0	N y ida	0	vuelta y N	7

Binomio	N <sub>1</sub> y N	Nº de tipos	N y N <sub>1</sub>	Nº de tipos	N <sub>2</sub> y N	Nº de tipos
moco y baba	moco y N	3	N y moco	0	baba y N	1
pan y agua	pan y N	54	N y pan	30	agua y N	107
pecho y espada	pecho y N	36	N y pecho	15	espalda y N	13
sangre y fuego	sangre y N	103	N y sangre	84	fuego y N	34
uñas y dientes	uñas y N	8	N y uñas	8	dientes y N	25
viento y marea	viento y N	19	N y viento	15	marea y N	2

**Tabla 3: Análisis de 30 binomios según el número de combinaciones con otros sustantivos (CDE de Mark Davies).**

La Tabla 3 muestra en cuántos tipos de colocaciones aparecen como nodos N<sub>1</sub> y N<sub>2</sub>. Por ejemplo en cuanto al primer binomio de la tabla, *acoso y derribo*, encontramos 5 combinaciones del N<sub>1</sub>, *acoso*, con otro sustantivo (*acoso y abuso*, *acoso y vigilancia*, *acoso y protección*, *acoso y persecución* y *acoso y derribo*); dos casos en que *acoso* aparece como segunda parte del binomio (*violación y acoso* y *persecución y acoso*). No se registran casos en que el N<sub>2</sub>, *derribo*, sea primera parte de otro binomio. Es decir que, según nuestros datos, parece que *acoso* se especializa como N<sub>1</sub> pues aparece predominantemente en esa posición. Si hacemos una comparación entre las columnas tres y cinco, tenemos que, de los treinta ejemplos tomados, hay solo 3 en que N y N<sub>1</sub> tiene más tipos de binomios que N<sub>1</sub> y N (se trata de los casos *capa y espada*, *cara y cruz* y *cuenta y riesgo*). En estos casos, *espada*, *cruz* y *riesgo* son nodos que se combinan más veces como N<sub>2</sub>. Hay también tres casos en los que la colocación N<sub>1</sub> y N, tiene los mismos tipos que N y N<sub>1</sub> (*cal y canto*, *ida y vuelta* y *uñas y dientes*). Excluidos estos 6 casos, el 80% de los casos el nodo más productivo es N<sub>1</sub>.

Si comparamos del mismo modo el número de tipo de colocaciones N<sub>1</sub> y N con N<sub>2</sub> y N (columnas 3 y 7), tenemos resultados muy similares: 83% de predominio de N<sub>1</sub> y solo 17% de predominio de N<sub>2</sub> (*capa y espada*, *cielo y tierra*, *ida y vuelta*, *pan y agua*, y *uñas y dientes*).

¿Se puede deducir pues que las palabras que están en la primera parte de un binomio tienen una potencia combinatoria mayor y esta ayuda a mantener en esa posición? La respuesta es sí, pero con salvaguardas: los casos en que no hay especialización clara (por ejemplo *valor*) y los de binomios que se especializan como N<sub>2</sub> como *confianza*, *director* y *libertad*.

## 4.5 Prototipicidad

En cuanto a las relaciones entre colocativos, un análisis somero realizada con los corpus más grandes arrojan resultados comparables a las relaciones de N<sub>1</sub> y N<sub>2</sub>.

Aparte de las relaciones semánticas esperables (hiponimia, sinonimia, hiperonimia, etc.) y las relaciones morfológicas lógicas (misma categoría gramatical), resulta llamativo hallar relaciones sobre la estructura léxica relacionadas en el sonido (*clang association*): con el número de sílabas, la posición de la sílaba tónica, similitudes fonéticas y el uso de afijos. Aquí nos referimos a relaciones entre colocativos, no entre elementos de un mismo binomio como en *troche y moche* y *tomo y lomo*.

Por ejemplo en la Tabla 4 vemos que de las 10 primeras colocaciones de *sumisión* ordenadas por su IM, hay 5 que son palabras llanas de cuatro sílabas 4 de las cuales tres terminan en *-ismo*, y hay dos parejas que son parecidas en su inicio y su final: *obsecuencia* y *obediencia*, y *abyección* y *abnegación*.

	<b>Binomio</b>	<b>Frecuencia</b>	<b>IM</b>
1	servilismo	42	8.24
2	obediencia	190	7.74
3	obsecuencia	17	7.15
4	vasallaje	12	6.9
5	entreguismo	10	6.64
6	docilidad	11	6.61
7	conformismo	15	6.33
8	pasividad	31	6.13
9	abyección	6	6.13
10	abnegación	10	5.97

**Tabla 4: Colocativos en binomios de *sumisión* y (Fuente: [www.SketchEngine.co.uk](http://www.SketchEngine.co.uk), EsTen-Ten11, American, TreeTagger).**

No es infrecuente hallar este tipo de situación. La Tabla 5 muestra que entre las diez colocaciones más frecuentes de *privaciones*, tiene gran parecido varias parejas: *austeridades* y *penalidades*, *escaseces* y *estrecheces*, *pobrezas* y *durezas*.

	<b>Palabra</b>	<b>Frecuencia</b>	<b>IM</b>
1	escaseces	161	9.98
2	fatigas	1338	9.8
3	abstinencias	49	9.72
4	austeridades	45	9.55
5	penalidades	641	9.51
6	sufrimientos	1379	9.44
7	estrecheces	172	9.37
8	pobrezas	56	9.15
9	miserias	979	9.07
10	penurias	282	8.94
11	durezas	43	8.91

**Tabla 5: Colocativos en binomios de *privaciones* y (Fuente: Googlebooks Spanish).**

## 5 Conclusiones

En primer lugar conviene recordar que los trabajos sobre binomios, y el nuestro no es una excepción, simplifican el valor del binomio pues en realidad esta estructura nunca deja de ser parte de una secuencia mayor con una función específica.

Hecha esta acotación, creemos que vale la pena explorar los binomios partiendo de principios básicos de la psicolingüística, tomando en cuenta la frecuencia de las palabras que aparecen como  $N_1$  y como  $N_2$ . Aunque haya motivos de toda índole, parece ser que la frecuencia de uso del  $N_1$  tiende a ser más alta que la del  $N_2$ . También hemos mostrado que hay nodos que se especializan en  $N_1$  o en  $N_2$ , y que normalmente los nodos que aparecen como  $N_1$  forman más binomios que los  $N_2$ .

El tamaño y la configuración del corpus son factores capitales en estudios cuantitativos porque no solo determinan las frecuencias, las estadísticas de asociación; las medidas de dispersión sino que nos van a señalar la variabilidad de relaciones conceptuales (secuencia de tiempo, de espacio, causa-efecto, jerarquía, etc). Asimismo, podemos observar fenómenos insospechados como la coincidencia en algunas colocaciones en el inicio y/o el final de la palabra (efecto tina/bañera) y la similitud del número de sílabas, del patrón consonántico, la posición de la sílaba acentuada, coincidencias fonéticas.

En definitiva, creemos que estos datos exploratorios respaldan la tesis de Hoey de que los hablantes registramos el uso de palabras, su contexto, su posición y más aspectos de los que no parecemos estar conscientes. Queda pues diluido el concepto de binomio tal como lo conciben Malkiel (1959) y Almela Pérez (2006), pasando a formar parte de un sistema dinámico de interacciones complejas e impredecibles en que se influyen mutuamente el uso, el procesamiento, el aprendizaje y la estructura de la lengua (Ellis & Frey 2009).

Este trabajo señala que los métodos semiautomáticos generados en una investigación guiada por datos sobre los binomios arrojan resultados ricos y complejos que obligan a replantearnos formas básicas generadas por la introspección.

## 6 Referencias

- Alderson, J. C. (2007). Judging the Frequency of English Words. *Applied Linguistics*, 28(3), 383-409 doi:10.1093/applin/amm024
- Almela Pérez, R. (2006). Binomios (irreversibles) en español. *LEA: Lingüística Española Actual*, 28(2), 135-160.
- Biber, D. (1999). Longman grammar of spoken and written English. Londres: Longman.
- Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263-286. doi:10.1016/j.esp.2006.08.003
- Cantos, P., & Sánchez, A. (2002). Lexical Constellations: What Collocates Fail to Tell. *International Journal of Corpus Linguistics*, 6(2), 199-228. doi:10.1075/ijcl.6.2.02can
- Corpas Pastor, G. (1996). *Manual de fraseología española*. Madrid: Gredos.
- Corrigan, R., Moravcsik, E. A., Ouali, H., & Wheatley (Eds.). (2009a). *Formulaic language* (Vol. 1). Amsterdam/ New York: John Benjamins Publishing Company.



- Corrigan, R., Moravcsik, E. A., Ouali, H., & Wheatley (Eds.). (2009b). *Formulaic language. (Vol. 2 Acquisition, loss, psychological reality, and functional explanations)*. Amsterdam/New York: John Benjamins Publishing Company.
- Cowie, A. P. (2006). Phraseology. *Encyclopedia of language & linguistics*. Amsterdam: Elsevier.
- Davies, M. (2002). *Corpus del Español: 100 million words, 1200s-1900s*. <http://www.corpusdelespanol.org> [12/032013]
- Davies, M. (2006). A frequency dictionary of Spanish: core vocabulary for learners. Abingdon: Routledge.
- Ellis, N. C. (1998). Emergentism, Connectionism and Language Learning. *Language Learning*, 48(4), 631-664. doi:10.1111/0023-8333.00063
- Ellis, N. C. (2008). The Dynamics of Second Language Emergence: Cycles of Language Use, Language Change, and Language Acquisition. *The Modern Language Journal*, 92(2), 232-249. doi:10.1111/j.1540-4781.2008.00716.x
- Ellis, N. C., & Frey, E. (2009). The psycholinguistic reality of collocation and semantic prosody (2). En R. Corrigan, E. A. Moravcsik, H. Ouali, & Wheatley (eds.), *Formulaic language (Vol. 2 Acquisition, loss, psychological reality, and functional explanations)*. Amsterdam/New York: John Benjamins Publishing Company.
- Evert, S. (2009). Corpora and collocations. En A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: an international handbook* (Vols. 1-2, Vol. 2, págs. 1212-1248). W. de Gruyter.
- Fauconnier, G., & Turner, M. (2003). *The Way We Think: Conceptual Blending And The Mind's Hidden Complexities*. Basic Books.
- García-Page, M. (1998). Binomios fraseológicos antitéticos. En G. Wotjak (Ed.), *Estudios de fraseología y fraseografía del español actual* (págs. 195-202). Madrid/Frankfurt am Main: Iberoamericana/Vervuert.
- García-Page, M. (2008). Introducción a la fraseología española. Estudio de las locuciones. Barcelona: Anthropos.
- van Geert, P. (2008). The Dynamic Systems Approach in the Study of L1 and L2 Acquisition: An Introduction. *The Modern Language Journal*, 92(2), 179-199. doi:10.1111/j.1540-4781.2008.00713.x
- Gries, S. T., & Stefanowitsch, A. (2007). *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis*. Berlin/New York: Walter de Gruyter.
- Hoey, M. (2005). *Lexical priming: a new theory of words and language*. Abingdon: Routledge.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hunston, S., & Francis, G. (2000). *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam/New York: John Benjamins Publishing Company.
- Johnson, M. (1987). *The body in the mind: the bodily basis of meaning, imagination, and reason*. Chicago: University of Chicago Press.
- Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, David Tugwell. *The Sketch Engine. Proc EURALEX 2004*, Lorient, France; Pp 105-116, <http://www.sketchengine.co.uk> [12/032013]
- Malkiel, Y. (1959). Studies in irreversible binomials. *Lingua*, 21, 142-155.
- Moon, R. (1998). *Fixed expressions and idioms in English: a corpus-based approach*. Oxford / New York: Clarendon Press.
- Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
- Real Academia Española: Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español actual*. <<http://www.rae.es>> [12/03/2013]
- Rodríguez Sánchez, I. (2013). Frequency and Specialization in Spanish Binomials N y N. *Procedia - Social and Behavioral Sciences*, 95, 284-292. doi:10.1016/j.sbspro.2013.10.649
- Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford / New York: Oxford University Press.
- Stubbs, M. (1996). *Text and corpus analysis*. Blackwell Publishers.
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science: A Multidisciplinary Journal*, 12(1), 49-100.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/New York: John Benjamins Publishing Company.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Wray, A. (2008). *Formulaic language: pushing the boundaries*. Oxford / New York: Oxford University Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge Mass.: Addison-Wesley Press

# Syntax and Semantics vs. Statistics for Italian Multiword Expressions: Empirical Prototypes and Extraction Strategies

Luigi Squillante  
Sapienza - Università di Roma  
Email: luigi.squillante@uniroma1.it

## Abstract

In this work we present an empirical analysis performed on Italian nominal multiword expressions (MWEs) of the form [noun + adjective] that aims at studying quantitatively their syntactic and semantic features in order to improve their automatic identification and collection. Three indices are proposed, which are able to measure syntactic and semantic frozenness of the expressions on empirical basis in a corpus of about 1.8 million words, composed of Italian texts concerning the domain of physics. The combination of the three indices can be used to create a global measure, that we call Prototypicality Index (PI), which appears to be useful in the automatic extraction of terminological MWEs. The performance of PI at extracting true positives out of a candidate list is compared to those of the well-known statistical association measures Log-likelihood and Pointwise Mutual Information. Our results show how the performance of PI can be comparable to those of association measures, although it does not involve statistical calculations. Thus, PI can be seen as a new option for lexicographers and terminologists to integrate the already available statistical methods when identifying MWEs from texts.

**Keywords:** multiword expressions; terminology; prototype; extraction; empirical tests

## 1 Introduction

Nowadays multiword expressions (MWEs) represent one of the most studied phenomena in phraseological and lexicographic studies. They include a great variety of entities lying on a *continuum* between lexicon and syntax, whose typical features include morpho-syntactic fixedness, semantic restrictions, semantic unpredictability, constructions which differ from standard syntax, conventionality, institutionalization, etc. Their interpretation generally crosses the boundaries between words (Sag et al. 2002), and one of the best definitions to refer to such entities is proposed by Calzolari et al. (2002:1934), according to whom a MWE is “a sequence of words that acts as a single unit at some level of linguistic analysis”.

Despite their apparent anomalous behavior, MWEs are a very important and frequent phenomenon in every language: in his famous *idiom principle* Sinclair (1991) states that idiomatic and morpho-syn-

tactically restricted combinations are as normal and natural in discourse as free combinations, while Jackendoff (1997) attests that the number of MWEs stored in the lexicon of any speaker is equal to that of simple words.

Throughout the twentieth century, linguists have developed a great amount of studies which examined the aspects of MWEs on a theoretical perspective, often leading to competing analyses, controversy on interpretations or overlapping terminology. In recent years, however, computational and corpus-based studies have become one of the dominant lines of research in this field, since quantitative features, such as the fact that MWE components tend to cooccur in text with higher frequencies, have proved to be very effective in the automatic treatment of MWEs, leading to the development and improvement of several association measures (AMs) in order to identify, study and automatically extract MWEs from texts (just to mention some works: Evert 2004; Evert 2008; Kilgarriff 2006; Ramisch et al. 2010; Seretan 2011).

When analyzing a corpus, by means of computational tools, linguists are usually able to create a list of candidate expressions of MWEs where each candidate has an association score assigned by AMs. In general, the primary goal is to identify the largest possible number of true positives (candidates that represent real MWEs) within a certain threshold of significance based on the score assigned to each candidate, e.g. to provide raw material for lexicography. In this process, AMs generally consider statistical quantities, such as the number of cooccurrences of the components, the number of occurrences of the single components, the size of the corpus, etc., often with no reference to any explicit linguistic behavior. Nevertheless, considering syntactic or semantic features of MWEs from a computational and corpus-linguistic point of view is useful to improve the performances of automatic extraction tools (as shown, for other languages, in Bannard 2007; Weller & Fritzingler 2010; Cap et al., 2013), as well as to develop a better understanding of the typical features of MWEs on empirical bases (cf. Squillante 2014), which are both aspects of preeminent interest for lexicographers dealing with multiword phenomena.

## 2 Motivations

Our work presents an empirical study conducted on the Italian language which, unlike other major languages like English or German, still lacks well-founded computational studies in lexicography dealing with complex expressions like MWEs. Although “GRADIT - Grande Dizionario Italiano dell’Uso” (De Mauro, 1999-2007), known as the most comprehensive lexicographic resource for Italian, has a highly corpus-oriented perspective and explicitly focuses on the quantitative presence of MWEs in Italian, no explicit computational methods were involved in identifying the expressions. Similarly, the most recently published Italian collocation dictionaries (Urzi 2009; Lo Cascio 2012; Tiberii 2012) still rely mostly on intuition and only partly replicate data collection strategies, without considering a defined and explicit methodology based on corpora. Thus, there is a need to investigate computational techniques for lexicographic analyses of Italian MWEs, especially because Italian morpho-syntax differs from those of the above-mentioned Germanic languages.

The nature of our study is twofold: on the one hand we focus on empirical evidences in order to study the prototypical concept of MWE; on the other hand we compare the traditional statistical measures with syntactic or semantic tests for the identification and the extraction of MWEs from texts.

Finally, our work is focused on terminology. In fact, especially in technical domains, MWEs appear in high number even in small corpora, since specialized languages are a powerful source of multiword terminology and we see it as a matter of importance that they are identified and collected so that they can be included in the respective dictionaries and multiword terminology collections.

### 3 Methodology

#### 3.1 Corpus and Prototype of MWE

As a first approach, in our study we opted to focus on the field of physics. The choice of physics is interesting since its lexicon, unlike other scientific domains such as that of medicine, is still primarily composed of highly polysemous every-day words which are put together in MWEs to form technical expressions, pursuing the established tradition started with Galileo Galilei in the seventeenth century, as recalled by Migliorini (1994:398).

In order to have an empirical base to perform our analysis, we built a corpus of about 1.8 million words collecting Italian texts concerning physics, including educational books (6,2% of the total), Wikipedia pages (34,5%), academic textbooks (20,7%), theses and dissertations (38,6%).

Our corpus was POS-tagged with TreeTagger (Schmid, 1994) and enhanced by means of a semi-automatic and manual post-tagging process in order to improve the tagging quality, e.g. to correct macroscopic systematic errors and include unrecognized technical lemmas in the dictionary. The final accuracy of the tagged corpus is evaluated at around 96% by manually checking 300 random sentences of the corpus.

We chose to analyze only nominal MWEs of the form [noun + adjective] in a first approach, representing the unmarked Italian nominal phrase. In physics, in fact, the use of nominal phrases is dominant and nominalization is often attested to be a standard feature of special languages. This is also supported by the fact that the majority of MWEs labeled by GRADIT as part of the special language of physics are nominal (2668), while only 9 belong to any other grammatical category.

Although MWEs can exhibit a great variability of behaviors, as it has been mentioned in the introduction, we chose to focus on features which could be investigated and tested on corpora, and we started with the initial hypothesis that the prototype of a MWE is an expression:

- that does not allow for interruptions or insertions of other words between its components;
- whose word order is not modifiable;
- whose components cannot be substituted by their synonyms.

The expression *relatività generale* ‘general relativity’ is a clear example of a terminological MWE which satisfies these three conditions, since it cannot be interrupted (cf. *\*relatività più generale* ‘more general relativity’), it does not allow a modification in the order of its components, although this is possible for Italian nominal phrases (cf. *\*generale relatività*) and it cannot be modified by substituting one of its components with a synonym (cf. *\*relatività universale* ‘universal relativity’ or *\*relatività totale* ‘total relativity’).

However, although these features involving fixedness are typically associated to nominal MWEs in Italian, they do not always appear together in all expressions. For example, interruptibility is allowed for *punto debole* ‘weak point’, which admits *punto più debole* ‘weaker point’; *infrarosso lontano* ‘far infrared’ is attested together with *lontano infrarosso*; while *gas ideale* ‘ideal gas’ can be substituted by *gas perfetto* ‘perfect gas’. Because of this, the concept of prototype is thought of just as a model which could help to order the expressions on a continuous scale from a maximum grade of fixedness on several levels (adhesion to the prototype) to more flexible expressions.

The reason for considering the hypothesis of such a prototype comes from studies like those of Masini (2009) and Squillante (2014), which show how the nucleus of the prototype seems to include those expressions that are generally referred to as *polirematiche* in the Italian lexicographic tradition and exhibit syntagmatic and paradigmatic frozenness, needing the cooccurrence of their components in order to acquire their specific meaning (e.g. *luna di miele* ‘honeymoon’; *essere al verde* ‘to have no money’, lit. ‘to be at green’). Terminological expressions are generally part of this group.

When fixedness becomes less strict and modification is allowed, the *continuum* of MWEs moves towards those expressions that we can call *lexical collocations*, which show only preference for the cooccurrence of their components (e.g. *capelli castani* ‘chestnut brown hair’ or *compilare un modulo* ‘to fill a form’), being «not fixed but recognizable phraseological units» (Tiberii, 2012).

### 3.2 Three Indices for the Measure of Empirical Frozenness

Following Squillante (2014), we implemented a computational tool that performs empirical tests concerning the above-mentioned features of modifiability for each candidate expression. Each of the features is quantified by an index whose value is computed on the basis of the comparison between the occurrences of the modified expression and those of the regular basic unmarked form in the corpus, i.e. the lemmatized form, regardless of inflection (which our analysis proved to be not a relevant feature in discriminating MWEs from standard expressions). All the queries are made on surface forms or POS categories, depending on the test, and do not involve syntactic structures as they would arise from parsing.

Given an expression, the index of interruptibility ( $I_i$ ) counts the number of the occurrences of the sequence in its basic form [noun + adjective], say  $n_i$ , and the occurrences of the same sequence with one word occurring between the two components ( $n_{bf}$ ), calculating the following ratio:

$$I_i = \frac{n_i}{n_{bf} + n_i}$$

In this way, a high number of interrupted expressions with respect to those which are not interrupted let the index acquire a high value. The sum in the denominator let the index be limited between 0 and 1.

In an analogous way, the index concerning the reverse order ( $I_o$ ) compares the number of occurrences of the inverted sequence [adjective + noun] ( $n_o$ ) with those of the basic form  $n_{bf}$  according to the formula:

$$I_o = \frac{n_o}{n_{bf} + n_o}$$

Finally, the index concerning the feature of substitutability compares the number of occurrences of the basic form with the occurrences of all the sequences in which one of the two components is replaced by one of its synonyms (if present). If the number of occurrences of the  $i$ -th synonym of the first and the second component are called respectively  $n_{s1,i}$  and  $n_{s2,i}$ , the total number of substituted sequences for the expression is:

$$n_s = \sum_i n_{s1,i} + \sum_i n_{s2,i}$$

and the index  $I_s$  is given by the formula:

$$I_s = \frac{n_s}{n_{bf} + n_s}$$

The calculation of  $I_s$  is subjected to the availability of an external synonym list. In our study, as a first approach, we chose the GNU-OpenOffice Thesaurus for the Italian language<sup>1</sup> for practical reasons, since it was immediately available, easily manageable and proved to be good enough for our purpose. However, one can integrate the tool with other more specific resources in the future, in order to improve the quality of the results.

The values of the three indices can be merged into a single function that we call Prototypicality Index (PI), representing the adherence of the expression to the hypothesized prototype. We consider the following formula:

$$PI = \frac{n_{bf}}{n_{bf}^{max}} \cdot \frac{1}{1 + I_i + I_o + I_s}$$

whose value increases when the values of the three indices decrease (thus, a high PI value means high fixedness), and in which the three features are weighted in the same way by the operation of

1 [http://linguistico.sourceforge.net/pages/thesaurus\\_italiano.html](http://linguistico.sourceforge.net/pages/thesaurus_italiano.html).

sum. In this way an expression with a very high value for just one of the indices can have a resulting PI value similar to that of an expression with average values distributed on all the three indices. Therefore, this structure is useful to take into account the flexibility of the nature of MWEs. Finally, the PI considers a correction factor, given by the normalized ratio between the frequency of the expression and that of the most frequent candidate expression  $n_{bf}^{max}$ . This correction factor, which is bounded between 0 and 1, is needed to take into account the fact that low occurrences for the basic form in the corpus reduce the reliability of the empirical tests, since the presence or the absence of modifications cannot be tested on a large set of expressions.

## 4 Analysis and Results

As a first analysis, we considered the whole set of nominal MWEs labeled as part of the lexicon of physics in GRADIT. The considered set consists of a total amount of 1.551 MWEs, 595 of which are attested to occur in our corpus.

The resulting values of the three indices (considered separately) indicate that 73% of the attested expressions are never interrupted, 93% never appear in reverse order and 64% do not attest any substitution of their components. The empirical evidence, hence, suggests that the syntactic fixedness, more than paradigmatic frozenness, seems to be relevant in outlining the prototype of nominal MWEs in physics Italian terminology. It must be underlined that the absence of modifications in the corpus does not mean that the expression does not allow them in general, nevertheless the empirical evidence can be considered a good approximation in our computational perspective.<sup>2</sup>

Since the list of physics-related MWEs extracted from GRADIT is supposed to include only terminological expressions with a completely definite phraseological status, we can consider them as a gold standard for further analyses.

In fact, the PI can be used as a new measure for the automatic extraction of MWEs from texts.

On the basis of the PI values, it is possible to assign each expression of a list of candidates a score and order the expressions according to it.

In order to have empirical evidence of the performance of the PI, we considered an input list from our corpus, composed of all the bigrams of the form [noun + adjective] which were extracted automatically, forming a set of about 22.700 expressions.

If we order the list according to PI we obtain results which appear analogous to those generally produced by statistical AMs, since PI is able to filter out most non-MWE candidates, which get very low scores and are pushed to the end of the list. At the same time, expressions appearing with very high

---

2 It must be said that some noise in this kind of approach is unavoidable, since it can happen that few expressions can exhibit modifications, but the modified expressions are not MWEs anymore, as in the case of *forza debole* 'weak force' meaning one of the four fundamental interactions, which is attested together with *debole forza*, meaning just that the intensity of a generic force is weak.



scores at the top of the list have high probability of representing true MWEs. Table 1 and Table 2 show, respectively, the top and the end of the list sorted according to PI.

Rank	MWE candidate	English translation	PI value
1	Campo magnetico	Magnetic field	0.9565
2	Campo elettrico	Electric field	0.6133
3	Momento angolare	Angular momentum	0.5717
4	Meccanica quantistica	Quantum mechanics	0.5205
5	Calorimetro elettromagnetico	Electromagnetic calorimeter	0.4748
6	Modello standard	Standard model	0.4259
7	Valore medio	Mean value	0.4206
8	Massa invariante	Rest mass	0.3683
9	Energia cinetica	Kinetic energy	0.3630
10	Campo gravitazionale	Gravitational field	0.3423
11	Campo elettromagnetico	Electromagnetic field	0.3314
12	Relatività generale	General relativity	0.3155
13	Buco nero	Black hole	0.2997
14	Meccanica classica	Classic mechanics	0.2591
15	Carica elettrica	Electric charge	0.2395

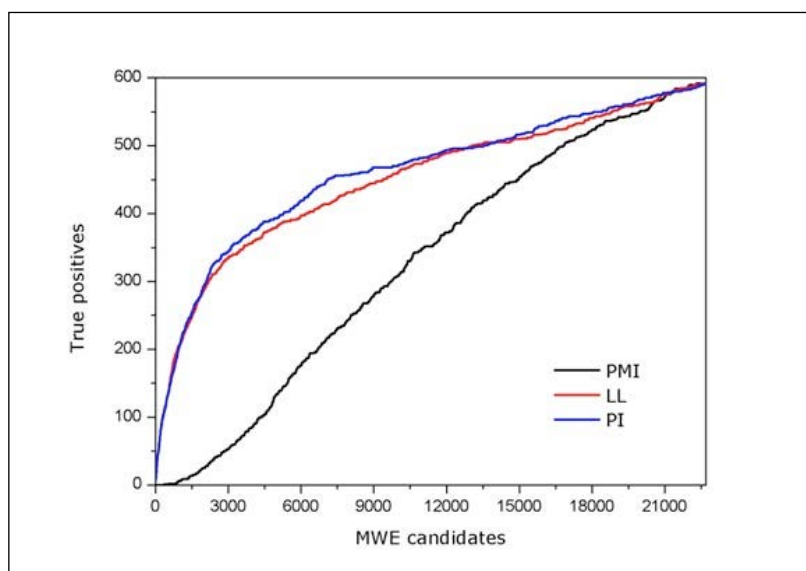
**Table 1: Top-15 of the candidate list made of all the [noun + adjective] bigrams attested in our corpus sorted according to the Prototypicality Index values.**

Rank	MWE candidate	English translation	PI value
22686	Entità indipendente	Independent entity	$7.0651 \cdot 10^{-6}$
22687	Caso tale	Case such	$6.8689 \cdot 10^{-6}$
22688	Fotone due	Photon two	$4.8725 \cdot 10^{-6}$
22689	Condizione fondamentale	Fundamental condition	$4.4757 \cdot 10^{-6}$
22690	Sistema vivente	Living system	$4.3961 \cdot 10^{-6}$
22691	Parte maggiore	Bigger part	$4.3766 \cdot 10^{-6}$
22692	Ambito magnetico	Magnetic range	$3.9057 \cdot 10^{-6}$
22693	Condizione finale	Final condition	$2.6518 \cdot 10^{-6}$
22694	Dimensione media	Average dimension	$2.5493 \cdot 10^{-6}$
22695	Forma standard	Standard shape	$2.2079 \cdot 10^{-6}$

**Table 2: End of the candidate list made of all the [noun + adjective] bigrams attested in our corpus sorted according to the Prototypicality Index values.**

In order to evaluate the performance of the PI, we chose to compare its results on our candidate list with two well-known statistical association measures, Log-likelihood (Dunning 1993), hereafter LL, and Pointwise Mutual Information (Church & Hanks 1990), hereafter PMI, which are widely used in corpus-linguistics to identify MWEs. Both AMs can be seen as representatives of two general groups of measures which quantify two different aspects of word combinations: LL measures how unlikely it is that the two words are independent while PMI investigates “how much the observed cooccurrence frequency exceeds expected frequency” as stated in Evert (2008: 1128). In this way, their use can provide two different perspectives on the statistical extraction of MWEs.

By means of the computational tool “mwetoolkit” (Ramisch et al. 2010), each bigram of our candidate list is assigned a LL and a PMI value, so that all the expressions can be ordered according to their statistical scores. The performance of PI and the two measures is evaluated on the basis of the rate of the retrieval of true positives in the lists: we compare how many true MWEs are detected while going through the lists, according to the ordering established by the scores of statistical measures and PI.



**Figure 1: Comparison between the extraction rates of true positives of Pointwise Mutual Information (black), Log-likelihood (red) and Prototypicality Index (blue).**

Figure 1 shows the curves representing the extraction rates for the three measures. As one can see, LL and PI had quite similar performances at identifying true positives, thus indicating that syntactic and semantic tests on empirical data can provide good results when used in extraction tasks. The poorer result of PMI can be justified by the fact that no frequency threshold was applied at the beginning and this AM is known for overestimating low-frequency expressions which are often false positives (Evert 2008).

We noted that for the first 1.800 candidates (corresponding to a 40% of true MWEs retrieved) LL obtained slightly better results with respect to PI, but for the remaining 20.900 candidates, the PI was almost always the better choice. This seems to indicate that on large scales the PI can be more useful

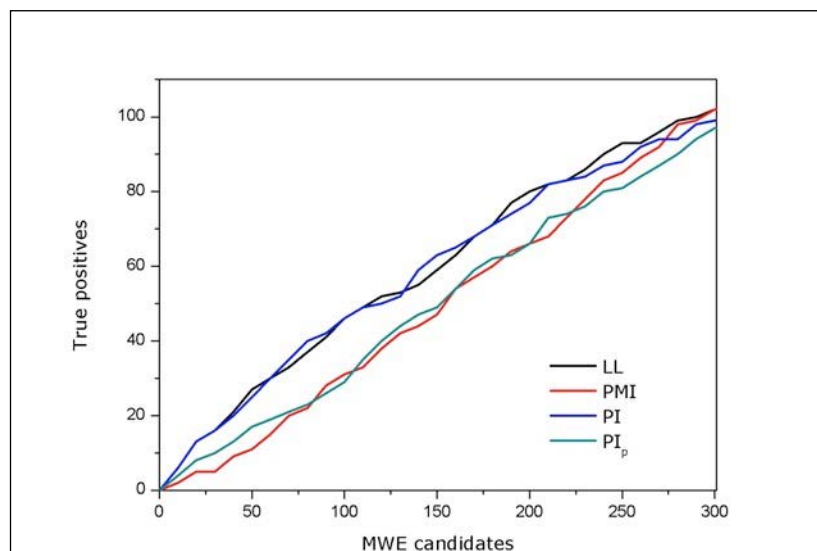
to lexicographers, who are generally interested in retrieving the largest possible number of MWEs and not only those in the first positions of the lists generated by statistics.

As an additional analysis, we considered also a frequency threshold on the input candidate lists, in order to minimize the problems related to low-frequency expressions, which especially affect PMI. Thus, we filtered our list, keeping only expressions with a frequency  $f \geq 30$  (for a total of 301 expressions) and performed the same procedure as above.

Since a frequency of at least 30 occurrences can provide a good empirical basis for the tests, we decided to consider in this case also a “pure” variant of the PI, which is not corrected by the frequency information and is given by the following formula:

$$PI_p = \frac{1}{1 + I_i + I_o + I_s}$$

Figure 2 shows the extraction rates for the four measures. Once again LL and PI are the best choices and their performances are almost equal. This time PMI appears to be more useful, as one could expect, although its extraction rate is less effective than LL or PI. Finally  $PI_p$  shows an extraction rate which is clearly better than that of PMI for the first 80 candidates, while for the remaining candidates its performance can be comparable to PMI. At the end of the process the number of true positives retrieved was 101 for LL and PMI, 99 for PI and 98 for  $PI_p$ .



**Figure 2: Comparison between the extraction rates of true positives for Log-likelihood (black), Pointwise Mutual Information (red), Prototypicality Index (Blue) and the pure version of PI (green) on a candidate list with a frequency threshold of 30 occurrences.**

## 5 Conclusions and future work

In this work we have shown how syntactic and semantic features can play an important role in studying MWEs from a computational perspective. In the case of Italian nominal MWEs of the form [noun + adjective] belonging to the special language of physics, empirical tests performed on a corpus of 1.8 million words suggested that syntactic and semantic frozenness are effective features when outlining the prototype of this kind of expressions, although semantic substitutions are more tolerated than syntactic modifications.

The three indices that quantify empirical frozenness considered in this work proved effectiveness in extraction tasks of MWEs when merged in a function that we called Prototypicality Index, which produced results that can be considered comparable to those of statistical association measures.

Such results show how our methodology can be seen as a new option for lexicographers and terminologists, to integrate the already available statistical methods when identifying MWEs from texts, thus providing one more perspective in the extraction task which can be useful to have a more complete and general overview of the phenomena as well as to create complete terminological dictionaries or resources.

Moreover, as mentioned above, the PI works better on larger scales and appears to be useful to lexicographers who are interested in retrieving more efficiently MWEs when considering a high coverage, thus dealing with expressions spanning throughout the candidate list and not focusing only on its top. This feature of PI can be explained by the fact that syntax and semantics, unlike statistical features, show more strength and reliability when dealing with less frequent expressions.

Nevertheless, the fact that a simplified version of the PI, which does not involve frequency information, produced worse results (but still similar to AMs) on a limited candidate list composed by expressions with more than 30 occurrences, shows that frequency inevitably plays a role in helping the retrieval of true positives.

However, the empirical results presented in this work must be tested on larger and more general corpora, as well as on corpora of other specialized domains, in order to evaluate the usefulness of the PI for general and specialized lexicography.

Future works must include the development of tools which can deal with other pattern of nominal MWEs as well as other grammatical categories, such as verbal or adverbial MWEs, where the above-mentioned features of modifiability are to be used in different ways when defining the prototype.

Lastly, the tools developed are to be made available, e.g. as a part of corpus research workbenches, for lexicographers and terminologists.

## 6 References

- Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions: Analysis, Acquisition and Treatment (ACL 2003)*. Sapporo, Japan.
- Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C. & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3<sup>rd</sup> International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Canary Islands.
- Cap, F., Weller, M. & Heid, U. (2013). Using a Rich Feature Set for the Identification of German MWEs. In *Proceedings of Machine Translation Summit XIV*. Nice, France.
- Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Computational Linguistics*, 16(1), pp. 22-29.
- De Mauro, T. (1999-2007). GRADIT, Grande Dizionario Italiano dell'Uso. Torino: UTET.
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. In *Computational Linguistics*, 19(1), pp. 61-74.
- Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD Thesis. University of Stuttgart.
- Evert, S. (2008). Corpora and collocations. In A. Lüdeling, M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, pp. 1212-1248.
- Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge: MIT Press.
- Kilgarriff, A. (2006). Collocationality (and how to measure it). In *Proceedings of the 12<sup>th</sup> EURALEX International Congress*. Torino: Dell'Orso, pp. 997-1004.
- Lo Cascio, V. (2012). *Dizionario combinatorio compatto italiano*. Amsterdam: John Benjamins Publishing Company.
- Masini, F. (2009). Combinazioni di parole e parole sintagmatiche. In M. Catricalà, P. Pietrandrea, E. Lombardi Vallauri, P. Di Giovine, D. Cerbasi, L. Mereu, L. Gaeta, G. Fiorentino, P. D'Achille, M. Grossmann, E. Jezeq, F. Masini, A. Pompei, E. Bonvino, F. Orletti, M. Frascarelli (eds.) *Spazi linguistici. Studi in onore di Raffaele Simone*. Roma: Bulzoni, pp. 191-209.
- Migliorini, B. (1994). *Storia della lingua italiana*. Milano: Bompiani [I. ed 1960].
- Ramisch, C., Villavicencio, A. & Boitet, C. (2010). Mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the 7<sup>th</sup> International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta.
- Sag, I., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3<sup>rd</sup> CICLing (CICLing-2002), vol. 2276/2010 of LNCS*. Mexico City, Mexico.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.
- Seretan, V. (2011). *Syntax-based Collocation Extraction*. Berlin: Springer.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Squillante, L. (2014). Towards an Empirical Subcategorization of Multiword Expressions. To appear in *Proceedings of the EACL 10<sup>th</sup> Workshop on Multiword Expressions*. Gothenburg, Sweden.
- Tiberii, P. (2012). *Dizionario delle Collocazioni. Le combinazioni delle parole in italiano*. Bologna: Zanichelli.
- Urzi, F. (2009). *Dizionario delle Combinazioni Lessicali*. Luxembourg: Convivium.
- Weller, M. & Fritzing, F. (2010). A hybrid approach for the identification of multiword expressions. In *Proceedings of the SLCT 2010 Workshop on Compounds and Multiword Expressions*. Linköping, Sweden.



# **Historical Lexicography and Etymology**





# Il DiVo (Dizionario dei Volgarizzamenti). Un archivio digitale integrato per lo studio del lessico di traduzione nell'italiano antico

Diego Dotto  
Opera del Vocabolario Italiano - CNR  
dotto@ovi.cnr.it

## Abstract

Il progetto *Divo* (*Dizionario dei volgarizzamenti*) si propone uno studio analitico del lessico di traduzione dei volgarizzamenti medievali e allo stesso tempo la costruzione di strumenti *ad hoc* per lo studio di questo lessico. Il progetto si compone di tre fasi: la prima è la compilazione di una bibliografia filologica secondo il modello *TLIon*; la seconda è la costruzione di un corpus lemmatizzato, in cui ciascun testo volgare è associato paragrafo per paragrafo all'originale latino; il terzo e ultimo punto è lo studio del lessico di traduzione, cioè il lessico identificato come una traduzione diretta dal latino.

**Keywords:** Lessicografia; Italiano antico

## 1 Lo studio lessicale dei volgarizzamenti dei classici dal cantiere del *DiVo*

Il progetto *DiVo*, ideato, promosso e diretto da Elisa Guadagnini e Giulio Vaccaro presso l'Opera del Vocabolario Italiano e la Scuola Normale Superiore di Pisa, ha l'obiettivo di studiare analiticamente il lessico dei volgarizzamenti italo-romanzi dei testi classici e tardo-antichi, con limiti fissati al VI secolo (in particolare all'opera di Boezio), per i testi di partenza, e al XIV secolo per i testi di traduzione (con l'inclusione, in casi particolari, di volgarizzamenti realizzati a cavallo tra il XIV e il XV secolo). Per fare questo, è in corso – ma si tratta di un lavoro già in larga parte disponibile alla consultazione da parte della comunità scientifica – la costruzione di tre strumenti digitali, gratuitamente e liberamente accessibili in rete:<sup>1</sup>

- la *Bibliografia filologica – DiVo DB*: un repertorio analitico di schede sulla tradizione dei testi latini e volgari oggetto dello studio, consultabile all'indirizzo <http://tliion.sns.it/divo/>;
- il *corpus DiVo*: un corpus interrogabile per forme, lemmi e iperlemmi che raccoglie esaustivamente i volgarizzamenti disponibili in edizione affidabile, con l'associazione paragrafo per paragrafo del

---

1 *DiVo DB* rientra nella rete *TLIon – Tradizione della letteratura online (TLIon DB)*, diretta da Claudio Ciociola, per cui cfr. *infra*. I *corpora* sono gestiti dal software lessicografico *GATTO 3.3*, ideato e sviluppato da Domenico Iorio-Fili presso l'Opera del Vocabolario Italiano, nella versione *Gattoweb*, realizzata dallo stesso Domenico Iorio-Fili con la collaborazione di Andrea Boccellari – si tratta dello stesso software che gestisce il corpus di riferimento dell'italiano antico, il *corpus OVI dell'Italiano antico* (in abbreviazione *corpus OVI*).

testo latino di partenza e con note filologiche sulla tradizione latina e volgare, consultabile all'indirizzo <http://divoweb.ovi.cnr.it/>;

- il *corpus CLaVo*: un corpus che permette ricerche a partire da latino per forme e lemmi, con l'associazione paragrafo per paragrafo del testo volgare d'arrivo, consultabile all'indirizzo <http://clavoweb.ovi.cnr.it/>.

Le analisi lessicali condurranno alla redazione di voci per il *TLIO* (*Tesoro della lingua italiana delle origini*) e a studi specifici di carattere onomasiologico e semasiologico sul lessico di traduzione.

I tre strumenti, nell'integrazione reciproca tra dati e meta-dati da un lato, tra linguistica, filologia e informatica dall'altro, puntano a porsi come un punto di riferimento nella ricerca lessicale sui volgarizzamenti, e più in generale sui testi dell'italiano antico, considerato il ruolo, da un punto di vista quantitativo e qualitativo, dei testi di traduzione nella documentazione italiano antica (cfr. *infra*). Nella lessicografia storica dell'italiano, questo ruolo era riconosciuto sin dal Vocabolario della Crusca (1612) e prima ancora tale consapevolezza era ben presente alle riflessioni che ne ispirarono l'elaborazione, nella fattispecie di Salviati e di Borghini: dalla valutazione dell'errore di traduzione in rapporto all'uso linguistico, da considerare positivamente come un'attestazione linguistica fede degna quando restituisce un uso possibile nell'architettura della lingua, a prescindere dal fatto che essa possa provenire da un'incomprensione totale o parziale del dettato del testo di partenza, all'effetto di trascinarsi del latino su numerosi lemmi, perlopiù crudi latinismi, che trovano nei volgarizzamenti un'attestazione prevalente o addirittura esclusiva, con, sullo sfondo, il problema della "naturalità", tema centrale per le teorie linguistiche del XVI secolo.<sup>2</sup>

Passeremo brevemente in rassegna le caratteristiche principali di questi tre strumenti, per poi portare alcuni esempi delle analisi lessicali, mostrando come con la loro interazione si possa ottenere un affinamento delle nostre conoscenze sul lessico dell'italiano antico.

Speciale attenzione sarà dedicata al rapporto tra linguistica e filologia, soprattutto in riferimento al problema dell'estrazione dei dati lessicografici da *corpora* informatici (e alla diversa e relativa affidabi-

---

2 Cfr. Guadagnini (2013: 62-65), che osserva però come la sensibilità per il problema della testimonianza dei volgarizzamenti come fonte lessicografica sia andata gradualmente perdendosi nel XVIII e XIX secolo. Riprende questo filo interrotto il *TLIO* con la *Bibliografia dei volgarizzamenti*, un repertorio sintetico di schede dedicato a tutti i volgarizzamenti presenti nel *corpus OVI dell'italiano antico*, finalizzato ad agevolare il redattore nel reperimento e nel confronto con il testo originale di partenza, in modo da fornire un'interpretazione corretta di un contesto o evidenziare un'accezione particolare di un lemma (cfr. Artale 2003: 299 e Beltrami 2010: 246-247). Naturalmente per "testo originale" va intesa un'approssimazione ricostruibile a partire dalle edizioni moderne di riferimento, dallo spoglio degli apparati delle stesse o dalle ipotesi formulabili integrando i dati disponibili - sono invece rari i casi in cui abbiamo un testo specificamente noto, e anche in questi casi, in realtà, occorrerebbe distinguere tra la lettura "reale" di una determinata lezione nel testo di partenza e la lettura mentale da parte del volgarizzatore.

lità delle edizioni su cui si fonda la costituzione di un corpus).<sup>3</sup> Si aggiunga che, tra tutte le discipline linguistiche, la lessicografia sconta una certa inerzia, legata in particolare all'effetto di trascinamento che caratterizza la tradizione lessicografica – a qualsiasi livello, ogni dizionario è inevitabilmente in dialogo con i dizionari che lo hanno preceduto – e alla scarsa o modesta affidabilità filologica dei dizionari, un problema ancora più grave e delicato quando si tratta di dizionari storici, che a loro volta sono oggetti storici.<sup>4</sup>

La prospettiva che ispira il progetto *DiVo* è così espressa da Elisa Guadagnini:

la circolarità ineliminabile fra qualità delle edizioni e qualità (vale a dire affidabilità) dei *corpora* testuali e degli studi che ne derivano è una tara che inficia la scientificità dei risultati soltanto per chi abbia la feticistica presunzione di estrapolare – dai *corpora* e dalle edizioni – dei dati di verità: la consapevolezza che qualunque testo restituito da qualunque tipo di edizione è di per sé un testo “ricostruito” consente invece, a nostro avviso, di preservare il valore, ma ancora prima il senso e la legittimità, di strumenti o di analisi che coprano vasti insiemi di materiali in una prospettiva ampiamente comparatistica, al netto del margine di oscillazione, di variabilità, di potenziale cambiamento nella lezione o nell'interpretazione, che è sempre postulabile per ogni dato testuale. (Burgassi e Guadagnini 2014, i.c.s.)

Corollario di questa impostazione è che il *corpus DiVo* si fondi su edizioni di testi con differenti gradi di affidabilità, posto che sono state escluse le edizioni inaffidabili. L'inclusione più problematica ha riguardato soprattutto le edizioni sette-ottocentesche che tengono ancora il campo nonostante il rinnovamento degli studi e dei metodi. La loro esclusione si sarebbe rivelata una soluzione facilmente percorribile, ma di fatto scarsamente funzionale alla necessità di uno studio ad ampio raggio del lessico dei volgarizzamenti: infatti, da un lato, in negativo, un corpus dei volgarizzamenti dei classici fondato in maniera esclusiva su edizioni critiche moderne avrebbe scontato una serie così ampia di lacune che il suo valore rappresentativo sarebbe stato quasi annullato; dall'altro lato, in positivo, è opportuna una parziale e ragionata rivalutazione di una parte di queste edizioni, le quali, fondate su un manoscritto unico, restituiscono di norma una lezione fede degna almeno sul piano della sostanza, ciò che più importa per chi sia interessato al lessico. Va da sé che, per esempio, un uso dell'intero

---

3 Con focalizzazione sul rapporto tra filologia e storia della lingua (cui è legittimo senz'altro sostituire linguistica, in particolare con riferimento alle varietà linguistiche antiche), questo rapporto “è spesso così vincolante da configurare un circolo vizioso: non abbiamo una buona edizione perché ci mancano sufficienti conoscenze storicolinguistiche perché non disponiamo di edizioni affidabili dei testi donde dovremmo attingere” (Stussi 1993: 214). Da un'altra angolatura, centrata invece sulla linguistica, cfr. i principi che fondano la *Grammatica dell'italiano antico* (Salvi e Renzi 2010: 7-16): “il circolo filologia-linguistica, per cui ognuno dei due punti di vista presuppone in realtà l'altro [...] si può alla volte spezzare emettendo ipotesi e provando ad applicarle”.

4 Cfr. Beltrami (2011) e Picchiorri (2013; 2014). Nella prospettiva del *TLIO*, cfr. la ricostruzione storica del dibattito sull'avvio dei lavori per il vocabolario in Vaccaro (2013). La condizione dell'esistenza stessa del *TLIO* è la scommessa di “fare” un dizionario storico fondato su spogli di prima mano, a loro volta fondati sulle edizioni esistenti disponibili, spezzando il circolo vizioso che altrimenti avrebbe costretto a una dilazione continua dell'elaborazione del vocabolario a favore della preparazione dei testi (Beltrami 2011: 342).

*corpus DiVo* per ricerche che investano la forma dei testi (i livelli della fonologia, della morfologia o alcuni aspetti della morfosintassi) è a carico della (ir)responsabilità di chi consulta il corpus, mentre è una responsabilità di chi lo costruisce mettere a disposizione tutti i dati per una corretta valutazione dell'edizione presente nel corpus, chiarendo le metodologie seguite dall'editore e verificandone l'affidabilità. Viceversa un utilizzo parziale anche per la forma è possibile, ma solo per sottocorpora, selezionando i testi editi secondo criteri rigorosi anche sul piano formale o particolarmente interessanti su questo fronte, com'è il caso, per esempio, dei volgarizzamenti testimoniati da un autografo.<sup>5</sup> Insomma alle spalle e davanti al lessicografo, come a chiunque interroghi un corpus informatizzato di testi, sta (o dovrebbe stare) sempre "una continua valutazione critica dei dati" (Beltrami 2011: 348).

## 2 La Bibliografia filologica del DiVo - DiVo DB

Premessa e allo stesso tempo risultato della valutazione delle edizioni da inserire nel corpus dei volgarizzamenti dei classici è la *Bibliografia filologica del DiVo (DiVo DB)*, repertorio di schede con la presentazione dei dati essenziali sull'opera: cenni biografici sull'autore del volgarizzamento, datazione, identificazione della coloritura linguistica, indicazione della tipologia testuale e del genere, catalogazione della tradizione diretta mediante il censimento dei testimoni manoscritti e delle stampe antiche, sintesi sulla storia della tradizione, identificazione dell'edizione di riferimento e panorama bibliografico articolato per punti<sup>6</sup>. Le schede sui testi di partenza (perlopiù opere latine, per cui cfr. Zago 2012) sono invece sintetiche, fornendo dati sull'autore, sulla datazione, sulla tipologia testuale, sul genere e sull'edizione di riferimento sulla base di una valutazione dello stato degli studi.

Come ha scritto Giulio Vaccaro (Guadagnini e Vaccaro 2014: 127), il lavoro alla base di *DiVo DB* "è stato foriero di robuste novità nel campo delle acquisizioni testuali", e di conseguenza anche per i dati lessicografici in uscita.

Si pensi per esempio al campo delle retrodatazioni: tradizionalmente il volgarizzamento delle *Collazioni dei santi Padri* di Giovanni Cassiano era datato in modo generico al XIV secolo (nel *GDLI* e come fuori corpus nel *TLIO*) sulla base dell'edizione ottocentesca fondata sul manoscritto 1637 della Biblioteca statale di Lucca, datato al 1442 (Bini 1854). Ma l'indagine sulla tradizione manoscritta di questo testo, peraltro non unitario, ma suddivisibile in due distinti volgarizzamenti per le collazioni I-X e XI-XXIV, ha portato al rinvenimento di un testimone senese del volgarizzamento A, databile su base paleografica alla fine del XIII secolo (Siena, Biblioteca degli Intronati, I V 8), che produrrà sicuramente numero-

---

5 Sui criteri che hanno guidato l'inclusione delle edizioni nel corpus, anche in rapporto al *corpus OVI dell'Italiano antico*, che ha criteri leggermente diversi, cfr. Dotto (2013: 75-78) e Guadagnini e Vaccaro (2014: 119-127). Nel *corpus DiVo*, per esempio, si leggono (o si leggeranno) le edizioni dei volgarizzamenti delle *Heroides* di Filippo Ceffi secondo il riconosciuto autografo Vaticano Palatino Latino 1644 (Zaggia 2009) o del *De brevitate vitae* di Seneca secondo la copia di mano di Andrea Lancia (Frullani 5 della Biblioteca Moreniana di Firenze), da riconoscere verosimilmente come il volgarizzatore (cfr. per ora De Robertis e Vaccaro 2013).

6 Il censimento dei volgarizzamenti dei classici è passato attraverso una versione a stampa: cfr. Artale, Guadagnini, Vaccaro 2010.

se retrodatazioni: per esempio la prima attestazione della voce *inquietudine* del *TLIO*, la cui più antica attestazione compare nel volgarizzamento fiorentino delle *Pistole* di Seneca, databili *ante* 1325, potrà essere retrodata proprio grazie a questo ritrovamento, sfruttando le potenzialità di uno strumento come il *TLIO*, pubblicato direttamente in rete e pertanto facilmente aggiornabile.<sup>7</sup>

### 3 Il corpus DiVo (Corpus del Dizionario dei volgarizzamenti)

Il *corpus DiVo* raccoglie le edizioni affidabili dei volgarizzamenti dei testi classici e tardo-antichi in una qualsiasi varietà italo-romanza con limite posto al XIV secolo. Esiste un vincolo all'eshaustività per i volgarizzamenti delle opere classiche. Sono inclusi nel corpus anche i testi che non sono volgarizzamenti diretti da latino, ma hanno un intermediario romanzo (di norma il francese), per esempio le *Pistole* di Seneca, che dipendono da un volgarizzamento francese, o in un'altra varietà italo-romanza (di norma il toscano), per esempio *l'Istoria d'Eneas* siciliana di Angilu di Capua, che dipende dalla compilazione fiorentina tradizionalmente attribuita ad Andrea Lancia. Inoltre, in casi motivabili con il valore storico-culturale o linguistico del testo, oltre ai volgarizzamenti veri e propri, con la resa generalmente puntuale del dettato latino, sono consultabili nel corpus anche alcune compilazioni che corrispondono al montaggio di volgarizzamenti distinti: è il caso dei *Fatti de' Romani*, vasta compilazione di storia romana di origine antico francese, che riusa Sallustio, Cesare, Lucano e Svetonio e che è circolata in Italia secondo differenti redazioni.

Attualmente, col rinnovo di marzo 2014, il corpus comprende 150 testi volgari, per complessive 5.941.061 occorrenze e 169.845 forme grafiche distinte. Per dare un'idea della sua entità e della ragione per cui esso potrà integrare utilmente il corpus di riferimento dell'italiano antico (*corpus OVI*), ricordo che quest'ultimo comprende 2316 testi per complessive 23.157.266 occorrenze e 467.098 forme grafiche distinte.

I punti di forza del corpus sono:

- l'associazione paragrafo per paragrafo del testo latino di partenza, in modo da rendere immediatamente conto del rapporto tra testo tradotto e corrispondente traduzione;
- un sistema di annotazione, che contempla la lemmatizzazione e l'iperlemmatizzazione sugli ambiti semantici sensibili per lo studio del lessico dei volgarizzamenti.<sup>8</sup>

---

7 Per questa ragione nel *corpus DiVo* sono e saranno consultabili un'edizione a uso interno del volgarizzamento A sulla base del manoscritto senese, a cura del progetto *DiVo*, e l'edizione Bini che si fonda sul manoscritto di Lucca, che accorpa entrambi i volgarizzamenti. Si rinvia alla relativa scheda in *DiVo DB*.

8 Sui criteri dell'associazione del latino al testo volgare, cfr. Burgassi 2013: l'associazione dei volgarizzamenti dei classici è completa, per cui il corpus ad oggi presenta 79 testi con latino associato, che coprono 3.172.419 occorrenze. La lemmatizzazione e l'iperlemmatizzazione sono invece ancora in costruzione (per un primo abbozzo dei principi e delle soluzioni tecniche che si seguiranno, cfr. Dotto 2012).

## 4 Il corpus CLaVo (Corpus dei classici latini volgarizzati)

Il *corpus CLaVo* è il corpus “gemello” del *corpus DiVo* in quanto raccoglie tutte le opere latine classiche contenute nel *corpus DiVo*, con l’associazione paragrafo per paragrafo di ciascun volgarizzamento. Al momento contiene 26 opere latine, associate a 45 volgarizzamenti, per complessive 913.656 occorrenze di 78.587 forme grafiche distinte, ma non è ancora completo; è consultabile per forme e in parte anche per “lemmi muti”, una speciale lemmatizzazione, curata da Anna Zago, che fornisce una prima griglia di coppie forma-lemma per agevolarne l’interrogazione. A regime esso conterrà più di 80 testi, per circa 1.300.000 occorrenze e oltre 120.000 forme grafiche distinte; grazie al lemmatizzatore semi-automatico di GATTO 4.0 potrà contare su una lemmatizzazione vera e propria largamente esaustiva.<sup>9</sup>

La funzione primaria del *corpus CLaVo* è la possibilità di recuperare agevolmente attraverso lo scaricamento dei contesti tutte le rese traduttive dello stesso lemma latino a partire dal testo latino di partenza, attraverso i meccanismi del prestito, del calco o della riformulazione volgare (un equivalente o una perifrasi).<sup>10</sup> Così per es., come vedremo in parte più avanti, è possibile recuperare i traduttori del *lat. lictor*: *littore, berroviere, giustiziere, masnadiere, messo, sergente*, ecc. Uno strumento simile ha fortissime potenzialità, per sondare sia il lessico tecnico e “storico” (cfr. *infra*), sia il lessico generico. In questo modo infatti è possibile tracciare relazioni onomasiologiche all’interno del volgare, per inventariare le relazioni sinonimiche e soprattutto per apprezzare le frequenze d’uso, i valori connotativi e denotativi dei diversi lemmi in sincronia e i cambiamenti che i medesimi lemmi hanno subito in diacronia, nella “breve” diacronia del XIII e XIV secolo con interesse centrato sull’italiano antico, come nella “lunga diacronia” per una migliore intelligenza di determinati lemmi, come nel caso dei “latinismi latenti” (cfr. Burgassi e Guadagnini 2014).

Da notare che l’ordinamento cronologico dei testi non si riferisce a quello dei testi latini, ma a quello dei testi volgari: questa soluzione intende agevolare lo studio di come siano cambiate le modalità traduttive nel XIII e XIV secolo, per verificare analiticamente l’ipotesi già di Cesare Segre (1953) dell’esistenza di una prima fase, in cui i volgarizzatori tendono a ricorrere a equivalenti volgari, e di una seconda, per certi aspetti preumanistica, benché l’attività del volgarizzare sia operazione schiettamente anti-umanistica, in cui i volgarizzatori privilegiano il prestito dal latino.

## 5 Esempi di analisi lessicali

Uno dei problemi più cospicui dei volgarizzatori impegnati nella traduzione dei classici era quello di rendere il lessico materiale o “storico”: si tratta di un lessico non-marcato in latino perché non indivi-

9 Su GATTO 4.0, destinato a sostituire GATTO 3.3, e in particolare sul lemmatizzatore semi-automatico, cfr. Iorio-Fili 2012.

10 Un’altra potenzialità del *corpus CLaVo*, che non si può discutere qui, è il suo riuso indipendente dall’analisi lessicale dei volgarizzamenti in quanto banca dati di testi classici in sé e per sé, interrogabile grazie agli strumenti del software lessicografico GATTO 3.3.

dua settori disciplinari specifici, ma risulta “speciale” solo in una prospettiva storica perché rinvia a referenti che sono tipici della società e della cultura antica e sono scomparsi in quella medievale. Ri-entrano in questa categoria i lemmi che si riferiscono a oggetti di uso quotidiano o alla misurazione del tempo, sostituita in séguito dalla cronologia cristiana, o ancora all’esistenza di popoli scomparsi e territori ridisegnati rispetto alla prospettiva e alle conoscenze geografiche dei volgarizzatori medievali. È il caso del lessico della cariche e degli uffici, che conteneva ampie di zone di discontinuità tra il mondo antico e quello medievale. In questi casi il volgarizzatore poteva oscillare tendenzialmente tra una traduzione orientata sulla lingua di partenza, il latino, ricorrendo al prestito diretto o una traduzione orientata sulla lingua d’arrivo, il volgare, cercando nel repertorio lessicale della propria lingua un equivalente che desse conto del significato del lemma latino. Possibilità collaterale, in realtà caratteristica di molti volgarizzamenti dei classici, specialmente nella prima metà del XIV secolo a Firenze, era quella di accompagnare il testo con apparati paratestuali in forma di chiose, marginali o interlineari, o di glossari, all’inizio o alla fine del testo.<sup>11</sup> È quanto avviene per il lat. *lictor* “ufficiale che accompagnava vari magistrati romani recando con sé fasci di verghe con una scure in mezzo”:<sup>12</sup>

(1) gli maggiori Romani con gradissima diligenza ritennero questa usanza: che alcuno non s’interponesse tra ’l consolo e ’ pressimani *lictori*, tutto ch’egl’andassero insieme per cagione d’oficio. (*Valerio Massimo* (Vb), a. 1326 [fior.] / cfr. Val. Max. II, 2, 24: “Maxima autem diligentia maiores hunc morem retinuerunt, ne quis se inter consulem et proximum *lictozem*, quamvis officii causa una progrederetur, interponeret”

– nel volgarizzamento parziale di Valerio Massimo, Vb, troviato un prestito diretto, ma nello stesso testo nelle chiose troviamo l’equivalenza tra “ufficiale” e “littore” (2) e in un glossario, peraltro presente in uno dei due testimoni che tramanda il volgarizzamento Vb, tra “sergente” e “littore” (3):

(2) xii imperialissimi onori erano li xii *officiali* de’ consoli, li quali si chiamavano ‘*lictori*’ e portavano le ’nsegne de’ consoli. (*Chiose a Valerio Massimo* (Vb), 1326 [fior.]

(3) Ciascuno consolo aveva xii *sergenti* li quali erano chiamati *littori*. (*Gloss. degli uffici romani* (red. Marc.), XIV pm. [tosc.]

L’equivalenza tra il termine generico “sergente” e “littore” è documentata anche nel volgarizzamento della *Deca quarta*:<sup>13</sup>

(4) sopra tutti con portamento eccelso rendere superbe leggi intorniato di *sergenti* chiamati *littori*; i quali sempre e sudditi stanno con le verghe del ferro al dosso, e con le securi sopra le teste; e questo avvi-

---

11 Le chiose, come i glossari, servivano a colmare la “distanza epistemica” tra la cultura antica e quella medievale: “Il commento è il termometro delle difficoltà della comunicazione. Il caso più ovvio è quello della distanza cronologica e geografica tra emittente e ricevente: sono i testi antichi o quelli in altre lingue ad essere fregiati più spesso di commento. Si potrebbe parlare, meglio, di distanza epistemica: si terrebbe così conto, oltre che della distanza cronogeografica, anche di quella culturale” (Segre 1992: 4).

12 I testi citati e le relative abbreviazioni bibliografiche corrispondono a quelle del *corpus DiVo*: per accedere alla bibliografia: *DiVoWeb* (o *CLaVoWeb*) > *Altre funzioni* > *Accesso ai dati bibliografici* (con i link a *DiVo DB* per ulteriori approfondimenti).

13 Se ne hanno riscontri anche in ambito francese antico: cfr. la *Base de civilisation romaine (XII<sup>e</sup>-XV<sup>e</sup> s.)*, s.v. *lictor*.



ene quanti anni ora l'uno ed ora l'altro signore rinnovando sortiscono. (*Deca quarta*, a. 1346 [fior.]) / cfr. Liv., XXXI, 39, 9: "Praetor Romanus conventus agit: eo imperio evocati conveniunt, excelso in suggestu superba iura reddentem, stipatum lictoribus vident, virgae tergo, securae cervicibus imminent"

Un caso interessante è discusso da Massimo Zaggia (1991: 611-613) in una sua recensione al volgarizzamento anonimo della prima Epistola di Cicerone al fratello Quinto (Piva 1989):

- (5) Per le quali cose Gneo Ottavio poco tempo fa fu reputato da tutti molto soave e benigno, nel cui reggimento il primo *littore* o *berroviere* tacette senza vietare la venuta <e non bisognava dire 'l tale vuole parlare>, ciascuno parlò quante volte gli piacque e quanto lungamente egli volle. (*Ep. a Quinto*, XIV sm. [tosca.]) / cfr. Cic. *Q. fr.*, 21: "apud quem primus *lictor* quievit, tacuit accensus"

Il manoscritto base seguito dall'editrice legge "littore o sergente", mentre il resto della tradizione ha "littore o berroviere". In questo caso Maria Antonia Piva opta per l'espunzione di "sergente" / "berroviere" ritenendo che si tratti di una glossa che ha così costituito una dittologia sinonimica, secondo un processo normale nella tradizione dei volgarizzamenti.<sup>14</sup> All'operazione della Piva, Zaggia (1991: 612) obietta che "se l'eliminazione delle esplicazioni sinonimiche sembra legittima quando queste risultano trasmesse da un solo ramo della tradizione [...], non pare autorizzata quando tutti i testimoni concordano nel trasmettere una dittologia". A ulteriore sostegno della propria obiezione, Zaggia riporta il seguente esempio, in cui tutta la tradizione concorda nella dittologia "berroviere e littore":

- (6) Sia ogni tuo *berroviere* e *littore* dimostratore non della sua benignità e dolcezza, anzi della tua, e quelli frusti e quelle scure o mannaie che portano più dimostrino segno della dignità dell'ufficio tuo che della signoria e forza. (*Ep. a Quinto*, XIV sm. [tosca.]) / cfr. Cic. *Q. fr.*, 13: "sit *lictor* non suae sed tuae lenitatis apparitor"

Nel *corpus DiVo*, si è scelto di seguire l'ipotesi formulata da Zaggia, che preferisce la dittologia "littore o berroviere" a causa del riscontro di (6), inserendo una nota che dà conto della divergenza rispetto all'edizione di riferimento e si giustifica l'intervento apportato.

Un altro caso che chiama in causa il delicato rapporto tra edizioni e dati lessicografici ricavabili da esse. A fronte del lat. *praesultor*, che evidentemente a causa della propria rarità non doveva risultare perspicuo, le tre redazioni del volgarizzamento toscano di Valerio Massimo (in bibliografia con le sigle Va V1 V2) presentano i seguenti esiti:<sup>15</sup>

- (7) Iove comandò a un popolare latino in sogno che dicesse al consolo che no· gli piaceva ne' prossimi giochi di Circe quello *prosentuoso vendicatore*... (Valerio Massimo (red. Va), a. 1336 [tosca.]) / cfr. Val. Max., I, 7, 4: "<T>. Latinio homini ex plebe Iuppiter in quiete praecepit ut consulibus diceret sibi *praesultorem* ludis circensibus proximis non placuisse"

14 Se "sergente" è termine generico, non così per "berroviere", che in italiano antico era ben attestato nel significato di "funzionario con mansioni esecutive al servizio di un ufficiale pubblico (capitano, podestà, priore ecc.) o di un signore" (cfr. *TLIO* s.v.).

15 Nel *corpus DiVo* le redazioni del volgarizzamento di Valerio Massimo si leggono per la prima volta integralmente grazie alle edizioni di lavoro approntate da Vanna Lippi Bigazzi, che ringraziamo di cuore qui. Cfr. la scheda in *DiVo DB*.



Da dove nasce la lezione “prosentuoso vendicatore”? Da un’erronea segmentazione, per cui il volgarizzatore deve aver letto “praes ultorem” in luogo di “praesultorem”. La lezione non è però confermata nella redazione V1, almeno secondo l’edizione DeVisiani (1867-1868) (8), mentre l’edizione di Vanna Lippi Bigazzi conferma per V1 la lezione di Va (9):

(8) Iove comandò a uno latino del popolo in sogno, che dicesse al consolo, che non li piaceva nelli prossimi giuochi circensi quello *presultore*. (Valerio Massimo (red. V1, ed. De Visiani), a. 1336 [fior.])

(9) Iove comandoe a uno latino del popolo in sogno che dicesse al consolo che no· lli piaceva <vedere> ne li prossimi giuochi di Circe quello *presumptuoso vendicatore*... (Valerio Massimo (red. V1, ed. Lippi Bigazzi), a. 1336 [fior.])

Spulciando l’apparato dell’edizione De Visiani (1867-1868: 80), si apprende che il suo manoscritto base (“St. e Cod.”) legge in realtà “presentuoso vendicato” e solo un manoscritto, il Palatino 27 della Biblioteca Palatina di Parma, legge il crudo latinismo, non altrimenti attestato, “presultore”.

Se passiamo all’ultimo anello delle redazioni del Valerio Massimo, V2, troviamo un’altra lezione, frutto probabilmente di un emendamento come nel caso di “presultore”:

(10) Iove in sogno comandò ad uno uomo latino de la minuta gente che dicesse al consolo che a lui non era piaciuto quello *antisaltatore* ne li prossimi giuochi Circesi... (Valerio Massimo (red. V2), c. 1346 [tosc.])

Da un punto di vista lessicografico, importa constatare che “presultore” andrà scalato cronologicamente, visto che la lezione genuina di V1 sarà certo l’errore di traduzione “presuntuoso vendicatore” per trascinamento da Va. I due emendamenti restituiti dalla tradizione corrispondono inoltre a due distinte tipologie di resa traduttiva: la prima per prestito, la seconda per calco.

Va notato infine che nel contesto dello stesso esempio il volgarizzatore della *Deca prima* di Tito Livio, Filippo da Santa Croce ricorre a una complessa perifrasi a dimostrazione della delicatezza della resa di *praesultor* o dell’affine *praesultator*:

(11) e fugli avviso che Giove gli dicesse, che *quegli che la prima danza aveva alla festa menata*, gli dispiacque... (Filippo da Santa Croce, *Deca prima*, 1323 [tosc.]) / cfr. Liv., II, 36, 2: “visus Iuppiter dicere sibi ludis praesultatorem displicuisse”

Gli esempi portati vorrebbero dimostrare l’assunto iniziale: una “continua valutazione critica dei dati” è condizione necessaria per l’integrazione di linguistica e filologia.

## 6 Riferimenti bibliografici

Artale, E. (2003). I volgarizzamenti del *corpus TLIO*. In *Bollettino dell’Opera del Vocabolario Italiano*, 8, pp. 299-377.

*Base de civilisation romaine (XII<sup>e</sup>-XV<sup>e</sup> s.)*, ed. F. Duval, CNRTL CNRS-ATILF. Accessed at: <http://www.cnrtl.fr/lexiques/civrom/> [02/02/2014].

- Artale, E., Guadagnini, E., Vaccaro, G. (2010). Per una bibliografia dei volgarizzamenti dei classici (il *Corpus DiVo*). In *Bollettino dell'Opera del Vocabolario Italiano*, 15, pp. 309-366.
- Beltrami, P.G. (2010). Lessicografia e filologia in un dizionario storico dell'italiano. In C. Ciociola (ed.), *Storia della Lingua Italiana e Filologia. Atti del VII Convegno ASLI (Pisa-Firenze, 18-20 dicembre 2008)*. Firenze: Franco Cesati, pp. 235-248.
- Beltrami, P.G. (2011). Il mito dell'edizione per lessicografi e il *Tesoro della Lingua Italiana delle Origini*. In A. Overbeck, W. Schweickard, H. Völker (eds.), *Lexicon, Varietät, Philologie. Romanistische Studien Günter Holtus zum 65. Geburtstag*. Berlin: De Gruyter, 2011, pp. 341-349.
- Bibliografia dei volgarizzamenti [del corpus TLIO]*, ed. E. Artale. Accessed at: <http://tlio.ovi.cnr.it/BibVolg/> [02/02/2014].
- Bini, T. (ed.). 1854. *Volgarizzamento delle Collazioni dei SS. Padri del venerabile Giovanni Cassiano*. Lucca: Giusti.
- Burgassi, C. (2013). Notizie dal *DiVo*. Teoria e pratica dell'associazione latino-volgare, In P. Larson, P. Squillacioti, G. Vaccaro (eds.), «Diverse voci fanno dolci note». *L'Opera del Vocabolario Italiano per Pietro G. Beltrami*. Alessandria: Edizioni dell'Orso, pp. 85-96.
- Burgassi, C. e Guadagnini, E. (2014). Prima dell'«indole». Latinismi latenti dell'italiano. In *Studi di lessicografia italiana*, XXXI, i.c.s.
- Corpus CLaVo. Corpus dei classici latini volgarizzati*, eds. C. Burgassi, D. Dotto, E. Guadagnini, G. Vaccaro. Accessed at: <http://clavoweb.ovi.cnr.it/> [02/02/2014].
- Corpus DiVo. Corpus del Dizionario dei Volgarizzamenti*, eds. C. Burgassi, D. Dotto, E. Guadagnini, G. Vaccaro. Accessed at: <http://divoweb.ovi.cnr.it/> [02/02/2014].
- Corpus OVI dell'Italiano antico*, ed. E. Artale e P. Larson. Accessed at: <http://gattoweb.ovi.cnr.it/> [02/02/2014].
- De Robertis, T. e Vaccaro, G. (2013). *Il Libro di Seneca della brevitade della vita humana* in un autografo di Andrea Lancia. In *Studi di filologia italiana*, 71, i.c.s.
- De Visiani, R. (ed.). 1867-1868. *Valerio Massimo, De' fatti e detti degni di memoria della città di Roma e delle strane genti*. Bologna: Romagnoli.
- DiVo DB. DiVo - Bibliografia filologica*, eds. E. Guadagnini e G. Vaccaro. Accessed at: <http://tlion.sns.it/divo/> [02/02/2014].
- Dotto, D. (2012). Note per la lemmatizzazione del corpus DiVo. In *Bollettino dell'Opera del Vocabolario Italiano*, 17, pp. 339-366.
- Dotto, D. (2013). Notizie dal *DiVo*. Un primo bilancio sulla costituzione del corpus. In P. Larson, P. Squillacioti, G. Vaccaro (eds.), «Diverse voci fanno dolci note». *L'Opera del Vocabolario Italiano per Pietro G. Beltrami*. Alessandria: Edizioni dell'Orso, pp. 71-83.
- GDLI*. (1961-2002). *Grande dizionario della lingua italiana*, ed. S. Battaglia, [poi G. Bàrberi Squarotti]. Torino: UTET.
- Guadagnini, E. (2013). Notizie dal *DiVo*. Parole tradotte e lessicografia dell'italiano. In P. Larson, P. Squillacioti, G. Vaccaro (eds.), «Diverse voci fanno dolci note». *L'Opera del Vocabolario Italiano per Pietro G. Beltrami*. Alessandria: Ed. dell'Orso, pp. 59-70.
- Guadagnini, E. e Vaccaro, G. (2014). Un contributo allo studio del "Volgarizzare e tradurre": il progetto *DiVo*. In *Lingue, testi, culture. L'eredità di Folena, vent'anni dopo*. Atti del XL Convegno Interuniversitario di Bressanone (12-15 luglio 2012). Padova: Esedra, pp. 113-127.
- Iorio-Fili, D. (2012). Il lemmatizzatore semi-automatico di GATTO4. In *Dizionari e ricerca filologica. Atti della Giornata di Studi in memoria di Valentina Pollidori (Firenze, Villa Reale di Castello, 26 ottobre 2010)*. Alessandria: Ed. Dell'Orso, pp. 41-56.
- Picchiorri, E. (2013). Sulla genesi di un errore nel Battaglia. In *Studi linguistici italiani*, 39(1), pp. 134-136.
- Picchiorri, E. (2014). Problemi filologici nei dizionari storici italiani dal *GDLI* al *TLIO*. In J.M. Brincat, R. Coluccia, F. Möhren (eds.), *Actes du XXVII<sup>e</sup> Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013)*. Section 5: Lexicologie, phraséologie, lexicographie. Nancy: ATILF, i.c.s.

- Piva, M.A. (ed.). 1989. Anonimo trecentesco. *Volgarizzamento della prima Epistola di Cicerone al fratello Quinto*. Bologna: Commissione per i Testi di lingua.
- Salvi, G. e Renzi, L. (eds.). (2010). *Grammatica dell'italiano antico*. Bologna: il Mulino.
- Segre, C. (ed.). 1953. *Volgarizzamenti del Due e Trecento*. Torino: UTET.
- Segre, C. 1992. Per una definizione del commento ai testi. In O. Besomi e C. Caruso (eds.), *Il commento ai testi*. Atti del Seminario di Ascona (2-9 ottobre 1989). Basel-Boston-Berlin: Birkhäuser, pp. 3-14.
- Stussi, A. (1993) *Lingua, dialetto e letteratura*. Torino: Einaudi.
- TLIO. *Tesoro della lingua italiana delle origini*, ed. P. Squillacioti. Accessed at: <http://tlio.ovl.cnr.it/TLIO/> [02/02/2014].
- TLion DB. *Tradizione della letteratura italiana online*, ed. C. Ciociola. Accessed at: <http://www.tlion.it/> [02/02/2014].
- Vaccaro, G. (2013). Veniamo da molto lontano e andiamo molto lontano. Documenti per la storia dell'Opera del Vocabolario Italiano dalle origini al 1992. In *Bollettino dell'Opera del Vocabolario Italiano*, 18, ic.s.
- Zaggia, M. 1991. Rec. a Piva 1989. In *Rivista di letteratura italiana*, IX.3, pp. 611-616.
- Zaggia, M. (ed.). 2009. Ovidio, «Heroides». Volgarizzamento fiorentino trecentesco di Filippo Ceffi. I. Introduzione, testo secondo l'autografo e glossario, Firenze, SISMEL-Ed. del Galluzzo.
- Zago, A. (2012). La bibliografia dei testi latini (e greci) inclusi nel *corpus DiVo*. In *Bollettino dell'Opera del Vocabolario Italiano*, 17, pp. 367-391.

### **Acknowledgements**

Questo contributo rientra nel progetto *DiVo* (*Dizionario dei Volgarizzamenti*), ospitato dall'Opera del Vocabolario Italiano - CNR e dalla Scuola Normale Superiore di Pisa, finanziato dal MIUR all'interno del programma FIRB - Futuro in Ricerca 2010.



# Informatiser le *Französisches etymologisches Wörterbuch*: la nécessaire prise en compte de l'utilisateur

Pascale Renders, Esther Baiwir  
F.R.S-FNRS/Université de Liège  
pascale.renders@ulg.ac.be, ebaiwir@ulg.ac.be

## Résumé

« Know your user » (Atkins & Rundell 2008 : 5). Ce conseil ne s'applique pas uniquement à la conception et la rédaction d'un nouveau dictionnaire. La transformation d'un dictionnaire imprimé en dictionnaire électronique peut également bénéficier d'une étude réévaluant les parcours de consultation suivis par les utilisateurs et les difficultés qu'ils rencontrent. Dans une étude préalable à l'informatisation du *Französisches Etymologisches Wörterbuch* de Walther von Wartburg, la prise en compte du point de vue des utilisateurs a mis en évidence deux facettes de l'ouvrage, vu tantôt comme un recueil de monographies, tantôt comme un thesaurus. Après un résumé de l'avancement du projet (qui est maintenant dans sa phase de production), cette communication expose la façon dont les deux visions ont influencé la modélisation informatique du discours lexicographique, permettant ainsi de résoudre une grande partie des problèmes rencontrés par les utilisateurs et ouvrant la voie à une mise à jour de l'ouvrage.

**Keywords:** FEW; digitalization; user perspective

## 1 Introduction

Il est aujourd'hui admis que le concepteur d'un dictionnaire doit d'abord définir précisément le public auquel il s'adresse:

[...] the most important single piece of advice we can give to anyone embarking on a dictionary project is : know your user. [...] This doesn't imply a superficial concern with 'user-friendliness', but arises from our conviction that the content and design of every aspect of a dictionary must, centrally, take account of who the users will be and what they will use the dictionary for. (Atkins & Rundell 2008 : 5)

Cette remarque s'applique-t-elle aussi pour un dictionnaire existant, déjà publié, qu'on voudrait transformer en dictionnaire électronique ? Si l'objectif de l'informatisation est d'augmenter les potentialités de consultation et de résoudre des problèmes d'utilisation, ne faut-il pas avoir une idée précise de l'identité des utilisateurs du dictionnaire, de ce qu'ils y cherchent, de la façon dont ils le consultent, des difficultés qu'ils rencontrent et des fonctionnalités qu'ils voudraient y trouver ? Plus générale-

ment, l'utilisation effective de l'ouvrage diverge-t-elle de ce qui était prévu lors de la conception du projet initial et peut-elle être améliorée par le changement de medium ?

Ces questions ont été le point de départ d'une étude qui s'interrogeait sur les possibilités d'informatisation du *Französisches etymologisches Wörterbuch* de Walther von Wartburg (ou FEW), dictionnaire de référence en linguistique historique française et romane. Cette étude a été achevée en 2011 ; l'informatisation du FEW est en cours depuis octobre 2012. Après un résumé du projet et de son avancement, nous proposons d'exposer comment le point de vue de l'utilisateur a modifié l'analyse des structures du dictionnaire et comment ce changement de perspective a été pris en compte dans l'informatisation proprement dite.

## 2 Le projet d'informatisation du FEW

L'idée d'informatiser le FEW vient de ses utilisateurs. Le premier fut Wooldridge, déjà en 1990 (1990 : 239) puis en 1998 : « [l]a seule façon de mettre au jour ce qui concerne le français du XVI<sup>e</sup> siècle dans le FEW serait d'informatiser les 25 volumes... puis de les interroger à partir de repères comme « fr. », « mfr. », « 16<sup>e</sup> s. », etc. » (Wooldridge 1998 : 211). Il suffit d'avoir consulté une fois l'ouvrage pour comprendre l'intérêt à la fois de celui-ci et de son informatisation. Le FEW, rédigé de 1922 à aujourd'hui (l'équipe de rédaction s'attelle depuis quelques années à la refonte du premier volume, cf. [www.atilf.fr/few](http://www.atilf.fr/few)), est un dictionnaire de référence en linguistique romane, contenant le lexique de tous les parlers galloromans (français, francoprovençal, occitan, gascon et dialectes). Il est néanmoins sous-exploité, en raison de la complexité de ses structures (cf. Büchi & Chambon 1995). Son informatisation, ardemment souhaitée par la communauté scientifique, est censée à la fois résoudre les problèmes d'accessibilité de l'ouvrage et permettre sa mise à jour ; toutefois, un contenu dense et l'existence de nombreux caractères spéciaux non Unicode (plus d'une centaine) rendaient, jusqu'il y a peu, le projet utopique. Après quelques tentatives partielles, une étude de faisabilité fut entamée, financée par le laboratoire ATILF à Nancy (où est hébergée la rédaction du FEW, cf. [www.atilf.fr](http://www.atilf.fr)), par la Fondation FEW (Suisse) et, en majeure partie, par l'Université de Liège en Belgique (bourse de doctorat).

Cette étude fut concluante : l'informatisation des 25 volumes du FEW était non seulement souhaitable, mais possible, sous la forme d'un balisage XML inséré a posteriori dans le discours lexicographique (Renders 2011). La réussite de l'entreprise nécessitait toutefois que soient pris en compte trois contraintes fortes, parmi lesquelles l'obligation de respecter les structures du dictionnaire, y compris dans leurs incohérences et leurs défauts, avec l'interdiction formelle de réécrire les articles pour normaliser le tout. La deuxième contrainte était la nécessité de pouvoir automatiser complètement l'insertion du balisage. La troisième contrainte, enfin, consistait à s'assurer que le résultat de l'informatisation répondrait effectivement aux attentes des utilisateurs, notamment en résolvant les divers problèmes d'accès auxquels ils se heurtaient. La prise en compte de ces trois contraintes mène en pratique à l'élaboration d'un compromis qui, seul, permet une conclusion positive.

Sur la base de la méthodologie mise au point dans cette étude, l'informatisation du FEW se déroule en trois phases :

- (1) l'acquisition du texte des 25 volumes, accompagné d'un balisage typographique minimal ;
- (2) le balisage XML complet des informations lexicographiques, de façon totalement automatisée, à l'aide d'un logiciel construit à cet effet ;
- (3) la mise en ligne des articles balisés et leur exploitation via une interface de consultation.

Ces trois phases (développées plus longuement dans Renders à paraître) comportent quelques particularités par rapport à d'autres projets d'informatisation. Pour la première phase, une saisie manuelle du texte a été préférée à une numérisation, cette dernière s'étant avérée peu fructueuse en raison des nombreux caractères spéciaux non reconnus par les logiciels OCR. Une solution de double saisie, déjà utilisée pour d'autres dictionnaires (par exemple le *Deutsche Wörterbuch* des frères Grimm, cf. [dwb.uni-trier.de/de/die-digitale-version](http://dwb.uni-trier.de/de/die-digitale-version)), a été proposée et apportée par le Center for Digital Humanities de Trèves, qui possède une expertise reconnue dans ce domaine (cf. [Kompetenzzentrum.uni-trier.de](http://Kompetenzzentrum.uni-trier.de)).

En ce qui concerne le balisage XML, il a pour particularité d'être pensé selon la perspective de l'utilisateur, contrairement par exemple au balisage du *Trésor de la Langue Française informatisé* qui fut automatisé selon les structures du dictionnaire uniquement (cf. Dendien & Pierrel 2003). Cette particularité le distingue également des dictionnaires conçus dès le départ dans une perspective électronique, ces derniers présentant généralement un balisage pensé pour les besoins de leur rédaction. L'automatisation du balisage et sa vérification s'effectuent à l'Université de Liège. Il est prévu que le balisage inséré soit plus tard converti au standard TEI (cf. [www.tei-c.org](http://www.tei-c.org)). Le résultat du processus est la création d'articles au format XML qui pourront être exploités sous la forme d'une base de données.

La mise en ligne finale des articles informatisés nécessite quant à elle l'affichage de caractères spéciaux non standards (non Unicode) et, de ce fait, la création d'une police de caractères spécifique comprenant la totalité de ceux-ci. Seule la phase de mise en ligne est directement concernée par cette police : les phases précédentes nécessitent certes la reconnaissance de tous ces caractères spéciaux (sous la forme de codes ou d'entités XML), mais pas leur affichage. L'Atelier National de Recherche Typographique ([www.anrt-nancy.fr](http://www.anrt-nancy.fr)) a proposé de créer cette police, tandis que l'ATILF se charge de développer l'interface d'interrogation.

Les trois phases sont successives, mais ne requièrent pas obligatoirement le traitement de la totalité du FEW à chaque étape. L'informatisation peut s'effectuer article par article. Actuellement (juillet 2014), trois des 25 volumes sont en cours de traitement. Ces volumes seront très certainement interrogeables en ligne avant que d'autres volumes ne soient saisis : la première phase est, en effet, la plus coûteuse et dépend donc fortement des financements apportés. En raison de cet obstacle financier à une informatisation rapide, il a parallèlement été décidé de mettre le FEW à la disposition de tous en mode image. Les 25 volumes sont accessibles depuis février 2014 à l'adresse <https://apps.atilf.fr/lecteurFEW>. Une possibilité d'interrogation minimale (par étymons et par lexèmes) de ces images est

prévue, en attendant l'interface d'interrogation complète qui accompagnera la mise en ligne du FEW en mode texte.

### **3 L'utilisation du FEW**

Les difficultés d'utilisation étant la raison principale du projet, il nous semblait évident que l'avis de l'utilisateur était à prendre en compte dès le départ, c'est-à-dire non seulement lors du développement de l'interface de consultation (phase 3), mais aussi dans la définition du balisage à insérer dans le dictionnaire (phases 1 et 2).

Afin de rencontrer au mieux les besoins des utilisateurs, il était d'abord nécessaire de les connaître, suivant le conseil donné par Atkins & Rundell (2008 : 5). Plusieurs questions se posaient, concernant d'abord l'utilité du dictionnaire, ensuite les parcours de consultation actuellement suivis dans la version imprimée du FEW et les problèmes rencontrés. Enfin, il s'agissait de s'interroger sur les parcours à mettre en place dans la version électronique pour répondre à ces problèmes. Les attentes des utilisateurs devaient, rappelons-le, dialoguer avec deux autres contraintes : l'obligation de respecter les structures du dictionnaire – c'est-à-dire le produit lexicographique tel qu'il a été pensé lors de sa rédaction et présenté dans la version imprimée – et la nécessité de pouvoir automatiser le balisage XML. Des demandes inconciliables avec ces contraintes seraient d'emblée soit rejetées, soit revues de façon à élaborer un compromis réaliste.

#### **3.1 Le FEW, son utilité, ses utilisateurs**

Si l'on en croit son titre, le FEW est un dictionnaire étymologique du français, ce qui pourrait faire croire qu'il est essentiellement utilisé pour connaître l'étymon des lexèmes de la langue française. En réalité, le titre de l'ouvrage est réducteur (cf. Büchi & Chambon 1995 : 947-948). D'une part, l'étymologie-histoire pratiquée par le FEW le mène à donner davantage d'informations que les autres dictionnaires étymologiques. Le FEW présente en effet une étymologie intégrante (cf. Malkiel 1976), c'est-à-dire que l'information étymologique représente le critère organisateur des données. La conséquence est une structure complexe à plusieurs niveaux (super-, macro-, micro- et infrastructure, cf. Büchi 1996 : 5-6). D'autre part, le domaine couvert par le FEW dépasse la langue française pour embrasser de façon presque exhaustive la totalité des lexèmes du domaine galloroman. Il s'agit donc d'un ouvrage de référence pour tous les parlers et dialectes concernés.

L'étendue du domaine linguistique pris en compte explique que chaque lexème soit associé dans le FEW à une étiquette géolinguistique, qui précise l'état de langue (ancien français, moyen français, français moderne ; ancien gascon etc.) ou le dialecte (lorrain, champenois, picard etc.) auquel il appartient. Des références bibliographiques complètent et précisent la chronologie suggérée par l'information géolinguistique. Le FEW sert donc prioritairement à étymologiser, localiser et dater un lexème



dans un sous-domaine linguistique, même si d'autres utilisations sont possibles, par exemple pour connaître le sens d'un mot, sa graphie, sa forme phonique ou sa catégorie grammaticale.

La nature des données contenues dans le FEW, à savoir le lexique des parlers du domaine galloroman, explique qu'il soit utilisé en linguistique historique, dans les études concernant le lexique du français et des autres langues ou dialectes du domaine galloroman. Il est systématiquement utilisé par les étymologistes des autres langues romanes, ainsi que des langues non romanes. De manière générale, le FEW constitue une référence pour l'étude historique de toute langue qui a été en contact étroit avec le français. Mais pour incontournable qu'elle soit, cette référence est toujours un simple outil au service du chercheur, dont l'objet d'étude n'est évidemment jamais le FEW en lui-même. Ainsi, dialectologues, philologues, éditeurs, lexicographes consulteront avidement le FEW, mais pour mieux construire leur objet propre — nous y reviendrons. Diverses catégories d'utilisateurs consultent donc l'ouvrage, avec des besoins variés et avec des ressources différentes face à la complexité du discours lexicographique. Ce sont majoritairement des spécialistes en leur domaine, mais une partie est constituée par les étudiants et par un public d'amateurs. Tous, y compris les spécialistes, rencontrent des difficultés de consultation et souhaitent une informatisation rapide de l'ouvrage.

### 3.2 Parcours de consultation et de lecture

L'avis des utilisateurs a pu être recueilli de diverses manières, d'abord via les publications scientifiques des disciplines concernées (voir par exemple Rey 1971 : 103-104 ; Roques 1991 : 94 ou encore Wooldridge 1998 : 211 ; cf. Renders 2011 : 8-15), ensuite via la diffusion d'un questionnaire au sein de la communauté internationale des chercheurs en linguistique française et romane lors du *XXVe Congrès International de Linguistique et Philologie Romanes* (cf. Renders 2010 ; pour les résultats, Renders 2011) et, enfin, via de nombreuses rencontres individuelles. Il a ainsi été possible d'obtenir un aperçu des pratiques actuelles d'utilisation du FEW et, dans un second temps, de connaître les souhaits des utilisateurs dans l'optique d'un FEW informatisé, souhaits en relation étroite avec les difficultés qu'ils rencontrent dans la consultation de la version imprimée. Les attentes exprimées sont révélatrices de la façon dont les utilisateurs voudraient exploiter le FEW et, plus généralement, de la façon dont ils perçoivent l'ouvrage et ses structures.

L'analyse des comportements d'utilisation du FEW a distingué deux activités distinctes et successives lors de l'utilisation du FEW : d'une part, la consultation, opération consistant à repérer dans le dictionnaire l'endroit où se trouve l'information que l'on recherche ; d'autre part, la lecture, opération consistant à s'approprier de façon complète l'information recherchée ainsi que l'analyse qu'offre le FEW en rapport avec cette information. Pour chacune de ces deux opérations, l'étude a montré des divergences entre les parcours actuels, permis par la version imprimée, et les parcours souhaités dans une version électronique.

En ce qui concerne la consultation, les souhaits des utilisateurs d'un futur FEW numérique induisent des itinéraires totalement nouveaux par rapport aux itinéraires traditionnels. En effet, dans la version

imprimée, les points d'entrée dans le dictionnaire sont réduits aux étymons-vedettes (à condition de connaître la langue de l'étymon, qui détermine la partie superstructurelle et donc le volume à consulter) et aux lexèmes (à condition soit d'avoir une idée de leur étymon, soit de trouver ces lexème – ou des lexèmes apparentés – dans les index du FEW qui ne sont nullement exhaustifs : cf. ATILF 2003, qui remplace les divers index situés en fin de volume). En pratique, l'utilisation du FEW s'apparente souvent à un jeu de piste.<sup>1</sup> Dans la perspective d'une version électronique, les points d'entrée ne seraient toutefois plus réduits aux seuls lexèmes et lemmes (qui deviendraient en outre plus facilement repérables), mais s'étendraient fructueusement à tout type d'information présent dans le discours lexicographique (étiquettes géolinguistiques, sources bibliographiques, dates etc.). Ce mode de consultation « transversale », qui mène à plusieurs endroits dans le dictionnaire, est impossible dans la version papier du FEW, mais très attendu dans l'optique de son informatisation.

En ce qui concerne la lecture, l'étude a mis en évidence la complexité des itinéraires traditionnels, due à la nécessité, pour s'approprier l'analyse approfondie des données fournies par l'ouvrage, de mettre ces données en relation et en contexte à plusieurs niveaux. Par ailleurs, un aller-retour est constamment requis entre le dictionnaire et son *Complément*, qui explicite les nombreuses abréviations géolinguistiques et bibliographiques propres au FEW. Il est intéressant de constater que ces parcours présentent des variantes selon l'expérience qu'a l'utilisateur des structures du dictionnaire, selon ses compétences en linguistique française et dialectale et selon son besoin d'explicitation des abréviations. Les difficultés d'accès de l'ouvrage expliquent que de nombreux souhaits soient émis dans l'optique d'une informatisation, tels que la résolution des nombreuses abréviations, l'explicitation des sigles bibliographiques et des sources, la traduction des termes allemands ou, encore, la mise en évidence du plan des articles longs. La plupart de ces besoins se résolvent par des mises en relation (avec le commentaire de l'article, avec le *Complément*, avec des outils externes) qui, certes, sont possibles dans la version imprimée du FEW, mais seraient grandement facilitées par une informatisation de son contenu. Il ne s'agit donc pas de modifier les itinéraires de lecture classiques de la version imprimée, mais de faciliter la mise en relation de données qui, dans le discours lexicographique, ne sont pas situées côte à côte. Ce faisant, on ouvre la voie à des parcours de lecture hypertextuels qui n'étaient pas identifiés comme tels dans l'analyse des comportements d'utilisation du FEW papier, mais qui étaient sous-jacents.

### 3.3 Deux visions du FEW

L'analyse des parcours effectifs et des parcours souhaités par les utilisateurs a fait apparaître deux modes a priori contradictoires de consultation et de lecture du dictionnaire. Les difficultés de lecture s'expliquent en effet par la vision de l'article du FEW comme un discours construit et structuré, dans lequel chaque information est à mettre en relation avec celles qui l'entourent. Rappelons que les ar-

---

1 Il faut ajouter à ces difficultés d'accès le problème des classements multiples, cf. Baldinger 1980.

ticles du FEW classent et hiérarchisent les données différemment, de façon à retracer l'histoire particulière de chaque famille lexicale, ce qui fait de chaque article une monographie à part entière :

L'ouvrage se présente, en fait, comme un ensemble structuré de monographies, dont la forme lexicographique n'est qu'un auxiliaire au service de la « visée globalisante » (Swiggers 1990 : 347) de Wartburg, qui l'anime et la domine. (Büchi & Chambon 1995 : 952)

Cette particularité explique que les articles du FEW se lisent davantage qu'ils ne se consultent. Il s'agit de ce que nous appelons la *dimension monographique* (ou dimension M) du FEW, qui n'est accessible que par l'opération de lecture. Les demandes, partagées par les utilisateurs, d'un plan de l'article et d'une traduction du commentaire sont tout à fait liées à cette dimension monographique du FEW : elles visent à atteindre plus aisément le classement des données et l'analyse qui en découle.

L'enthousiasme des utilisateurs pour l'informatisation du FEW est toutefois davantage à expliquer par une autre vision de l'ouvrage, que nous appelons la *dimension thesaurus* (ou dimension T) du FEW. Dans cette vision du FEW comme un thesaurus, les utilisateurs sont intéressés par la masse de données qui s'y trouve et par les informations qui sont associées à chaque lexème. C'est le lexème, et non plus l'article, qui constitue leur centre d'intérêt. Le FEW est en effet le seul dictionnaire où se trouvent rassemblés tous les lexèmes des langues et dialectes du domaine galloroman, ce qui en fait un ouvrage des plus précieux dans de nombreuses sous-disciplines linguistiques. Cette dimension thesaurus est à l'œuvre lorsque l'utilisateur imagine des modes de consultation transversale, qui permettraient d'accéder directement à un groupe de lexèmes partageant un point commun malgré leur dispersion lexicographique due à leur appartenance à des familles lexicales différentes. Les besoins de mise en relation avec le *Complément* et avec d'autres dictionnaires, qui permettraient d'accéder directement à l'intégralité des références bibliographiques associées à un lexème, font partie de cette vision du FEW comme thesaurus. Enfin, c'est cette dimension qui explique que le FEW se consulte généralement à partir des lexèmes, alors que les entrées de la nomenclature sont des étymons.

Les deux dimensions dégagées ci-dessus ne sont pas inconciliables, mais étroitement imbriquées et complémentaires dans la construction du discours lexicographique. Si la consultation gagne à être envisagée dans une dimension thesaurus où chaque lexème est individualisé et accessible séparément, la lecture, quant à elle, ne peut s'effectuer que dans une dimension monographique, c'est-à-dire dans une mise en relation et une contextualisation des données.

Ces besoins, exprimés par les utilisateurs, ne sont pas résolus par les itinéraires de consultation et de lecture permis par le discours lexicographique sous sa forme actuelle. Les difficultés d'accès aux données du FEW ont toujours été attribuées à la présentation condensée et hautement structurée du discours lexicographique ; l'étude montre que ces difficultés proviennent également, si pas davantage, du fait que les utilisateurs veulent consulter le FEW dans une optique plus large que celle pour laquelle il a été conçu.

## 4 La prise en compte de l'utilisateur dans la version électronique

Partant de ce constat, il s'est avéré essentiel de résoudre l'inadéquation entre la conception initiale du dictionnaire et l'utilisation qui en est souhaitée, à la fois en tant que recueil de monographies et en tant que thesaurus. En dimension M, il s'agit de faciliter la lecture de l'article en tant qu'ensemble structuré et autonome et d'optimiser les itinéraires de mise en relation entre les informations (à la fois au sein d'un article et hors article). En dimension T, il s'avère nécessaire d'ouvrir l'accès aux nouveaux itinéraires de consultation attendus par la communauté (consultation transversale). Enfin, les deux dimensions sont concernées par la nécessité de permettre la mise à jour du FEW (intégration des ajouts et corrections apportés ailleurs) et sa mise en réseau avec d'autres ressources lexicographiques. Ces nouveaux parcours doivent être pris en compte non seulement au moment de créer l'interface de consultation, mais aussi dans la modélisation préalable et la formalisation XML du discours lexicographique.

### 4.1 Dimension monographique : itinéraires de lecture

Rétablir l'autonomie de l'article en dimension M nécessite tout d'abord de permettre son extraction hors du volume imprimé, tout en conservant les informations qui étaient données par le contexte physique et la situation de l'article au sein de l'ensemble. Rappelons que le FEW possède une superstructure divisant l'ouvrage en fonction de la langue d'origine des lexèmes et que les étymons-vedettes ne sont dès lors pas nécessairement partout classés par ordre alphabétique. La solution informatique consiste à « redescendre » au niveau de l'article les informations données dans les niveaux supérieurs (essentiellement le numéro de volume ainsi que la page où débute l'article). En pratique, ces informations sont automatiquement explicitées au début de l'article, dans les attributs des balises XML identifiant l'article et la colonne. Par exemple, la version XML de l'article MASCŪLĪNUS (FEW 6/1, 424b) commence ainsi :<sup>2</sup>

```
<art book="1" ici="1" id="0" lang="german" type="doc-com" volume="6">
<col pg="424" s="b"/>
```

Afin de faciliter la lecture de l'article en tant qu'ensemble structuré, il a été décidé de proposer au lecteur un plan résumant cette structure. Ce résumé ne consiste pas en une réécriture (impossible à automatiser et donc inconcevable, conformément au respect des trois contraintes précitées), mais, plus simplement, en l'affichage automatique de la première unité lexicale de chaque paragraphe, précédée du marquage alphanumérique situant le paragraphe dans la numérotation microstructurale de l'article (réexplicitée si nécessaire) et, lorsqu'il existe, du marqueur textuel (titre de section) explicitant le

2 L'attribut *ici* (*in-column index*) indique l'ordre de l'article dans la colonne, *lang* la métalangue utilisée dans le commentaire; l'attribut *type* indique si l'article est divisé en une partie documentaire (*doc*) et une partie de commentaire (*com*) (cf. Büchi 1996 : 78).

critère de regroupement des lexèmes au sein du paragraphe. Le plan de l'article CHOCOLATL (FEW 20, 63b-64a) se présente par exemple ainsi dans sa version XML :

```
<!--article map
```

```
1 Mfr. chocholate m. „breuvage fait avec des amandes de cacao“ (1598)
```

```
1 Ablt. — Nfr. chocolatière f. „vase où l'on prépare, où l'on sert le chocolat en boisson“ (seit 1680)
```

```
2 Nfr. chicolate f. „chocolat“ (1658)
```

```
-->
```

Ce plan, affiché en tête d'article, donne immédiatement au lecteur une vision synthétique de la structure de l'article (classement des lexèmes au sein des différentes subdivisions) : ici, trois paragraphes structurés en deux parties numérotées, la première partie reprenant des dérivés (*Ablt.* pour *ableitung*). Ce faisant, il aide à saisir d'un coup d'oeil l'organisation générale de la famille lexicale traitée.

Enfin, les itinéraires de lecture sont également facilités par la création de liens hypertextuels internes et externes. Les liens internes concernent d'une part les notes et appels de note (identifiés et associés de façon à faciliter le passage de l'un à l'autre), d'autre part les références renvoyant, dans le commentaire, au marquage alphanumérique structurant les matériaux. Par exemple, toujours dans l'article CHOCOLATL, le commentaire explique l'origine étymologique de chacune des deux parties numérotées. Tant les marqueurs alphanumériques que les références à cette numérotation (*<pref>*) ont été automatiquement reconnus et balisés, ce qui permettra la création de liens hypertextuels :

```
<pref id="1">1</pref> ist aus <lang>sp.</lang> <form><i>chocolate</i></form> entlehnt [...]. Unklar ist auch das verhältnis von <pref id="2">2</pref> zu <pref id="1">1</pref>. Der erste beleg von <pref id="2">2</pref> kommt von den kleinen Antillen, wodurch wahrscheinlich gemacht wird, dass diese form sich selbständig verbreitet hat.
```

Les liens externes concernent d'une part les renvois à d'autres articles du FEW (mis en oeuvre par le balisage automatique des étymons, volumes et pages), d'autre part les renvois externes au dictionnaire. Le balisage des sigles bibliographiques permet en effet de créer un lien hypertextuel vers leur explicitation (fournie dans la base de données contenant le *Complément* au FEW) et vers la ressource électronique si cette dernière existe.

Ces trois nouveautés apportées par la version électronique (autonomie de l'article, résumé de sa structure, liens hypertextuels) permettent de faciliter et d'optimiser les parcours de lecture du FEW dans sa dimension monographique.

## 4.2 Dimension thesaurus : itinéraires de consultation

En dimension T, il s'agit en priorité de permettre une consultation du FEW via les lexèmes (unités lexicales) et les informations qui y sont associées. L'autonomie de chaque unité lexicale est rétablie en balisant au sein d'un même élément XML les informations qui la composent (étiquette géolinguistique, signifiant, catégorie grammaticale, définition, références) et en rétablissant celles de ces informations qui seraient implicites, cas de figure très fréquent dans le FEW puisque certaines informa-

tions (étiquette géolinguistique, catégorie grammaticale, signifiant, définition) ne sont jamais répétées si elles ont déjà été citées dans l'unité précédente (cf. Büchi 1996 : 117 et Renders 2011 : 76-81). Le principe de « redescende » des informations s'applique donc également ici. Dans l'article ACCUSATIVUS par exemple (FEW 24, 94b), qui comporte uniquement deux lexèmes *accusatif*, l'un substantif masculin (« cas auquel on met le complément direct »), l'autre adjectif (« qui concerne l'accusatif »), l'étiquette géolinguistique et le signifiant associés au deuxième lexème n'ont pas été répétées. La version XML rétablit ces informations grâce à l'insertion automatique d'une balise <imp>, identifiant le type d'information manquant et son contenu implicite :

```
<unit><imp contents="Mfr." type="geoling"/><imp contents="accusatif" type="form"/><gram>adj.</gram> <def>„qui concerne l'accusatif“</def>
```

Ce balisage permet d'extraire de chaque article toutes les unités lexicales qu'il contient et de leur rendre leur autonomie, indépendamment de leur insertion dans le discours monographique. Une liste de toutes les unités est systématiquement créée pour chaque article ; cette liste attribue en outre à chaque lexème son étymon (étymon-vedette de l'article) et sa référence FEW. Toujours pour l'article ACCUSATIVUS, la liste créée est la suivante :

```
<fiche etymon="accusativus" lang="Mfr." lang="nfr." forme="accusatif" gram="m." def="„cas auquel on met le complément direct““ ref="(seit ca. 1170, EdConf, FrMod 21, 217)" N="FEW 24/1, 94b, ici 1, §1, u1"></fiche>
```

```
<fiche etymon="accusativus" lang="(imp.) Mfr." forme="(imp.) accusatif" gram="adj." def="„qui concerne l'accusatif““ ref="(1380, Aalma 98; Pom 1671-1700; Lar 1866-1948)" N="FEW 24/1, 94b, ici 1, §1, u2"></fiche>
```

Le balisage rend ainsi possible une consultation transversale du FEW via les informations associées aux lexèmes : consultation par parler (étiquette géolinguistique), par catégorie grammaticale, par élément de définition, par signifiant, par référence bibliographique, par étymon. La consultation par critères chronologiques nécessite en outre que les dates correspondant à chaque sigle bibliographique soient explicitées, ce qui s'effectue via le *Complément* au FEW.

Enfin, des consultations à cheval entre la dimension monographique et la dimension thesaurus, à savoir selon la langue d'origine des lexèmes (souvent implicite dans le FEW, car associée à la partie superstructurelle où se trouve l'article) ou selon le type de descendance (emprunts ou lexèmes héréditaires) sont également rendues possibles par le balisage.

### 4.3 Sortir du FEW

Comme évoqué au point 3.1., les utilisateurs du FEW sont divers, de même que leurs objets d'étude. Le parcours dialectique entre le FEW et ces diverses entreprises peut être illustré par un exemple qui nous est familier, celui de l'*Atlas linguistique de la Wallonie* (ALW). Tout au long de l'analyse des matériaux de l'ALW, les rédacteurs tissent un dialogue entre leur ouvrage et le formidable outil qu'est le FEW

: d'une part, en extrayant de celui-ci les données servant à éclairer les matériaux wallons, d'autre part, en amenant divers compléments, amendements ou ajustements. Au fil de ce cheminement se développe « une véritable réévaluation de l'état de l'art représenté par le FEW » (Chauveau & Buchi 2011 : 12). Si les spécialistes s'accordent à reconnaître un certain intérêt aux compléments apportés par l'ALW au dictionnaire de Wartburg, ceux-ci sont de toute évidence difficilement accessibles. Cette conclusion peut être étendue à d'autres ouvrages, qu'ils soient atlantographiques, lexicographiques ou philologiques. Dès lors, il nous semble que prendre en compte l'utilisateur du FEW passe également par une intégration, à quelque niveau que ce soit, de ses apports à la construction d'un savoir commun. Le balisage de l'œuvre selon les modalités exposées ci-dessus permettra une navigation optimisée et des consultations transversales internes. Ce balisage pourrait ensuite être exploité pour développer des liens externes vers divers projets apparentés, tels que l'ALW pour des apports ponctuels ou des entreprises telles que DEAF, Godefroy, TLFi etc. pour des apports plus systématiques. Ces divers ouvrages intégrant déjà dans leur programme lexicographique des références au FEW, leur mise en réseau nécessite uniquement que le FEW informatisé dispose d'url pérennes qui puissent servir à la fois de référence et de lien hypertextuel. Toutefois, l'intégration des apports externes pourrait aller plus loin qu'une simple mise en réseau et conduire à un FEW évolutif. Cette idée encore utopique est en phase avec la dimension T du FEW, puisqu'elle permettrait des consultations basées sur des critères (géolinguistiques, chronologiques, étymologiques ou autres) qui correspondraient à l'état actuel de la science et non à une version périmée. Le balisage du FEW en dimension T, et plus particulièrement le rétablissement de l'autonomie des unités lexicales, devrait permettre ces consultations sans que la dimension M du FEW n'en soit affectée.

## 5 Conclusion

Loin de constituer une modélisation théorique et gratuite du dictionnaire, les deux dimensions décelées dans la conception et l'utilisation effective du FEW d'une part, mais également dans l'exploitation que rêvent d'en faire ses utilisateurs, ont très concrètement guidé la réflexion dans le cadre du projet. Nous espérons avoir montré comment ces deux dimensions s'articulent et devront continuer à le faire, de la façon la plus explicite possible, dans la future version électronique.

En outre, nous avons montré en quoi les fonctionnalités permettant l'exploitation du dictionnaire dans ses deux dimensions ne concernaient pas seulement la création de l'interface utilisateur, mais devaient être prévues dès la phase de modélisation du discours lexicographique. Aborder les structures du dictionnaire selon le point de vue de l'utilisateur est la condition nécessaire pour pouvoir rendre compte de la façon dont le FEW est, non plus conçu et rédigé, mais perçu et utilisé.

Bien entendu, c'est au niveau de l'interface, toujours en développement à l'heure où nous écrivons ces lignes, que l'utilisateur prendra pleinement conscience des potentialités des nouveaux parcours qui



lui sont offerts. Gageons cependant que cet avant-goût aiguisera encore davantage l'appétit des chercheurs!

## 6 Bibliographie

- ALW = Remacle, L. et al. (1953-). Atlas linguistique de la Wallonie. Tableau géographique des parlers de la Belgique romane d'après l'enquête de Jean Haust et des enquêtes complémentaires. Liège: Vaillant-Carmagne.
- ATILF (2003). *Französisches Etymologisches Wörterbuch. Index A-Z*. Paris: Champion.
- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Baldinger, K. (1980). Etymologies doubles dans le FEW. In H. J. Izzo (éd) *Italic and Romance: Linguistic Studies in Honor of Ernst Pulgram*. Amsterdam: Benjamin, pp. 189-194.
- Büchi, E., Chambon, J.-P. (1995). Un des plus beaux monuments des sciences du langage : le FEW de Walter von Wartburg (1910-1940). In G. Antoine, R. Martin (éd.) *Histoire de la langue française, 1914-1945*. Paris: CNRS Editions, pp. 935-963.
- Büchi, E. (1996). *Les structures du Französisches Etymologisches Wörterbuch. Recherches métalexigraphiques et métalexicologiques*. Tübingen: Niemeyer.
- Chauveau, J.-P., Buchi, E. (2011). État et perspectives de la lexicographie historique du français. In *Lexicographica. International Annual for Lexicography*, 27, pp. 101-122.
- Complément = Chauveau, J.-P., Greub, Y. & Seidl, C. (2010). *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. Supplement zur 2. Auflage des Bibliographischen Beiheftes*. Bâle: Zbinden.
- DEAF = Baldinger, K. et al. (1974- ). *Dictionnaire Étymologique de l'Ancien Français*. Québec/Tübingen/Paris: Presses de l'Université Laval/Niemeyer/Klincksieck. Site internet : <http://deaf-server.adw.uni-heidelberg.de>.
- Dendien, J., Pierrel, J.-M. (2003). Le trésor de la langue française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence. In *Traitement automatique des langues (TAL)*, 43 (2), pp. 11-37.
- FEW = von Wartburg, W. et al. (1922-2002). *Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes* (25 vol.). Bonn/Heidelberg/Leipzig-Berlin/Bâle: Klopp/Winter/Teubner/Zbinden.
- Godefroy = Godefroy, F. (1881-1895). *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle* (8 vol.). Paris: Vieweg.
- Malkiel, Y. (1976). *Etymological dictionaries. A tentative typology*. Chicago-London: The University of Chicago Press.
- Renders, P. (2010). L'informatisation du *Französisches Etymologisches Wörterbuch* : quels objectifs, quelles possibilités ? In M. Iliescu, H. Siller-Runggaldier & P. Danler (eds.) *Actes du XXVe Congrès International de Linguistique et de Philologie Romanes* (Innsbruck, 3-8 septembre 2007). Berlin/New York: De Gruyter, vol. 6, pp. 311-320.
- Renders, P. (2011). Modélisation d'un discours étymologique. Prolégomènes à l'informatisation du *Französisches Etymologisches Wörterbuch*. PhD Thesis. Université de Liège, Liège, BE.
- Renders, P. (à paraître). Mise en ligne, mise à jour et mise en réseau du *Französisches Etymologisches Wörterbuch*. In D. Trotter, A. Bozzi & C. Fairon (éds.) *Actes du XXVIIe Congrès international de linguistique et de philologie romanes* (Nancy, 15-20 juillet 2013). Nancy: ATILF.
- Rey, A. (1971). Le dictionnaire étymologique de W. von Wartburg : structures d'une description diachronique du lexique. In *Langue française*, 10, pp. 83-106.



- Roques, G. (1991). L'articulation entre étymologie et histoire de la langue. In *Travaux de linguistique*, 23, pp. 91-95.
- Swiggers, P. (1990). Lumières épistolaires sur l'histoire du F.E.W. : Lettres de Walther von Wartburg à Hugo Schuchardt. In *Revue de linguistique romane*, 54, pp. 347-358.
- TLFi = CNRS/Université Nancy2/ATILF (2004). *Trésor de la langue française informatisé* (cédérom). Paris: CNRS Éditions (site internet : <http://www.cnrtl.fr/definition>).
- Wooldridge, T. R. (1990). Le FEW et les deux millions de mots d'Estienne-Nicot : deux visages du lexique français. In *Travaux de linguistique et de philologie*, 28, pp. 239-316.
- Wooldridge, T. R. (1998). Le lexique français du XVIe siècle dans le GDFL et le FEW. In *Zeitschrift für romanische Philologie*, 114, pp. 210-257.



# A Morphological Historical Root Dictionary for Portuguese

João Paulo Silvestre, Alina Villalva  
Centro de Linguística da Universidade de Lisboa  
jpsilvestre@fl.ul.pt, alinavillalva@campus.ul.pt

## Abstract

The project of a Morphological Historical Root Dictionary (MHRD) for Portuguese aims to build a specialized dictionary containing a critical selection of lexical units: the first stage is devoted to adjectives, namely those that can be found in Figueiredo (1913). Its main goals are the clarification of morphological and semantic issues in the evolution of the lexicon, and the assessment of the communicative adequacy of the words that are registered in current Portuguese dictionaries, particularly by signaling unused or seldom used words.

The methodology we established is three-sided: it first relies on the lexical analysis of the selected words; then, it seeks for lexicographic information in old Portuguese dictionaries (16<sup>th</sup> to 19<sup>th</sup> centuries), in Portuguese textual corpora and in etymological dictionaries; finally, it contrasts the Portuguese data with data from other romance languages, searching for lexical and semantic loans.

To conclude, we propose a prototype dictionary entry, applying the above-described methodology to the adjective *bravo*.

**Keywords:** word roots; morphology; Portuguese; historical lexicography

## 1 Preliminary remarks

The entry list of most contemporary Portuguese dictionaries still echoes lexicographic approaches that most of the other European languages have discarded along the past century. The main problem resides in the fact that they give room to the vast majority of the contents of previous dictionaries, thus accumulating, with an identical status, words that make part of the contemporary lexicon and a huge amount of unused words. This abundance of entries renders other problems usually found in dictionaries, such as graphic alternates, inadequate meanings and wrong etymological information, to name a few.

Furthermore, in the last two decades, mainstream dictionary publishing houses have rendered their workforce to the un compelling issue of the orthographic ‘entente’ between Portugal, Brazil and (eventually) all other Portuguese-speaking countries, which motivated ‘new’, ‘updated’ paper editions. The remaining workforce of dictionary companies is fully devoted to the introduction of ‘neologisms’ that will haunt future editions.

There is an urge, then, to review the wordlists, to clean each entry, to correct the information given, to expunge old unused words. A dictionary may, of course, be cumulative about the selection of entries, but it must mark those that are unused, although they can be found in old literary texts.

## 2 Diachronic incoherence in Portuguese contemporary lexicography

Take, for instance, the case of the verb *abundar* ‘to abound’. *Infopedia* is an online dictionary that claims to be the most complete dictionary of European Portuguese, covering general, technical and scientific vocabularies. Surprisingly, it registers words like *bondar*, *abondar* and *avondar* that are certainly not part of those vocabularies:

- (1) Avondar. Verbo transitivo e intransitivo. Ver abundar. Do latim abundāre, «abundar» (*Infopédia*)
- (2) Abondar. Verbo intransitivo. Regionalismo. Ser suficiente; bastar; bondar (Do latim abundāre, «idem» (*Infopédia*))
- (3) Bondar. Verbo intransitivo. Popular. Bastar; ser suficiente (Do latim abundāre, «trasbordar; abundar» (*Infopédia*))

Probably, the form *avondar* is the oldest in Portuguese – it can be found in 14<sup>th</sup> to 17<sup>th</sup> century textual sources; *abondar* can also be found between the 14<sup>th</sup> century and the 19<sup>th</sup>; *abundar* starts in the 16<sup>th</sup> century and is the only form in contemporary usage<sup>1</sup>. Notice that the third form, i.e. *bondar*, marked in *Infopedia* as a ‘popular’ form has very few registers in the *Corpus do Português* database<sup>2</sup>. In fact, it has only two, one of which comes from an oral corpus, and it is quite difficult to understand.

A search in non-contemporary lexicographic sources helps to consolidate the hypothesis above: *avondar* occurs in 16<sup>th</sup> and 17<sup>th</sup> century dictionaries, already marked as peripheral; *abondar* occurs in the 16<sup>th</sup> century, and in an 18<sup>th</sup> century it is considered as an error. The contemporary form, i.e. *abundar*, which is graphically closer to the spelling of the Latin verb (i.e. *abundare*), appears in the 17<sup>th</sup> century. Curiously, the recovery of all these variants began in 19<sup>th</sup> century dictionaries, such as Morais and Figueiredo. *Infopedia* replicates them, particularly Figueiredo.

- (4) Auondar. Vide Abondar (Cardoso 1569)
- (5) Avondar. Vide Abundar (Pereira 1697)
- (6) Abundar. Ter abundancia. Erro, Abondar (Feijó 1734)
- (7) Avondar, n. Abastar, ser bastante em numero. Antiq. [antiquate] (Silva 1831)
- (8) Bondar v. i. Prov. Sêr bastante, sufficiente: mas isso não bonda. Alter. de abundar. (Figueiredo 1899)<sup>3</sup>

---

1 Examples were taken from the database *Corpus do Português*.

2 Only 7 matches found in texts, all from the 19<sup>th</sup> century (*Corpus do Português*).

3 Examples were taken from the database *Corpus Lexicográfico do Português*.

Filtering a general dictionary such as *Infopedia* is a desirable, but difficult task that requires highly trained manpower. This task is out of range for individual good will and low budget.

As we mentioned before, general contemporary Portuguese dictionaries either accumulate information from previous dictionaries, regardless of the errors they perpetuate and the real usage of the words, or they are produced on the basis of modern *corpora*, that integrate limited amounts of data, thus ignoring all the words that are not represented there. Lexical *corpora*, apart from coverage limitations, also frequently lack morphological tagging.

### 3 Planning a specialized historical dictionary

The MHRD project was designed in order to give a constructive response to such a negative perspective in the field of contemporary dictionary making. This project aims to build a specialized dictionary, which will contain a critical selection of the lexicon of Portuguese. Its main goals are:

- the clarification of the process of morphological and semantic evolution of the lexicon;
- the assessment of the communicative adequacy of the words that are registered in current Portuguese dictionaries, mostly by signalling unused or seldom used words.

MHRD will, thus, include simple and complex lexicalized roots, documented in a set of selected early lexicographic sources for Portuguese.

#### 3.1 Lexicographic sources

Revisiting old dictionaries is thus obligatory, but instead of aiming to consider all of them, in the preliminary stage of this project, we decided to make a selection based on an analysis of lexicography in Portugal. The set of dictionaries that form this *canon* was selected for qualitative reasons, since they all played an important role either for the quality of the information they provided or for the normative role that they assumed. The selection, ranging from the 16<sup>th</sup> century to the end of the 19<sup>th</sup> century, includes:

- Jerónimo Cardoso (1569) *Dictionarium Latinolusitanicum / Lusitanicolatinum* — The first printed dictionary with extended word list in Portuguese (about 12 thousand entries). It is a testimony of ancient lexical choices and word forms, prior to the systematic imitation of Latin in neologisms and spelling.
- Bento Pereira (1697) *Prosodia in Vocabularium Bilingue, Latinum, et Lusitanum* — Extensive Latin-Portuguese learners dictionary, which represents the increase of lexical variety in the 17<sup>th</sup> century, by adapting many Latin words into Portuguese. In the corpus, there are about 50 thousand Portuguese word forms.

- Raphael Bluteau (1712-28) *Vocabulário Português e Latino* — The first dictionary with examples from literary Portuguese texts, with more than 40 thousand entries. It makes a systematic collection of terminology and neologisms resulting from loans.
- António Morais Silva (1789) *Diccionario da Língua Portuguesa* — It is the first monolingual dictionary of the Portuguese, with a modern lexicographical technique. We also considered the fourth revised and extended edition of this dictionary, published in 1831. It was an authoritative dictionary throughout the nineteenth century. As a general rule, it notes old or unused words.
- Cândido de Figueiredo (1899) *Novo Dicionário da Língua Portuguesa* — A cumulative dictionary, which collects ancient and modern words without consistently noting their effective use in contemporary language. Served as lexical corpus to dictionaries in the 20<sup>th</sup> century, which reproduced the word list with little critical review.

Finally, in order to ascertain the usage of words in contemporary Portuguese, we consulted the *Corpus de Referência do Português Contemporâneo*. Completion and crosschecking of lexical analysis relies on the consultation of etymological dictionaries (Corominas and Pascual 1991) and historical dictionaries (*Le Trésor de la Langue Française Informatisé* and *El Nuevo Tesoro Lexicográfico de la Lengua Española*).

### 3.2 Root identification

Since feasibility is one of our main concerns, we decided to limit our research to simple (unanalysable) roots. We believe that, once we have achieved to isolate the core set of roots, we will be better equipped to identify and describe derived and compound words. Simple words, those that are projected from single roots, have a supplementary advantage: they are usually old words and old words tend to accumulate or to change meanings. Those moments, which are not easy to detect, are seldom documented.

The identification of the core set of roots is certainly crucial for a better understanding of the Portuguese lexicon, but no existing general or specialised Portuguese dictionary or lexical corpus provides this information. Two specialized dictionaries deserve to be mentioned, however. The first one is the *Dicionário de Raízes e Cognatos* [cf. Goes (1921)]. Apart being based on 19<sup>th</sup> century sources, it mainly deals with the small subset of neoclassical roots. The second one is the most important Portuguese morphological dictionary. It was made in Brazil, by gathering lexical information from unspecialized sources, such as general language dictionaries [cf. Heckler, Back, Massing (1984-1988)]. Both of them offer interesting data, but their consultation also needs extensive critical reading.

Most Portuguese words, irrespectively of their longer or shorter existence in the Portuguese lexicon, come from a relatively stable set of roots, which can be documented in morphologically simple words (as free roots) or in complex words (as free roots in compositional words and as bound roots in lexicalized words). This typology of roots (based on Villalva and Silvestre [in print]) also foresees cases of bound roots in compositional complex words:

Root	Head		Complement		Modifier	
	Free root	Bound root	Of a derivational suffix	Of another root	Of a simple word	Of a complex word
Type 1 e.g. rat-	rat-o 'mouse'		rat-ice 'cunning'	rat-icida 'rat poison'		
Type 2 e.g. gastr-			gástr-ico 'gastric'	gastr-onomia 'gastronomy'		
Type 3 e.g. graf-	graf-ar 'to write'	pluvió-graf-o 'rain gauge recorder'	gráf-ico 'graphic'	graf-ologia 'graphology'		
Type 4 e.g. super-					super-amigo 'super-friend'	super-confortável 'super-comfortable'

**Table 1: Typology of roots.**

Therefore, a specific methodology had to be established. Bearing feasibility in mind, we decided to devote our initial research to adjectives. The approach we decided to take is three-sided: it relies on the lexical analysis of the selected words; it seeks lexicographic information in old Portuguese dictionaries (16<sup>th</sup> to 19<sup>th</sup> centuries) and Portuguese textual *corpora* and in etymological dictionaries; it contrasts the Portuguese data with data from other romance languages, searching for lexical and semantic loans.

The first stage of the project is devoted to adjectives, the second to nouns and the third one to verbs.

### 3.3 Adjectives

The lexical analysis of adjectives considers their morphological, syntactic and semantic properties. Notice that, from a morphological point of view, Portuguese adjectives are not significantly different from nouns, which raises a practical problem for the selection of roots.

Adjectives and nouns, they both require number inflection (cf. Table 2, i) and they both comprise a subset that allows for gender variation (cf. ii) and a subset of invariable forms (cf. iii).

i)	ii)	iii)
casa <sub>Nsingular</sub> 'house' casas <sub>Nplural</sub> 'houses' leve <sub>ADJsingular</sub> 'light' leves <sub>ADJplural</sub> 'light'	gato <sub>Nmasculine</sub> 'male cat' gata <sub>Nfeminine</sub> 'female cat' novo <sub>ADJmasculine</sub> 'new' nova <sub>ADJfeminine</sub> 'new'	casa <sub>Nfeminine</sub> 'house' carro <sub>Nmasculine</sub> 'car' leve <sub>ADJ</sub> 'light'

**Table 2: Adjectives - morphology.**

There are, nevertheless, differences that can be spotted. As far as number is concerned, although its specification is compulsory for nouns and adjectives alike, in nouns it has semantic relevance (singular refers one entity, plural refers more than one), in adjectives it is semantically irrelevant: adjectives have no quantifiable meaning - number inflection is merely relevant for syntactic agreement (*casa*<sub>sin-</sub><sub>gular</sub> *nova*<sub>singular</sub> ‘new house’; *casas*<sub>plural</sub> *novas*<sub>plural</sub> ‘new houses’).

Gender is even more diverse. All nouns have to have a gender value, which is lexically determined, irrespective from their possibility to participate in gender contrasts (cf. *gato* / *gata*; *carro* and *casa*, in table 3, i). In general, animate nouns can participate in gender contrasts either by thematic alternation (cf. i), by a morphological resource (cf. ii) or lexically (cf. iii), but some animate nouns do not, which eventually creates a mismatch between grammatical gender and the gender of the referent (cf. iv):

i)	ii)	iii)	iv)
gato ‘male cat’ gata ‘female cat’ aluno ‘male student’ aluna ‘female student’	galo ‘rooster’ galinha ‘hen’ marquês ‘marquis’ marquesa ‘marchioness’	cavalo ‘horse’ égua ‘mare’ homem ‘man’ mulher ‘woman’	testemunha <sub>feminine</sub> ‘witness (male or female)’ cônjuge <sub>masculine</sub> ‘spouse (husband or wife)’ águia <sub>feminine</sub> ‘eagle (male or female)’ águia-macho <sub>feminine</sub> ‘male eagle’ águia-fêmea <sub>feminine</sub> ‘female eagle’ rinoceronte <sub>masculine</sub> ‘rhino’ rinoceronte-macho <sub>masculine</sub> ‘male rhino’ rinocerente-fêmea <sub>masculine</sub> ‘female rhino’

**Table 3: Adjectives – gender contrast.**

Inanimate nouns are never allowed to participate in gender contrasts. It is possible to find pairs of words that apparently share the same root, although they belong to different thematic classes. They are not in a gender contrast - they are different words (*casa*<sub>feminine</sub> ‘house’; *caso*<sub>masculine</sub> ‘case’).

For adjectives, gender is as irrelevant as number. It may be syntactically important, for agreement, but a great deal of adjectives is invariable, so the syntactic relevance is also questionable – it is probably just a vestige from Latin declension.

Apart from these morphosyntactic properties, adjectives and nouns also share the possibility to undergo evaluative affixation. It is particularly relevant to notice that the most productive suffix (i.e. *-in-**h*{*o, a*}(s)) is equally available, but its semantic effect on nouns differs from its semantic outcome in adjectives. In the first case, it is typically a diminutive or valuative (cf. table 4, i); in the second case its reading is ambiguous – typically, it can either be an attenuative or a superlative (cf. Table 4, ii). On the other hand, the superlative forming suffix (i.e. *-íssimo*) only adjoins to adjective bases, but a large set of unquestionable adjectives are not scalable and thus they do not allow the adjunction of this suffix (cf. Table 4, iii).



i)	ii)	iii)
casinha ‘small house’ carrinho ‘small car’ gatinho ‘small (dear) cat’	novinho ‘pretty/very young’	*casíssima ‘very+house’ novíssimo ‘very+new’ *teatralíssimo ‘very+theatrical’

**Table 4: Adjectives - affixation.**

In sum, morphology can provide some clues to help setting adjectives apart from nouns, but it fails to draw a neat borderline. Syntactic distribution has, thus, to be considered as well, in order to characterise adjectives as a word class; it is also relevant to establish adjective subclasses. The next set of examples comprises a subset of words that can only occur in an adjective position and another subset of nouns that never occur as adjectives (*cabelo<sub>N</sub> fino<sub>ADJ</sub>* ‘thin hair’); a third subset includes words that occur in adjectival and nominal contexts (*professor<sub>N</sub> assistente<sub>ADJ</sub>* ‘assistant professor’; *assistente<sub>N</sub> do produtor<sub>N</sub>* ‘producer’s assistant’):

In order to get a better understanding of adjectives on the basis of syntactic criteria, probably the most relevant are those that concern word order and the possibility to occur in predicative positions as well as in non-predicative positions. Colour adjectives, for instance, can never occur in a prenominal position (*vestido vermelho* ‘dress red’; \**vermelho vestido* ‘dress red’), but other adjectives can (*velho hábito* ‘old habit’), although this is a marked word order, except for ordinal adjectives (*primeiro dia* ‘first day’; \**dia quinto* ‘fifth day’):

The predicative vs. non-predicative distinction also fails to clearly set adjective subclasses: most adjectives can have both distributions (*o vestido vermelho* ‘the red dress’; *o vestido é vermelho* ‘the dress is red’) and small sets of adjectives have exclusive distribution (*o primeiro dia* ‘the first day’; *o vestido é vermelho* ‘the dress is red’).

Finally, we need to consider the semantics of adjectives, which is probably the most difficult aspect to deal with. Several ontologies have been suggested in the literature, but none of them is able to avoid very specific world knowledge constraints.

We will also use a set of criteria (somehow in parallel with morphological and syntactic criteria above considered) that will apply to each form. The first condition concerns gradability, measured on the basis of *-íssimo* affixation and also on the basis of syntactic comparative constructions. This condition allows us to identify three subsets of adjectives: those that respond positively to both tests (cf. table 5, i), those that respond positively just to *-íssimo* (cf. ii)<sup>4</sup> and those that respond negatively to both of them (cf. iii). Notice that this condition has to be tested in a specific syntactic context, since the result is not always the same (cf. iv):

4 This contrast is probably due to the fact that the evaluative suffix is closer to a rhetoric resource than to the setting of a degree.

i)	preço alto ‘high price’ preço altíssimo ‘very+high price’ preço muito alto ‘very high price’
ii)	casa enorme ‘enormous house’ casa enormíssima ‘very+enormous house’ *casa muito enorme ‘very enormous house’
iii)	sais minerais ‘mineral salts’ *sais mineralíssimos ‘very+mineral salts’ *sais muito minerais ‘very mineral salts’
iv)	os cavalos estão cansados ‘the horses are tired’ os cavalos cansados não ganham corridas ‘tired horses don’t win races’ os cavalos estão cansadíssimos ‘the horses are very+tired’ *os cavalos cansadíssimos não ganham corridas ‘very+tired horses don’t win races’

**Table 5: Adjectives – *íssimo* affixation.**

The second condition concerns the possibility to relate an adjective to another adjective, usually by opposition. Rasken and Nirenburg distinguish binary oppositions of non-gradable, complementary antonyms (cf. table 6, i), polar oppositions of gradable antonyms (cf. ii) and multiple non-gradable oppositions (cf. iii):

i)	ii)	iii)
morto vs. vivo ‘dead’ vs ‘alive’	( <i>muito frio</i> ) <i>frio</i> ( <i>pouco frio</i> ) vs. ( <i>pouco quente</i> ) <i>quente</i> ( <i>muito quente</i> ) ‘very cold’ ‘cold’ ‘bit cold’ vs. ‘bit hot’ ‘hot’ ‘very hot’	<i>análise histórica</i> ‘historical analysis’ ..... <i>económica</i> ‘economic’ ... ..... <i>social</i> ‘social’ ... ..... <i>política</i> ‘political’ ...

**Table 6: Adjectives – sense relations.**

Finally, we need to consider the semantic features of the antecedent, since this information is of crucial importance to circumscribe the meaning of adjectives. In particular, it is necessary to identify the value of animacy, humanness, countability, concreteness. Notice, for instance, that the adjective *bravo* (described in some detail in Villalva & Silvestre 2011) has accumulated different meanings, ranging from a negative pole to a positive pole. In the first case, it means ‘ferocious’ if it applies to animals (wild animals), but it will mean ‘angry’ when it gets to be applied to people.

(9) N<sub>[-human]</sub> *bravo* ‘ferocious’ (Cardoso 1569)

(10) N<sub>[+human]</sub> *bravo* ‘courageous’ (Bluteau 1712-1728)

(11) N<sub>[+human]</sub> *bravo* ‘angry’ (Silva 1789)

### 3.4 *-mente* adverbs

The classification of deadjectival adverbs may also bring some problems. Consider the case of *alto*, which is a good representative of a set of adjectives that are used to measure dimension, in its various features (height, length, weight, etc.). This is a word of Latin origin (i.e. *altus*), originally a participle of the verb *alo* ‘to feed’, thus meaning ‘fed, grown’. As an adjective, *altus* had two basic meanings<sup>5</sup>, related to the perception of height (A. Seen from below upwards, *high* and B. Seen from above downwards, *deep*). Both of them allowed a figurative, non-physical interpretation, respectively ‘elevated, distinguished’ and ‘profound’.

The semantic interpretation of the Portuguese adjective *alto* (as well as the Spanish cognate, *alto*, the French *haut* or the Italian *alto*) replicates the semantics of the A. interpretation of the Latin *altus*. Derived nouns, such as *altura* ‘height’ and *alteza* ‘highness’, as well as derived verbs such as *altear* ‘to heighten’ and *enaltecer* ‘to praise’, help to consolidate that conclusion.

If no other reason existed, the adjective *alto* would not deserve much more comment, but that is not the case if we also consider the derived adverb *altamente*. All romance languages share an adverb forming resource which is based in the grammaticalization of a Latin noun (*mens, mentis*) that took place prior to the individuation of Romance languages, thus explaining its vitality still in contemporary strata. In Latin, the ADJ + *mente* sequence had an adverbial usage. Probably, *-mente* adverbs (or adverbs to be) in Romance languages were initially strictly manner adverbs. Contemporary usage is a bit more complex: apart from manner adverbs (e.g. *elegantemente* ‘elegantly’, *injustamente* ‘unfairly’), *-mente* also forms temporal locatives such as those in table 7, generally equivalent to a locution that includes the base adjective (i.e. *anteriormente* = num momento anterior; *imediamente* = num momento imediato; *futuramente* = num momento futuro).

Past	Present	Future
anteriormente	actualmente	brevemente
antigamente	presentemente	futuramente
inicialmente		seguidamente

**Table 7: Adverbs - Temporal locatives.**

Although the formation of *-mente* adverbs is considered as a very productive word formation process, some restrictions have been identified. According to Scalise (1990), Italian *-mente* cannot be attached to several types of adjectives. Unsurprisingly, the same negative constraints apply in Portuguese, so *-mente* cannot apply to possessive (cf. *\*minhamente*), demonstrative (cf. *\*estamente*), indefinite (cf. *\*qualquermente*), or numeral adjectives (cf. *\*umamente*), nor can it be attached to qualifying adjectives denot-

5 <http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0059%3Aentry%3Daltus1>

ing physical qualities (cf. *\*gordamente*), colour adjectives (cf. *\*vermelhamente*), superlatives (cf. *\*melhormente*) and adjectives modified by evaluative suffixes (cf. *\*ligeirinhamente*).

Considering these negative constraints, the adjective *alto* should not yield a *-mente* adverb, but *altamente* is a quite frequently used word. Scalise (1990) considers that polysemic adjectives that refer a physical property and a psychological property can derive a *-mente* adverb from the second meaning<sup>6</sup>. In fact, *altamente* is never related to the physical meaning of *alto* – it never means ‘in a high manner’; but it not related to the psychological meaning of *alto* – it never means ‘in an elevated manner’.

Evidences from ancient dictionaries confirm that there are only occurrences of polysemous meanings of *altamente* and the same applies to other adjectives indicating size or extension, such as *baixo*, *largo*, *estreito*, *comprido*, *longo*, *curto*. Should be noted that in the same sources the adverbs *largamente* and *longamente* are considered as synonyms (cf. 21).

(12) Profundè. aduer. *Alta* & fundamente. (Cardoso 1569)

(13) Tragice loqui Falar *altamente*. como em tragedia. (Cardoso 1569)

(14) Eminenter, adv. Excellentemente, *altamente*. (Pereira 1697)

(15) *Altamente*. Alte. Sublimiter. (Pereira 1697)

(16) *Baixamente*. Abjecte. Ignobilater. (Pereira 1697)

(17) Spatiose, Adv. *Larga*, espaçozamente. (Pereira 1697)

(18) Largiter, Adv. *Larga*, liberal, *abundantemente*. (Pereira 1697)

(19) *Estreitamente*. Anguste. arcte. (Cardoso 1569)

(20) Conjunctissime, adv. superl. Muito junta, & *estreitamente*, muito amigavelmente. (Pereira 1697)

(21) Prolixe, Adv. *Larga*, *longa*, *liberalmente*. (Pereira 1697)

(22) Pleniter. Adv. Plena, cheia, copioza, perfeita, *compridamente*. (Pereira 1697)

(23) Perlonge, Adv. Mui longa, & *compridamente*; muito longe. (Pereira 1697)

(24) *Curtamente*. Timide. (Pereira 1697)

CRPC returns more than 8.000 matches for *altamente* and two conclusions become self-evident: it is always a modifier of an adjective and it is always a quantifier adverb, equivalent to *muito* ‘very’.

(25) situação **altamente** benéfica ‘highly beneficial situation’ (CRPC)

(26) factores **altamente** estimulantes ‘highly stimulating factors’ (CRPC)

In fact, most *-mente* adverbs that occur as adjective modifiers has exactly the same meaning as *altamente*, which demonstrates that the meaning of the adverb is not related to the meaning of the base adjective:

(27) a praia continua **abençoadamente** vazia ‘the beach continues blessedly empty’ (CRPC)

(28) convite tão **abertamente** amoroso ‘such openly loving invitation’ (CRPC)

(29) coisas **abissalmente** diferentes ‘abysmally different things’ (CRPC)

---

6 Scalise (1990) presents the example of *aridamente* that is related to the meaning ‘boring’ and not to the meaning ‘dry’ of the adjective *arido*.

### 3.5 Entry structure: the case of *bravo*

The structure of articles aims to gather information about the roots and derived words, revealing the diachronic sequence and semantic relations. The lexical research that precedes the compilation of an article can be exemplified in previous works about the semantic evolution of the adjective *bravo* (Villalva, Silvestre 2011) and about adjectives that have undergone a severe meaning shift, like *esquisito* (Silvestre, Villalva, in print). In this paper, we propose a prototype entry, applying the data collected on *bravo* (see table 8).

The entry headword is the root. The first category is the information on the simple word, indicating word class, etymology and a summary of documented semantic evolution. Complex words formed directly from the base (such as -mente adverbs) have a separate description. The other groups are the complex adjectives, verbs and names. For each of the words identified, we provide the date of first attestation in the selected dictionaries, as well as the semantic equivalence. Based on the occurrence in lexical corpora (especially CRPC), we finally evaluated the word frequency. The result is an assessment on the contemporary use of words (unused forms are explicitly marked), with which we intend to contribute to the review of dictionary wordlists and to improve word formation descriptions.

<b>root entry</b>		<p><b>BRAV-</b></p> <p><b>BRAVO/A</b>                  Adjetivo variável em género                  G. βαρβαρος; L. BARBARU-; LV ibérico *barbru &gt; *brabu  <i>feroz</i> 1569; <i>valeroso</i> 1712-28; <i>irado</i> 1789</p> <p>↳ <i>bravamente</i> <i>ferozmente</i> 1569; <i>com bravura</i> 1789                  ↳ <i>bravozinho</i> 1643-47</p> <p><i>bravio</i> <i>não cultivado</i> 1643-47; <i>não domesticado, não civilizado</i> 1712-28  <i>bravinho</i> 1569  <i>bravissimo</i> 1789  <i>bravoso</i> = <i>bravo</i> 1789 DES.                  ↳ <i>bravosidade</i> = <i>bravura</i> 1643-47 DES.</p> <p><i>bravaria</i> = <i>bravata</i> 1899 DES.  <i>braveza</i> <i>ferocidade</i> 1569; <i>fúria</i> 1712-28; <i>impetuosidade</i> 1789  <i>bravura</i> <i>ferocidade</i> 1569; <i>bravosidade</i> 1643-47; <i>coragem</i> 1899</p> <p><i>bravejar</i> <i>embravecer-se</i> 1643-47; = <i>esbravejar</i> 1712-28 DES.                  ou <i>bravejar</i> = <i>esbravejar</i> 1899 DES.  <i>desbravar</i> v. int. <i>amansar</i> 1789; <i>arrotear</i> 1899                  ↳ <i>desbravamento</i> 1899  <i>embravear-se</i> = <i>embravecer-se</i> 1789 DES.                  ↳ <i>embraveamento</i> 1569 DES.  <i>embravecer</i> <i>tornar bravo</i> 1569; <i>irritar</i> 1899                  ↳ <i>embravecido</i> 1643-47                  ↳ <i>embravecimento</i> 1643-47 DES.                  ↳ <i>desembravecer</i> <i>amansar</i> 1569 DES.                  ↳ <i>desembravecido</i> 1643-47  <i>esbravecer</i> <i>esbravejar, embravecer</i> 1789 DES.                  ↳ <i>desbravecer</i> = <i>desembravecer</i> 1899 DES.  <i>esbravejar</i> <i>estar furioso</i> 1643-47 <i>gritar</i> 1789                  ou <i>esbravear</i> <i>gritar</i> 1789; = <i>esbravecer</i> 1899</p>	<b>simple word</b>	<p><b>Information on the simple word</b></p> <p><b>G</b> Greek  <b>L</b> Latin  <b>LV</b> Vulgar Latin</p> <p>* hypothetical form                  &gt; diachronic change                  ↳ morphological relationship                  = semantic equivalence                  ou alternate form  <b>DES</b> unused word in contemporary Portuguese</p>
<b>complex words (base=simple word)</b>			<p><b>Lexicographic sources (examples)</b></p> <p>1569 CARDOSO, J., <i>Dictionarium latinolusitanicum</i></p> <p>1643-47 PEREIRA, B., <i>Prosodia in vocabularium bilingue, Latinum, et Lusitanum</i></p> <p>1712-28 BLUTEAU, R., <i>Vocabulario Portuguez e latino</i></p> <p>1789 SILVA, A. M., <i>Diccionario da Lingua Portuguesa</i></p> <p>1899 FIGUEIREDO, C. de <i>Novo Dicionário da Lingua Portuguesa</i></p>	
<b>base=root</b>	<b>complex adjectives</b>			
	<b>complex nouns</b>			
	<b>complex verbs</b>			

Table 8: Root *brav* - Prototype entry.

## 4 References

- Bluteau, R. (1712-1728). *Vocabulário português e latino*. Coimbra-Lisboa, Collegio das Artes da Companhia de Jesu
- Cardoso, J. (1569-1570). *Dictionarium latinolusitanicum & vice versa lusitanico latinum*. Conimbricæ, Joan. Barrerius
- Corominas, J., Pascual, J. A. (1981). *Diccionario crítico etimológico castellano e histórico*. Madrid, Gredos.
- Corpus de Referência do Português Contemporâneo*. Accessed at: <http://www.clul.ul.pt/pt/recursos/183-crpc#cqp>
- Corpus do português: 45 million words, 1300s-1900s*. Accessed at: <http://www.corpusdoportugues.org> [10/03/2014].
- Corpus Lexicográfico do Português*. Accessed at: <http://clp.dlc.ua.pt> [10/03/2014].
- Figueiredo, C. de (1913). *Novo dicionário da língua portuguesa*. Porto, Typ. da Empr. Litter. e Typographyc
- Goes, C. (1921). *Dicionário de raízes e cognatos da língua portuguesa*. Belo Horizonte, Paulo Azevedo & Cia.
- Heckler, E., Back, S., Massing, E. R. (1984-1988). *Dicionário morfológico da língua portuguesa*. São Leopoldo, Unisinos.
- Infopédia. Dicionário da Língua Portuguesa da Porto Editora*. Accessed at: <http://www.infopedia.pt/lingua-portuguesa/> [10/03/2014].
- [10/03/2014].
- Le Trésor de la langue française informatisé*. Accessed at: <http://atilf.atilf.fr/tlf.htm> [10/03/2014].
- Nuevo tesoro lexicográfico de la lengua española*. Accessed at: <http://ntlle.rae.es/ntlle/SrvltGUILoginNtlle> [10/03/2014].
- Pereira, B. (1697). Prosodia in vocabularium bilingue, latinum et lusitanum. Eboræ, Typographia Academiae
- Raskin, V., Nirenburg, S. (1995). *Lexical Semantics of Adjectives: A Microtheory Of Adjectival Meaning*. Memoranda in Computer and Cognitive Science M CCS-95-288. Las Cruces - New Mexico, New Mexico State University.
- Scalise, S. (1990) Constraints on the Italian suffix -mente. In W. U. Dressler (ed.), *Contemporary Morphology*. Berlim, Mouton de Gruyter.
- Silva, A. M. (1789). *Diccionario da lingua portugueza*. Lisboa, na Of. de Simão Thaddeo Ferreira
- Silvestre, J. P., Villalva, A. (in print). Mutations lexicales romanes: esquisito, bizarro et comprido. In *InnTrans: Innsbrucker Beiträge zu Sprache, Kultur und Translation*. Peter Lang Verlag.
- Villalva, A., Silvestre, J. P. (2011). De bravo a brabo e de volta a bravo: evolução semântica, análise morfológica e tratamento lexicográfico de uma família de palavras. In *ReVEL* v. 9, n. 17. Accessed at: [http://www.revel.inf.br/files/artigos/revel\\_17\\_de\\_bravo\\_a\\_brabo.pdf](http://www.revel.inf.br/files/artigos/revel_17_de_bravo_a_brabo.pdf) [10/03/2014].
- Villalva, A., Silvestre, J. P. (in print). *Introdução ao estudo do léxico. Descrição e análise do Português*. Petrópolis-Rio de Janeiro, Vozes.

# **Lexicological Issues of Lexicographical Relevance**





# What can Lexicography Gain from Studies of Loanword Perception and Adaptation?

Mirosław Bańko, Milena Hebal-Jeziarska

Institute of Polish Language, Institute of West and South Slavic Studies, University of Warsaw

m.banko@uw.edu.pl, milena.hebal-jeziarska@uw.edu.pl

## Abstract

In normative dictionaries and usage guides, many loanwords are dismissed as ‘unnecessary’ on the grounds that they have a native synonym, which should be favoured instead. If there is no native equivalent of a loanword in the recipient language, one is often invented in order to eradicate the unwanted loan. However, more detailed studies, in particular, those referred to in this paper, suggest that there is no such thing as fully equivalent words: even if two words have the same designative meaning, they differ in other respects, which makes them mutually unexchangeable except for contexts where the difference between them is inessential.

The goal of this paper is to demonstrate that the meaning potential of lexical loans is different from that of their native synonyms, just because their form is different and differently perceived by language users. The different perception of loanwords can in turn affect their semantic development, thus causing a loanword and its native synonym to diverge. The authors of normative dictionaries and language guides should, therefore, give more consideration to lexical borrowings before they condemn them as ‘unnecessary’ or ‘snobbish’.

**Keywords:** loanwords; purism; synonymy; variance

## 1 Introduction

Lexical loans can be divided into two groups: those which are motivated by nominative needs, i.e. the necessity to name a new object or a new phenomenon of foreign origin, and those which appear for expressive needs, because they seem to introduce certain stylistic values, emotional overtones, etc. Polish *smartfon* and Russian *смартфон*, both coming from English *smartphone*, are examples of the former group, while the English interjection *wow*, now used in many other languages, can exemplify the latter group. There are obviously more ways to fill lexical gaps, as the French alternative names for smartphone – *ordiphone* and *téléphone intelligent* – show, but we will not be dealing with them here. Instead, we will restrict ourselves to borrowings proper, i.e. words taken from a donor language with possible alterations in their pronunciation, spelling, morphological and syntactic features, sometimes also in their meaning.

By definition, loanwords borrowed for nominative needs have no commonly known synonyms in the recipient language, not at least at the moment they enter it. Loanwords adopted for expressive needs, again by definition, do have some synonymous words and that puts them in a disadvantageous position whenever purist attitudes or specific concern about the ‘economy’ of language come into play. Many such loanwords are dismissed as ‘unnecessary’, criticized as ‘overused’, pointed out as examples of bad taste and snobbery. Negative assessments of them are expressed in the popular press, scholarly literature, academic textbooks, language guides and dictionaries alike.

It is not our intention to question the efforts of normatively oriented linguists and lexicographers who aim at reducing the number of loanwords in a language. Critical assessment of word borrowing – and of particular borrowings – is something needed, if only because there are readers waiting for such criticism. Dictionaries have to account not only for how words are really used, but also for what the language users think about their correct use. On the other hand, lexicographers should be aware that cases of full equivalence between a loanword and its native synonym are practically non-existent. Even if two words have the same designative meaning, they differ in other respects and evoke different associations in the minds of the language users. Over the years, such associations may stabilize and become part of the designative meaning, thus making the originally synonymous words diverge.

The goal of this paper is not so much to remind us of these relatively simple truths, as to demonstrate that the adaptation of loanwords in the recipient language is guided, at least to a certain extent, by the interplay between their form and meaning. A tendency can be observed to maintain harmony between the form and meaning of loans, which can manifest itself, *inter alia*, in how the original spelling of a loan is assimilated in the recipient language and how its meaning is shaped in relation to its native synonyms. The tendency will be illustrated with a number of examples later on in the paper. Let us begin, however, with examples of purist attitudes from Polish, German and Czech lexicography and linguistics.

## 2 Examples of purist attitudes

In a textbook for students of Polish language and literature, Andrzej Markowski (2005), a prominent linguist, chairman of the Council of the Polish Language, gives long tables of ‘overused words’ and ‘vogue words’, usually of foreign origin, and demonstrates, by means of invented examples, how they can be replaced by other words, most of them native or borrowed so long ago that their foreign origin is no longer recognizable. The same or similar loanwords were reviled earlier in a standard dictionary of Polish usage, edited by the same author (Markowski 1999), in some of his other books, and in many popular dictionaries and usage guides, compiled by others. It is worth stressing that Markowski’s position is far from extreme purism. His judgments are of a ‘better/worse’, not ‘yes/no’ type, yet his decisions are clearly not based on detailed analyses of the meaning and use of the particular words he

paired. Had he looked at them more carefully, he would have found distinctions which make the words mutually unexchangeable, except for contexts where the difference between them is inessential.

The attitude to loanwords varies depending on the political and sociolinguistic situation, as well as the normative tradition in particular countries. In Germany, around 300 dictionaries of loanwords were published between 1801 and 1945, around half of them belonging to the class known as *Verdeutschungswörterbücher*, literally 'Germanizing dictionaries'. They were not intended to explain the meaning and illustrate the use of borrowed words, but rather to demonstrate how these could be replaced by native words, some of them specifically invented for this purpose (Lipczuk 2007, 2011). No doubt the regional disintegration of Germany before 1871 favoured the development of national purism, but purist attitudes developed within German society even after the unification of the country, because the rising power of the state created favourable conditions for German nationalism. Also the romantic tradition of treating the language as the embodiment of the spirit of a nation caused many Germans to believe that loanwords posed a threat not only to their language, but also to their national identity.

The example of Germany, where even international words became the object of purifying actions (cf. *Rundfunk* and *Fernsprecher*, coined to replace *Radio* and *Telephon*, respectively), is an extreme one, but similar 'nativizing' dictionaries are known from the lexicographic tradition of many countries. In Poland, which from 1795 to 1918 was partitioned among Russia, Prussia and Austria, the concern about the language was steadily expressed at the time and took on different forms. At one end of the scale was Linde's (1807-1814) six-volume dictionary of the Polish language, based on citations from about 800 sources, an attempt to save the treasures of the language and help the nation to survive the difficult time (Adamska-Sałaciak 2001). At the other end there were a number of much smaller dictionaries and usage guides whose aim and content made them similar to German *Verdeutschungswörterbücher*. Among them, Kortowicz (1891) is a good example, see Leszczyński (2000) and Czesak (2007) for information about his dictionary.

In the history of the Czech language, purist attitudes appeared in the times of Jan Hus and have been present continuously thereafter, up to the present day (Engelhardt 2001: 235). The purist trends were particularly strong at the end of the 19th century and between 1920s and 1940s. At the end of the 19th century purists tried to eliminate words of foreign origin, particularly Latin and Greek internationalisms, as well as German and French loans. A number of neologisms were formed on the basis of native words, but the new coinages often replicated the structure of the words they were supposed to replace (Engelhardt 2001:237). For instance, the Czech lexical innovation *pololetí*, patterned on the German word *Halbjahr* (literally 'half year'), was invented to substitute for the Czech internationalism *semestr* (cf. Latin *semestris* 'six-monthly'). As was often the case, the substitution failed, with both *semestr* and *pololetí* being used in present-day Czech.

At the beginning of the 20th century the purist tendencies in Czech linguistics became even stronger. Linguists aimed to eliminate not only proper loans, but also lexical and syntactic calques. Many usage

guides warning language users against loans of different kinds were published. Purists were particularly eager to identify German loans everywhere, even in native constructions (Král 1917), and they formed bizarre neologisms to replace them. Many of the new coinages had a short life, but some have survived to the present. A good example is the word *rozhlás*, which was introduced with the intention to replace *broadcasting* and *radio* (from the verb *hlásit* ‘report’ and the suffix *roz-*, denoting the spread of something from one place).

After the communists came to power in the Czech state, especially after 1948, English loans were fought most vigorously and replaced with native words or their spelling was changed to conceal their western origin. Many native neologisms were created at that time, e.g. *silostroj* (a compound word of the meaning ‘power and machine’) was introduced to replace *motor*, and *samoňyb* (another compound combining the meanings of ‘itself’ and ‘move’) was intended to take the place of *auto*, cf. Svobodová (2009: 33). Nowadays the tendencies to purify the Czech language are not so strong, but protective attitudes can still be observed, because some linguists are afraid that the increasing presence of foreign words poses a threat to the Czech language.

### 3 Why are native equivalents never fully equivalent to lexical loans?

Many linguists, philosophers and literary historians have claimed that no two words can be fully equivalent with respect to their linguistic function. One can find the same opinion among lexicographers, cf. the often-quoted passage from Urdang’s introduction to *The Synonym Finder*:

Those who work with language know that there is no such thing as a true ‘synonym’. (...) Even though the meanings of words may be the same – or nearly the same – there are three characteristics of words that almost never coincide: frequency, distribution, and connotation. (Urdang 1978)

Ullmann takes a less extreme position on this point and explains why cases of absolute synonymy are very rare:

(...) it is perfectly true that absolute synonymy runs counter to our whole way of looking at language. When we see different words we instinctively assume that there must also be some difference in meaning, and in the vast majority of cases there is in fact a distinction even though it may be difficult to formulate. Very few words are completely synonymous in the sense of being interchangeable in any context without the slightest alteration in objective meaning, feeling-tone or evocative value. (Ullmann 1964: 142)

As ‘completely synonymous’ he mentions technical terms, e.g. *caecitis* and *typhlitis* can both be used with reference to the inflammation of the blind gut. However, even such names differ with respect to non-designative features, e.g. they evoke different associations in the minds of the language users.

One often hears that language does not tolerate fully equivalent words and differentiates them, thus eliminating cases of full equivalence. The tendency to avoid redundant means of expression is said to be evidence that language is governed by laws of economy (Nagórko 2004: vii). However, such propositions do not explain the inner mechanism of linguistic economy and, in particular, they do not explain why there should be a difference in meaning, broadly understood, between loanwords and their native synonyms. Our position is that this has something to do with the word form itself. The unfamiliar forms of lexical loans are perceived differently from the familiar forms of their native synonyms and the difference in perception may result in different semantic development of such words. For example, shortly after *kurort*, a 19th-century borrowing from German, appeared in Polish, a native term *uzdrowisko* (literally 'health resort') was coined with the intention to relegate the unwanted loan from the language. However, this effort failed to have the desired effect: instead of disappearing, *kurort* changed its meaning to 'popular and snobbish holiday place'. The change was very likely directed by the connotations of the word: some pre-war dictionaries (e.g. *Słownik wyrazów obcych* of 1921) informed that *kurort* was most often used with reference to health resorts in Germany and the preference for foreign places in its use is still visible in modern texts. Furthermore, the collocation image of *kurort* includes such features as exclusiveness (strangely enough, not in conflict with popularity), modernity, reputation and elegance, whereas in the collocation image of *uzdrowisko* it is tradition and aesthetic values that are best seen (see Bańko 2013a for a more detailed analysis of both words).

The influence of word form on word meaning can be studied in texts and other cultural artifacts (e.g. by collocation analysis or Google image inspection), but it can also be brought to light by means of experiments. The results published by Song and Schwarz (2010) are worth quoting here. They experimented with nonce words, some of them familiar in shape, some strange, and observed a correlation between the familiarity of a word and its perception. For instance, fictitious food additives with names difficult to pronounce were evaluated as more harmful than food additives with easy names. Similarly, roller-coaster rides in a fictitious amusement park were judged as more risky and more exciting when their names were strange and difficult. Song and Schwarz explain the effect with a mistaken projection of the difficulties the subjects experienced in processing the unfamiliar words onto the referents of the words: unaware of the source of difficulty, the subjects attributed it to the referents, judging them as more risky, more dangerous, more harmful, etc. (for a critical review of these studies see Rączaszek 2013).

It would be premature to claim that Song and Schwarz's findings can explain all the distinctions observed between a synchronic loan and its native synonyms. More experiments are needed and they should be done on real language data, not on invented words. We will next briefly describe a project designed to perform more systematic research in this regard, using both linguistic and psycholinguistic methods.

## 4 About the APPROVAL project

The aim of the APPROVAL project is to search for various factors bearing on the psychological perception, social reception and linguistic adaptation of loanwords.<sup>1</sup> Among the possible factors, the relation between word meaning and word form is of particular importance, because we assume that the form of a word is not irrelevant (contrary to the widely accepted view of the arbitrary nature of linguistic signs, a foundation stone of Saussurian linguistics). The form of a word can influence its meaning (cf. *kurort* above), but also the meaning of a word can affect its form, e.g., by hindering the process of a loanword's adaptation (the word *jazz* can be a case in point: though borrowed to Polish and Czech almost a hundred years ago, it still appears mainly in its original spelling in both languages, very likely because the foreign spelling reflects better the symbolic values associated with jazz music in Poland and the Czech Republic, see Bańko and Hebal-Jeziarska 2012 for details).

We also assume that fully equivalent words do not exist, so we focus on comparative analysis of word pairs (sometimes triples, quadruples, etc.) in which one element is of foreign origin, the other native, or in which one element has the original spelling, while the other is graphically adapted to the recipient language. Fifty Polish word pairs have been subjected to in-depth analysis, based on language corpora and other data, not excluding the evidence in the language itself (e.g. we treat the frequency of a word as indicative of its importance and we pay attention to secondary uses of a word, its derivatives and idiomatic expressions, because such data reveal some of the typical associations the word calls up in the minds of its users and help to draw the stereotypical image of its referent). As this paper is being prepared, most of the 50 synonym and variant pairs have been already inspected and the results are available on the project website, see [http://www.approval.uw.edu.pl/en\\_GB/start.pl](http://www.approval.uw.edu.pl/en_GB/start.pl).

In order to make the observations more credible, the same research is being done on corresponding Czech word pairs which serve as a control group, e.g. the Polish pair *absurdalny – niedorzeczny* 'absurd, nonsensical' corresponds to the Czech pair *absurdní – nesmyslný*, in which the first element comes from the same root as the first element of the Polish pair. By composing the research material this way, it became possible to study the adaptation processes in two cognate languages on the basis of comparable examples, using the same methodology, the same kind of data, and even the same description format.

In total, 100 pairs in two languages will have been analyzed by the end of the project, using corpora, as well as dictionaries, web archives, digital libraries, library catalogues, Google images and other sources of language relevant data (e.g., library catalogues are being used to check frequencies of words in book titles). In addition, psycholinguistic experiments are being carried out on the Polish material to enrich and verify the observations based on language corpora and other textual and non-textual sources by means of linguistic methods (see the project website for details).

---

1 The name of the project comes from 'Adaptation, Psychological Perception and Reception of Verbal Loans'. It is also meant as a reminder that in normal circumstances loanwords do not pose a threat to a language; to the contrary, they add to its wealth.

The results gained so far are encouraging and they largely support the hypothesis of there being a relationship between form and meaning in the adaptation of lexical loans. For limits of space, a few examples from Polish will have to suffice here. We will focus on selected details, with no intention to account for a full analysis of any of the words mentioned below.

#### 4.1 strofa and zwrotka

The Polish words *strofa* and *zwrotka* both mean 'stanza', but in technical literature usually the former word is used, while in the general language the latter one is more common. The likely reason is not only that *strofa* is of Greek origin (borrowed via Latin), but also that *zwrotka* contains a familiar suffix *-k-* which in many other nouns (though not in *zwrotka*) has a diminutive function.

The difference between these two words was first observed in corpus analysis, especially in their collocation images. For example, *zwrotka*, but not *strofa*, is used in reference to popular songs and children's poems, *strofa* can be the subject of aesthetic evaluation (cf. *piękne strofy* 'beautiful stanzas') and artistic activity (cf. *pisać, układać strofy* 'write, arrange stanzas'), while *zwrotka* is less frequent in such contexts. In addition, *strofa* can be recited, but *zwrotka* is sung. Only *zwrotka* collocates with the word *refren* ('refrain'), which confirms its connection to songs.

A study of free associations, based on Osgood's semantic differential, was next carried out. Twenty-two subjects took part in it, each asked to mention up to three associations for one word, so the maximum number of associations for a word was 66. Among the associations noted more than once, *refren* 'refrain', *rymy* 'rhymes', *śpiewanie* 'singing', *muzyka* 'music' and *ognisko* 'camp-fire' were given only for *zwrotka*, while *szkoła* 'school', *poezja* 'poetry', *literatura* 'literature' and *Mickiewicz* (the best known Polish poet) were given only for *strofa*. In addition, though *piosenka* 'song' and *wiersz* 'poem' were mentioned for both words, *piosenka* had a frequency of 19 with *zwrotka* and 3 with *strofa*, while *wiersz* appeared 16 times with *strofa* and only 3 times with *zwrotka*. As can be seen, the results of corpus analysis are in line with the study of free associations.

#### 4.2 helikopter and śmigłowiec

Though *helicopter* and *śmigłowiec* both mean 'helicopter' in Polish, their stylistic distribution is different. *Helicopter* is common in spoken language and in many other language varieties, while *śmigłowiec* tends to be used in technical literature. This is probably the reason why in the domain *lego.com/pl-pl*, belonging to the producers of Lego bricks, the Google search engine finds far more occurrences of *helikopter* than *śmigłowiec*. As far as book titles are concerned, *helikopter* can be found on the covers of children's stories, while *śmigłowiec* appears in the titles of books on aeronautical technology and military science. Among the Google images indexed with the word *helikopter*, toys and miniature models are more frequent than among images indexed with the word *śmigłowiec*.



However, a more interesting and more surprising observation about *helikopter* and *śmigłowiec* can be made when comparing their relative frequencies in certain contexts. Though on Polish-language websites *helikopter* is several times more frequent than *śmigłowiec*, the quantitative advantage of *mały helikopter* ‘small helicopter’ and *szybki helikopter* ‘fast helicopter’ over *mały śmigłowiec* and *szybki śmigłowiec*, respectively, is significantly lower. Moreover, *lekki helikopter* ‘light helicopter’ is less frequent than *lekki śmigłowiec*. Apparently, small, light and fast machines of this type are more often referred to with the word *śmigłowiec* than its relative frequency to the word *helikopter* would suggest. This may be due to the fact that the name *śmigłowiec* is related to the words *śmigło* ‘propeller’, *śmigły* ‘swift’ and *śmigać* ‘move quickly, zip (around)’.

However, the overall picture is not quite clear yet, partly because the relative frequencies of *helikopter* and *śmigłowiec* in two reference corpora of Polish – Narodowy Korpus Języka Polskiego and Korpus Języka Polskiego PWN – are opposite to those found on the Internet and partly because the study of free associations has yielded different results for these two words than obtained in corpus analysis. Further research into the psychological perception of *helikopter* and *śmigłowiec* is planned within the APPROVAL project with the intention to confirm or refute the conjectures made on the basis of corpus data. Whatever the results of the research, there is no doubt a difference between *helikopter* and *śmigłowiec* in their semantic content, if only non-designative components of the word meaning are allowed.

### 4.3 eksplozja and wybuch, kuriozalny and osobliwy

*Eksplozja* ‘explosion’ and *wybuch* ‘explosion, outbreak, outburst’ may refer to the same kind of events, but the phenomena referred to by the former word are perceived as stronger and more violent. The difference is so distinct that it has even been noted in the definitions for these two words in some dictionaries. There are more synonym pairs in which the referents of a loanword seem larger and more powerful than the referents of its native synonym, cf. *dewastować* and *niszczyć*, both meaning ‘destroy’. Here the first element, cognate with English *devastate*, denotes a purposeful or mindless activity, especially against public property or natural environment.

However, sometimes the difference between a loanword and its native synonym lies not in the referents themselves, but in the way they are talked about, e.g. in the values conveyed. The adjective *kuriozalny* ‘peculiar, bizarre’, related to Latin *curiosum*, is half as frequent in Polish as its native counterpart *osobliwy* ‘peculiar’, but in parliamentary reports the former word prevails overwhelmingly. A closer inspection shows that Polish MPs need it to criticize their political opponents, e.g. *Pana poglądy są kuriozalne, panie pośle* ‘Your views are bizarre, Mr. X’.

The tendency to use difficult and erudite words for hyperbolic effects, whether to make a phenomenon look more powerful or just to convey negative attitudes, can be well explained in the context of Song and Schwarz’s (2010) experiments discussed above.



#### 4.4 dealer and diler

The last of our examples is different from the previous ones, because it is not concerned with a synonym pair. It deals with a pair of spelling variants of the same word, a recently new Polish borrowing from English. However, the situation is much the same as before, because one element of the pair is foreign while the other one is 'nativised' (rather than native), and the two elements therefore exhibit the same 'unfamiliar - familiar' opposition as in the case of synchronic loans and their native synonyms. Thus, the necessary conditions are met for different associations to evolve around the different words.

As all variants, *dealer* and *diler* have the same designative meaning, but their stylistic distribution and the areas of their application are not identical. In the press, *dealer* is almost twice as frequent as *diler* and used mainly with reference to car vendors, whereas *diler* has equal frequency in automobile and drug-related contexts. In literary texts, on the other hand, *diler* is twice as frequent as *dealer* and applied usually to drug sellers, whereas *dealer* is equally often used in automobile contexts. Apparently, the foreign variant is more prestigious and better suited to name the job of authorized vendors in car showrooms; the domestic variant, on the other hand, is unpretentious and corresponds well with the dubious job of drug peddling. The difference can well be observed in the Google image galleries, too, which once more confirms the usefulness of Google images in linguistic analysis (Bańko 2013b).

### 5 Conclusions for practical lexicography

Our intention was not to question the need for normative assessments in dictionaries, nor to claim that cumulative synonym dictionaries have no *raison d'être*. The conclusions from our work are relevant to the theoretical foundations of lexicography, but also to lexicographic practice. It is important for lexicographers to be aware that distinctions often may lie where similarity seemingly prevails (in a way, all of language is based on distinctions, and here we are in complete agreement with de Saussure). Furthermore, it is important to realise that in the adaptation of lexical loans, form and meaning are interdependent: the form of a word can affect its meaning, but it can also be influenced by it. In many cases, a tendency to maintain harmony between a word form and word meaning can be observed in the process of loanword adaptation.

Better recognition of differences between near synonyms is essential for both monolingual and bilingual lexicography. Adequate definitions, especially in production dictionaries, should explain differences between synonymous words. Adequate translation equivalents are dependent, among other things, on how well near synonyms of the source language and the target language are discriminated. This is not to say that each dictionary ought to be equally specific in its treatment of word meanings. For example, decoding dictionaries need not focus so much on distinctions between words as may be expected from encoding dictionaries. Lexicographers should consider for themselves to what extent

the observations made in this paper may be useful in their work. In any event, more caution is advised before assessing a loanword as unnecessary and more thoroughness is needed in how new words borrowed from other languages are treated. It is not enough to blame those who use them of snobbery.

A separate question, not to be dealt with here, is how our findings can be incorporated in what are called distinctive synonym dictionaries, which are intended to account for differences among synonyms rather than to gather as many words close in meaning as possible (cf. *Dystynktywny słownik synonimów* by Nagórko et al. 2004 as an example of works of this type). Another area where it is more important to show differences between words than to identify similarities is the so called synonym discussions, known from some dictionaries (cf. special paragraphs, headed *Synonyms*, in *The American Heritage Dictionary*).

## 6 References

- Adamska-Sałaciak, A. (2001). Linde's Dictionary: A landmark in Polish lexicography. In *Historiographia Linguistica*, 28(1-2), pp. 65-83.
- American Heritage Dictionary. Second College Edition. Boston: Houghton Mifflin Company, 1982.
- Bańko, M. (2013a). Normatywista na rozdrożu. Dwugłos w sprawie tzw. kryterium narodowego. In J. Migdał, A. Piotrowska-Wojaczyk (eds.) *Cum reverentia, gratia, amicitia... Księga jubileuszowa dedykowana Profesorowi Bogdanowi Walczakowi*, vol. 1. Poznań: Wydawnictwo Rys, pp. 141-148.
- Bańko, M. (2013b). Obrazy Google jako źródło informacji lingwistycznej. In W. Chlebda (ed.) *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole: Wydawnictwo Uniwersytetu Opolskiego, pp. 73-84.
- Bańko, M. & Hebal-Jeziarska, M. (2012). Proč jazz, nikoliv džez? Harmonie grafické podoby lexému a obsahu – jako jeden z činitelů ovlivňujících adaptaci cizojazyčných přejímek? In S. Čmejrková, J. Hoffmannová, J. Klímová (eds.) *Čeština v pohledu synchronním a diachronním*. Praha: Karolinum, pp. 371-375.
- Czesak, A. (2007). *Oczyszcziciel mowy polskiej* E. S. Kortowicza, Poznań 1891 – idee i zawartość. In J. Kamper-Warejko, I. Kaproń-Charzyńska (eds.) *Z zagadnień leksykologii i leksykografii języków słowiańskich*. Toruń: UMK, pp. 79-85.
- Engelhardt, G. (2001). Český a německý purismus z konce 19. století. In *Naše řeč* 84(5), pp. 235-242.
- Korpus Języka Polskiego PWN [PWN Corpus of Polish]. Online at <http://korpus.pwn.pl>.
- Kortowicz, E. S. (1891). *Oczyszcziciel mowy polskiej, czyli Słownik obcośłów, składający się z blisko 10,000 wyrazów i wyrażen z obcych mów utworzonych a w piśmie i w mowie polskiej niepotrzebnie używanych, oraz z wyrazów gminnych, przestarzałych i ziemskich w różnych okolicach Polski używanych z wysłowieniem i objaśnieniem polskiem*. Poznań: czcionkami drukarni „Dziennika Poznańskiego”.
- Král, J. (1917). Naše brusy I. In *Naše řeč*, 1(4). Online at <http://nase-rec.ujc.cas.cz/archiv.php?art=64> [05/04/2014]
- Leszczyński, Z. (2000). Krótka relacja o puryście sprzed wieku. In *Prace Filologiczne*, XLV, pp. 347-352.
- Linde, S. B. (1807-1814). *Słownik języka polskiego*, 6 vols. Warszawa.
- Lipczuk, R. (2007). Geschichte und Gegenwart des Fremdwortpurismus in Deutschland und Polen. Frankfurt am Main: Peter Lang.

- Lipczuk, R. (2011). O słownikach wyrazów obcych, słownikach zniemczających i spolszczających. In B. Afeltowicz, J. Ignatowicz-Skowrońska, P. Wojdak (eds.) *In silva verborum. Prace dedykowane Profesor Ewie Pajewskiej z okazji 300-lecia pracy zawodowej*. Szczecin: Volumina.pl, pp. 205-216.
- Markowski, A. (1999) (ed.). *Nowy słownik poprawnej polszczyzny*. Warszawa: PWN.
- Markowski, A. (2005). *Kultura języka polskiego. Teoria. Zagadnienia leksykalne*. Warszawa: PWN.
- Nagórko, A., Łaziński, M. & Burkhardt, H. (2004). *Dystynktywny słownik synonimów*. Kraków: Universitas.
- Narodowy Korpus Języka Polskiego [National Corpus of Polish]. Online at <http://nkjp.pl>.
- Rączaszek, J. (2013). Studying the semantics of loanwords which have near synonyms in the host language: Psycholinguistic and multidimensional corpus representation methods. Accessed at [http://www.approval.uw.edu.pl/en\\_GB/publikacje](http://www.approval.uw.edu.pl/en_GB/publikacje) [05/04/2014].
- Słownik wyrazów obcych. 25.000 wyrazów, wyrażań, zwrotów i przysłów cudzoziemskich, używanych w mowie potocznej i w prasie polskiej, 9th ed. Warszawa: Wydawnictwo M. Arcta, 1921.
- Song, H. & Schwarz, N. (2010). If it's easy to read, it's easy to do, pretty good, and true. In *The Psychologist*, 23(2). Accessed at [http://www.thepsychologist.org.uk/archive/archive\\_home.cfm?volumeID=23&editionID=185&ArticleID=1629](http://www.thepsychologist.org.uk/archive/archive_home.cfm?volumeID=23&editionID=185&ArticleID=1629) [05/04/2014].
- Svobodová, D. (2009). *Aspekty hodnocení cizojazyčných přejímek: mezi módností a standardem*. Ostrava: Universitas Ostraviensis.
- Ullmann, S. (1964). *Semantics. An Introduction to the Science of Meaning*. Oxford: Basil Blackwell.
- Urdang, L. (1978). Introduction. In J. I. Rodale (ed.) *The Synonym Finder*. Emmaus, Pa.: Rodale Press.

### **Acknowledgements**

This project has been supported by a grant from the National Science Centre, Poland, no DEC-2011/03/B/HS2/02279.



# Pejorative Language Use in the Satirical Journal “Die Fackel” as documented in the “Dictionary of Insults and Invectives”

Hanno Biber  
Austrian Academy of Sciences  
hanno.biber@oeaw.ac.at

## Abstract

Satirical literary texts have certain properties that are highly interesting for the study of pejorative language use. The language of the satirical journal “Die Fackel” published and almost entirely written by Karl Kraus is the text basis for a text-lexicographic exploration into the field of pejorative language and its specific lexicographic units. The „Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift „Die Fackel“. Alphabetisches, Chronologisches, Explikatives” was published in 2008. The three volumes document the usage of invectives in the journal, in alphabetical and in chronological order, and in explicative form explained through the example of the last article of the journal. The alphabetical part consists of 2,775 examples of pejorative phrases and related indices. The chronological part presents 555 of these pejorative phrases arranged in chronological order providing expanded contexts. The third volume contains explicatory texts as well as “Wichtiges von Wichten”, the final article of “Die Fackel”, where pejorative phrases were marked up and accompanied by commentaries. This source is representing a literary genre that offers a variety of different forms of pejorative language to be studied from various perspectives. The lexicographic insights offered by the text dictionary into the use of pejection by Karl Kraus will be presented in this paper.

**Keywords:** text lexicography; literary studies; pejorative language

## 1 Text Dictionary of “Die Fackel”

“Die Fackel” (“The Torch”) is the name of the satirical magazine of 22.586 pages which was published by Karl Kraus in 922 issues in Vienna from 1 April 1899 until February 1936. The work of the satirist, language critic and social critic Karl Kraus, who was born in Bohemia in 1874 and died in Vienna in 1936, is an abundant and highly interesting source not only for the history of his time, but for the language spoken and written at the time and above all for the moral transgressions which the satirist observed by interpreting the words and phrases of his time and which he pointed out in his satirical and polemical texts. His very influential literary journal comprises a great variety of articles, essays, glosses, notes, commentaries, aphorisms, poems, songs, advertisements, and many other literary forms. The main method of his satire is the method of quotation, whereby Karl Kraus wittingly com-

ments upon the quotations he finds in the newspapers, journals and magazines as well as in the literature and in the political speeches of his time. He criticises in numerous satirical and polemical articles of his magazine the acts and the words of his contemporaries who were active in various intellectual fields, not only in the media, but also in the theatre, the university, the church, in politics, the economy, the military, and so on. Karl Kraus covers in his typical style in thousands of texts the themes of journalism and war, of politics and corruption, of literature and lying. The language of the satirical journal “Die Fackel” - for several decades since its start in 1899 almost entirely written by Karl Kraus, and from 1911 on without external contributions, - is the text basis for a unique text lexicographic exploration which has been carried out at the Austrian Academy of Sciences for several years. Nowhere else in German literature, to mention but one aspect, is there such an extensive documentation of the socio-political idiom of the time as there is in “Die Fackel” by Karl Kraus. The idea of compiling a text-dictionary of the “Fackel” derives from our interest in language and how it is used in “Die Fackel”. The Fackel-Dictionary is a selective text dictionary research project initiated by Werner Welzig. The original plan was to develop three different types of dictionary. First, a “Dictionary of Idioms”, the *Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‘Die Fackel’* (Welzig 1999), published on the occasion of the 100th anniversary of “Die Fackel” in 1999, a monumental scholarly publication that has won several international and national prizes, among them the Prix Logos by the French association of linguists and the Golden Letter at the Leipzig book fair as the most beautiful book of the year 2000 for its design worked out by the designer Anne Burdick in cooperation with the Fackellex working group (Hanno Biber, Evelyn Breiteneder, Susanne Buchner, Heinrich Kabas, Karlheinz Mörth, Christiane Pabst, Franco Schedl, Adriana Vignazia, Werner Welzig). In order to thoroughly analyse and interpret the more than 9000 idiomatic units of the text dictionary, of which 144 idioms are described in great detail in individual entries, which are longer than common dictionary entries are, it has proved reasonable to search large volumes of texts for comparison, a procedure that has also been made use of for the dictionary that followed. It is a lexicographic necessity to have large text corpora available, in particular when searching for lexical units, for example for idioms or for pejorative phrases, so that these text dictionary projects can be regarded as examples and applications of corpus based textual studies. The second example within this context of text lexicography is the “Dictionary of Insults and Invectives”, *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‘Die Fackel’* (Welzig 2008), which was also worked out by the Fackellex working group (Hanno Biber, Evelyn Breiteneder, Gerald Krieghofer, Karlheinz Mörth) and will be introduced in this short presentation. Third, an “Ideological Dictionary” had been originally planned, which later in the course of the development of the program plans has been decided to be transformed into a special edition of the posthumously published text “Third Walpurgis Night” (*Dritte Walpurgisnacht*) by Karl Kraus, written in May 1933, constituting a manifestation the most important contemporary text of German literature dealing with the early time of National Sozialism and the issue how the intellectuals reacted when this most violent regime came to power.

## 2 Pejorative Language Use

In this paper a short presentation of the main aspects of the second text dictionary, the “Dictionary of Insults and Invectives”, *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‘Die Fackel’* (Welzig 2008) will be given, offering an exploration into the field of pejorative language use and its specific lexicographic units as selected for this dictionary. “Die Fackel” can be regarded as an ideal text basis for such a dictionary in that it has no equal in the German literature of the twentieth century either in terms of form and content or in the use of language. The *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‘Die Fackel’* (Welzig 2008) published in 2008 consists of three parts: „Alphabetisches“, „Chronologisches“, „Explikatives“. The three volumes document the usage of the invectives and pejorative phrases in the journal, in alphabetical order (ALPHA), in chronological order (CHRONO), and explained through the example of the last article of the journal in a volume (EXPLICA), that shows by thoroughly analysing one short, but semantically very dense text, the full potential of a text lexicographic documentation of its pejorative forms.

Adjektivschmuser		ALPHA 1
<b>A</b>		
die <b>Aasgeier</b> des Interessanten	🔍	adg.ac.at/F/437_090
die <b>Ab- und Zufaller</b> meiner Region	🔍	adg.ac.at/F/890_248
Abhub des bürgerlichen Geisteslebens	🔍	adg.ac.at/F/820_050
der <b>übelste Abhub</b> der Wiener Geistigkeit	🔍	adg.ac.at/F/351_054
schwacher <b>Abklatsch</b> von Lustspielfiguren	🔍	adg.ac.at/F/336_021
<b>Abkömmlinge</b> der Bürgerwelt	🔍	adg.ac.at/F/890_297
Abkömmlinge meines Lebenskreises	🔍	adg.ac.at/F/445_142
diese <b>Abkürzer</b> der Sprache und des Lebens	🔍	adg.ac.at/F/501_026
diese <b>Abkürzung</b> eines Rezensenten	🔍	adg.ac.at/F/457_024
<b>Abonnenten</b> der ‚Neuen Freien Presse‘	🔍	adg.ac.at/F/071_001
<b>Abonnenten</b> der großen Tagespresse	🔍	adg.ac.at/F/118_002
<b>Abschaum</b> von einer Creme	🔍	adg.ac.at/F/697_140
<b>Abschaum</b> der Wiener Advokatie	🔍	adg.ac.at/F/170_009
Alterston des <b>Abschiednehmers</b>	🔍	adg.ac.at/F/423_049
einer der beliebtesten <b>Abschilderer</b> von Verwundetenzügen	🔍	adg.ac.at/F/531_184
Franz Blei, ein <b>Abt</b> der roten Garde	🔍	adg.ac.at/F/568_035
<b>Achilles</b> des Wiener Feuilletons	🔍	adg.ac.at/F/368_025
glaubt das Volk, ein <b>Achtundvierziger</b> sei die Rufnummer eines Fickers, und ein Unnummerierter ist doch mehr	🔍	adg.ac.at/F/251_032
<b>Adabei</b> und <b>Niedabei</b>	🔍	adg.ac.at/F/622_146
<b>Adept</b> der Neuen Freien Mythologie	🔍	adg.ac.at/F/121_012
<b>Adepten</b> des Ritualmordglaubens	🔍	adg.ac.at/F/059_001
<b>Adepten</b> seiner Schminke	🔍	adg.ac.at/F/349_008
<b>Adjektivkünstler</b>	🔍	adg.ac.at/F/387_002
<b>Adjektivschmuser</b>	🔍	adg.ac.at/F/857_033

Fackel Schimpfwörterbuch

Figure 1: ALPHA.

The alphabetical volume of the text dictionary (ALPHA) consists of 2,775 examples of pejorative phrases and several related indices. In this alphabetical part the lemmatized entries of the pejorative forms are marked and each of the entries is given a short context. Only one reference is given with one entry and those entries which are represented in full detail of the page in the chronological volume of

the text dictionary are printed in red color. The indices at the end of the alphabetical part, referring to the pages and the lines indicated by the marginal letters, comprise a complete index of the documented word forms, second a selective index of inverted word forms and also a few indices of names (an index of personal names, placenames and other names).

The chronological volume (CHRONO) of the *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift 'Die Fackel'* (Welzig 2008) presents 555 selected examples of the overall 2,775 selected pejorative phrases arranged in chronological order as they appear in the magazine, thereby providing expanded contexts by printing the full page of the original journal in a graphically transformed facsimile where the quoted passage is highlighted. In all three volumes of the text dictionary the references to the digital edition of the journal, the AAC-Fackel - published within the framework of the corpus research enterprise "AAC - Austrian Academy Corpus" (AAC-FACKEL 2007) - are given by means of short URLs of the individual pages, so that the user of the text dictionary has always the full context available and can at the same time make use of the print representations of the highlighted pejorative expressions. The larger context given in the chronological volume of the text dictionary allows the reader to evaluate the textual dynamics and the effects how the pejorative expression is constituted, which in many cases is gradually intensified and accompanied by other related expressions in the context of the satirical or polemical text of "Die Fackel". In many cases the pejorative intensifications are made possible by word formation processes or syntactical effects, which is one of the reasons why the particular use of compounds is documented to a larger extent as well as certain significant pejorative collocations are taken into consideration for this text dictionary of insults and invectives. In many cases words that are not commonly used pejoratively are used in this way by the satirist, who is also reflecting upon this particular satirical procedure in his texts.



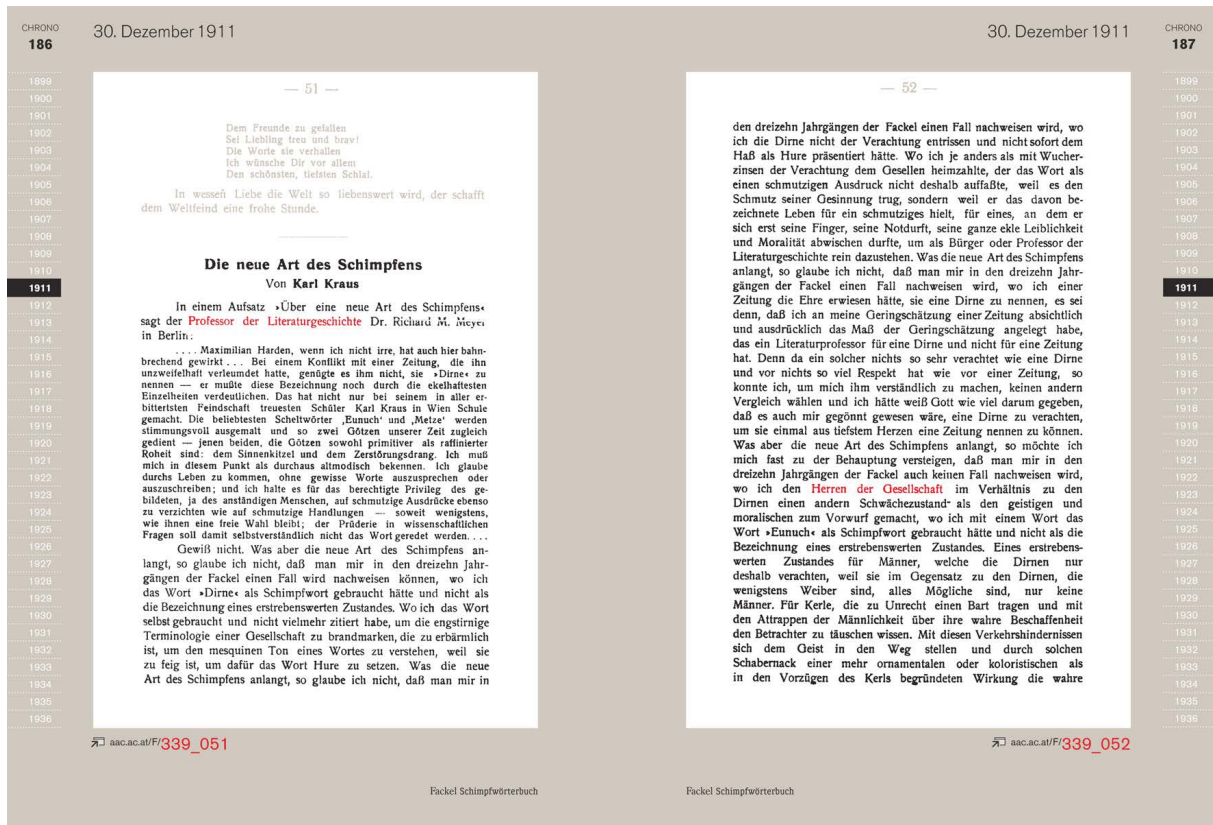


Figure 2: CHRONO.

The third volume (EXPLICICA) contains explicatory texts as well as “Wichtiges von Wichten”, the final article of “Die Fackel”, where pejorative phrases were marked up in this specific text and have been accompanied by detailed commentaries. This last text published in “Die Fackel” in February 1936 is treated as a source text for the analysis of pejorative terms in order to exemplify the difficult and ambitious task of getting to terms with the high level of pejoration in the satirical and polemical texts written by Karl Kraus. The politically interesting polemical note “Wichtiges von Wichten” from 1936, referring to the political situation at the time and how to react to it by means of writing, is reproduced in the explicatory volume of the dictionary in a plain form first, giving the readers an uninterrupted chance to read a large piece of textual evidence first and then it is given in a form in which, according to the text lexicographers’ interpretations, pejoratively used expressions are highlighted in the text and provided with their alphabetical list. In a third part of this volume the same text is reproduced again, this time dashed out with only those expressions left visible which are commented upon by the editors and compilers of the text dictionary, documenting the intensive need and necessity for detailed commentary in order to understand and fully assess the pejorative qualities of the expressions selected.

This source text in particular as well as the whole journal in general is representing a literary genre that offers a variety of different forms of pejorative language to be studied from various perspectives. The various use of the pejorative expressions not only in the last text, but above all in all texts chosen

are, as has been observed, dominated to a large extent by the creative transformations and configurations performed by the writer. These creative adaptations and modifications are not only important for the documentation and the analysis of the idiomatic expressions as represented in the text lexicographic project of the “Dictionary of Idioms”, but also to a large extent these creative adaptations are most relevant for the way, in which pejorative expressions are formed, which can be studied in great detail in the “Dictionary of Insults and Invectives”. In the case of the idioms, a more or less normal form is creatively transformed into other forms, which cannot easily be detected by standard automated corpus query systems, only systematic annotation and possible semi-automatic methods could provide the scholar with reasonable results to be gained from larger corpora. In the case of the “Dictionary of Insults and Invectives” the corpus of the whole text has been made use of for purposes of systematization. This extensive literary text is full of a great variety of satirical forms in the context of pejorative expressions. For this reason it can be used as a research object following a certain combined interest in the study of pejorative language and in phenomena related to the more methodological questions of satirical and polemical language as a research topic for text lexicography as well as for corpus research.

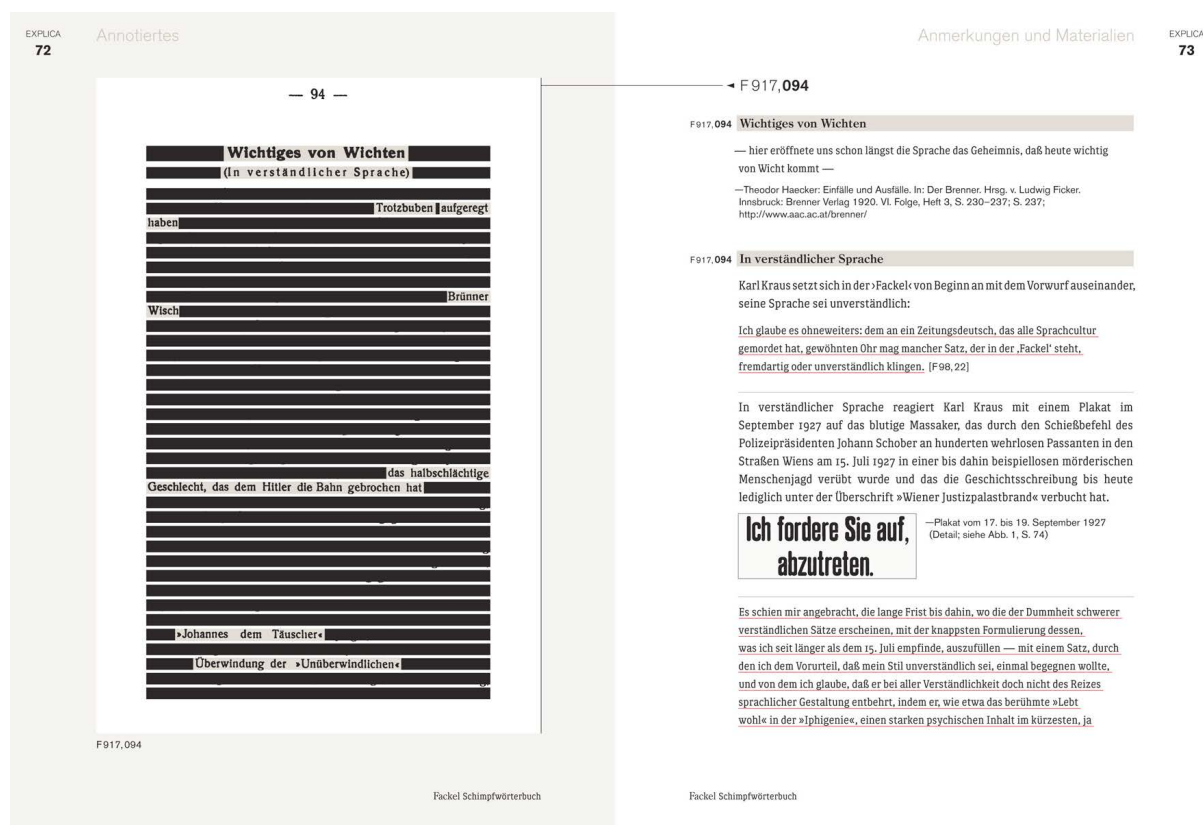


Figure 3: EXPLICA.

The research initiatives concerning the magazine published by Karl Kraus from April 1899 until February 1936 has offered the lexicographers at the Austrian Academy of Sciences in Vienna a unique opportunity to study and to document the language of this important writer in great detail. No author of the 19th and 20th centuries has thought or written about the language of his contemporaries in Vienna, Berlin, or Prague as precisely, as continuously and as passionately as Karl Kraus. No other author is showing such a productive and such an effective use of pejorative expressions as Karl Kraus in his satirical and polemical texts in “Die Fackel”.

### 3 References

- AAC - Austrian Academy Corpus: AAC-FACKEL, Online Version: *Die Fackel*. Herausgeber: Karl Kraus, Wien 1899-1936. AAC Digital Edition No 1 (ed. Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, Karlheinz Mörth), 2007, Accessed at: <http://www.aac.ac.at/fackel> [01/01/2007].
- Welzig, W. (ed.) (1999). Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‚Die Fackel‘. Austrian Academy Press, Vienna.
- Welzig, W. (ed.) (2008). *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift ‚Die Fackel‘* (3 volumes: Alphabetisches, Chronologisches, Explikatives). Austrian Academy Press, Vienna.



# The Presence of Gender Issues in Spanish Dictionaries

Ana Costa Pérez  
Universidad Carlos III de Madrid  
Acosta.hum@uc3m.es

## Abstract

Dictionaries are ideological creations as they are but a reflection of society itself. A dictionary sets a standard for language; makes an authority, a cultural product, and builds a lexical *encyclopaedia* and a social reference. This analysis is built upon the idea of an undeniable overlap between ideology and dictionary and the role of the latter as a mechanism to transmit the limited sights of everything around us; talking is a world-defining act by individuals who forces themselves to adapt to a code which is seemingly open and closed at the same time; a code imposed by the society to which they belong and which will be enforced on future generations. Therefore, the present work will highlight the catalogue of definitions that challenge the descriptive neutrality of current lexicographical work, turned into dictionaries which should mirror an equal society, without discrimination.

This requires defining the concept from different scopes: linguistic, anthropological, sociolinguistic, philosophical or cognitive. We aim at showing how grammar, with its two (and even up to three) *genera*, provides us with the perfect field to focus on sexes, at both biological and social senses, on Nature and Culture, without favouring the existence of two different sexes nor any individual powers and decisions. The *Academia* notes that the Spanish language foresees the possibility to refer to mixed groups through the grammatical masculine gender, possibility in which there is no discriminatory intent, but the application of the linguistic law based in the expressive economy. Only when the opposition of the sexes is a relevant factor in the context, the Academy considers necessary the explicit presence of both genders

**Keywords:** Dictionaries; Gender; Ideology

## 1 Preliminary Issues

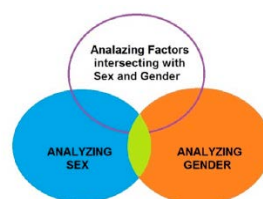
Gender is linguistically reserved for words in which sex evidences a sexed condition of human beings. But if such a distinction is so shockingly clear, why does it continue to emerge almost cyclically? why the same arguments about this pair of concepts in specialized areas such as grammar or lexicography?

Taking a diachronic perspective, we can check how talking about gender from a feminist perspective, rather than a more objective one, sets a more objective scope, that of women-centered movements.

The objective, as such, arises when we try to return gender to unmarked meanings, free from criticism, politics or claims.

When defining one or several related categories based on what has been historically common to them, it is essential to analyze the gaps where they cease to be common. For example, categorically man or woman can not conceive based on specific characteristics, but the problem is that some concepts contain a complex network of variables that unravel some concepts of our language, such as gender, mother, sex, marriage, education, fatherhood, female, manhood, religion or science. Though taking the form of descriptive definitions, they are actually conditioned acts. This is due to an attempt to isolate, to “numb” the emotional charge that for years has been lodged in certain socially marked terms.

That is why a big part of what has been traditionally charged to both man and woman have depended on an intervened meaning. Therefore, we have chosen a descriptive approach based on the analysis of the definitions provided by the academic dictionary related to various *nuclea* of meaning: the major professions, body differences, adjectives related to personality or maternity / paternity will be the main ones. Through comparative analysis (masculine and feminine terms will be opposed), we will try to highlight the ideology underlying the definitions of the Spanish Academy dictionary. ( Figure1)



**Figure 1: Analyzing Gender.**

## 2 Methodology

For the last 20 years, treating languages by meanings of databases (linguistic *corpora*) has turned into a challenge that has been imposed to all the linguists and lexicographers. The structure of information in IT bases presents big advantages. The constant number of systems of meaning (semantic fields) allows describing the lexical elements in an unified and coherent way. Thus, it is guaranteed, for all the lexical units of a certain type, a common methodological response. The comparisons among related terms are possible and objectivable and we can proceed to controls which guarantee coherence, regularity and uniformity.

*Corpus* linguistics stands as a method for carrying out linguistic analyses. As it can be used for researching many kinds of linguistic questions and, as it has been shown to have the potential to yield

highly interesting and new insights about language and relationships between men and women, it has become one of the most widespread methods of linguistic research in the last years.

In order to elaborate complete and systematic definitions, I hereby make a proposal to establish a standard of definition for every category included in meaning systems (1 to 6). Then, we will decide the characteristics that must appear in the concept definitions belonging to the systems that I specify in the appendix. At present, different strategies or methods of definition coexist, and then the lexicographer chooses, yet never openly opting for a single one, combining different methods.

For these definitions we have to bear in mind the following:

- (1) Context: general information of the word in question in relation to its frequency of use.
- (2) Situation: how people deal with, define or perceive the term. This point is based on the topics the study is based upon.
- (3) Perspectives: possible ways of defining a term (end)
- (4) I process (try): it (he, she) sequences of semes, flow of information, changes of meaning in the time.
- (5) Activities and events: difficulties to find definitions related the obsolescence or frequency of the different meanings.
- (6) Strategies: ways of managing information.
- (7) Relations and social structure: ideology presents in the standards of definitions.

After extracting the terms from the CREA (Sincronic Spanish Database) and CORDE (Diacronic Spanish Database) databases, we have proceeded to arrange the results from the definitions of the dictionaries. They are generally accessible *corpora*, accessible to everyone *via* Spanish Academy Web (<http://www.rae.es>).

Coming across the definitions and contexts, we will be able to observe the features describing men and women

The definitions which dictionaries offer about correlative terms (in masculine and feminine) do not coincide at informative levels, thus offering a diverse kind of *seme* in every case.

The aim of this project is to analyze the definitions that refer to terms related to women and men in different semantics systems and them to arrange when recounted to men and to analyze possible faults so much to macrostructure as well as microstructure level. We pretend to give a fully readable account of how dictionaries represent women and men.

The result of this study goes towards a standard of definition for the concepts included in the systems listed in the appendix.

### 3 Discussion

Gender, as we have seen, is globally understood as the set of beliefs, prescriptions and attributions that are socially constructed taking sexual difference as a base. This social construction works sometimes as a kind of cultural “filter”, one through which reality is interpreted, due to that tendency of every society to define what is proper for women and what is proper for men, and above this cultural framework, setting out the obligations of each sex.

From childhood we perceive representations of what is proper to each sex through language, and the materiality of culture (objects, images, etc.). It has been that children between two and three years old, know how to refer to themselves in feminine or masculine, but do not necessarily have a clear notion of the actual biological differences.

Today, the notion of ideology, born linked to *bourgeois* society in which a set of values and ideals, driven by political and social pluralism, led to today’s modern society. Social representations are symbolic constructions that give powers to the objective and subjective behavior of people. The social environment is more than a territory, a symbolic space defined by the imagination, and decisive in the construction of each person self-image, consciousness is inhabited by social discourse. Lucien Goldmann states in this regard that

“The overall vision of human relations between man and the universe implies, this type of collective consciousness, the possibility, and often the actual presence of an ideal man and this leads us to differentiate the type of collective consciousness that we call ideology, called view of the world “(Goldman, 1969: 210).

**Sociolinguistics:** Although the behavioral differences between men and women are explained in a general way as a product of sex (vocal cords, tonal range), gender has been linked to the position in society and the complex network of relationships that are developed within it.

**Anthropological:** The word gender attempts to rebuild each and every one of the areas of significance that have been superimposed for decades, unraveling the network of relationships and social interactions that are constructed from the symbolic division into sexes. In the psychological field, we also emphasize the neither natural nor spontaneous character of the categories male and female,

**Linguistic:** We aim at showing how grammar, with its two (and even up to 3) genera, provides us with the perfect field to focus on sexes, at both biological and social senses, on Nature and Culture, without favoring the existence of two different sexes nor any individual powers and decisions. The Academy notes that the Spanish language foresees the possibility to refer to mixed groups through the grammatical masculine gender, possibility in which there is no discriminatory intent, but the application of the linguistic law based in the expressive economy. Only when the opposition of the sexes is a relevant factor in the context, the Academy considers necessary the explicit presence of both genders.



## 4 Conclusion

Human beings symbolize a basic material, which is identical in every society: bodily difference, specifically sex. Although apparently biology shows that human beings are in both sexes, more combinations arise from the five physiological areas. Of these five areas depends on what, in general terms and in a very simplistic way, has been called the "biological sex" of a person: genes, hormones, gonads, internal reproductive organs and external reproductive organs (genitals).

Although the multitude of cultural representations of biological facts is very large and has varying degrees of complexity, sexual difference has some basic persistence and is the source of our image of the world, as opposed to some other. The body is the first uncontrollable evidence of human difference. The culture marks human beings with gender and gender marks perception of everything else: social, political, religious and quotidian.

The dictionary is an ideological creation. It reflects the society and the dominant ideology. As indisputable authority, as a cultural tool, the dictionary acts as a fixing element and intends to the conservation, not only of language but also the attitudes and ideology behind it.

## 5 References

- Arias Barredo, A. (1995): *De feminismo, machismo y género gramatical*, Valladolid, Universidad.
- Blecua, J.M. (1990): «Análisis provisional de una muestra aleatoria en el DRAE», en *El vocabulari i l'escrit*, Barcelona: Universidad de Barcelona, 1990.
- Calonge, J. (1981): "Implicaciones del género en otras categorías gramaticales". *Logos Semantikos, Studia Linguistica in honorem Eugenio Coseriu*, Madrid, Gredos, IV, 1991, pp. 19-28.
- Casares, J. (1992 [1950]): *Introducción a la lexicografía moderna*, Madrid, CSIC, 1992.
- Demonte, V. (1982a): *Lenguaje y sexo, ideología y papeles sociales*, Madrid, Akal.
- Goldman, Lucien. 1969. *The Human Sciences and Philosophy*. London: Cape.
- López García, Á. y Morant, R. (1991): *Gramática femenina*, Madrid, Cátedra.
- López García, Á. (1992): *Lenguaje y discriminación sexista en los libros escolares*, Murcia, Universidad
- Lozano Domingo, I. (2005): *Lenguaje femenino, lenguaje masculino*, Minerva: Madrid.
- Pascual J.A. y Olaguíbel, M.C. (1992): «Ideología y diccionario», en I. Ahumada Lara (ed.): *Diccionarios españoles: contenido y aplicaciones. Lecciones del I Seminario de Lexicografía Hispánica*, Facultad de Humanidades, Jaén, 21 al 24 de enero de 1991, *El Estudiante Facultad de Humanidades*, Jaén, 1992, pp. 73- 89.

### 5.1 Dictionaries

- Alvar Ezquerro, Manuel (dir.) (2000): *Diccionario para la enseñanza de lengua española*, Barcelona, Bibliograf / Universidad de Alcalá de Henares.
- Gutiérrez Cuadrado, Juan (dir.) (1996): *Diccionario Salamanca de la lengua española*, Madrid, Santularia/ Universidad de Salamanca.
- Moliner, María (1998): *Diccionario de uso del español*, Madrid, Gredos.

Real Academia Española (2001): Diccionario de la lengua española, Madrid, Espasa- Calpe, vigésima segunda edición.

## **Appendix**

Corpora: List of Words used in the research (Spanish Language).

### System 1: Women social respect

Distinguished: dama, damisela, dona, dueña, gran señora, madama, madamisela, madona, maestra, matrona, ricadueña, ricahembra, señora, señorita, señora principal, señorita.

Esteem: mujercilla, mujeruca, mujerzuela, mujeruca, pingo, prójima

### System 2: Women relationships

Lawful: dama, barragana, cara mitad, conyuge, consorte, costilla, esposa, desposada, media naranja, mujer, mujer velada, mujer de bendición, oponente, pareja, señora.

Unlawful: barragan, coima, combleza, compañera, concubina, daifa, entretenida, manceba, pretendida

### System 3: Women appearance

Seme +beuty: Beldad, belleza, bombón, gachí, gachona, hembra, hembra de bandera, hermosura, monumento, preciosidad, sílfide, venus.

Seme stocky: buena moza, moza, mujerona, real moza

Seme masculinity : machota, machirulo, marimacho, marota, varona, varonesa, virago.

### System 4: Women procreation

seme sterile: horra, mañera, machorra

seme fertile: descinta, embarazada, empuñada, encinta, gestante, gravida, madre, malparida, multipara, mulipara, parida, paridera, paridora, parturienta, parturiente, preñada, primeraza, primípara, puépera, recién parida.

### System 5: Women sexuality

seme virginity: doncella, doncellueca, escosa, entera, poncela, prematura, virgen

seme homosexuality: bollera, fricadora, lesbiana, tortillera, tribada

seme "sexual desires": cachonda, salida, ninfómana.

Seme " sex trade": prostituta, andorra, ave nocturna, bagasa, baldonada, bacanera, burraca, buscona...

### System 6: Women personality

Seme "dishonest" corralera, escaldada, facilona, farota, galante, mujer fatal, piruja, tigresa, tragona, vampiresa, ventanera,

Seme "gossip": alcahueta, celestina, comadre, lagarta, pécora, tercera, trotaconventos, víbora.

Seme "bad temper": arpía, mujerota, sargenta, sargentona

Seme "boastful": bachillera, coqueta, lechugina, marisabidilla, petimetra.

# Reflexive Verbs in a Valency Lexicon: The Case of Czech Reflexive Morphemes

Václava Kettnerová, Markéta Lopatková  
Charles University in Prague  
kettnerova@ufal.mff.cuni.cz, lopatkova@ufal.mff.cuni.cz

## Abstract

In this paper, we deal with Czech reflexive verbs from the lexicographic point of view. We show that the Czech reflexive morphemes *se* and *si* constitute different linguistic meanings: either they are formal means of the word formation process of the so called reflexivization, or they are associated with the syntactic phenomena of reflexivity, reciprocity, and diatheses.

All of these processes are associated with changes in the valency structure of verbs. We formulate a proposal for their lexicographic representation for the valency lexicon of Czech verbs, VALLEX. We make use of the division of the lexicon into a data component and a grammar component which represents a part of the overall Czech grammar. The data component stores information on valency structure of verbs in unmarked (active) structures. The grammar component consists of formal rules describing regular changes in the valency structure of verbs; these rules allow for the derivation of valency frames underlying the usages of verbs in marked structures (reflexive, reciprocal, deagentive and dispositional) from the valency frames corresponding to unmarked structures (non-reflexive, unreciprocal, and active).

Czech reflexive verbs thus represent an illustrative example of the lexical-grammar interplay: we demonstrate that a close interaction between the lexicon and the grammar is necessary for a representation of these verbs and they both are indispensable if such a representation is to be adequate and economical.

**Keywords:** reflexive verb; reflexive morpheme; valency lexicon; Czech

## 1 Introduction

In this paper, the possibility of a lexicographic representation of the reflexivity of Czech verbs is described in detail. In Czech, the reflexives *se* and *si* are on the one hand formal means of word formation process of so called reflexivization; on the other hand, they are associated with syntactic phenomena of reflexivity (in the narrow sense, also called “true reflexives”), reciprocity, and diatheses. According to their function, the reflexives *se* and *si* represent clitic morphemes corresponding either (ia) to components of verb lemmas (*se* and *si*), or (ib) to a component of verb form (only *se*), or (ii) to the personal pronoun *se* (with its inflected variant *si* and corresponding non-clitic variants *sebe* and *sobě*,

respectively). As examples (1), (2), (3) and (4) with the verb *zabít* “to kill” show, both types of reflexives can occur with a single verb, constituting different linguistic meanings: in example (1), *se* is a component of the verb lemma *zabít se* “to kill (oneself)” (type (ia)); in examples (2) and (3), *se* is interpreted as the reflexive personal pronoun (however, expressing different meanings, true reflexivity and reciprocity, respectively, type (ii)); and example (4) illustrates *se* as a component of a verb form of the verb *zabít* “to kill” (type (ib)).

- (1) *Zabil se pádem ze střechy.* (CNC, SYN2006pub)  
 “killed - SE<sub>morph</sub> - by falling - from roof.”  
 Eng. He killed himself (unintentionally) by falling from the roof.
- (2a) *Zabil se vlastní zbraní ...* (CNC, SYN2006pub)  
 “killed - SE<sub>pron</sub> - own - weapon ...”  
 Eng. He killed himself with his own weapon ...
- (2b) *Zabil sebe vlastní zbraní.*  
 “killed - SEBE<sub>pron</sub> - own - weapon ...”  
 Eng. He killed himself with his own weapon ...
- (3a) *Zabíli se navzájem.*  
 “killed - SE<sub>pron</sub> - each other.”  
 Eng. They killed each other.
- (3b) *Zabíli sebe navzájem.*  
 “killed - SEBE<sub>pron</sub> - each other.”  
 Eng. They killed each other.
- (4) *... zabila se dvě vykrmená prasata a pečínka provoněla celý dům.* (CNC, SYN2005)  
 “... killed - SE<sub>morph</sub> - two - fattened - pigs - and - roast meat - scented - the whole - house.”  
 Eng. Two fattened pigs were killed and roast meat scented the whole house.

We show that the reflexives *se* and *si* belong to several different language phenomena which involve specific changes in the valency structure of verbs. As a consequence, they require to be represented in a lexicon in different ways. Here we describe the representation of these phenomena in the Valency Lexicon of Czech Verbs, VALLEX.

## 1.1 Related Work

Reflexivity has been extensively studied in the theoretical linguistics since the 1980s. The research has focused on linguistic means encoding reflexivity, their interpretations and ambiguities in individual languages. Recently, this linguistic phenomenon has received considerable attention even from the cross-linguistic perspective (König & Gast, 2008), (Nedjalkov, 2007). Numerous analyses show that linguistic means expressing reflexivity are usually ambiguous as they fulfill diverse functions in in-

dividual languages and that drawing clear distinctions between these functions represents a tricky task. For these reasons, developing a satisfactory lexicographic representation of reflexivity – despite being highly beneficial esp. for natural language processing and foreign learners – remains rather challenging (Renau & Battaner, 2012).

In Czech, reflexivity encoded by the reflexives *se* and *si* represent widely debated phenomenon from both theoretical (Oliva, 2001; Panevová, 1999, 2007) and computational point of view (Oliva, 2003; Petkevič, 2013). Two primary functions of the Czech reflexives are determined: (i) the reflexives as components of verb lemmas or verb forms and (ii) the reflexives as the personal pronoun, see Section 1. However, despite the plenitude of studies focused on Czech reflexivity, testable criteria for their distinction have not yet been established. In most cases, the substitutability of the reflexives *se* and *si* with *sebe* and *sobě*, respectively, can be applied as an operational test for distinguishing the reflexive personal pronoun from *se* and *si* as the components of verb lemmas or verb forms. However, this test can fail esp. in cases where the substitution leads to stylistically unacceptable sentences or in cases of haplology, see Section 2.1. In such cases, we adopt solutions taking economy and systematicity of the lexicographic representation into account.

As we attempt a lexicographic representation of reflexivity in a lexicon, let us introduce several lexical resources providing the information on these phenomena. First, *LexIt*, a large-scale lexical resource providing the automatically derived information on subcategorization and semantic properties of Italian verbs, nouns and adjectives stores the information on reflexivity as well (Lenci et al., 2012). Second, this type of information is also covered in *Diccionario de enseñanza del español como lengua extranjera*, *DAELE*, a Spanish learner's dictionary (Renau & Battaner, 2012). Third, *FrameNet* records the information on reciprocity of frame elements (linguistic phenomenon closely related to reflexivity) by adding special semantic frames indicating reciprocity (Ruppenhofer, et al., 2010).

For Czech, *PDT-VALLEX*, a valency lexicon linked with word occurrences in the Prague Dependency Treebank 2.0 (PDT) (Hajič, et al., 2006), provides the information on valency behavior of verbs, nouns, adjectives and adverbs (Hajič, et al., 2003). Although the information on reflexivity and reciprocity of verbs is not explicitly recorded in this lexicon, it can be easily extracted from PDT (if reflexive or reciprocal usages appear in the corpus). In addition, a fully automatically derived *Czech Syntactic Lexicon* (which is however not publically available) was designed, providing the information on possible reciprocity, reflexivity and diatheses of verbs (Skoumalová, 2001).

## 1.2 VALLEX

The valency lexicon of Czech verbs, *VALLEX*,<sup>1</sup> is a collection of linguistically annotated data and documentation (Žabokrský & Lopatková, 2007; Lopatková et al., 2008). It provides the information on valency structure of Czech verbs in their particular meanings / senses, possible morphological forms of

---

1 <http://ufal.mff.cuni.cz/vallex/>

their valency complementations and additional syntactic information accompanied with glosses and examples. In VALLEX, version 2, there are roughly 2,730 lexeme entries containing together around 6,460 lexical units ('senses'). Verb lexemes were selected according to their frequency in the Czech National Corpus.<sup>2</sup> The lexicon has been developed for both human users and NLP applications, and is therefore in three different formats: HTML, XML and printable versions.

In VALLEX, the valency theory developed within the theoretical framework of the Functional Generative Description (henceforth FGD) is used as the theoretical background for the description of valency of verbs, see esp. (Sgall et al., 1986), (Panevová, 1994). According to this theory, valency complementations are divided into arguments (inner participants) and free modifications (adjuncts). They both can be obligatory or optional. The types of (verbal) arguments are distinguished mainly on the basis of the syntactic behavior of verbs. Five types of arguments have been determined – 'Actor' (ACTor, label ACT), 'Patient' (PATient, PAT), 'Addressee' (ADDRessee, ADDR), 'Origin' (ORIGin, ORIG), and 'Effect' (EFFect, EFF). In contrast to the arguments, free modifications are semantically distinctive, being identified on the basis of their syntactico-semantic functions.

In VALLEX, the key information on the valency structure of a given lexical unit is encoded in the form of valency frames. A valency frame is formed as a sequence of slots; each slot stands for one valency complementation and consists of its type ('ACTor', 'ADDRessee', etc.), possible morphemic forms and its obligatoriness (obligatory or optional). Further, each lexical unit can be characterized by additional syntactic information on, e.g., syntactico-semantic class membership, diatheses, reciprocity of valency complementations, reflexivity. This information is provided in special attributes attached to individual lexical units.

The lexicon is divided into the data and the grammar component; the latter stores rules describing regular syntactic properties of verbs and it represents a part of the overall grammar of Czech, see esp. (Kettnerová et al., 2012a). The close interplay of these two parts of the lexicon is demonstrated on the representation of the reflexives *se* and *si* in the following sections. First, the reflexives *se* and *si* as components of verb lemmas (i.e., as formal means of the word formation process of reflexivization) are discussed in detail, esp. the syntactic properties of reflexivization are described and their lexicographic description is outlined in Section 2. Second, the reflexive pronoun *se* as a formal means of reflexivity (in the narrow sense) and reciprocity is surveyed and its representation in the lexicon is introduced in Section 3. Finally, the reflexive *se* as a component of reflexive verb forms that is involved in two types of Czech diatheses is debated in Section 4. In conclusion, the lexical entry of the verb *zabít* "to kill" (illustrated in Section 1) is displayed.

---

2 <http://www.korpus.cz/>

## 2 Czech Morphemes *se* and *si* as Components of Verb Lemmas

In Czech, the reflexive morphemes *se* and *si* can represent *freestanding components of verb lemmas* of the so called *reflexive verbs*. In Czech, there are two types of reflexive verbs: (i) reflexive tantum verbs (Section 2.1) and (ii) derived reflexive verbs (Section 2.2).

### 3 Reflexive Tantum Verbs

The first type is represented by the so-called reflexive tantum verbs, i.e., the verbs that have no non-reflexive counterparts, e.g., *bát se* “to be afraid”, *smát se* “to laugh”, *stěžovat si* “to complain”, *zapamatovat si* “to remember”, *domnívat se* “to assume”, *chlubit se* “to boast”, *líbit se* “to like”, *ptát se* “to ask”, *zamilovat se* “to fall in love”, see examples (5) and (6). In the case of reflexive tantum verbs, the reflexive morpheme *se* or *si* is a part of the verb lemma representing the respective verb lexeme in the data component of the lexicon.

Moreover, there are cases in Czech where a reflexive verb seemingly has a non-reflexive counterpart but these reflexive and non-reflexive verbs are not related by any derivational relation: the lexical meanings of these verbs are completely different, e.g. *dit se* “to happen” vs. *dit* “to tell”, *dopustit se* “to commit” vs. *dopustit* “to fill (with water)”, *hodit se* “to match” vs. *hodit* “to throw”, see examples (7)-(8). These are represented as separate verb lexemes (in separate lexical entries) in the valency lexicon.<sup>3</sup>

(5a) *Jan se bojí zkoušky.*

“John – SE<sub>morph</sub> – is afraid – of the exam.”

Eng. John is afraid of the exam.

(5b) \**Jan bojí zkoušky.*

“John – is afraid – of the exam.”

(6a) *Hosté si stěžovali na špatnou stravu v hotelu.*

“guests – SI<sub>morph</sub> – complained – of bad food – at the hotel”

Eng. The guests complained about bad food at the hotel.

(6b) \**Hosté stěžovali na špatnou stravu v hotelu.*

“guests – complained – of bad food – at the hotel”

(7) *Co se děje?*

“what – SE<sub>morph</sub> – happens.”

Eng. What is happening?

---

3 In case where a sentence contains more than one reflexive tantum verbs, the reflexive *se* can be subject to haplogy: a single occurrence of the reflexive can be associated with two verbs, see the following example where both the verb *pokusit se* “to try” and *usmát se* “to smile” are reflexive tantum verbs:

*Jan se pokusil usmát.*

“John – SE<sub>morph</sub> – tried – smile.”

Eng. John tried to smile.

- (8) *“Pravdu díš,” odpověděl Petr.*  
 “the truth – you are telling – replied – Peter.”  
 Eng. “You are telling the truth,” replied Peter.

## 4 Derived Reflexive Verbs

Reflexive verbs of the second type are derived from non-reflexive verbs by adding the freestanding morpheme *se* or *si*. This process is called reflexivization, see esp. (Dokulil, 1986). In Czech, the reflexivization is a productive word formation process, which is largely syntactically motivated. Basically, two types of changes in valency structure of verbs are associated with this process (Sections 2.2.1 and 2.2.2). Further, in rare cases, reflexivization does not involve any change in valency structure (Section 2.2.3).

### 4.1 Reflexivization Applied to Transitive Verbs Resulting in Reflexive Intransitive Verbs

When reflexivization is applied to transitive verbs, it results in reflexive intransitive verbs associated with specific shifts in the lexical meaning of verbs: whereas non-reflexive transitive verbs express intentional acts (9a), reflexive intransitive verbs prototypically indicate non-intentional acts (9b):<sup>4</sup> the argument corresponding to the direct object (expressed by the accusative) of the transitive non-reflexive verb maps onto the subject (expressed by nominative) of the derived intransitive reflexive verb.

- (9a) *Maminka vaří brambory.*  
 “mother – cooks – potatoes<sub>Dobj-acc.</sub>”  
 Eng. The mother is cooking potatoes.
- (9b) *Brambory se vaří.*  
 “potatoes<sub>Subj-nom</sub> – SE<sub>morph</sub> – cook.”  
 Eng. Potatoes are cooking.

### 4.2 Reflexivization Applied to Verbs Implying Reciprocity

Reflexivization is also involved in the derivation of reflexive reciprocal verbs, i.e., verbs indicating reciprocity in their lexical meaning (Panevová & Mikulová, 2007). These verbs are derived by the reflexive morphemes *se* or *si* from verbs that imply (at least two) semantically homogeneous arguments, typically structured as ACTor (in nominative) and PATient or ADDRessee (expressed either by the

4 The act expressed by verbs denoting movement can be conceived as intentionally or unintentionally performed, e.g., *Petr opřel kolo o zeď*. Eng. Peter leaned the bike against the wall. (intentional act) – *Petr se opřel o zeď*. Eng. Peter leaned against the wall. (un/intentional act).



accusative or by the dative), see examples (10a) and (11a), respectively. (As for reflexive pronouns in reciprocal constructions, see esp. Section 3.2.)

Reflexive verbs indicating reciprocity are associated with specific changes in their valency structure: the argument of the non-reflexive verb that is expressed in the accusative or dative is expressed by a prepositional group with the reflexive verb indicating reciprocity, see examples (10b) and (11b), respectively.

(10a) *Petr potkal Marii.*

“Peter – met – Mary<sub>PAT-acc</sub>.”

Eng. Peter met Mary.

(10b) *Petr se potkal s Marií.*

“Peter – SE<sub>morph</sub> – met – with Mary<sub>PAT-s+instr</sub>.”

Eng. Peter met with Mary.

(11a) *Dědeček vypráví dětem pohádky.*

“the grandpa – tells – the children<sub>ADDR-dat</sub> – fairy tales.”

Eng. The grandpa is telling the children fairy tales.

(11b) *Dědeček si vypráví s dětmi pohádky.*

“the grandpa – SI<sub>morph</sub> – tells – with the children<sub>ADDR-s+instr</sub> – fairy tales”

Eng. The grandpa and the children are telling each other fairy tales.

### 4.3 Reflexivization without Changes in Valency Structure

For a limited number of verbs, reflexivization does not result in any changes in either the valency structure or the meaning. The derivation by the morphemes *se* or *si* without clear syntactic or semantic motivation can be illustrated by the following examples (12) and (13).

(12a) *Myslím, že je to dobře.*

“I think – that – is – it – good.

Eng. I think that it is good.

(12b) *Myslím si, že je to dobře.*

“I think – SI<sub>morph</sub> – that – is – it – good.”

Eng. I think that it is good.

(13a) *Zítřa začíná ve městě festival vína.*

“tomorrow – starts – in the town – a festival of wine.”

Eng. Tomorrow a wine festival starts in the town.

(13b) *Zítřa se ve městě začíná festival vína.*

“tomorrow – SE<sub>morph</sub> – in the town – starts – a festival of wine.”

Eng. Tomorrow a wine festival starts in the town.

## 5 Representation of Reflexive Tantum and Derived Reflexive Verbs in the Lexicon

In the case of both *reflexive tantum verbs* and *derived reflexive verbs*, the reflexive morphemes *se* and *si* are represented in the *data component* of the lexicon as a part of their verb lemmas (Section 2.1, 2.2.1 and 2.2.2, respectively). Derived reflexive verbs and their non-reflexive counterparts are recorded as separate verb lemmas (and thus separate lexical entries). Only derived reflexive verbs without syntactic changes (Section 2.2.3) are handled as variants of the respective non-reflexive verbs.

## 6 Czech Morphemes *se* and *si* as a Reflexive Pronoun

The reflexive *se* can also represent a *personal pronoun* (with the morphemic form *se* for accusative and *si* for dative, and their non-clitic variants *sebe* and *sobě*, respectively). The reflexive pronoun expresses reflexivity (in the narrow sense, Section 3.1) and reciprocity (Section 3.2).

### 6.1 Reflexivity

In cases where ACTor performs an action that is focused on himself/herself (also called “true reflexivity”), the reflexive pronoun *se* is used in Czech as a formal means of grammatical coreference, see esp. (Hajičová, et al., 1985, 1986, 1987): in these cases, the reflexive pronoun *se* stands for an argument of the verb that is referentially identical with ACTor in the subject position, examples (14) and (15). The form of the reflexive pronoun depends on the morphemic case of the argument (*se* in the accusative and *si* in the dative). In the case of reflexivity, the clitic forms of the reflexive pronoun *se/si* can be replaced by their non-clitic variants *sebe/sobě*:<sup>5</sup>

(14a) *Petr se myje.*

“Peter<sub>ACT-Subj</sub> – SE<sub>pron-acc</sub> – washes.”

Eng. Peter is washing himself.

(14b) *Petr myje sebe (ale ne dítě).*

“Peter<sub>ACT-Subj</sub> – washes – SEBE<sub>pron-acc</sub> – (but not the child).”

Eng. Peter is washing himself (but not the child).

(15a) *Marie si koupila k obědu sendvič.*

“Marie<sub>ACT-Subj</sub> – SI<sub>pron-dat</sub> – bought – for lunch – a sandwich.”

Eng. Mary bought herself a sandwich for lunch.

---

5 The use of clitic and non-clitic variants of the reflexive pronoun is affected esp. by the topic-focus articulation – thus the possibility to replace clitic forms by non-clitic forms of the reflexive pronoun in a sentence is often conditioned by changes in word order; however, this issue is not addressed in this paper as it goes beyond its scope.

- (15b) *Sobě k obědu Marie koupila sendvič, dětem hranolky.*  
“SOBĚ<sub>pron-dat</sub> - for lunch - Marie<sub>ACT-Subj</sub> - bought - a sandwich, - to the children - French fries”  
Eng. Mary bought a sandwich to herself and French fries to children for lunch.

Reflexivity is represented in the lexicon by a special attribute -rfl attached to relevant lexical units. In this attribute, the information about the possibility of the reflexive usage of some arguments is provided by the value cor3 (for arguments in the dative, example (15)) and cor4 (for arguments in the accusative), example (14)). Other forms (e.g., prepositional groups) are not explicitly marked in the lexicon as they are expressed only by long variants of the reflexive personal pronoun (which are not ambiguous).

## 6.2 Reciprocity

Further, the reflexive pronoun *se* can express reciprocity. Reciprocalization is a syntactic operation on two (or three) arguments of a verb which puts the involved arguments in the symmetry. The main conditions imposed on such arguments are (i) their semantic homogeneity and (ii) same status with respect to topic-focus articulation. Reciprocalization leads to specific changes in the valency structure of a verb: the involved argument expressed in a less prominent surface syntactic position is shifted to the more significant position (subject or direct object) of the other symmetrically used argument, see (Panevová, 1999, 2007) and (Panevová & Mikulová, 2007). The resulting surface syntactic structure is characterized by a “multiplied” subject (or direct object) which is filled by a coordination, example (16), morphological, example (17), or semantic plural (e.g., the collective noun in example (18)). The syntactic position of the shifted (less significant) argument is typically formally filled by the reflexive pronoun *se* (expressed in the appropriate case), see below.

- (16a) *Petr a Pavel se bijí.*  
“Peter - and - Paul - SE<sub>pron-acc</sub> - beat.” “  
Eng. Peter and Paul are beating each other.
- (16b) *Petr a Pavel bijí sebe navzájem.*  
“Peter - and - Paul - beat - SEBE<sub>pron-acc</sub> - each other.”  
Eng. Peter and Paul are beating each other.
- (17a) *Děti se bijí.*  
“children - SE<sub>pron-acc</sub> - beat.”  
Eng. Children are beating each other.
- (17b) *Děti bijí sebe navzájem.*  
“children - beat - SEBE<sub>pron-acc</sub> - each other.”  
Eng. Children are beating each other.
- (18a) *Celá rodina si pomáhá.*  
“whole - family - SI<sub>pron-dat</sub> - help

Eng. The whole family helps each other.

(18b) *Rodina pomáhá sobě (navzájem – a ne jim).*

“family – helps – SOBĚ<sub>pron-dat</sub> (each other – and not them).”

Eng. The family helps each other.

In Czech, reciprocal constructions are created by two different types of verbs, by non-reciprocal verbs, i.e., by verbs that do not imply reciprocity in their lexical meaning (Section 3.2.1), and by inherently reciprocal verbs (Section 3.2.2).

### 6.2.1 Verbs Not Implying Reciprocity

Many verbs in Czech can potentially express reciprocity although reciprocity is not implied in their lexical meaning,<sup>6</sup> e.g., *děkovat* “to thank”, *obviňovat* “to accuse”, *hrozit* “to threaten”, *pomáhat* “to help”, *vydírat* “to blackmail”, examples (19) and (20). In such cases, the reciprocal constructions (as described above) are optionally accompanied with the lexical expressions *vzájemně*, *navzájem*, *jeden druhý* “each other”, and *spolu* “together”, emphasizing the reciprocal meaning.

(19) *Manželé se (vzájemně) obviňují z nevěry.*

“husband and wife – SE<sub>pron-acc</sub> – (each other) – accuse – of infidelity.”

Eng. Husband and wife accuse each other of infidelity.

(20) *Otec a syn si (vzájemně) lhali, aby si neublížili.*

“father and son – SI<sub>pron-dat</sub> – (each other) – lied – in order to – SI<sub>pron-dat</sub> – not-to-hurt.”

Eng. Father and son lied (to each other) in order not to hurt each other.

The lexical expressions (explicitly) indicating reciprocity are, however, obligatory in reciprocal constructions created by reflexive tantum verbs that do not imply reciprocity. For instance, although the verbs *smát se* and *vysmívat se* “to laugh at” do not imply reciprocity, their ACTor and ADDressee can be put in the symmetrical relation. In reciprocal constructions with these verbs, *se* represents the morpheme that is a component of the verb lemmas, not the reflexive pronoun (as it cannot be substituted with the non-clitic form *sebe*). The reciprocity therefore must be expressed by lexical means, see example (21a) and (22a). If no such lexical means is present, the construction is either not reciprocal (21b), or even not grammatical (22b):

(21a) *Petr a Pavel se smáli jeden druhému.*

“Peter and Paul – SE<sub>morph</sub> – laughed – at each other.”

Eng. Peter and Paul were laughing at each other.

(21b) *Petr a Pavel se smáli.*

“Peter and Paul – SE<sub>morph</sub> – laughed.”

Eng. Peter and Paul were laughing.

6 Compare also with Section 2.2.2 describing derived reflexive verbs that imply reciprocity in their lexical meaning.

(22a) *Petr a Pavel se vysmívali jeden druhému.*  
 “Peter and Paul – SE<sub>morph</sub> – laughed – at each other.”  
 Eng. Peter and Paul were laughing at each other.

(22b) *\*Petr a Pavel se vysmívali.*  
 “Peter and Paul – SE<sub>morph</sub> – laughed.”

### 6.2.2 Verbs Implying Reciprocity

In addition, reciprocalization can be also applied to inherently reciprocal verbs. There are basically two types of such verbs: (A) *verbs indicating reciprocity in their lexical meaning* that might undergo the derivation of reflexive verbs indicating reciprocity (see Section 2.2.2) and (B) *non-reflexive verbs that imply reciprocity* of some of their arguments *in their lexical meaning* (e.g., *soupeřit* “to fight”, *sousedit* “to neighbor”).

In the case of (A), in addition to the derivation of reflexive reciprocal verbs (with the change of an accusative or dative complement into a prepositional group (see Section 2.2.2 and example (10), here repeated as (23)), the verbs can undergo “standard” reciprocal derivation, as in example (24a).

Here the question arises whether reciprocal constructions are derived from the non-reflexive *potkat* “to meet” (23a) or the reflexive verb *potkat se* “to meet” (23b). From the theoretical point of view, this question remains still open. The reflexive *se* in these constructions can be seen as the reflexive pronoun (as it can be replaced by its non-clitic variant *sebe* (24b)); however, the possible haplogy of the morpheme *se* (Petkevič, 2013) makes the interpretation complicated and the theoretical interpretation of the derivation of these reciprocal constructions can differ, see also (Panevová, 2007).

For the practical implementation in the lexicon, we propose to derive the reciprocal constructions as in (24a) from the non-reflexive verbs (*potkat* for this case), not from its reflexive counterpart (*potkat se*) since this proposal allows us to use a single derivational rule for both types of verbs allowing for reciprocity (Section 3.2.1 and 3.2.2A).

(23a) *Petr potkal Marii.*  
 “Peter – met – Mary<sub>Pat-acc</sub>.”  
 Eng. Peter met Mary.

(23b) *Petr se potkal s Marií.*  
 “Peter – SE<sub>morph</sub> – met – with Mary<sub>PAT-s+instr</sub>.”  
 Eng. Peter met with Mary.

(24a) *Petr a Marie se potkali (navzájem).*  
 “Peter and Mary – SE<sub>pron-acc</sub> – met (each other)”  
 Eng. Peter and Mary met (each other).

(24b) *Petr a Marie potkali sebe navzájem i další přátele.*  
 “Peter and Mary met – SEBE<sub>pron-acc</sub> – each other – and other friends.”  
 Eng. Peter and Mary met (each other) as well as other friends.

In the case of (B), *non-reflexive verbs*, changes in the valency structure (i.e., the “multiplication” of subject) are sufficient markers of reciprocity, and the reflexive pronoun is not present in the reciprocal structures (25).

(25) *Týmy ČR soupeří o postup do finále. (= tým s týmem // mezi sebou/spolu)*

Eng. The teams of the Czech Republic fight for the finals. (= each team with other teams)

### 6.3 Reciprocity in the Lexicon

Reciprocity of arguments is described in the data component of the lexicon in a special attribute -rcp providing the list of the arguments that can enter the symmetrical relation. Changes in the valency structure of verbs (including the use of lexical means for expressing reciprocity) are regular enough to be captured by formal rules. These rules are stored in the grammar component of the lexicon and they make it possible to automatically derive valency frames underlying reciprocal constructions, see (Kettnerová et al., 2012b).

## 7 Czech Morpheme *se* as a Component of a Reflexive Verb Form

Finally, the reflexive *se* can represent a freestanding *morpheme* that is a component of a reflexive verb form in two types of diatheses in Czech: (i) the deagentive diathesis (Section 4.1) and (ii) the dispositional diathesis (Section 4.2). Diatheses are relations between syntactic structures of a verb which differ in the grammatical category of voice, i.e., they are associated with specific morphological meanings of a verb.

In Czech, five specific morphological meanings are determined: passive, deagentive, resultative, dispositional, and recipient-passive meanings (Panevová et al., in print). The surface structure of a verb with active voice is considered to be the unmarked member of a diathesis, whereas the structure with the given verb characterized by some of the five above given meanings constitutes its marked member. Whereas the passive, resultative and recipient-passive meanings of verbs are formed by the auxiliary verbs *být* (passive and resultative d.), *mít* (resultative d.), and *dostat* (recipient-passive d.), respectively, plus past participle of a lexical verb, the deagentive and dispositional diatheses are associated with the reflexive verb form. This form is constituted by the active form of a verb and the freestanding morpheme *se*.

## 8 Deagentive Diathesis

Marked members of the deagentive diathesis prototypically imply agentive ACTor of the event expressed by the verb; however, the ACTor is never expressed in the surface structure. The use of the deagen-

tive meaning of a verb results in specific changes in its valency structure. In Czech, deagentive meaning can be applied to both transitive and intransitive verbs; the reflexive verb form is limited to 3<sup>rd</sup> person.<sup>7</sup> For a transitive verb, (i) ACTor is shifted from the subject position (expressed in the nominative) and (ii) the subject is filled with the argument of the verb corresponding to the direct object in the unmarked construction (prototypically an accusative object), as in (26). For an intransitive verb, the shift of ACTor away from the subject position results in a subjectless surface structure in which the verb has prototypical form of 3<sup>rd</sup> sg neutrum, as in (27).

(26a) *Dělníci opravují silnici.*

“workers<sub>ACT-Subj-nom</sub> – repair – the road<sub>PAT-Dobj-acc</sub>.”

Eng. Workers repair the road.

(26b) *Silnice se opravuje.*

“the road<sub>PAT-Subj-nom</sub> – SE<sub>morph</sub> – repairs.”

Eng. The road is being repaired.

(27a) *Lidé na večírku tančili.*

“People<sub>ACT-Subj-nom</sub> – at the party – danced.”

Eng. People danced at the party.

(27b) *Na večírku se tančilo.*

“at the party – SE<sub>morph</sub> – danced<sub>3rd-sg-neutr</sub>.”

Eng. At the party, there was some dancing.

## 9 Dispositional Diathesis

As in the case of the deagentive diathesis, the marked members of the dispositional diathesis indicate human ACTor that is shifted from the subject position (in the nominative). This position is filled by the argument corresponding to the direct object position of a transitive verb, see example (28); in the case of an intransitive verb, dispositional meaning of the verb results in a subjectless surface structure (with the 3<sup>rd</sup> sg neutrum verb form), see example (29). In contrast to the deagentive diathesis, ACTor can be optionally expressed as an indirect object expressed by the dative. The marked members of

7 The status of several constructions with 2nd person (i) and 1st person (ii) is rather unclear, as they can be interpreted as either deagentive (with the grammatical morpheme *se*) (i), or reflexive (with the reflexive pronoun *se*) (ii).

(i) *Odsuzujete se k pěti letům vězení.*

“Sentence<sub>2nd-pl-masc/fem</sub> – SE<sub>morph/pron</sub> – to five year’s imprisonment.”

Eng. You are sentenced to five year’s imprisonment. // You sentence yourself to five year’s imprisonment.

(ii) *Léčím se u doktora Nováka.*

“Treat<sub>1st-sg</sub> – SE<sub>morph/pron</sub> – at the doctor Novák.”

dispositional diathesis are characterized by the presence of evaluative adverbs; thus if ACTor is expressed in the surface structure, it can be interpreted as an evaluator, see examples (28) and (29).

(28a) *Petr četl tuto knihu.*

“Peter<sub>ACT-Subj-nom</sub> - read - this book<sub>PAT-Obj-acc</sub>”

Eng. Peter read this book.

(28b) *Tato kniha se (Petrovi) dobře četla.*

“this book<sub>PAT-Subj-nom</sub> - SE<sub>morph</sub> - (Peter<sub>ACT-IObj-dat</sub>) - well - read.”

Eng. This book read well.

(29a) *Já jsem tam spal.*

“I<sub>ACT-Subj-nom</sub> - am - there - slept.”

Eng. I slept there.

(29b) *Spalo se (mi) tam dobře.*

“slept<sub>3rd-sg-neutr</sub> - SE<sub>morph</sub> - (me<sub>ACT-IObj-dat</sub>) - there - well.”

Eng. I slept well there.

## 10 Representation in the Lexicon

The changes in the valency structure of verbs in both deagentive and dispositional diatheses involve changes in morphemic forms of the arguments affected by surface shifts. These changes are regular enough to be captured by formal rules, which are stored in the grammar component of the lexicon. In the data component, only valency frames corresponding to the unmarked (active) uses of a verb are recorded. The optional attribute *-diat* is attached to each relevant valency frame in the data component (see Figure 1); it provides the information on applicability of the specific morphological meaning(s). On the basis of the formal rules (see Figure 2), the valency frames corresponding to marked structures of diatheses can be automatically derived (Kettnerová et al., 2012b).

- **lemma:** *zabíjet*<sup>impf</sup>, *zabít*<sup>pf</sup> 'to kill'  
 - **gloss:** impf: *usmrcovat* pf: *usmrtit* 'to cause death'  
 - **frame:** ACT<sub>1</sub><sup>obl</sup> PAT<sub>4</sub><sup>obl</sup>  
 - **example:** impf: *zabíjet někoho nožem; zabijeli mi manžela před očima*  
 pf: *zabít někoho nožem; zabili mi manžela před očima*  
 'to kill sb with a knife; they killed my husband before my eyes'  
 - **rfl: cor4:** impf: *zabíjel se z nešťastné lásky několikrát do roka*  
 'he used to kill himself several times a year'  
 pf: *zabíjel se z nešťastné lásky*  
 'he killed himself'  
 - **rep: ACT-PAT:** impf: *zabíjeli se navzájem* pf: *zabili se navzájem*  
 'they killed each other'  
 - **diat: Deagent:** impf: *o masopustu se každoročně zabíjí prase*  
 pf: *o masopustu se zabílo*  
 'during the carnival a pig is killed yearly'  
**Disp:** impf: *prase se řezníkovi špatně zabíjelo*  
 'the pig killed bad'



<p>- <b>lemma:</b> <i>zabíjet<sup>impf</sup> se, zabít<sup>pf</sup> se</i> 'to kill oneself, to die'                  - <b>gloss:</b> impf: <i>umírat; usmrcovat se</i> pf: <i>zemřít; usmrtit se</i> 'to die'                  - <b>frame:</b> ACT<sub>1</sub><sup>obl</sup>                  - <b>example:</b> impf: <i>každý rok se na hřištích zabije několik dětí</i>                                    'each year several children die (by accident) at the playground'                                    pf: <i>zabil se na kole při havárii</i>                                    'he died during the car accident with his bicycle'</p>
---

**Figure 1: Example of two lexical units of two lexemes *zabíjet<sup>impf</sup>*, *zabít<sup>pf</sup>* “to kill” and *zabíjet<sup>impf</sup> se, zabít<sup>pf</sup> se* “to kill oneself, to die” in the data component of the Valency Lexicon of Czech Verbs, VALLEX.**

## 11 Conclusion

We have discussed the possibility of a lexicographic representation of Czech reflexive verbs. We have shown that reflexivity should be described in different ways, depending on the function of the reflexive morphemes *se* and *si*: (i) Reflexive tantum verbs and derived reflexive verbs (where *se/si* is a part of a verb lemma) are stored as separate verb lexemes (represented by separate verb lemmas) in the data component of the lexicon. (ii) The possibility of a verb to be used in reflexive constructions (in the narrow sense) and in reciprocal constructions (where *se* is the personal pronoun coreferring to the subject) is marked in the special attributes *-rfl* and *-rcp*, respectively, assigned to relevant lexical units in the data component. For reciprocal constructions, formal rules stored in the grammar component make it possible to automatically derive respective valency frames. (iii) Similarly, the possibility of a verb to undergo deagentive or dispositional diathesis (where *se* is part of the verb form) is marked in the attribute *-diat* assigned to each relevant lexical unit in the data component; formal rules stored in the grammar component enable the derivation of the valency frames underlying the marked members of the diatheses.

Type:			Commentary
Deagent			
Action	verbform	replace (active vf → reflexive vf)	(1)
	ACT	delete (nom → ∅)	(2)
	PAT	replace (acc → nom)	(3)

Commentary:

- (1) The verb form changes from active to reflexive (adding the reflexive *se*).
- (2) ACTor cannot be expressed in the surface syntactic structure.
- (3) The morphemic expression of PATient changes from the accusative into thenominative (and shifts to the subject position).

**Figure 2: Example of a formal rule in the grammar component of the Valency Lexicon of Czech Verbs, VALLEX – the rule describing the deagentive diathesis.**

## 12 References

- Dokulil, M. (1986). *Mluvnice češtiny I*. Praha: Academia.
- Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová, V. & Pajas, P. (2003). PDT-VALLEX: Creating a Large-Coverage Valency Lexicon for Treebank Annotation. In Proceedings of the Second Workshop on Treebanks and Linguistic Theories, 15-15 November. Vaxjö, Sweden, pp. 57-68.
- Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J. & Mikulová, M. (2006). Prague Dependency Treebank 2.0. Philadelphia, PA: Linguistic Data Consortium. LDC2006T01.
- Hajičová, E., Panevová, J. & Sgall, P. (1985, 1986, 1987). Coreference in the Grammar and in the Text. Part I. In *The Prague Bulletin of Mathematical Linguistics*, 44, pp. 3-22. Part II. In *The Prague Bulletin of Mathematical Linguistics*, 46, pp. 1-11. Part III. In *The Prague Bulletin of Mathematical Linguistics*, 48, pp. 3-12.
- Kettnerová, V., Lopatková, M. & Bejček, E. (2012a). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In *Proceedings of the 15th Euralex International Congress 2012, 7-11 August 2012*. University of Oslo, Norway, pp. 434-443.
- Kettnerová, V., Lopatková, M. & Urešová, Z. (2012b). The Rule-Based Approach to Czech Grammaticalized Alternations. In *Proceedings of the 15th International Conference Text, Speech, Dialogue 2012, 3-7 September 2012*. Masaryk University, Czech republic, pp. 158-165.
- Lenci, A., Lapesa, G. & Bonansinga, G. (2012) LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of LREC 2012*, pp. 3712- 3718.
- Lopatková, M., Žabokrtský, Z., & Kettnerová, V. (2008). *Valenční slovník českých sloves*. Praha: Nakladatelství Karolinum.
- König, E., Gast, V. (2008). *Reciprocals and Reflexives: Theoretical and Typological Explorations*. Berlin, New York: Mouton de Gruyter.
- Nedjalkov, V. (2007). *Typology of Reciprocal Constructions*. Amsterdam: Benjamins.
- Oliva, K. (2001). Reflexe reflexivity reflexive. In *Slovo a slovesnost*, 57, pp. 200-207.
- Oliva, K. (2003). Linguistics-based PoS-tagging of Czech: disambiguation of *se* as a test. In *Contributions of the 4<sup>th</sup> European Conference on Formal Description of Slavic Languages, 28-30 November 2001*. Postdam University, Germany, pp. 299-314.
- Panevová, J. (1994). Valency Frames and the Meaning of the Sentence. In P.A. Luelsdorff (ed.) *The Prague School of Structural and Functional Linguistics*. Amsterdam, Philadelphia: John Benjamins Publishing Company, pp. 223-243.
- Panevová, J. (1999). Česká reciproční zájmena a slovesná valence. In *Slovo a slovesnost*, 60, pp. 269-275.
- Panevová, J. (2007). Znovu o reciprocitě. In *Slovo a slovesnost*, 68, pp. 91-100.
- Panevová, J., Mikulová, M. (2007). On Reciprocity. In *The Prague Bulletin of Mathematical Linguistics*, 87, pp. 27-40.
- Panevová et al. (in print). *Mluvnice současné češtiny. Část 2: Syntax češtiny na základě anotovaného korpusu*. Praha: Nakladatelství Karolinum.
- Petkevič, V. (2013). Formal (Morpho)Syntactic Properties of Reflexive Particles *se, si* as Free Morphemes in Contemporary Czech. In *Proceedings of the 7<sup>th</sup> International Conference 2013, 13-15 November 2013*. Slovenská akadémia vied, Slovakia, pp. 206-216.
- Renau, I., Battaner, P. (2012). Using CPA to Represent Spanish pronominal Verbs in a Learner's Dictionary. In *Proceedings of the 15th Euralex International Congress 2012, 7-11 August 2012*. University of Oslo, Norway, pp. 350-361.
- Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, Ch.R. & Scheffczyk, J. (2010) *FrameNet II: Extended Theory and Practice*. <https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf>
- Sgall, P., Hajičová, E. & Panevová, J. (1986). *The Meaning of the Sentence in Its Pragmatic and Semantic Aspects*. Dordrecht: Reidel.
- Skoumalová, H. (2001). *Czech syntactic lexicon*. PhD thesis. Charles University, Prague, Czech Republic.

Žabokrtský, Z., Lopatková, M. (2007) Valency Information in VALLEX 2.0: Logical Structure of the Lexicon.  
In *The Prague Bulletin of Mathematical Linguistics*, 87, pp. 41-60.

### **Acknowledgements**

The research reported in this paper has been supported by the Czech Science Foundation GA ČR, grant No. P406/12/0557. This work has been using language resources developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).



# Polysemous Models of Words and Their Representation in a Dictionary Entry

Tinatin Margalitadze

Lexicographic Centre at Ivane Javakhishvili Tbilisi State University

tinatin@margaliti.ge

## Abstract

The paper deals with one of the universal models of polysemous adjectives and verbs, namely one-dimensional model and examines the ways of its representation in a dictionary entry. Polysemy is connected with the human perception and cognition of the world. It is determined by the process of perceiving not only particular objects and phenomena, but also the similarities existing between them or seen as such by members of the given language community. It is also connected with the ability of the language to reflect the new, yet un-cognized objects and phenomena by means of their associations and relations with already known, cognized objects and phenomena, i.e. to translate diversity of the world into linguistic unity. This makes polysemy an extremely interesting linguistic phenomenon but also leads to controversies concerning the interpretation of different issues connected with it. The paper touches upon some debatable issues connected with polysemy, such as: boundaries between senses, meaning and context, the role of context in the process of realization of meanings of polysemous words, meanings and sub-meanings, sense-numbering in a dictionary entry, etc. The paper also discusses some peculiarities of lexical meanings of adjectives and verbs.

**Keywords:** one-dimensional; general semantic component; subsume

## 1 Introduction

As early as in the 1<sup>st</sup> century AD, Marcus Fabius Quintilian, a theoretician in oratorical skills, explains in his textbook on rhetoric *Institutio Oratoria* the concepts of metaphor and metonymy, which are important mechanisms of semantic changes and development of transferred meanings of words. The compilation of comprehensive explanatory dictionaries in the 17<sup>th</sup> century, first in Italy (1612) and then in France (1694), in the 18<sup>th</sup> century's England (*A Dictionary of the English Language* by Samuel Johnson, 1755), made a significant contribution to the description and study of semantics as a field of knowledge, and particularly to that of polysemy. In the early 20th century, a German linguist Hermann Paul distinguishes between usual (*usuelle Bedeutung*) and occasional meanings (*okhasionelle Bedeutung*), drawing the attention of linguists to context as an important tool for the realization of polysemous meanings of a word (Paul 1920: 75). Since then, much has been written on polysemy and context and the issue is still under discussion. Although it is not the aim of the present paper to dis-

cuss all problems connected with polysemy, still we cannot avoid touching upon some important issues, namely: What is the meaning of a polysemous word? Does it exist on the systemic level of language, or the meaning of a polysemous word is entirely determined by the context in which the word is used? Does a polysemous word have one abstract / general meaning, activated differently in different contexts, or does its structure represent a set of interconnected lexical units (LU) arranged into a single whole by means of various semantic relationships?

These issues aroused bitter controversy not only in 1950s and 1960s (Firth 1968; Antal 1963; Ulmann 1964 and others). Over the last 10-15 years, many scholars yet again discuss these issues and, as it seems, the answer is still not unequivocal (Stock 1984; Atkins 1993; Kilgariff 1997; Hanks 2000; Rundell 2002; Kosem 2008; Trap-Jensen 2010 and others). “I don’t believe in word senses”, states Sue Atkins (Atkins 1993). According to P. Hanks, a word does not have separate meanings, but rather a set of meaning potentials, which may be activated in a particular context (Hanks 2000). M. Rundell distinguishes between senses of polysemous words with clearly distinct meanings which, in this respect, “conform quite well to the conventional dictionary model, and much fuzzier *meaning-clusters*, where a basic semantic core is elaborated, in real text, in a variety of ways” (Rundell 2002: 148). More and more questions arise on how to present word meanings? How to find boundaries between senses? Lumping or Splitting? How to deal with the issue of meaning clines? How are meanings of polysemous words activated in a context?

Such discussions led some scholars even to the questioning of lexicographic practice of division of words’ meanings into senses, particularly for NLP purposes (Kilgariff 1997; Hanks 2000).

The study of above issues is important not only from the theoretical point of view, in order better to perceive the phenomenon of polysemy, but also from the viewpoint of representing a polysemous word in a dictionary entry, also for NLP purposes. The present paper expounds on one of the universal models of polysemous words studied in adjectives (Margalitadze 1982), later in verbs and nouns (Margalitadze 2006), namely the one-dimensional model. The study of this model shed some light on specific debatable issues of polysemy.

## 2 General Semantic Component and Subseme

As mentioned above, one-dimensional model is characteristic of adjectives and verbs. For the description of the present model two semantic components of a word’s lexical meaning are to be introduced: general semantic component, and subseme. ‘General semantic component’ (GS) denotes that semantic component of word, which serves as the basis for the development of a number of LUs of polysemous adjectives and verbs.

By the term ‘subseme’ we denote the semantic component which serves to differentiate LUs of a polysemous word. The subseme concretizes the abstract meaning of GS in each particular LU, thus activating meanings of polysemous adjectives and verbs<sup>1</sup>.

In order to illustrate the GS (marked in blue colour in examples given below) and the subseme (marked in red colour in examples given below), let us examine the adjective STRAIGHT. LUs of the adjective concerned are based on the GS – being free from deviation / bending /.

(1) STRAIGHT

1. Direct, not crooked (a straight street, a straight edge, a straight railway line) –

being free from deviation / bending / in *direction*

2. Erect, not crooked or stooping (a straight back) –

being free from deviation / bending / in *deportment*

3. Direct, shortest, uninterrupted (a straight flight, a straight road, a straight path) –

being free from deviation in *course*

4. Straightforward, frank, open (a straight answer, a straight question, straight talks) –

being free from deviation in *truth, openness, frankness*

5. Fair, virtuous, honest (a straight woman) –

being free from deviation in *dealings / rectitude /*

6. Consistent, logical, clear (a straight thinker) –

being free from deviation in *some method*

7. Conventional; respectable (she looked straight, a straight play) –

being free from deviation from *conventional, accepted, traditional behaviour / norms / views*

and so on.

## 2.1 Systemic Context and Subseme

The first example represents the following meanings of the adjective straight:

1. Direct, not crooked;
2. Erect, not crooked or stooping;
3. Direct; shortest; uninterrupted;
4. Straightforward, frank, open;
5. Fair, virtuous, honest;
6. Consistent, logical, clear;
7. Conventional; respectable

---

1 The term ‘general semantic component’ is not an adequate translation of the Georgian name given to this semantic component, which is well rendered by its Russian equivalent *сквозная сема* ‘skvoznaia sema’ (literally ‘through-going semantic component’). *Сквозная сема* ‘skvoznaia sema’ has its origin in the theatrical term introduced by Constantin Stanislavski – “through-going action” (*сквозное действие*). In fact, the term aptly expresses the essence of the semantic component identified in the semantic structure of verbs and adjectives.

and so on.

Adjectives and verbs denote an important logical category of 'feature': a feature, quality of an object – in adjectives, and a feature of an action or a state in verbs. The feature denoted by these words is singled out from different classes of objects or actions / states. Accordingly, adjectives and verbs contain in their meanings various expressions of this or that feature in objects or actions of different classes. As a result of this, along with each meaning of an adjective or a verb, there is constantly implied a systemic context denoting components of one and the same category of the objective reality.

Systemic context of the first meaning of the adjective STRAIGHT are the nouns denoting objects having linear shape (*a street, an edge, a railway line, etc.*). Out of this group of nouns, the GS 'being free from deviation / bending /' singles out the common seme '*direction*', which becomes included in the semantic structure of the given meaning of the adjective as its component, and concretizes the general meaning of the GS 'being free from deviation / bending /' – 'being free from deviation / bending / in *direction*', thus enabling the realization of the meaning 'straight'.

Systemic context of the second meaning of STRAIGHT are the nouns denoting back, shoulders, *etc.* Out of this group of nouns, the GS 'being free from deviation / bending /' selects the common seme '*deportment*' which becomes included in the semantic structure of the given meaning of the adjective as its component, and concretizes the general meaning of the GS – 'being free from deviation / bending / in *deportment*', thus enabling the realization of the meaning 'erect, not crooked or stooping'.

Systemic context of the third meaning of STRAIGHT are the nouns denoting flight, road, way, *etc.* Out of this group of nouns, the GS 'being free from deviation' selects the common seme '*course*', which concretizes the general meaning of the GS – 'being free from deviation in *course*', thus enabling the realization of the meaning 'direct, shortest'.

Systemic context of the fourth meaning of STRAIGHT are the nouns denoting conversation, question, answer, *etc.* Out of this group of nouns, the GS 'being free from deviation' selects the common seme '*truth, openness*', which concretizes the general meaning of the GS – 'being free from deviation in *truth, openness*', thus enabling the realization of the meaning 'straightforward, frank, open'.

Systemic context of the fifth meaning of STRAIGHT are the nouns denoting human beings. Out of this group of nouns, the GS 'being free from deviation' selects the common seme '*dealings, rectitude*', which concretizes the general meaning of the GS – 'being free from deviation in *dealings, rectitude*', thus enabling the realization of the meaning 'fair, honest'.

Systemic context of the sixth meaning of STRAIGHT are the nouns denoting thinking, thinker, *etc.* Out of this group of nouns, the GS 'being free from deviation' selects the common seme '*method*', which concretizes the general meaning of the GS – 'being free from deviation in *some method*', thus enabling the realization of the meaning 'consistent, logical', and so on.

Thus GS 'being free from deviation / bending /' singles out the common semes – subsemes from systemic contexts: '*direction*' from nouns, denoting objects with linear shape; '*deportment*', from nouns denoting back, shoulders, *etc.*; '*course*' from nouns denoting flight, road, way, and so on. These subsemes enter the semantic structure of LUs of the adjective STRAIGHT as their component, and concretize



the abstract meaning of the GS – ‘being free from deviation / bending / in *direction*’, in *deportment*’, in *course*’, etc, thus activating the realization of LUs: ‘straight’, ‘erect, not crooked or stooping’, ‘direct, shortest’ and so on.

GS and subseme can be illustrated on the example of other adjectives and verbs.

(2) LUs of the adjective CROOKED are based on the GS – having deviation in / from /.

1. not straight, bent, twisted (crooked streets, a crooked road, a crooked blade) –

having deviation *in direction*

2. deformed; bent (an aged man with a crooked frame, yellow and crooked teeth) –

having deviation *from normal form*

3. dishonest, not straightforward (crooked politicians, crooked dealings) –

having deviation *in rectitude*

4. fraudulent; illegal (crooked business, crooked business deal) –

having deviation *from legal frame*

and so on.

(3) LUs of the adjective LOW are based on the GS – being below the average level.

1. of small upward extent (a low wall, a low hill) –

being below the average level *in upward extension*

2. not elevated in position (low bridges, Low Countries) –

being below the average level *in elevation from the ground or some other downward limit*

3. not tall, short (a low man, a man of low stature) –

being below the average level *in statute*

4. not high in amount (low price, low wages) –

being below the average level *in amount*

5. deficient in degree of intensity (low redness, low colour) –

being below the average level *in degree of intensity*

6. not loud (low voice, low laugh) –

being below the average level *in volume*

7. of humble rank, position (low birth, low life) –

being below the average level *in social rank*

8. wanting in elevation, of inferior quality (low art, low standard) –

being below the average level *in quality*

9. wanting in decent breeding, vulgar, coarse (low person, low company) –

being below the average level *in social “respectability”*

and so on.

(4) LUs of the verb ESCAPE are based on the GS – breaking / getting / away from.

1. to get away, to get free (to escape from prison, to escape from the army) –

breaking / getting / away from *(physical) confinement*

2. to avoid or retreat from the realities of life (to escape reality) –

- breaking / getting / away from *unpleasant realities of life*
  - 3. to avoid or elude an evil that threatens (to escape poverty, to escape punishment) –  
breaking / getting / away from *misfortune of any kind*
  - 4. to avoid psychological problems (to escape television addiction) –  
breaking / getting / away from *mental / psychological / problems*
  - 5. to elude notice or recollection (to escape one’s mind, to escape smb.’s eyes) –  
breaking / getting / away from *notice / mental grasp /*
  - 6. to leak from a container (of a gas, liquid, etc) –  
(as if ) breaking / getting / away from *some confining envelope or enclosure*
- and so on.

## 2.2 Mechanism of the Interrelation between Adjective / Verb and Noun

Figure 1 demonstrates the underlying mechanism of the interrelation between adjective or verb and noun in their semi-automated syntagms, where dotted line represents GS, ellipse represents a LU of a polysemous adjective or verb, and a circle – the systemic context of the given LU. GS acts from adjective or verb to noun ( $A / V \rightarrow N$ ). GS determines the choice of nouns, from them selects common feature – subseme, which concretizes its abstract meaning. Whereas subseme acts in the opposite direction, from noun to adjective / verb ( $N \rightarrow A / V$ ). Subseme enters the semantic structure of adjectives or verbs, concretizes meaning of GS and activates individual LUs of polysemous adjectives and verbs.

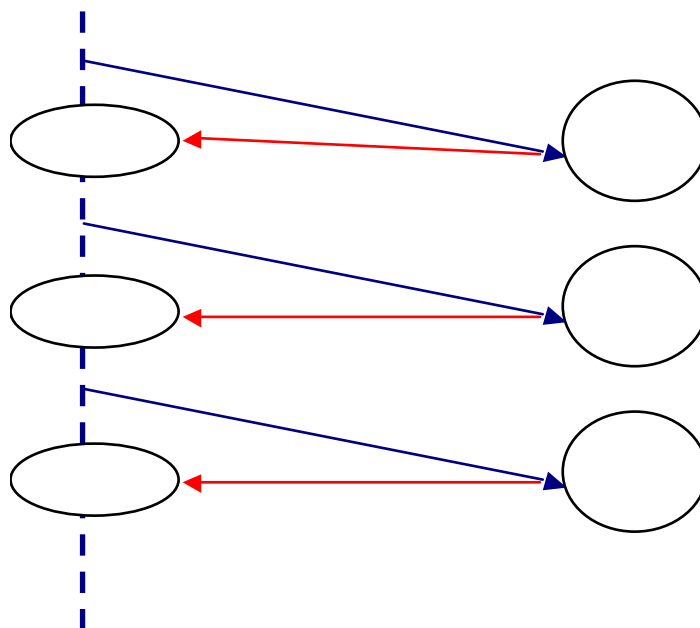


Figure 1: Semi-automated Syntagms of Adjectives / Verbs and Nouns.

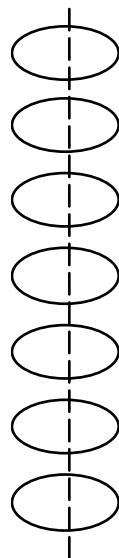
- GS is a generative semantic component, providing the basis for the development of a number of LUs;
- GS is a common semantic component, a kind of semantic “thread” uniting several LUs of a polysemous word;
- GS is an integrating seme, by means of which a polysemous verb and adjective can function as one word;
- GS is generated in paradigmatic, that is, vertical section. Its existence is revealed through the comparison of several meanings of a polysemous word;
- By its structural and semantic status, GS is more abstract, than differential and potential semes making up the lexical meaning of a word, as far as it governs several LUs of verb and adjective;
- GS inherently implies the idea of the classes of objects, which may be characterized by a given verb or adjective. Accordingly, it motivates or blocks the selection of a systemic context, wherewith a given verb or adjective may liaise.

GS differs from common semantic component, archeseme or hyperseme by being a generative semantic component. Not only is it the common semantic component of several LUs, but it is also the basis of the generation of polysemous meanings of verb and adjective and does govern them.

- Subseme is a differential seme, on the basis of which a concrete LU of verb and adjective is generated. Like GS, we regard subseme as a generative seme. While GS generates several LUs of polysemous verb and adjective, subseme serves as the basis for the creation of one specific LU;
- Subseme is singled out from an entire class of objects, which is represented by a definite group of nouns. Consequently, it implies the idea of the given class of objects and, accordingly, that of the definite area of denotation;
- Upon the syntagmatic axis, subseme is generated in the course of interrelationship between verb and adjective on the one hand, and semantic structures of noun on the other hand;
- As a result of the existence of subseme, for each LU in the semantic structure of verb and adjective there is generated a systemic context united by the given subseme.

### **3 One-Dimensional Model**

‘One-dimensional’ are termed such polysemous verbs and adjectives, all LUs of which are generated on the basis of a single GS (see Figure 2).



**Figure 2: One-Dimensional Model.**

All the examples discussed above represent one-dimensional models. One-dimensional is not the only model of polysemous adjectives and verbs. There are more models described for these parts of speech (Margalitadze 1982; 2006) but this model is quite universal and many adjectives and verbs develop their LUs on the basis of one GS. Below are given more examples of one-dimensional adjectives and verbs.

- (5) Polysemous meanings of the English verb 'to break' are based on the GS - 'destroying / violating / the completeness, wholeness, continuity'. Its subsemes in different LUs may be *a bone, a plate, a surface, skin, a performance, a lecture, spiritual, moral or financial state, silence, etc.*
- (6) The verb 'to kill' has the GS - 'depriving of some essential quality', which is concretized by different subsemes in different LUs: *life, vitality, activity, feeling, desire, etc.*
- (7) The GS of the verb 'to erupt' is - 'bursting forth from natural or artificial limits', concretized by the following subsemes: *volcano, water, fire, air, soldiers, etc.*
- (8) The one-dimensional adjective 'dull' has its polysemous meanings generated on the basis of the GS - 'wanting some essential quality'. Its subsemes in different LUs are: *wit, sensibility or keenness of perception, motion or action, vivacity or cheerfulness, colour, intensity, etc.*
- (9) All polysemous meanings of the adjective 'small' are developed on the basis of the GS - 'being less than average' (being less than average in *size, in statute, in number, in duration, in importance, in amount, in rank or condition, in scale, etc.*).
- (10) All polysemous meanings of the adjective 'great' are developed on the basis of the GS - 'being more than average' (being more than average in *size, in number, in duration, in importance, in rank or condition, in scale, etc.*).
- (11) All polysemous meanings of the adjective 'high' are developed on the basis of the GS - 'being above the average level' (being above the average level in *upward extension, in elevation from the ground or some other downward limit, in stature, in amount, in social rank, etc.*).

## 4 Discussion

As it has been shown by the above analysis, one-dimensional words have meanings of equal status. What may seem an abstract / general meaning, meaning potential or a semantic core activated differently in real contexts, is in fact a semantic component, the GS which is very general, very abstract, as far as it contains in itself the idea of those classes of words wherefrom it is singled out. Whenever the GS is concretized by a subseme, there appears an individual meaning of a polysemous word, an individual LU. Subseme shows the boundaries between senses of a polysemous adjective and verb. Thus, one-dimensional model has an extremely abstract GS and meanings of equal status.

The role of context must be mentioned specifically. Context is always necessary for the actualization of the meanings of a polysemous word, but the GS, as we have seen, is an active semantic component that can select or block nouns in question. Context can not trigger any meaning of one-dimensional adjective or verb which is not present in their semantic structure. Consequently, context reveals what is already present in the semantic structure of adjective or verb, it does not motivate meaning, it actualizes existing meanings. This shows the relative independence of adjectives and verbs within the language system.

Feature does not exist independently in the objective reality. It is present inside object and is unimaginable without the latter. However, the feature translated into a linguistic category appears as a separate category, as that of adjective and verb, thus acquiring a different linguistic status. Within the system of language, feature acquires relative autonomy, which results in complex interrelations between adjective and verb and their systemic context in their lexical syntagms. The role of systemic context in the process of realization of meanings of verb and adjective consists in conveying particular information in the form of subseme to the semantic structure of verb and adjective. This information concretizes general meaning of the GS and breaks it up into concrete variants, thus enabling the differentiation and realization of separate LUs of verb and adjective. On the other hand, verb and adjective, as independent parts of speech, contain such semes within the semantic structure of their meaning, which not only generate a number of LUs, but also determine the selection of a definite, rather than any noun. Within the language system, the interaction between objects / actions and their features is formed on a completely different level of generalization.

## 5 One-Dimensional Words in a Dictionary Entry

Dictionaries of the English language give different interpretation to the polysemous words of the described model and represent them accordingly in a dictionary entry. E.g. MEDAL, Oxford Dictionary of English treat the following meaning of 'escape' - leak from a container (of a gas, liquid, etc) - as a full-fledged meaning of this verb, while OED, Shorter Oxford English Dictionary, Webster's Third New International Dictionary interpret it as a sub-meaning of 'escape'. Likewise, LU - avoid capture, punis-

hment, or something unwelcome – are meanings according to Shorter Oxford English Dictionary, Webster's Third New International Dictionary, while MEDAL, Oxford Dictionary of English treat them as sub-meanings, meaning-clusters. Such examples may be cited *ad infinitum*.

LUs of one-dimensional model have equal status. Semantic relationships between LUs, generated on the basis of GS, are equipollent and not that of dependence. Consequently, LUs should be represented as full-fledged meanings in a dictionary entry and they should be numbered in the same manner by Arabic numerals. GS may be given at the beginning of an entry as a general description of the feature. Each LU should be supplied with its systemic context, i.e. with nouns denoting one and the same category of the objective reality. Below are given some examples of entries.

(12) Straight *adjective*

[being free from deviation / bending]

1. Direct, not crooked (*used with nouns denoting objects, having linear shape*);  
a straight street, a straight edge, a straight railway line;
2. Erect, not crooked or stooping (*used with nouns back, shoulders, etc*)  
a straight back;
3. Direct, shortest, uninterrupted (*used with nouns denoting travelling on foot or by other means*)  
a straight flight, a straight road, a straight path;
4. Straightforward, frank, open (*used with nouns denoting talking*)  
a straight answer, a straight question, straight talks;
5. Fair, virtuous, honest (*used with nouns denoting human beings*)  
a straight woman;
6. Consistent, logical, clear (*used with nouns denoting thinking*)  
a straight thinker;

and so on.

(13) Escape *verb*

[breaking / getting / away from usually smth. unpleasant]

1. to get away, to get free (*used with nouns denoting places of physical confinement*)  
to escape from prison, to escape from the army;
2. to avoid or retreat from the realities of life (*used with nouns denoting unpleasant realities of life*)  
to escape reality;
3. to avoid or elude an evil that threatens (*used with nouns denoting any misfortune*)  
to escape poverty, to escape punishment;
4. to avoid psychological problems (*used with nouns denoting different addictions*)  
to escape television addiction;
5. to leak from a container (*used with nouns denoting gas, liquid, etc*)

and so on.

Another alternative of a dictionary entry may be GS+subseme descriptions in each LU of the one-dimensional adjective or verb (see example 14).

(14) Straight *adjective*

[being free from deviation / bending]

1. Direct, not crooked (*used with nouns denoting objects, having linear shape*);

a straight street, a straight edge, a straight railway line;

[being free from deviation / bending / in direction]

2. Erect, not crooked or stooping (*used with nouns back, shoulders, etc*)

a straight back;

[being free from deviation / bending / in deportment]

3. Direct, shortest, uninterrupted (*used with nouns denoting travelling on foot or by other means*)

a straight flight, a straight road, a straight path;

[being free from deviation / bending / in course]

4. Straightforward, frank, open (*used with nouns denoting talking*)

a straight answer, a straight question, straight talks;

[being free from deviation / bending / in truth, frankness]

5. Fair, virtuous, honest (*used with nouns denoting human beings*)

a straight woman;

[being free from deviation / bending / in dealings / rectitude /]

6. Consistent, logical, clear (*used with nouns denoting thinking*)

a straight thinker;

[being free from deviation / bending / in some method]

and so on.

## 6 Conclusion

The study of the deep structure of interrelation between adjectives / verbs and nouns in their semi-automated syntagms has revealed the active generating semantic component – GS in the semantic structure of polysemous adjectives and verbs. On the one hand, GS generates several LUs and governs them, on the other hand, GS inherently has the knowledge of the classes of objects wherefrom it is singled out, thus motivating or blocking the selection of nouns wherewith adjectives and verbs may liaise. GS selects the subseme from the systemic context, which enters the semantic structure of LU and is present there. As a result of this, along with each meaning of adjective or verb, there is constantly implied a systemic context denoting components of one and the same category of the objective reality.

The interrelation between GS and subseme and the presence of subseme in the semantic structure of LU indicates that context reveals existing meaning of adjective and verb and does not motivate it.

Subsemes mark the boundaries between senses of polysemous adjectives and verbs.

One-dimensional adjectives and verbs have meanings of equal status, which should be numbered in the same manner in a dictionary entry, as full-fledged meanings.

GS may be given in a dictionary entry, as a general description of the feature, expressed by adjective and verb (see examples 12, 13).

Each LU should be supplied with its systemic context, specifying the group of nouns used with the respective LU (see examples 12, 13).

Unlike identifying words such as nouns which, depicting objects and phenomena, comprise multiple semantic components in the semantic structure of their meanings, verbs and adjectives denote feature. Accordingly, their lexical meaning is “scarce” of semantic components and thus it is natural that polysemous structure of adjectives and verbs should be characterized by linear development and one feature, one semantic component should become the basis for the formation of multiple meanings.

## 7 References

- Antal, L. (1963) *Questions of Meaning*. The Hague: Mouton Co.
- Atkins, B.T.S. (1993) “Theoretical Lexicography and its relation to Dictionary-making”. *Dictionaries* 14:4-43.
- Firth, J.R. (1958) *Papers in Linguistics*. Oxford University Press.
- Hanks, P. (2000) Do Word Meanings Exist? *Computers and the Humanities* 34: 205-215.
- Kilgariff, A. (1997) I Don’t Believe in Word Senses. *Computers and the Humanities* 31(2): 91-113.
- Kosem, I. (2008). Dictionaries for University Students: A Real Deal or Merely a Marketing Ploy? Proceedings of the XIII EURALEX International Congress. Barcelona.
- Margalitadze, T. (1982) Strukturno-semanticheskaia Kharakteristika Mnogoznachnykh Prilagatel’nykh, kak Nominativnykh Edinits v Sovremennom Angliiskom Iazyke. Candidate’s Thesis. Tbilisi : Tbilisi University Press.
- Margalitadze, T. (1982) The Main Models of the Semantic Structure of Adjectives in Modern English. In: *Bulletin of the Academy of Sciences of the Georgian SSR*. 105, 3 : 181 – 184.
- Margalitadze, T. (2006) *Meaning of a Word and Methods of its Research*. Tbilisi State University.
- Paul, H. (1920) *Prinzipien der Sprachgeschichte*. Halle: Niemeyer.
- Rundell, M. (2002) Good Old-fashioned Lexicography: Human Judgement and the Limits of Automation. *Lexicography and Natural Language Processing*. EURALEX : 138-155.
- Stock, P. (1984) Polysemy. *Lexeter ’83 Proceedings*. Tübingen: Max Niemeyer.
- Trap-Jensen, L. (2010) One, Two, Many: Customization and User Profiles in Internet Dictionaries. *Proceedings of the XIV Euralex International Congress*. Fryske Akademy, Leeuwarden.
- Ullmann, St. (1964) *Semantics. An Introduction to the Science of Meaning*. Oxford: Basil Blackwell.
- Dictionaries:
- Hanks P. et al (2005). *Oxford Dictionary of English*. Second Edition, Revised. Oxford University Press.



*Macmillan English Dictionary for Advanced Learners* (MEDAL). Accessed at: <http://www.macmillandictionary.com> [05.09.2013]

*Oxford English Dictionary on Historical Principles* (1989). Second edition on CD-ROM. Version 2.0. Oxford University Press.

Stevenson A. et al (2007). *Shorter Oxford English Dictionary*. Sixth Edition (SOED). Oxford University Press.

Webster's Third New International Dictionary (Unabridged). Merriam Webster Inc., 1981.



# One Lexicological Theory, two Lexicographical Models and the Pragmatemes

Lena Papadopoulou  
Hellenic Open University  
papadopoulou.lena@gmail.com

## Abstract

Generally little attention has been paid to pragmatics in most dictionaries. The present paper focuses on the concept of pragmatemes and their lexicographical treatment within the frame of Explanatory Combinatorial Lexicology. First, necessary preliminary notions closely related with pragmatemes are considered, by briefly reviewing the mel'čukian global model of human linguistic behaviour, linguistic sign and phrasemes typology. Second, the definition of pragmatemes, that is of phrasemes used in given extralinguistic situations, and the central acting part of the conceptual representation of the communicative situation are presented. Following, the structure of *Explanatory Combinatorial Dictionary* (ECD) and *PragmatLex* are outlined and an illustration of the Greek pragmatemes *Συγχαρητήρια* 'Congratulations' and *Συλλυπητήρια* 'Condolences' is provided within the simplified versions of ECD - *Dictionary of Collocations* and *Lexique actif du français*- and the lexicographical model for pragmatemes *PragmatLex*. Finally, we conclude our paper with a brief discussion on which lexicographical approach is preferable.

**Keywords:** Meaning Text Theory; phraseology; pragmatemes

## 1 Introduction

Pragmatics generally is an area of great importance. However, it is relatively poorly treated in the majority of dictionaries, so there is scope for work on this subject. The concept of pragmatemes, that is expressions that are used in specific extralinguistic situations, and the developed models for their lexicographical treatment represent a significant step towards addressing that challenge.

This paper aims to present two lexicographical models in which pragmatemes can be processed; the ECD and the PragmatLex. To do so, first our theoretical framework (Meaning $\Leftrightarrow$ Text Theory) will be set, then the definition of pragmatemes will be provided and, following, the dictionaries' structure will be described and illustrated by processing the Greek pragmateme *Συγχαρητήρια* 'Congratulations' and its antonym *Συλλυπητήρια* 'Condolences'. Finally, criteria for model selection will be proposed.

## 2 Preliminary notions

Our work is framed within Meaning $\leftrightarrow$ Text Theory (MTT), the main aim of which is to build models of natural languages (among others, Mel'čuk 1988a; Mel'čuk 1997; Mel'čuk 2001b; Polguère 1998; Milićević 2001; Milićević 2006; Kahane 2001).

Concept-Sound Model (CSM) is the global model of human linguistic behavior which is developed within MTT:

$$\{\text{WORLD}\} \leftrightarrow \{\text{SemR}_i\} \leftrightarrow \{\text{SPhonR}_i\} \leftrightarrow \{\text{LINGUISTIC SOUNDS}\}$$

**Figure 1: Concept-Sound Model (Mel'čuk 2012: 170-181).**

Conceptics, Meaning-Text Model (MTM) and Phonetics/Graphics are the three major models of CSM which represent the production of an utterance. First, Conceptics model captures the construction of a semantic representation (SemR) based on the conceptual representation (ConceptR) of the given extralinguistic situation (SIT). Second, MTM describes the construction of the Phonological Representation (PhonR) of the given SemR. The third model - Phonetics/Graphics- represents the construction of the corresponding sound/letter string for the given PhonR.

A typology of linguistic signs has been established within MTT based on the transitions between these three models and the applied restrictions. A simple linguistic sign within MTT corresponds to the triplet of  $X = \langle 'X'; /X/; \Sigma X \rangle$ , where 'X' is the signified, /X/ the signifier and  $\Sigma X$  the combinatorial properties of the linguistic sign X. Simple linguistic signs are combined into complex linguistic signs. The notions of unrestrictedness and regularity are implied by such combination, that is freedom in the selection of meanings and lexical units and the compositionality, respectively.

Free phrases are complex linguistic signs whose signified and signifier are constructed both unrestrictedly and regularly, while on the contrary phrasemes, or non free phrases, are not. Phrasemes are classified into semantic phrasemes and pragmatic phrasemes, or pragmetemes, (Mel'čuk 1995; Mel'čuk 1998). On the one hand, pragmetemes are restrictedly constructed by the ConceptR(SIT). On the other hand, the signified 'X' of a semantic phraseme is unrestrictedly constructed by the ConceptR(SIT) but its signifier /X/ is constrained by the selected SemR. Semantic phrasemes are non compositional and they are categorized into three types on the basis of their semantic opacity: (i) full idioms, (ii) semi-idioms, or collocations, and (iii) quasi-idioms.

## 3 Pragmatemes

Pragmatemes are compositional phrasemes whose signified is restrictedly constructed by the Conceptual Representation of the given extralinguistic situation (Mel'čuk 1998). Blanco (to appear) points out that this definition concerns the prototypical pragmatemes. On the one hand, a pragmateme can

be both constrained by the ConceptR and the SemR, i.e. the idiom/pragmateme *break a leg* [to wish good luck to actors and musicians before they go on stage to perform]. On the other hand, a lexeme whose signified is bound by the Concept(SIT) is considered as a pragmateme, i.e. *Congratulations*.

The ConceptR(SIT) plays the lead role in pragmatemes definition. Although, Mel'čuk recognizes the inherent difficulties in defining the extralinguistic reality, he proposes that ConceptR is based on three main models (2001a, p. 90): (i) the speaker's model, (ii) the speaker's model of the addressee and (iii) the situation's model. The ConceptR(SIT) of the pragmateme *break a leg* will be based on that 'I am addressing to an actor which is going on stage to perform. I wish (s)he will have a successful presentation. If I were (s)he I would like to be encouraged. I will wish him/her good luck, as I should do' (speaker's model), '(S)he is thinking that (s)he is going on stage to perform, that (s)he is stressed, that (s)he expects to be encouraged' (speaker's model of the addressee) and on that 'the speaker wants to encourage a performer before going on the stage by wishing him good luck' (situation's model).

## 4 Lexicological processing of pragmatemes

Once the pragmatemes have been defined and before moving to the presentation of the two lexicological models for pragmatemes, which are both framed within Explanatory Combinatorial Lexicology, some preliminary remarks upon pragmatemes have to be made. Although pragmatemes are linguistic signs, they are not considered to be LUs, because they dispose of an internal argumental structure, so as they are ordered within the keyword(s) that phraseologically bind(s) them, that is within the LU(s) that can define the SIT of the pragmateme. It has to be also pointed out that pragmatemes and specifically their SIT is described by non standard lexical functions (LFs) (Mel'čuk, 1995); (Blanco, 2010).

### 4.1 Explanatory Combinatorial Dictionaries

Explanatory Combinatorial Lexicology is developed within the MTT (among others, Mel'čuk & Zholkovsky 1984; Mel'čuk 1988b; Mel'čuk 1995; Mel'čuk, Clas, & Polguère 1995; Mel'čuk 2006b; Mel'čuk & Polguère 2007) and Explanatory Combinatorial dictionaries (ECD) are compiled within it.

ECDs are highly formal theoretical lexicons, whose entries are exhaustively described on the basis of explicitness and consistency. The macrostructure of an ECD is structured by super-entries, entries and sub-entries. Vocables constitute the super-entries, which are sets of lexical units (LUs) that share the same signifier and they are linked by a semantic bridge, LUs are the entries, which can correspond to lexemes, idioms or quasi-idioms, and collocations and pragmatemes are considered to be sub-entries. As far as microstructure is concerned, it is structured in four zones: (i) the semantic, (ii) the phonological/graphematic zone, the (iii) syntactics zone and (iv) illustrative zone.

Due to the theoretical basis of ECD and the its subsequent high lengthiness, the Dictionary of Collocations (DiCo) and Lexique actif du français (LAF) have been developed as simplified versions. DiCo (Dictionary of Collocations) is the formalized version of the purely “theoretical” ECD. DiCo is sort of a “simplified” and more formalized ECD and in which the lexical units are structured as a series of eight main fields: (i) Name of the unit, (ii) grammatical properties, (iii) semantic formula, (iv) government pattern, (v) synonyms, (vi) semantic derivations and collocations, (vii) examples and (viii) full idioms that include the LU (Polguère 2000: 519) and LAF is the “popularized” version of the ECD which attempt to bridge the gap between “theoretical” and “commercial” lexicography with regard to explanatory combinatorial lexicology in order to be as much as possible accessible to a public of non-specialists (Polguère 2000: 522-3).

Following an illustrative example of processing the Greek pragmateme *Συλλυπητήρια* ‘Condolences’ (Figure 2 and 3) (Papadopoulou to appear) within the keyword-LU, respectively:

<p><b>a ΠΕΝΘΟΣ</b>          nom, neutr.          sentiment négatif : ~ του <b>ατόμου X</b> για το γεγονός <b>Z</b> του <b>ατόμου Y</b> με <b>W</b>          XII YIII, XII YIII ZI          {QSyn} <b>θλίψη</b>          {A0 expression of sympathy for Y on Z} <b>συλλυπητήριος</b>          {A0} <b>πένθιμος</b>          {A0Locin a nation} <b>εθνικό</b> ~          {AntiVer.A1} <b>βαρυπενθών</b> (ironic)          {CausMagnFact0} <b>βυθίζομαι στο</b> ~          {expression of sympathy for Y on Z} <b>συλλυπητήρια</b>          {FinV0} <b>βγάζω τα μαύρα</b>          {Magn expression of sympathy for Y on Z} <b>βαθιά, θερμά&lt;ολόθερμα συλλυπητήρια</b>          {Magn.A1} <b>βουτηγμένος στο</b> ~, <b>βαρυπενθών</b> (literary) &lt; <b>βουτηγμένος στα μαύρα</b>          {Magn} <b>βαρύ</b> ~          {MagnA0Locin Greek nation} <b>πανελλήνιο</b> ~          {MagnV0} <b>βαρυπενθώ</b>          {Oper expression of sympathy for Y on Z} <b>εκφράζω, απευθύνω, δίνω, στέλνω, λέω συλλυπητήρια</b>          {to support X throught ~} <b>συμπαραστέκομαι στο</b> ~          {to sympathize with X in ~} <b>συμμετέχω στο</b> ~, <b>συλλυπούμαι</b>          {V0} <b>πενθώ, κρατάω</b> ~          {Ver expression of sympathy for Y on Z} <b>ειλικρινή&lt;εγκάρδια&lt;ολόψυχα συλλυπητήρια</b>          {X=Y’s husband} <b>χήρος</b>          {X=Y’s wife who AntiV0} <b>εύθυμη</b> ~ (ironic)          {X=Y’s wife} <b>χήρα</b>          {Z= death} <b>θάνατος</b>  <i>Ολο το έθνος πενθεί (για) το θάνατο του ηγέτη.</i></p>
--

Figure 2 LU a ΠΕΝΘΟΣ in DiCo (Papadopoulou to appear).

<p><b>a ΠΕΝΘΟΣ</b> noun, neutral Negative emotion: ~ του <b>ατόμου X</b> για το γεγονός <b>Z</b> του <b>ατόμου Y</b> με <b>W</b></p> <p>☞ <b>θλίψη</b> Adjective for the expression of sympathy for Y for the <b>Zσυλλυπητήριοις</b> Adjective <b>πένθιμος</b> Adjective for the ~ expressed in a nation <b>εθνικό ~</b> Adjective for X who do not have deep ~ as (s)he should have <b>βαρυπενθών</b> (ironic) To make someone to get involved in deep ~ <b>βυθίζομαι στο ~</b> expression of sympathy for Y on Z <b>συλλυπητήρια</b> To finish having ~ <b>βγάζω τα μαύρα</b> Adjective for the expression of sympathy for Y on Z to a high degree <b>βαθιά, θερμά&lt;ολόθερμα συλλυπητήρια</b> Adjective for X who has deep ~ <b>βουτηγμένος στο ~, βαρυπενθών</b> (literary)&lt; <b>βουτηγμένος στα μαύρα</b> Adjective for deep ~ <b>βαρύ ~</b> Adjective for the ~ expressed in extensively in Greece <b>πανελλήνιο ~</b> To have deep ~ <b>βαρυπενθώ</b> To express the sympathy for Y on Z <b>εκφράζω, απευθύνω, δίνω, στέλνω, λέω συλλυπητήρια</b> To support someone who has ~ <b>συμπαραστέκομαι στο ~</b> To participate to the ~ of the X <b>συμμετέχω στο ~, συλλυπούμαι</b> To have ~ <b>πενθώ, κρατάω ~</b> Adjective for the expression of sincere sympathy for Y on Z <b>ειλικρινή&lt;εγκάρδια&lt;ολόψυχα συλλυπητήρια</b> Noun for X who is husband of Y <b>χήρος</b> Noun for X who is wife of Y and do not have ~ <b>εύθυμη ~</b> (ironic) Noun for X who is wife of Y <b>χήρα</b> Noun for Z <b>θάνατος</b> <i>Όλο το έθνος πενθεί (για) το θάνατο του ηγέτη.</i></p>
---

Figure 3 LU aΠΕΝΘΟΣ in LAF (Papadopoulou to appear).

## 4.2 PragmatLex

Please note that there must always be at least two level 2 and level 3 headings if you need to use these in your paper (e.g. at least 4.1 and 4.2 Blanco (2010; to appear<sub>a</sub>; to appear<sub>b</sub>) obviously based on MTT and recognizing the lack of dictionaries of ECD type in the majority of languages proposed the PragmatLex, which is designated as a lexicographical model for the processing of pragmatemes. PragmatLex is highly formal and it provides an exhaustive description for each pragmateme, which is structured in thirteen fields. It is worth pointing out that PragmatLex is written in XML in order to be applicable to NLP systems.

PragmatLex is a dictionary of monolingual coordinated dictionary type (Blanco 2001), considering that the translation equivalence of each pragmateme is provided linearly according to the overall micro-structure information, so as the description of pragmatemes is enterassigned within the language indication: (<ARTICLE language=" ">description of pragmatemes</ARTICLE language = " ">. First, the canonical form of the pragmateme is indicated **Lemma>canonical form of the pragmateme</Lemma>**. Second, the morphosyntax of the pragmateme is annotated based on the six deep-syntactic parts of speech (Mel'čuk 2006a), Third, the translation equivalence is provided in the target language

according to the corresponding structure of the L2 PragmatLex. Following, the LU-keyword(s), the definition of the SIT, the performing Speech act, the semantic structure and the lexical functions of the pragmateme are indicated. Afterwards, the coda, that is pragmatemes extensions which with no semantic addition complement the pragmatemes, the synonyms and the antonyms of the pragmateme and, finally, the decomposition of the local grammar that may the lemma disposes.

In the following figures the pragmatemes *Συλλυπητήρια-Condolencias* ‘Condolences’ (Papadopoulou to appear) and *Συγχαρητήρια-Felicidades* ‘Congratulations’ are shown within PragmatLex in Greek and Spanish language:

```

<ARTICLE language="el">
  <Lemma>Συλλυπητήρια</Lemma>
  <Morphosyntax>N</Morphosyntax>
  < TRANSLATION language="es">condolencias</ TRANSLATION language="es">
  <Keyword>πένθος, κηδεΐα</Keyword>
  <SIT>expresión escrita u oral de compasión hacia alguien en duelo</SIT>
  <SPEECH ACT>compadecerse</SPEECH ACT>
  <SS>~ X[=of X, Aposs, Adj (p.ej. προεδρικά συλλυπητήρια)] a Y por Z</SS>
  <LF>
    <Magn>βαθιά, θερμά<ολόθερμα</Magn>
    <Ver>ειλικρινή<εγκάρδια<ολόψυχα</Ver>
    <Oper>εκφράζω, απευθύνω, δίνω, στέλνω, λέω</Oper>
    <V0>συλλυπούμαι</V0>
    <A0>συλλυπητήριος</A0>
  </LF>
  <CODA>
    <01>Γεροί να είστε να τον θυμάστε</01>
    <02>Να ζήσετε να τον θυμάστε</02>
    <03>Ζωή σ' εσάς</03>
    <04>Ζωή σε λόγου σας</04>
  </CODA>
  <SYNONYM>-</SYNONYM>
  <ANTONYM>συγχαρητήρια</ANTONYM>
  <PARADIGM>-</PARADIGM>
</ARTICLE language="el">

```

Figure 4: *Συλλυπητήρια* in PragmatLex (Papadopoulou to appear).



```

<ARTICLE language="es">
  <Lemma>condolencias</Lemma>
  <Morphosyntax>N</Morphosyntax>
  < TRANSLATION language="el"> συλλυπητήρια</ TRANSLATION language="el">
  <Keyword>duelo, funeral</Keyword>
  <SIT>expresión escrita u oral de compasión hacia alguien en duelo</SIT>
  <SPEECH ACT>compadecerse</SPEECH ACT>
  <SS>~ X[=of X, Aposs, Adj (p.ej. Condolencias presidenciales)] a Y por Z</SS>
  <LF>
    <Magn>mayores<profundas</Magn>
    <Ver>sentidas<sinceras<cordiales</Ver>
    <Oper>expresar, dar, manifestar, enviar,</Oper>
    <V0>condoler</V0>
  </LF>
  <CODA>Siempre lo recordaremos</CODA>
  <SYNONYM>pésame</SYNONYM>
  <ANTONYM>congratulaciones, felicitaciones</ANTONYM>
  <PARADIGM>.</PARADIGM>
</ARTICLE language="es">

```

Figure 5: *Condolencias* in PragmatLex (Papadopoulou to appear).

```

<ARTICLE language="el">
  <Lemma> Συγχαρητήρια</Lemma>
  <Morphosyntax>N</Morphosyntax>
  < TRANSLATION language="es">felicidades</ TRANSLATION language="es">
  <Keyword>γάμος</Keyword>
  <SIT>expresión escrita u oral para expresar felicitación o enhorabuena a la pareja recién casada
    en una boda </SIT>
  <SPEECH ACT>felicitarse</SPEECH ACT>
  <SS>~ X[=of X, Aposs, Adj] a Y por Z</SS>
  <LF>
    <Magn> πολλά<θερμά<ολόθερμα</Magn>
    <Ver>ειλικρινή<εγκάρδια<ολόψυχα</Ver>
    <Oper>εκφράζω, απευθύνω, δίνω, στέλνω, λέω</Oper>
    <V0>συγχαίρω</V0>
    <A0>συγχαρητήριος</A0>
  </LF>
  <CODA>
    <01>να ζήσετε</01>
    <02>και καλούς απογόνους</02>
  </CODA>
  <SYNONYM>.</SYNONYM>
  <ANTONYM>συλλυπητήρια</ANTONYM>
  <PARADIGM>.</PARADIGM>
</ARTICLE language="el">

```

Figure 6: *Συγχαρητήρια* in PragmatLex.

```

<ARTICLE language="es">
  <Lemma>felicidades</Lemma>
  <Morphosyntax>N</Morphosyntax>
  <TRANSLATION language="el"> Συγχαρητήρια</TRANSLATION language="el">
  <Keyword>boda</Keyword>
  <SIT>expresión escrita u oral expresión escrita u oral para expresar felicitación o enhorabuena a
    la pareja recién casada en una boda</SIT>
  <SPEECH ACT> felicitar</SPEECH ACT>
  <SS>~ X[=of X, Aposs, Adj] a Y por Z</SS>
  <LF>
    <Magn>muchas<profundas</Magn>
    <Ver>sinceras<honestas</Ver>
    <Oper>dar, enviar, decir</Oper>
    <V0>felicitar</V0>
  </LF>
  <CODA>enhorabuena</CODA>
  <SYNONYM> enhorabuena </SYNONYM>
  <ANTONYM>condolencias</ANTONYM>
  <PARADIGM>.</PARADIGM>
</ARTICLE language="es">

```

Figure 7: *Felicidades* in PragmatLex.

## 5 ECD or PragmatLex?

ECD, or PragmatLex, that is NOT the question, as two different types of dictionaries are concerned, which are based on the same lexicological theory, yet; ECD is a dictionary of lexical units and PragmatLex is a dictionary of pragmatemes. However, we could answer the question in three different rounds from three different points of view.

First, the ideal lexicographical treatment of pragmatemes is within ECD, given that ECD's structure provides a global description of pragmatemes within their semantic frame (lexical units' links). However, there are no available complete ECD dictionaries for all languages. Second, PragmatLex' structure is proper for pragmatemes processing, as it focuses only on pragmatemes. Third, we propose a parallel processing of ECD and PragmatLex, that is the lexicographer elaborates pragmatemes which are associated with a keyword within PragmatLex and (s)he incorporates these data into the structure of ECD, i.e. the information of pragmateme *condolences* can be introduced as subentries into the structure of the lexical unit MOURNING (Papadopoulou, to appear) or *congratulations* into WEDDING.

## 6 References

- Blanco, X. (2001). Dictionnaires électroniques et traduction automatique espagnol-français. In *Langages* 143 (pp. 49-70).
- Blanco, X. (to appearb). Équivalents de traduction pour les pragmatèmes dans la lexicographie bilingue Français-Espagnol.
- Blanco, X. (2010). Los frasemas composicionales pragmáticos. In S. Mejri, & P. Mogorrón, *Opacité, Idiomaticité, Traduction*. Universitat d'Alacant.
- Blanco, X. (to appear<sub>a</sub>). Microstructure Évolutive pour un Dictionnaire de Pragmatemes. In *Actes des Journées Lexicologie, Lexicographie et Traduction*. Paris: AGence Universitaire de la Francophonie.
- Kahane, S. (2001). *Grammaires de dépendance formelles et théorie Sens-Texte*. Tours: Tutoriel, TALN 2001.
- Mel'čuk, I. (2006a). *Aspects of the Theory of Morphology*. Berlin/New York: Mouton de Gruyter.
- Mel'čuk, I. (1998). Collocations and Lexical Functions. In A. P. Cowie, *Phraseology: Theory, Analysis, and Applications* (pp. 23-53). Oxford: Clarendon Press.
- Mel'čuk, I. (2001a). *Communicative Organization in Natural Language. The Semantic-Communicative Structure of Sentences*. Amsterdam/Philadelphia: Benjamins.
- Mel'čuk, I. (1993). *Cours de morphologie générale*. (Montréal — Paris: Les Presses de l'Université de Montréal — CNRS.
- Mel'čuk, I. (1988a). *Dependency syntax: theory and practice*. Albany, NY: State Univ. of New York Press.
- Mel'čuk, I. (2006b). Explanatory Combinatorial Dictionary. In G. Sica, *Open Problems in Linguistics and Lexicography* (pp. 225-355). Monza: Polimetrica.
- Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner, *Lexical Functions in Lexicography and Natural Language Processing* (pp. 37-102). Amsterdam/Philadelphia: Benjamins.
- Mel'čuk, I. (2006a). Parties du discours et locutions. *Bulletin de la Société de Linguistique de Paris* 101:1, págs. 29-65.
- Mel'čuk, I. (2006c). Parties du discours et locutions. In *Bulletin de la Société de Linguistique de Paris* 101:1 (pp. 29-65). Paris.
- Mel'čuk, I. (1995). Phrasemes in Language and Phraseology in Linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder, *Idioms. Structural and Psychological Perspectives* (pp. 167-232). Hillsdale, N.J.—Hove, UK: Lawrence Erlbaum Associates.
- Mel'čuk, I. (2013). Phraseology: its place in the language, in the dictionary, and in natural language processing. In Z. Gavriilidou, A. Efthymioy, E. Thomadaki, & P. Kambakis-Vougiouklis, *Selected Papers of the 10th I.C.G.L.* (pp. 62-67).
- Mel'čuk, I. (1988b). Semantic Description of Lexical Units in an Explanatory Combinatorial Dictionary: Basic Principles and Heuristic Criteria. In *International Journal of Lexicography*, 1 : 3 (pp. 165-188).
- Mel'čuk, I. (2001b). *Semantics and the Lexicon in Modern Linguistics*. Unpublished Article.
- Mel'čuk, I. (2012). *Semantics. From Meaning to Text*. Amsterdam/Philadelphia: Benjamins Publishing Company.
- Mel'čuk, I. (to appear). Tout ce que nous voulions savoir sur les phrasèmes, mais .... In *Cahiers de lexicologie, revue internationale de lexicologie et de lexicographie*. Paris: Classiques Garnier.
- Mel'čuk, I. (1982). *Towards the Language of Linguistics*. München: Wilhelm Fink.
- Mel'čuk, I. (1997). *Vers une linguistique Sens-Texte. Leçon inaugurale (given on Friday January 10th 1997)*. Collège de France, Chaire internationale.
- Mel'čuk, I., & Polguère, A. (2007). *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérives sémantiques et collocations du français*. Bruxelles: De Boeck.
- Mel'čuk, I., & Zholkovsky, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian*. Vienna: Wiener Slawistischer Almanach.

- Mel'čuk, I., Arbatchewsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.-N., et al. (1988). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexicosémantiques II*. Montréal: Les Presses de l'Université de Montréal.
- Mel'čuk, I., Arbatchewsky-Jumarie, N., Elnitsky, L., Iordanskaja, L., & Lessard, A. (1984). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexicosémantiques I*. Montréal: Les Presses de l'Université de Montréal.
- Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., & Mantha, S. (1992). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexicosémantiques III*. Montréal: Les Presses de l'Université de Montréal.
- Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S., & Polguère, A. (1999). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexicosémantiques IV*. Montréal: Les Presses de l'Université de Montréal.
- Mel'čuk, I., Clas, A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.
- Melchuk, I., & Zholkovsky, A. (1984). Explanatory Combinatorial Dictionary of Modern Russian. In *Sonderband 14*. Vienna: Wiener Slawistischer Almanach.
- Milićević, J. (2001). A short guide to the Meaning-Text linguistic theory. In A. Gelbukh, *Intelligent Text Processing and Computational Linguistics*. Mexico: Colección en Ciencias de Computación, Fondo de Cultura Económica - IPN - UNAM.
- Milićević, J. (2006). A short Guide to the Meaning-Text Linguistic Theory. *Journal of Koralex* (Vol. 8), pp. 187-233.
- Papadopoulou, L. (to appear). "My deepest condolences": Lexical functions of Greek pragmatemes [in a funeral]. In *Proceedings of the 11th International Conference on Greek Linguistics, 26 - 29 September 2013, Rhodes, Greece*.
- Polguère, A. (1998). La théorie Sens-Texte. *Dialangue*, Vol. 8-9, pp. 9-30.
- Polguère, A. (2000). Towards a Theoretically-Motivated General Public Dictionary of Semantic Derivations and Collocations for French. In *Proceedings of EURALEX 2000* (pp. 517-527). Stuttgart.

# Analyzing Specialized Verbs in a French-Italian-English Medical Corpus: A Frame-based Methodology

Anna Riccio  
University of Naples “L’Orientale”  
ariccio@unior.it

## Abstract

The aim of this study is to investigate the semantics and syntax of verbs in French, Italian, and English medical discourse by exploring the relationship between verb semantics and argument realization. The verbs under consideration are common lexical units which have acquired the status of a term through their specialization of meaning, such as *affect*, *involve*, etc. Unlike terminological verbs (e.g. *keratinize* or *lyophilize*), they have a lower level of technicalness, and co-occur with arguments (usually terms) in syntagmatic units. The data are extracted from the parallel EMEA corpus including documents published by the European Medicines Agency. The description of the verbs is based on the theoretical model of Frame Semantics (Fillmore1977a-b, 1982, 1985; Fillmore and Atkins 1992) and on the FrameNet methodology (Ruppenhofer et al. 2010). The resultant analysis of the collected data reveals a sentence-level scenario (i.e., the Damaging frame) which groups together verbal forms which share similar syntactic and semantic valence patterns, both within and across languages.

**Keywords:** specialized verbs; medical domain; parallel corpus; intra-/cross-linguistic equivalents; Frame Semantics; FrameNet methodology

## 1 Introduction

The special status of the verb in terminological resources, rather than the noun, is an issue which has been widely discussed in literature over the last fifteen years. See, among others, Picht (1987), L’Homme (1995, 1998), Lorente and Bevilacqua (2000), Valente (2000), Costa and Silva (2004), and more recently De Vecchi and Estachy (2008), Tellier (2008), Pimentel (2012) and Pettersson (2013). Verbs, like nouns, tend to have particular usages within situational communication between experts of specific fields.

The initial stage of this study focuses on the analysis of certain verbs that can have an unusual significance, or a meaning which is specific to the medical field, such as *affect*, *involve*, *enhance*, etc. Unlike terminological verbs (e.g. *keratinize* or *lyophilize*), they have a lower level of technicalness.

The observation of the behaviour of verbal forms in a corpus of medical texts has been explored by Tellier (2008) and Pettersson (2013) in French. In this work, verbs are examined in French, Italian and

English. The data are extracted from the parallel (translation) EMEA corpus from the EMA (European Medicines Agency). The corpus-based analysis of specialized verb equivalents (lexical units which have the same meaning and usage intra- and cross-linguistically) may be useful for the elaboration of a multilingual terminological resource which covers the subject field of medicine. This could be useful for translators, the teaching of specialized translation and terminological or technical writers.

The description of the verbs in question is based on the theoretical model of Frame Semantics (Fillmore 1977a-b, 1982, 1985; Fillmore and Atkins 1992) and on the FrameNet methodology (Ruppenhofer et al. 2010), because verbs are “frame-evoking” or “frame-bearing” words par excellence (Pimentel 2012: 5). Each specialized verb evokes a semantic frame representing a sentence-level scenario which groups together verbal forms that share similar syntactic and semantic valence patterns. The study also tests the hypothesis that semantic frames can function as “interlingual representations” in the organization of a multilingual lexicon (Boas 2005).

This paper is organized as follows. Section 2 provides a brief description of the research methodology: the instruments used for the data collection (2.1., 2.2.) and the theoretical model of Frame Semantics as well as the FrameNet methodology (2.3.). Section 3 illustrates the frame of the specialized verbs examined in this study (The Damaging frame) and their morphological and syntactic patterns. Section 4 follows with some concluding remarks.

## **2 Methodology**

### **2.1 Corpus**

The data are extracted from the multilingual parallel (translation) corpus EMEA from the European Medicines Agency (available in 22 European official languages). The corpus is made up of PDF documents which are representative of a genre of written medical discourse, specifically, package leaflets for medicinal products (Tiedemann 2009). The leaflets are specialized texts that make use of one of the different types of communication between experts and non-experts, such as doctor-patient interactions.

The corpus includes over 311,65 million tokens in all, 14,9 million of which are in French, 14,1 million in Italian, and 12,1 million in English. The corpus is available through the OPUS site (<http://opus.lingfil.uu.se/>) and can also be accessed through the Sketch Engine interface (Kilgarriff et al. 2004). The verbal items are collected and organized using the Sketch Engine to facilitate their quantitative and qualitative analysis.

Table 1 illustrates the verbal word-types and word-tokens in each language (i.e. French, Italian and English):

Language	Type-frequency	Token-frequency
French	1862	1,836,737
Italian	1855	1,256,154
English	1843	1,328,560

**Table 1: Type and token frequency of verbs in EMEA.**

The Type- / Token-frequency lists also include verbs which do not have any kind of specific value in medical discourse. Thus, specialized verbs have to be selected from the list of concordances for each language. The contexts which have been examined thus far show a large number of specialized verbs among the three languages. This article simply presents the preliminary results on 8 verbal lexical items (see table 2 below), which actually allow us to observe their special status within medical terminology and lexicology.

## 2.2 Data

The specialized verbs in question are those which Lorente (2000) calls *verbos fraseológicos* (Eng. ‘phraseological verbs’), which are different from the *verbos terminológicos* (Eng. ‘terminological verbs’). The former are predicative verb units that appear in specialized texts in order to express states, actions and processes. When isolated, their meaning is similar to the meaning of the verbs in non-specialized contexts, e.g. (Fr.) *administrer*, (It.) *somministrare*, (Eng.) *administer*. However, when they co-occur with arguments (usually terms) in syntagmatic units they acquire a specialized value. For example, we usually say (Fr.) *administrer un médicament*, (It.) *somministrare un farmaco*, (Eng.) *administer a medicine*, but not (Fr.) *donner un médicament*, (It.) *dare un farmaco*, (Eng.) *give a medicine*, even if their respective meanings in such contexts could justify the alternation of verbal forms. See examples in (1a-c):

- (1) a. Lorsqu’il est nécessaire d’administrer des produits radiopharmaceutiques chez la femme en âge de procréer, [...].  
 b. Quando è necessario somministrare un prodotto radioattivo ad una donna potenzialmente gravida, [...].  
 c. Where it is necessary to administer radioactive medicinal products to women of childbearing potential, [...].

Phraseological verbs include verbs that appear in collocations (strict lexical selection), in fixed phrases and also in support verb constructions.

The verbs examined in the first stage of this study are listed in Table 2:

French	Italian	English
affecter	coinvolgere	affect
atteindre	interessare	involve
intéresser		
toucher		

**Table 2: Verbs examined in French, Italian and English.**

Consider the Italian verb *interessare* and its French equivalent *intéresser*. Such verbs are mainly used by experts in the field, and they can be substituted by other words related to the general language (see table 2) without affecting the ‘scientific’ meaning which is given to them (see forward Section 3). Seriani (2005) labels the verbs *interessare/intéresser* as “tecnicismi collaterali” (subtechnical terms), i.e., words (nouns, adjectives, verbs and phrases) which are used to maintain a high, formal register in specialized languages.

Unlike the phraseological verbs, the terminological verbs correspond to those units whose meanings are specifically related to the specialized field, as in (2a-c):

- (2) a. Des études *in vitro* ont montré que l’irbésartan est oxydé principalement par l’isoenzyme CYP2C9 du cytochrome P450 [...].
- b. Studi *in vitro* indicano che irbesartan viene principalmente ossidato tramite il citocromo P450-enzima CYP2C9 [...].
- c. *In vitro* studies indicate that irbesartan is primarily oxidised by the cytochrome P450 enzyme CYP2C9 [...].

These verbs often have deverbal nouns, which are terms themselves and should be included in terminological resources, e.g. (Fr.) *oxidation*, (It.) *ossidazione*, and (Eng.) *oxidation*.

### 2.3 Theoretical framework

Over the last few years, some researchers have proposed frame-based organizations of specialized fields, in other languages as well as English, such as environmental science (see Faber et al. 2005, among others), law (see Alves et al. 2005, among others), soccer (see Schmidt 2006 and his following writings), molecular biology (Dolbey et al. 2006 and his following writings), computing and the Internet (see L’Homme 2008).<sup>1</sup>

Frame Semantics (Fillmore 1977, 1982, 1985; Fillmore and Atkins 1992) is a theory of language understanding based on the principle that the meaning of a linguistic item (Lexical Unit, LU) interacts with the scene which it has activated (“meanings are relativized to scenes”, Fillmore 1982). Thus, Fra-

1 For a full bibliography on the application of Frame Semantics within LSPs, see Pimentel (2012).



me Semantics contributes towards understanding the significance of the verbal syntactic patterns, as well as the understanding of the components (Frame Elements, FEs) that form them semantically. For instance, defining the verbal lexical unit *learn* presupposes an educational teaching strategy (i.e., Education\_teaching frame). Specialized verbs are often accompanied by other information (non-core Frame Elements, non-core FEs) that may be optionally added to a sentence.

The methodological approach applied to the analysis of specialized verbs is both bottom-up and top-down: the verbs are analyzed and grouped into frames for each language separately, and the use of specialized dictionaries and other reference resources provides helpful background information (Faber et al. 2009: 6). Thus, the analysis of text corpus allows us to observe how the arguments (core FEs and non-core FEs), the organization of syntax and the semantic connection between words put together specialized verbs and their suitable equivalents intra- and cross-linguistically.

The possibility of creating a multilingual specialized lexicon using the FrameNet database of its English-specific lexical descriptions is considered by Boas (2005), since semantic frames are conceptual structures independent of language. In this study, frames are assumed to be “interlingual representations” that can group together not only verbs in one language but also across several languages (French-Italian-English), by transferring semantic annotations from one language to another (Padó 2007; Baker 2009). Thus, frames can group together intra-linguistic and cross-linguistic equivalents (synonyms, near-synonyms, hyponyms, related LUs), as described in the next section.

### 3 Results

All the verbs examined in this study (see table 2) can be grouped together in the Damaging frame, since they all mean ‘to have a strong effect on something or someone’, or ‘causing physical damage to something or someone’, as shown in Table 3 (below). However, the Lexical Unit *to affect* is semantically identified with a general meaning in the FrameNet database, and it is linked to the Objective\_influence frame.<sup>2</sup> This frame is the Parent frame of the Transitive\_action frame from which the Damaging frame originates. Therefore, the Damaging frame is a Child frame which inherits from more than one Parent frame (multiple inheritance). Unlike *to affect*, the verb *involve* is not listed as a Lexical Unit in the database. Only the adjective *involved* and the noun *involvement* are included, and both belong to the Participation frame.<sup>3</sup> The corpus shows that the verb *involve* is used frequently as a synonym of the

---

2 The definition of the Objective\_influence frame is as follows: “An Influencing\_variable, an Influencing\_situation, or an Influencing\_entity has an influence on a Dependent\_entity, Dependent\_variable, or a Dependent\_situation”.

3 The definition of the Participation frame is as follows: “An Event with multiple Participants takes place. It can be presented either symmetrically with Participants or asymmetrically, giving Participant\_1 greater prominence over Participant\_2. If the Event is engaged in intentionally, then there is typically a shared Purpose between the Participants. It is, however, possible that an expressed Purpose only applies to Participant\_1.”

verb *affect* in medical discourse, and therefore it can be considered as a Lexical Unit of the Damaging frame.

Frame	Damaging
Definition	An AGENT affects a PATIENT in such a way that the PATIENT (OR some SUBREGION of the PATIENT) ends up in a non-canonical state. Often this non-canonical state is undesirable, and some lexical units (marked with the Negative semantic type) specifically indicate that the PATIENT is negatively affected.
Core FEs	AGENT [Agt] The conscious entity, generally a person, that performs the intentional action that results in the damage to the Patient. CAUSE [cau] An event which leads to the damage of the Patient. PATIENT [Pat] The entity which is affected by the Agent so that it is damaged.
Non_core FEs	CHARACTER_OF_END_STATE, DEGREE, INSTRUMENT, MANNER, MEANS, PATIENT, PLACE, PURPOSE, REASON, RESULT, SUBREGION, TIME
Contexts	<p>Selon le NCI-CTC, les réactions cutanées de grade 2 sont caractérisées par une éruption <b>intéressant</b> jusqu'à 50 % de la surface corporelle, alors que les réactions de grade 3 affectent 50 % ou plus de la surface corporelle.</p> <p>Secondo i criteri NCI-CTC, le reazioni cutanee di grado 2 sono caratterizzate da rash che <b>interessa</b> fino al 50 % della superficie corporea, mentre quelle di grado 3 <b>interessano</b> il 50 % o più della superficie corporea.</p> <p>According to NCI-CTC, grade 2 skin reactions are characterized by rash up to 50 % of body surface area, while grade 3 reactions <b>affect</b> equal or more than 50 % of body surface area.</p> <p>Cette nécrose peut <b>atteindre</b> fascias musculaires ainsi que le tissu adipeux et peut par conséquent provoquer la formation d'une cicatrice.</p> <p>Questa può essere estesa e può <b>interessare</b> lo strato muscolare così come lo strato adiposo causando quindi la formazione di cicatrici.</p> <p>It can be extensive and may <b>involve</b> muscle fascia as well as fat and therefore can result in scar formation.</p> <p>Sintomi che coinvolgono il cervello e i nervi che si sono <b>manifestati</b> nell'arco di un mese [...]</p> <p>Réactions <b>touchant</b> le cerveau et les nerfs apparues dans le mois suivant la vaccination [...]</p> <p>Symptoms <b>affecting</b> the brain and nerves that have occurred within one month after vaccination [...]</p> <p>Les cas les plus graves ont été rapportés chez des patients prenant d'autres médicaments ou atteints de maladies pouvant <b>toucher</b> le foie (exemple alcoolisme, infection sévère).</p> <p>I casi più gravi sono stati osservati in pazienti trattati anche con altri medicinali o affetti da disturbi che possono <b>interessare</b> il fegato (ad es. abuso di alcolici, infezioni gravi).</p> <p>The most serious were reported in patients taking other drugs or who were suffering from diseases that can <b>affect</b> the liver (e.g. alcohol abuse, severe infection).</p> <p>S'ils ne sont pas <b>atteints</b>, la main et le pied doivent être protégés par une bande d'Esmarch, un garrot doit être placé au niveau proximal du membre.</p> <p>Mano e piede, se non <b>interessati</b>, devono essere protetti da bendaggi Esmarch (espulsione).</p> <p>Hand and foot, if not <b>affected</b>, should be protected by Esmarch (expulsion) bandages.</p>

Table 3: The Damaging frame.

The Damaging frame groups together 8 candidate equivalents, i.e. 4 French verbs, 2 Italian verbs and 2 English verbs, more specifically 16 likely combinations of equivalents, as shown in more detail in Table 4:

French			Italian			English		
Cause	target	Patient	Cause	target	Patient	Cause	target	Patient
éruption	intéresser	surface corporelle	rash	interessare	superficie corporea		-	
réaction cutanée	affecter	surface corporelle	reazione cutanea	interessare	superficie corporea	reaction	affect	surface area
nécrose	atteindre	fascia musculaire	questa (necrosi)	interessare	strato muscolare	it (necrosis )	involve	muscle fascia
-	atteint(e)	osseuse	tumore maligno	interessare	osso	malignancie	involve(ing)	bone
maladie	toucher	foie	disturbo	interessare	fegato	disease	affect	liver
affection parodontale	toucher	gencive	disturbo parodontale	interessare	gengiva	periodontal	affect	gum
réaction	toucher	cerveau	sintomo	coinvolgere	cervello	symptom	affect	brain
	affect(ion)	moelle osseuse	patologia	coinvolgere	midollo osseo	conditions	affect(ing)	bone marrow

**Table 4: Cross-linguistic comparison of verb (or *noun*) equivalents and FEs.**

Most of the FEs in Table 4 are synonyms (or semantic equivalents) because they have similar meanings and distributions (uses). In a few cases, the corpus presents transcategorization phenomena from the verbal to the nominal form, as exemplified in (3a-c):

- (3) a. des patients atteints de pathologie maligne à un stade avancé avec atteinte osseuse  
 b. tumori maligni allo stadio avanzato che interessano l'osso  
 c. in patients with advanced malignancies involving bone

According to L'Homme (2004), the presence of deverbal nouns, such as (Fr.) *atteinte* (<*atteindre*), (It.) *interessamento* (<*interessare*), (Eng.) *involvement* (<*involve*), establishes the specialized value of these verbs (see Section 2.2. for nouns derived from terminological verbs).

French is the language with the most verbal equivalents, since it distinguishes 4 items, whereas Italian and English contexts show 2, respectively. The English verbs *affect* and *involve* have peculiar features that characterize them as synonyms. In Italian as well as in French, the verbs can also be defined as hyponyms or hyperonyms. For instance, the Italian verb *coinvolgere* is a hyponym of *interessare*. All the verbs in Table 4 are equivalents because no particular differences have been observed: they have the same number of arguments (NP/Subject, NP/Object), the semantic nature of the arguments does

not differ (CAUSE and PATIENT) (they refer to the same kind of entities), and their syntactic patterns are similar (see Pimentel 2012 for the criteria identifying equivalents). Only in a few cases, are the verbs not translated because of a syntactic change, e.g. (Fr.) *réaction allergique sévère touchant le corps entier*, (It.) *una grave reazione allergica dell'intero organismo*, and (Engl.) *a severe, whole-body allergic reaction*.

The analysis of the data allows us to identify the typical verbal features (tense, person, number, voice, and mood) which characterize the leaflets, and medical discourse in general. The grammatical persons are the singular and plural third-persons. The realis mood (indicative) is obviously the commonest, whereas the unrealis moods, such as the conditional, imperative and subjunctive forms are less frequent. For instance, the use of the conditional form in Italian and in French are 0,87% and 0,63%, respectively. In relation to the grammatical tense, the present is the most common tense when the indicative is used: (Fr.) 34,56%, (It.) 33,74%, (Engl.) 17,57%. A further note deserves to be made for the grammatical voice: the use of the passive construction placing the thema of a sentence at the beginning of the clause and the rhema at the end is very common in medical discourse, and generally allows one to omit the agent, e.g. *COX-2 is also thought to be involved in ovulation*.

## 4 Concluding remarks

The exploitation of specialized parallel corpora makes it easy to identify the repertoire of both intra-linguistic and cross-linguistic verb equivalents which acquire specialized value when used in medical texts. The Frame Semantics analysis of each verb pattern as well as the FrameNet methodology allow us to make a description of the interaction of the lexeme, syntax and conceptual background frame. All the verbal items evoke the same frame (Damaging) describing physical damage to something or someone. Thus, the lexicological findings could be useful for the development of a multilingual lexicographical resource specialized in the medical field which could give support with L2 writing. This of course involves a comprehensive and systematic investigation.

## 5 References

- Alves, I., Chishman, R. & Quaresma, P. (2005). Verbos do domínio jurídico: uma proposta de organização ontológica com vistas ao PLN. In *Revista de Estudos Linguísticos da Universidade Federal de Juiz de Fora*, 9(1/2), pp. 123-137.
- Baker, C. F. (2009). La sémantique des cadres et le projet FrameNet: une approche différente de la notion de «valence». In *Langages*, 4, pp. 32-49.
- Bertoldi, A., Chishman, R. (2012). Frame Semantics and Legal Corpora Annotation: Theoretical and Applied Challenges. In *Linguistic Issues in Language Technology*, 7(9), pp. 1-15.
- Boas, H. (2005). Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. In *International Journal of Lexicography*, 18(4), pp. 39-65.

- Costa, R., Silva, R. (2004). The verb in the terminological collocations. Contribution to the development of a morphological analyser Morphocomp. In *Proceedings of the IV International Conference on Language Resources and Evaluation, May 26-28, 2004*. Lisbon, Portugal, pp. 1531-1534.
- De Vecchi, D., Eustachy, L. (2008). Pragmaterminologie: les verbes et les actions dans les métiers. In *Actes des conférences Toth 2008, Annecy 5-7 June 2008*, pp. 35-52.
- Dolbey, A., Ellsworth, M. & Scheffczyk, J. (2006). BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies. In O. Bodenreider (ed.), *Proceedings of KR-MED*, pp. 87-94.
- Faber, P., Márquez Linares, C. & Vega Expósito, M. (2005). Framing Terminology: A Process-Oriented Approach. In *Meta*, 50(4), CD-ROM.
- Faber, P., León, P. & Prieto, J. A. (2009). Semantic relations, dynamicity and terminological knowledge bases. In *Current Issues in Language Studies*, 1(1), pp. 1-23.
- Fillmore, C. (1977a). Scenes-and-Frames Semantics, Linguistic Structures Processing. In A. Zampolli (ed.), *Fundamental Studies in Computer Science*, 59, pp. 55-88.
- Fillmore, C. (1977b). The Case for Case Reported. In P. Cole, J. Sadock (eds.), *Syntax and Semantics, Volume 8: Grammatical Relations*. New York: Academic Press.
- Fillmore, C. (1982). Frame Semantics. In *Linguistics in the Morning Calm* (ed.), Seoul: Hanshin Publishing Co, pp. 111-137.
- Fillmore, C. (1985). Frames and the Semantics of Understanding. In *Quaderni di Semantica*, 6(2), pp. 222-254.
- Fillmore, C., Atkins, S. (1992). Towards a Frame-based Lexicon: The semantics of RISK and its Neighbors. In A. Lehrer, E. Kittay (eds.), *Frames, Fields, and Contrast: New Essays in Semantics and Lexical Organization*, Hillsdale: Lawrence Erlbaum Associates, pp. 75-102.
- FrameNet: <https://framenet.icsi.berkeley.edu/fndrupal/home> [September-October 2013; February-March 2014]
- Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In W. Geoffrey, S. Vessier (eds.), *Proceedings of the XI Euralex International Congress, July 6-10, 2004, France: Lorient*, pp. 105-111.
- L'Homme, M. C. (1995). Définition d'une méthode de recensement et de codage des verbes en langue technique: applications en traduction. In *Traduction, terminologie, rédaction*, 8(2), pp. 67-88.
- L'Homme, M. C. (1998). Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie*, 73(2), pp. 61-84.
- L'Homme, M. C. (2004). *La terminologie: principes et techniques*. Montréal: Les Presses de l'Université de Montréal.
- L'Homme M. C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. In *Traduire*, 271, pp. 78-103.
- Lorente, M. (2000). Tipología verbal y textos especializados. In M. González Pereira, M. Souto Gómez (eds.), *Cuestiones conceptuales y metodológicas de la lingüística*. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 143-153.
- Lorente, M., Bevilacqua, C. (2000). Los verbos en las aplicaciones terminográficas. In *Actas del VII Simposio Iberoamericano de Terminología RITerm 2000*. Lisboa: ILTEC.
- OPUS Corpus : <http://opus.lingfil.uu.se/> [September-October 2013]
- Padó, S. (2007). Cross-lingual Annotation Projection Models for Role-Semantic Information. Ph. D. thesis. Saarland University.
- Pettersson, Å. (2013). Les syntagmes participiaux et les verbes spécialisés dans un texte médical: Une étude contrastive entre le français et le suédois. Ph. D. thesis. Linnaeus University.
- Pimentel, J. (2012). Identifying the equivalents of specialized verbs in a bilingual corpus of judgments: A frame based methodology'. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul (Turkey), 21-27 May 2012, pp. 1791-1798.
- Picht, H. (1987). Terms and their LSP Environment - LSP Phraseology. In *Meta*, 32(2), pp. 149-155.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R. & Scheffczyk, J. (2010). *FrameNet II: Extended Theory and Practice*. ICSI Technical Report.

- Schmidt, T. (2006). Interfacing Lexical and Ontological Information in a Multilingual Soccer FrameNet. In Proceedings of OntoLex 2006, Interfacing Ontologies and Lexical Resources for Semantic Web Technologies, Genoa, Italy, May 24-26, 2006.
- Serianni, L. (2005). Un treno di sintomi. I medici e le parole: percorsi linguistici nel passato e nel presente. Milano: Garzanti.
- Tellier, C. (2008). Verbes spécialisés en corpus médical: une méthode de description pour la rédaction d'articles terminologiques. PhD thesis. Université de Montréal.
- Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, R. Mitkov (eds.), Recent Advances in Natural Language Processing (vol V). Amsterdam/Philadelphia: John Benjamins, pp. 237-248.
- Valente (2000). Peut-on considérer que le verbe est une unité lexicale spécialisée? Description de verbes spécialisés portugais. In *TradTerm*, 6, pp. 171-187.

# Neoclassical Formatives in Dictionaries

Pius ten Hacken, Renáta Panocová

Leopold-Franzens-Universität Innsbruck, P.J. Šafárik University Košice

Pius.ten-Hacken@uibk.ac.at, Renata.Panocova@upjs.sk

## Abstract

Neoclassical formatives are elements that occur in neoclassical word formations such as *ortho* and *paed(o)* in *orthopaedic*. We can be sure that a separate system of neoclassical word formation is in place when we find words such as *orthopaedic* that use classical elements but cannot have been borrowed from a classical language, because the concept they refer to did not exist yet. However, it must be taken into account that such words can also have been borrowed from a modern language that has such a system. In the lexicographic treatment of neoclassical word formation, a central question is whether neoclassical formatives should be treated in separate entries. We investigated how different English and Russian dictionaries treat them. In order to arrive at an unbiased sample of formatives, we used a Catalan word formation dictionary.

The results of our investigation support the hypothesis that neoclassical word formation constitutes a separate system in English, but not in Russian. This means that in English neoclassical formatives should have entries. In electronic dictionaries they can usefully be connected to the full class of words they appear in. In Russian, neoclassical formations are borrowings from languages such as English or French and their internal structure belongs to the domain of etymology.

**Keywords:** neoclassical word formation; neoclassical formatives; combining forms; English; Russian

In this paper we investigate the optimal treatment of neoclassical word formation in dictionaries. The main problem of neoclassical word formation in lexicography is to determine how to treat the internal structure and the components that are not words by themselves. First, we give a general introduction to the phenomenon of neoclassical word formation (section 1) and present some considerations as to the treatment of word formation in dictionaries (section 2). In order to investigate the coverage of neoclassical word formation in dictionaries, we present a sampling method that is not directly dependent on the languages and dictionaries under investigation (section 3). Then we compare the realization of neoclassical word formation and its treatment in some standard general dictionaries in two languages, English (section 4) and Russian (section 5). These languages were chosen because, as we will argue, the status of neoclassical word formation in them is different in interesting ways. On the basis of our observations in the preceding sections, we conclude in section 6 that there are good arguments for treating neoclassical formatives in separate entries in English, but not in Russian.

## 1 Neoclassical Word Formation

The discussion of neoclassical word formation has a long history. In many cases, it is connected to the opposition between learned and non-learned word formation. Bloomfield (1933: 153-54) discusses the opposition between unmarked and learned forms in relation to the way the lexicon of a language is organized and gives examples for various languages. He discusses the phenomenon of learned language mainly as a matter of register. Here, however, we will be interested only in the specific type of word formation involving Greek and Latin stems that is found in a wide range of European languages. An example from English is *orthopaedic*.

An issue that arises immediately in the context of neoclassical word formation is the boundary between word formation and borrowing. This issue arises in two shapes. The first can be illustrated on the basis of the contrast between *orthogonal* and *orthopaedic*. Both of these are based on Ancient Greek, but not in the same way. In Ancient Greek, the word ὀρθόγωνος ([orthógonos] ‘rectangular’) is attested and can be analysed as formed from ὀρθός ([orthós] ‘straight’) and γωνία ([gonía] ‘angle, corner’). Therefore, it is possible to classify *orthogonal* in English as a borrowing from Ancient Greek. In the case of *orthopaedic*, there is no corresponding word in Ancient Greek. However, in Ancient Greek we have the words ὀρθός ([orthós] ‘straight’) and παῖς (stem παιδ- [paid-], ‘child’) which can be analysed as the basis for *orthopaedic*. This is the first realization of the issue of borrowing.

However, there is a second form in which this question arises. In fact, the non-existence of a corresponding Ancient Greek word does not necessarily show that the word was formed in English. In this case, it is possible to trace the origin to French. The French physician Nicolas Andry de Boisregard (1658-1742) created the term *orthopédique* in 1741. From French it was subsequently adopted in other European languages. In this case, it is possible to trace the formation of the word to a single person and a single publication, because the concept was invented by this person. In many cases it is much more difficult to trace the origin so precisely. If a concept is ‘in the air’ and arises in a context in which different people might think of the same name independently, or spread it very rapidly once someone starts using it, it is almost impossible to reconstruct which language is at the origin of a particular word.

In order to determine the significance in lexicography of the question of borrowing or formation in any particular case, we can choose between two perspectives. One is to aim for a description of the historical development of language use. In this perspective it is important to distinguish the origin of *orthogonal* and *orthopaedic*, stating that the former is borrowed from Ancient Greek and the latter from French where it was constructed from Ancient Greek components. The second possible perspective concentrates rather on the vocabulary as known by the speakers. Here the question is whether these words are structured or not. Most speakers will not be aware of the Ancient Greek origin of *orthogonal* or the French origin of *orthopaedic*. The difference between words formed in Ancient Greek and those formed in modern languages is not one that determines the structure of the mental lexicon of a contemporary speaker of English.



Following ten Hacken (2012), we will assume that in the case of genuine neoclassical word formation, there must be a system of neoclassical formatives in the lexicon. Such a system can emerge as the result of the reanalysis of borrowings. After a sufficient number of words with similar components have been borrowed, speakers may notice the regularity. In the first instance, both a new system and a new set of formatives have to be set up. It is quite likely that *ortho-* belonged to the initial set of formatives, because there are quite a number of old loanwords from Ancient Greek that contain it, e.g. *orthogonal*, *orthodox*, *orthography*. Once the system exists, it is easier to extend it. A single item may be sufficient. Thus, for *paedo-*, the origin may be a word such as *paedagogical*. We even find cases in which a neoclassical formative is borrowed directly from Ancient Greek, e.g. *psepho-* in *psephological*. This was not a formative in English when the word *psephological* was created in the 1950s. The first attestation in the OED is from 1952.

Systems of neoclassical word formation have come into existence in a number of languages, probably in the course of the 18th century. In fact, *orthopédique* is a very early example of a neoclassical formation that cannot have been borrowed directly from Ancient Greek. A much larger set of such words start appearing in the 19th century. Because languages such as French, English, and German all went through the same process, it is often difficult to determine which of these languages is at the origin of a particular word. However, the question of which modern language is at the origin of a particular formation is only relevant if we want to describe the history of a individual words. Speakers of such languages who have the neoclassical subsystem of word formation in their mental lexicon will analyse the relevant words independently of whether the original source is in their own language or in another one. Therefore, it is less relevant which of these languages is the historical origin of a particular word. Once it exists in one language it is quickly adopted in the others.

This is especially the case for words such as *orthopaedic*, because they belong to a very specialized register of language. Neoclassical words are part of the language of science. Not many speakers of English will have known them in the 18th or 19th century, but those who did are more likely to be in contact with and competent in other languages, such as French and German. As languages such as English, French and German only exist as entities when they are rational constructions associated with a conscious aim for codification, the decision to attribute the origin of a particular word to one of them is ultimately arbitrary.

## 2 Word Formation in Dictionaries

As argued in ten Hacken (2009), dictionaries do not describe the vocabulary of a language, but provide information about words to dictionary users. The aim of a description is not realistic because there is no suitable empirical object to describe. What we have is a number of speakers and a set of texts and utterances. The idea of a language to be described in a dictionary requires the classification of the spe-

akers and texts/utterances and is in this sense a constructed entity, not one empirically found as an entity.

The role of word formation in dictionaries is significantly enhanced by the insight that dictionaries are no more or less than sources of information for their users. Ulsamer (2013) gives an overview of current practice and insights as to the representation of word formation in dictionaries of various types. Booij's (2003: 254) suggestion that the place of word formation in dictionaries is mainly to state "if a possible morphologically complex word actually exists" is unnecessarily restrictive. First of all, Booij's wording suggests that there is an independent sense in which a word should exist. In the same way as for languages, there is no empirical entity corresponding to the word. What the lexicographer has is a number of speakers and a corpus of texts and utterances. In addition, Booij's suggestion reduces the role the representation of word formation can play in supporting the dictionary user.

On the basis of the available corpus and linguistic knowledge, for each morphologically complex word, lexicographers have to take decisions such as the following:

- whether a particular word should be represented in the dictionary,
- whether it should be represented as morphologically complex,
- which information should be given about the word, and
- how this information should be presented.

It is crucial to understand that these are decisions based on judgements. The available data in a corpus can serve to support the judgement, but they do not result in a decision in an algorithmic sense. There is also no objective truth to be discovered as to the existence or the nature of the words considered. Svensén (2009: 131-133) sketches some of the traditional techniques that have been used, including the so-called *run-on* entry that Booij seems to allude to.

In electronic dictionaries, the possibilities for representing and accessing information are greatly enhanced compared to paper dictionaries. However, it is by now well-known that these additional possibilities depend on an appropriate encoding of the information. It is not sufficient to present information in the way it is encoded in a print dictionary and just make it available in electronic form. This argument was made quite systematically by ten Hacken (1998) and has now become commonplace. Ulsamer (2013: 35-50) can refer to a number of dictionaries and lexical resources that were developed with this idea in mind and exploit the specific strengths of the electronic medium.

Ten Hacken (1998) distinguishes three types of lexicographic representation of word formation. One type focuses on the word formation analysis of individual words. If we take as an example *happiness*, this means that the information that this word was formed by adding to the suffix *-ness* to the adjective *happy* is given in the dictionary entry for *happiness*. Giving this information is not as commonplace in paper dictionaries as one might expect. Most general dictionaries only give a run-on entry and in learners' dictionaries we often find a separate entry without this information. In both cases, this

leaves it to the user to work out the structure of the word. In electronic dictionaries, there is of course no excuse based on space limitations to motivate such a decision.

A second type of information that dictionaries can give about word formation is the grouping of words that belong to the same word formation class. Here a word formation class is interpreted as a set of words that have some word formation properties in common. Ideally, it would be up to the user to choose which properties are most relevant for their question. In the case of *happiness*, the most obvious class would be that of nouns formed by suffixation of *-ness* to an adjective. This is a very large class and in a paper dictionary it is not normally presented, but in an electronic dictionary it is feasible to do so. Most users will only want to consider some examples, but linguistic researchers (a minority user group, but heavy dictionary users) may well be interested in the entire set. Other word formation classes that *happiness* is a member of include the class of words formed with *happy* as a basis, the class of nouns formed from adjectives, or the class of nouns formed by suffixation. Especially for the larger classes in these examples, any representation in an alphabetically ordered print dictionary will be too unwieldy, both to print it and to use it, but as Ulsamer (2013) shows, in electronic dictionaries there is an advantage in the proper representation of some of the possible classes.

The third type of representation concentrates on the rule as such. In the case of *happiness*, this may mean, for instance, an entry for *-ness* describing the word formation process. This is actually a common type of information also in print dictionaries. COED (2011) gives an entry for *-ness* with the different senses and an example for each sense. Atkins & Rundell (2008: 180) mention such entries as possible lemmata under the category of “partial words” and Svensén (2009: 132-133) presents it as one of the main representations of word formation in the dictionary. Of course, the representation in an electronic dictionary can be much improved by making a hyperlink available between the entries for the individual words (e.g. *happiness*) and the entry explaining the word formation rule, as well as by linking entries such as for *-ness* to the corresponding word formation class (i.e. all nouns formed with *-ness*).

The questions we want to address in relation to neoclassical word formation are then the following:

- which of these possible representations are used in current dictionaries?
- which improvements could be made in these representations, especially in electronic dictionaries? and
- to what extent do the choices depend on properties of different languages?

In order to provide a proper basis for the discussion of the last of these questions, we will consider two languages, English and Russian, in which it can be argued that the mechanism of neoclassical word formation does not work in the same way. First, however, we have to find a way to collect data for the first of these questions.

### 3 The Lexicographic Representation of Neoclassical Word Formation

In general, the decision how to treat a particular word formation process in a dictionary is determined at least in part by the productivity of the process. As shown by Bauer (2001), *productivity* is a notion that can be interpreted in different ways and has therefore raised a lot of discussion. A very useful analysis of the different notions of productivity is the one by Corbin (1987: 176-178). She distinguishes three aspects of productivity that in obvious cases will point in the same direction, but also give the conceptual vocabulary for a discussion about the choice of relevant properties in more controversial cases. The first is *régularité* ('regularity'). This is the extent to which the form and meaning of a particular formation can be predicted on the basis of the input (base) and the rule. The second is *disponibilité* ('availability'). This is a binary feature, indicating whether or not the rule is available for application to new bases. Finally, there is *rentabilité* ('profitability'). This is the degree to which the rule is applied to many new bases, yielding new formations. In her own work, Corbin (1987) concentrates only on availability. This is understandable because it is the underlying condition that has to be met in the linguistic competence before the other aspects can apply at all. However, in lexicography, the other two concepts are also relevant. Thus, regularity of new formations clearly influences to what extent run-on entries can be used and the profitability of a process will determine how important it is to treat it in the dictionary at all.

Much of the discussion of word formation in the context of lexicography concerns affixation. In the case of affixation it makes sense to consider each particular affix as a rule for which productivity (in its different aspects) can be calculated. In the case of *-ness*, the availability for new formations and the high degree of profitability make it a good choice for a separate entry. The regularity of many of the individual formations, such as *happiness*, makes it attractive to treat these as run-on entries.

If we want to extend this approach to productivity from affixation to neoclassical word formation, we encounter the problem that many neoclassical formations are more like compounding than like affixation. In the example of *orthopaedic*, there is a suffix *-ic*, but the central piece of the formation of this word is the combination of *ortho* with *paed(o)*. That neoclassical formatives such as *ortho* and *paedo* are not affixes is obvious from their distribution as well as from their contribution to the meaning of the resulting word. In lexicography, they are often called *combining forms*, e.g. by Svensén (2009:133). The variation in form is sometimes accounted for by different entries for the initial combining form and the final combining form. Thus, COED (2011) has different entries for *-phone* and *phono-*. This is not optimal for the insight that they represent the same formative, because there is no link from one entry to the other and a user not actively looking for the two variants will not find the connection, as they are separated by nine entries in the alphabetic order.

Returning now to productivity as a criterion to determine how to treat neoclassical word formation in a dictionary, we are faced by the situation that we have to decide for individual neoclassical formatives whether they deserve an entry and in which other ways they might be referred to. In this deci-

on, affixation is not a good parallel, because neoclassical formatives are not affixes, but compounding is not a good model either. In compounding, the lexicographer has to decide whether or not to devote an entry to the compound, but not whether the components of the compound should be treated. In the case of *apple juice*, the only question is whether it deserves an entry of its own, not whether *apple* and *juice* should have an entry. We do not attempt to determine the productivity of *apple* in compounding as the basis for any lexicographic decision. In the case of *ortho* and *paedo*, we have to consider the number of different formations they appear in and the frequency of those formations in some way to determine whether they deserve being treated in a separate entry.

In order to explore the way neoclassical formations are covered in English dictionaries, we selected a sample of neoclassical formations. In many discussions of neoclassical word formation, a very limited set of examples is discussed time and again. However, there is no indication to what extent these examples are representative of the phenomenon. We considered that a sample based on any specific English dictionary would be biased, in particular when we want to compare the coverage in English dictionaries with the coverage in dictionaries of another language. Therefore we used a source from a language that is not in the scope of the study, Bruguera i Talleda's (2006) Catalan dictionary of word formation.

There are several reasons why Bruguera i Talleda (2006) is a good source for our study of neoclassical word formation. First, Catalan has neoclassical word formation in a way similar to other European languages. As a Romance language, it is not biased to English or Russian and it has a more direct link to Latin than either of these languages, so that we can expect that the set of neoclassical formations tends to be larger. Secondly, the dictionary offers a type of access to neoclassical word formation that is convenient for our purposes. The lemmata of the dictionary are affixes and neoclassical formatives. The entries contain basic information about the use of the formatives and a full list of words formed with them. Where appropriate, entries are divided into different senses. In addition, the introduction (2006: 9-50) contains a detailed discussion of the different types of word formation and a classification of the lemmata. Therefore, it was easy to select a randomized sample of relevant formatives. Specialized dictionaries of this type are quite rare, in particular published as paper dictionaries.

Bruguera i Talleda (2006: 31-48) treats neoclassical word formation as *formaciò culta* ('learned formation'), which connects with Bloomfield's (1933) category of learned word formation as treated in section 1. Neoclassical formatives are listed in the section on neoclassical compounding. This is not ideal in principle, because neoclassical word formation is not restricted to compounding, as evidenced by formations such as *ethnic* or *morpheme*. However, compounding is much more prominent in neoclassical word formation than derivation, so it can be expected that in practice any formative that appears in derivation will also appear in compounding. Neoclassical word formation as it is used, for instance, in medicine is based on Greek formatives that in many cases passed through Latin. Bruguera i Talleda (2006: 38-47) gives separate tables for formatives of Greek and of Latin origin. We only considered the former. In addition, the formatives are divided into initial and final combining forms. There is a large degree of overlap between the two, but the former list is significantly longer. An additional practical

advantage of initial combining forms is that it is immediately evident when looking them up in an alphabetically ordered dictionary which and how many entries use them. Therefore we based our sample of neoclassical formatives on a random selection from the list of initial combining forms of Greek origin. The only adjustment we had to make is to adapt the spelling from Catalan to English and Russian.

## 4 Neoclassical Word Formation in English Dictionaries

For English, we took as our dictionaries CED (2000) and COED (2011). In Béjoint's (2000: 42-91) overview of dictionaries of English, it is obvious that British and American dictionaries follow different patterns. Therefore it would not be possible to generalize from one type of dictionary to the other. However, within British dictionaries, CED (2000) and COED (2011) fall into different subtypes. Béjoint (2000: 57-58) classifies the Oxford dictionaries as traditional, where the Collins dictionary belongs to an innovative trend that started in the 1970s. Both belong to the type van Sterkenburg (2003) identifies as 'the' dictionary, i.e. general-purpose dictionaries of a size big enough to give a fairly comprehensive overview of the vocabulary without giving a full scholarly account of its development. Therefore, especially when the findings of the two dictionaries coincide, we can safely draw conclusions for British dictionaries of this type in general.

The first sample we took was a randomized set of items from Bruguera i Talleda's (2006: 40-44) list of initial combining forms of Greek origin. We found that almost all of them had at least two examples of formations that were described in both English dictionaries. This means that the basic condition for identifying the item as a neoclassical formative would be fulfilled. However, only about a third of the formatives is described in separate entries. Thus, CED (2000) gives *thanatology* and *thanatopsis*, but no separate entry for *thanato-*. The structure of the words is only addressed in the section on etymology, where we find the following:

thanatology: [C19: from Greek *thanatos* death + -LOGY]

thanatopsis: [C19: from Greek *thanatos* death + *opsis* a view]

The difference in presentation suggests that *-logy* and *-opsis* are treated differently in the dictionary. In fact, both have an entry as a combining form. The difference between CED (2000) and COED (2011) is small. There are a few cases where COED (2011) gives the etymology and CED (2000) does not, e.g. *mammography*. Conversely, CED (2000) gives slightly more separate entries for combining forms. Thus, only CED (2000) has an entry for *noso-*. However, on the whole the two dictionaries have a remarkably similar treatment of the formatives in the sample.

We were not fully satisfied with our first sample, because for many of the formatives there were so few items that were listed in the dictionary that it was not clear whether lexicographic decisions concerning the inclusion of particular words or the lexicographic approach to neoclassical formati-

ves were responsible for the treatment we found. Thus, for *thanato-*, CED (2000) gives the two entries listed above, but COED (2011) gives only the first. Therefore, we took a second sample, which only included formatives that were relatively profitable in Corbin's (1987) sense.

For this second sample, we took as a selection criterion the length of the entry in Bruguera i Talleda (2006). As this dictionary gives all attested words with the relevant formative in Catalan, the length of an entry gives a measure of the profitability of the formative. From our first sample, we discarded each item for which the entry in Bruguera i Talleda (2006) is shorter than ten lines. Where the same formative occurs as an initial as well as final combining form, we combined the length of both entries. We replaced the discarded items with formatives that fulfilled this length criterion. The results of looking up this second sample in CED (2000) and COED (2011) showed that these formatives are generally more often described in a separate entry. In COED (2011) more than two thirds of the second sample appears as an entry and in CED (2000) we found almost all of them. This difference is in line with the tendency we observed in the first sample that CED (2000) gives more separate entries than COED (2011). As an example of an entry, COED (2011) gives the following for *diplo*:

**diplo-** ▶ **comb. form 1** double: *diplococcus*. **2** diploid: *diplotene*.

- ORIGIN from Gk *diplous* 'double'.

The corresponding entry from CED (2000) is as follows:

**diplo-** or before a vowel **diplo-** combining form. double: *diplococcus*. [from Greek, from *diploos*, from DI<sup>1</sup> + *-ploos* -fold]

The entries in the two dictionaries are very similar, but it is interesting to note the differences. Only CED gives the variant form. Only COED gives the second sense. This second sense is the result of a shortening of one of the complex forms the formative appears in. It is similar to the use of *gastro-* in the sense of 'gastronomic' rather than 'stomach'. In the etymological information, COED gives the classical form as it is likely to have been at the origin of the borrowing, whereas CED gives the oldest attested form in Greek as well as its word formation origin.

On the basis of the samples of neoclassical formatives we considered, we can conclude that British general dictionaries tend to include separate entries for neoclassical formatives when they appear in many different words. In the entries for these formatives, but also in the ones for neoclassical formations, the connection to the Ancient Greek forms is made explicit. There is a slight difference between CED (2000) and COED (2011) in the sense that the former has more separate entries for neoclassical formatives whereas the latter is more systematic in giving etymologies for neoclassical formations.

## 5 Neoclassical Word Formation in Russian Dictionaries

The position of neoclassical word formation in Russian is not directly comparable to that in English or other Germanic and Romance languages. Neoclassical forms are generally much rarer in Russian



and in many cases, a competing form with Slavic roots is preferred. Therefore, our hypothesis is that Russian dictionaries will give fewer neoclassical formations and will not analyse them as readily as English dictionaries.

In exploring this hypothesis, we used the dictionaries by Ušakov (1946-47) and Efremova (2000). They are roughly comparable in size to the dictionaries we used for English. In the Russian lexicographic tradition, the monolingual dictionary by Ušakov (1946-47) is categorized as a normative dictionary of contemporary standard Russian (Šanskij, 1972: 286; Ožegov, 1974: 171; Germanovič, 1979: 264). The dictionary includes nearly 90,000 entries. Germanovič (1979: 265) points out it is the first dictionary which gives separate entries for productive word formation elements such as prefixes or affixoids. The dictionary by Efremova (2000) is more recent and lists around 140,000 entries. The introductory material to the online version of the dictionary gives the information that prefixes, suffixes, initial elements of complex words and final elements of complex words are described in separate entries.

On the basis of this policy description, one might expect that the findings of checking our samples of formatives in Russian dictionaries would be closer to the results in CED (2000) and COED (2011) for English than originally thought. However, a closer look at the coverage of our first sample of formatives from Bruguera i Talleda (2006) in these two dictionaries confirmed our original hypothesis that only few neoclassical formatives are described in separate entries. The two dictionaries often differ in the treatment of the formatives in the sample. For instance, the dictionary by Ušakov (1946-47) gives этнолог ([etnolog] 'ethnologist'), этнологический ([etnologičeskij] 'ethnological'), этнология ([etnologija] 'ethnology') but does not have a separate entry for этно- ([etno-] 'ethno-'). A large part of the formatives in the first sample follow a similar pattern. For some neoclassical formations, Ušakov (1946-47) gives etymologies explaining the components of words as based on Ancient Greek.

ЭТНОГРАФИЯ, этнографии, мн. нет, ж. (от греч. ethnos - народ и grapho - описываю)

[etnografija, etnografii, mn. net, ž. (ot greč. ethnos - narod i grafo - opisivaju)]

'ethnography, ethnography<sub>GEN</sub>, no plur., fem. (from Greek ethnos - nation and grapho - writing)'

In principle, such etymological information would enable the user to add a system of neoclassical word formation to their mental lexicon, but there is no indication that many speakers of Russian do this. On the other hand, the dictionary by Efremova (2000) classifies the same formative as an initial part of complex words:

этно-

Начальная часть сложных слов, вносящая значение сл.: народ (этногенез, этнолингвистика, этнопсихология и т.п.).

[etno-]

[načalnaja čast' složnyx slov, vnosjaščaja značenie sl.: narod (etnogenenez, etnolingvistika, etnopsichologija i t.d.)]

'ethno-'

'initial part of complex words having the meaning nation (ethnogenesis, ethno-linguistics, ethnopsychology, etc.)'



Efremova (2000) does not give information about the etymology of neoclassical formatives, but the introductory material mentions that entries are included for around 900 combining forms. Of course not all combining forms are neoclassical formatives.

Checking our second sample of formatives from Bruguera i Talleda (2006) in these two dictionaries, we found that our initial expectations were largely confirmed again. Only for just over half of the formatives do the Russian dictionaries give any examples of formations, the number of formations per formative is lower than in English dictionaries, sometimes just one example, and only very rarely is the neoclassical formative described in a separate entry. Despite the generous coverage of combining forms announced in Efremova's (2000) introductory material, we only found very few separate entries for formatives that are part of our second sample. For *phago-*, Efremova (2000) gives the following entry:

фаго

Начальная часть сложных слов, вносящая значения: 1) поедание, пожирание чего-л. (фагоциты); 2) связанный с бактериофагом (фагодиагностика, фагопрофилактика).

[fago]

[Načal'naja časť složnych slov, vnosjaščaja značenija: 1) pojedanije, požiranije čego-l- (fagocity); 2) svjazannyj s bakteriofagom (fagodiagnostika, fagoprofilaktika).]

'phago'

'initial part of complex words having the meanings: 1) eating of something (phagocytes); 2) related to bacteriophage (phagodiagnosis, phagoprophylaxy)'

The second sense given in this entry is similar in nature to the second sense of *diplo* in section 4. It results from the shortening of a full neoclassical formation. Interestingly, Efremova (2000) lists separately also the formative *-phag* occurring in final position. The entry provides the following information:

фаг

Конечная часть сложных существительных, вносящая значение: поедающий, поглощающий то, что указывается в первой части слова (ихтиофаг, фитофаг и т.п.).

[fag]

[Konečnaja časť složnych suščestvitel'nych, vnosjaščaja značenie: pojedajuščij, pogloščajuščij to, čto ukazyvajetsja v pervoj časti slova (ichtiofag, fitofag i t.p.)]

'phag'

'final part of complex nouns having the meaning: eating, eating what is denoted by the initial part of the word (ichtiophag, phytophag, etc.)'

Ušakov (1946-47) does not give a separate entry for the initial neoclassical formative *phago*, but gives фагоцит ([fagocit] 'phagocyte'), and фагоцитоз ([fagocitoz] 'phagocytosis'). Similarly, the final neoclassical formative *-phag* is not given as an independent entry, but can only be traced, at least in principle, in entries such as фитофаг ([fitofag] 'phytophag').

An interesting case is represented by the formative *diplo*. The meaning corresponding to the examples from English in section 4 is found in the formations диплоккок ([diplokok] ‘diplococcus’), but the formative itself does not appear in a separate entry. In Efremova (2000), an initial part of complex words дип ([dip] ‘dip’), is treated as an independent listed item. The meaning is, however, related to дипломатический ([diplomatičeskij] ‘diplomatic’). It is a clipping found in formations such as дипкупе ([dipkupe] ‘diplomatic compartment’), дипкорпус ([dipkorpus] ‘diplomatic corpus’) or диппочта ([dippočta] ‘diplomatic mail’). This pattern of forming words is very productive in Russian and it is referred to as *stump compounds* or *abbreviated compounds* (cf. Benigni and Masini, 2009, Comrie and Stone, 1978, Molinsky, 1973).

On the basis of our sampling, we may conclude that in Russian monolingual general dictionaries the coverage of neoclassical word formation is much less extensive than in English. First of all there are far fewer entries for neoclassical formations. Moreover the way they are treated reflects to a much smaller extent a system of neoclassical word formation. Ušakov (1946-47) only gives etymologies for neoclassical formations, which is hardly sufficient to retrieve the use of neoclassical formatives. Efremova (2000) gives many entries for combining forms, but from our samples only few neoclassical formatives are actually covered in them.

## 6 Conclusion

In this paper we investigated the treatment of neoclassical formatives in English and Russian dictionaries. The basis of our research was a sample of initial combining forms. In order to exclude a direct bias towards English or Russian in our sampling, we used a word formation dictionary for a third language, Catalan, as the basis for our samples. The choice of initial combining forms was based on the consideration that where neoclassical formatives appear as the base of a derivation, they also occur as initial combining forms in a neoclassical compound. Almost all final combining forms also appear as initial combining forms. Initial combining forms are a much larger set and they can be easily retrieved also in an alphabetically ordered paper dictionary.

As our dictionaries, we selected CED (2000) and COED (2011) for English and Ušakov (1946-47) and Efremova (2000) for Russian. These dictionaries represent more traditional and more modern trends in British and Russian lexicography in such a way that we can consider our results typical of British and Russian dictionaries in general.

We found that English dictionaries give more neoclassical formations than their Russian counterparts, which suggests that they are more numerous in English. In addition, individual formatives are in many cases described in a separate entry in English dictionaries, but this is very rare in Russian dictionaries. These findings are in line with the hypothesis that for a significant proportion of speakers of English there is a system of neoclassical word formation, whereas this is not the case for speakers of Russian.

To the extent that this hypothesis is correct, the lexicographic policy on the inclusion of entries for neoclassical formatives adopted in English and Russian dictionaries can be justified by properties of the languages they cover. In English, having the information as to what a particular formative means will enable the user to decode new formations that are not in the dictionary, help the user build up a system of neoclassical word formation as part of their mental lexicon, and thus support the acquisition and retention of new neoclassical formations.

In Russian, the situation is different. New neoclassical formations will be borrowings, not the result of applying a neoclassical formation rule. If such a borrowing combines two neoclassical components, an entry for one of these components will often be of little help. In many cases, the other component will not exist yet, so that there is no obvious background structure into which a new formation could be incorporated.

In electronic dictionaries, representing neoclassical formatives is to be recommended for English. In an optimal representation, an entry for a formative will give access to the class of all formations it is part of. This includes both the use as an initial and as a final combining form, because a formative that occurs in both roles (e.g. *paedo* in *orthopaedic* and in *paedagogical*) is basically the same formative. For Russian, there is no similar level of support for setting up such a system.

## 7 References

### 7.1 Dictionaries

Bruguera i Talleda, Jordi (2006), *Diccionari de la formació de mots*, Barcelona: Enciclopèdia Catalana.

CED (2000), *Collins Dictionary of the English Language*, 5th edition, Glasgow: Collins.

COED (2011), *Concise Oxford English Dictionary*, 12th edition, Angus Stevenson & Maurice Waite (eds.), Oxford: Oxford University Press.

Efremova, Tatjana F. (2000), *Новый словарь русского языка* [New dictionary of the Russian language], Moscow: Russkij Jezik, <http://www.efremova.info>.

OED (2014), *Oxford English Dictionary*, Third edition, edited by John Simpson, [www.oed.com](http://www.oed.com).

Ušakov, Dmitrij N. (1946-47), *Толковый словарь русского языка* [Explanatory dictionary of the Russian language], online edition <http://www.dict.t-mm.ru/ushakov>.

### 7.2 Other works

Atkins, B.T. Sue & Rundell, Michael (2008), *The Oxford Guide to Practical Lexicography*, Oxford: Oxford University Press.

Bauer, Laurie (2001), *Morphological Productivity*, Cambridge: Cambridge University Press.

Béjoint, Henri (2000), *Modern Lexicography: An Introduction*, Oxford: Oxford University Press.

Benigni, Valentina and Masini, Francesca, (2009), 'Compounds in Russian', in *Lingue e linguaggio* 2/2009, pp. 171-194.

Bloomfield, Leonard (1933), *Language*, London: Allen & Unwin.

- Booij, Geert (2003), 'The codification of phonological, morphological, and syntactic information', in van Sterkenburg, Piet (ed.), *A Practical Guide to Lexicography*, Amsterdam: Benjamins, pp. 251-259.
- Comrie, Bernard and Stone, Gerald (1978), *The Russian language since the revolution*, Oxford: Clarendon Press.
- Corbin, Danielle (1987), *Morphologie dérivationnelle et structuration du lexique*, Tübingen: Niemeyer (2 vol.).
- Germanovič, Ivan Klimovič (1979), 'Лексикография' [Lexicography], in Šuba, Pavel Pavlovič, Современный русский язык I [The contemporary Russian language I], Minsk: BGU, pp. 256-312.
- ten Hacken, Pius (1998), 'Word Formation in Electronic Dictionaries', *Dictionaries* 19:158-187.
- ten Hacken, Pius (2009), 'What is a Dictionary? A View from Chomskyan Linguistics', *International Journal of Lexicography* 22:399-421.
- ten Hacken (2012), 'Neoclassical word formation in English and the organization of the lexicon', in Gavrilidou, Zoe; Efthymiou, Angeliki; Thomadaki, Evangelia & Kambakis-Vougiouklis, Penelope (eds.), *Selected papers of the 10th International Conference of Greek Linguistics*, Komotini: Democritus University of Thrace, pp. 78-88.
- Molinsky, Steven J. (1973). Patterns of ellipsis in Russian compound noun formations. The Hague/Paris: Mouton.
- Ožegov, Sergej Ivanovič (1974), Лексикология, лексикография, культура речи [Lexicology, lexicography, speech culture], Moskva: Vyššaja škola.
- van Sterkenburg, Piet (2003), 'The' dictionary: Definition and history', in van Sterkenburg (ed.), *A Practical Guide to Lexicography*, Amsterdam: Benjamins, pp. 3-17.
- Svensén, Bo (2009), *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*, Cambridge: Cambridge University Press.
- Šanskij, Nikolaj Maksimovič (1972), Лексикология современного русского языка [Lexicology of the contemporary Russian language], Moskva: Prosveščeniye.
- Ulsamer, Sabina (2013), 'Wortbildung in Wörterbüchern – Zwischen Anspruch und Wirklichkeit', in Klosa, Annette (ed.), *Wortbildung im elektronischen Wörterbuch*, Tübingen: Narr, pp. 13-59.

# **Reports on Lexicographical and Lexicological Projects**



# Revision and Digitisation of the Early Volumes of *Norsk Ordbok*: Lexicographical Challenges

Sturla Berg-Olsen, Åse Wetås  
Norsk Ordbok 2014, University of Oslo  
sturla.berg-olsen@iln.uio.no, ase.wetas@iln.uio.no

## Abstract

2014 will see the work on the 12<sup>th</sup> and final volume of the academic dictionary *Norsk Ordbok* (NO) finished. Still, the dictionary will remain heterogeneous due to variation in editorial practice throughout the project and incomplete in the sense that its early volumes are not digitally available. The online version of NO currently only covers the alphabet from the letter *i*. This paper describes the present state of the different parts of NO and argues that the early volumes of the dictionary must be revised and digitised to bring them up to the standards of the rest of the work. The revision and digitisation will not only give the dictionary a unitary profile but also make it possible to use it for a number of other purposes and facilitate the continuous process of keeping the dictionary up to date. The paper discusses some of the lexicographical challenges involved in the planned revision project and displays examples of the changes that must be made to the structure of the early material. It also touches upon questions concerning project organisation and funding.

**Keywords:** Academic dictionaries; Digitisation; Project planning

## 1 The history and present status of *Norsk Ordbok*

*Norsk Ordbok* (NO) is an academic dictionary covering Norwegian Nynorsk and all Norwegian dialects. The dictionary will provide a scholarly and exhaustive account of spoken Norwegian and of texts written in Nynorsk from 1860 up till today, and is to be completed during 2014, the year of the bicentenary of the Norwegian constitution. From 2002 the dictionary work has been organised in the time-limited project organisation Norsk Ordbok 2014 (NO 2014). The project owner is the Department of Linguistics and Scandinavian Studies at the University of Oslo. In 2014, the finished work will include more than 300,000 entries, published in 12 volumes.

When NO was conceived in the late 1920s, Nynorsk was still a written language in the making, and the standard was continuously fed by Norwegian dialect words. The proponents of Nynorsk wanted to make a comprehensive scholarly dictionary building on the works of the famous Norwegian 19<sup>th</sup> century linguist Ivar Aasen. The immediate goal behind the dictionary was to develop Nynorsk further, and to raise the prestige of the new written standard. The combination of dialects and written standard in one dictionary – somewhat unusual in a wider European context – was considered a natural

choice, given the crucial role dialect data had played in the codification of Nynorsk from the outset. Even today the editors of NO regularly write entries entirely based on dialect material. This process often includes codifying the spelling and inflection of these words according to the Nynorsk standard. The collection of data for a new and comprehensive dictionary of Nynorsk started in 1930. A dictionary board of trained lexicographers instructed and supervised more than 550 volunteers, who during these early years collected dialect data from all over the country and made it possible for the dictionary board to build up a huge slip archive. The learned dictionary board also supervised the extraction of literary excerpts from Nynorsk literature, both fiction and non-fiction. In addition, they compiled a draft manuscript combining Ivar Aasen's dictionaries (Aasen 1850 and 1873) with a range of other canonical dictionaries dating from 1870 to 1910, also adding data from glossaries and local dictionaries dating from 1600 to 1850 (Skard 1932). This draft manuscript for the new, academic dictionary was finished by 1940.

The editing of the dictionary started in 1946, and the first volume of NO was published 20 years later, covering the alphabet from the letter *a* to the adjective *doktrinær*. The original plan was to make a 2-3 volume dictionary, but in 1966 the chief editor estimated that 8-9 volumes would be needed to cover the whole alphabet (Hellevik 1966). During the first 50 years the editing of the dictionary progressed slowly. At the same time the source material grew, and so did the dictionary entries in volumes 2, 3 and 4. All the work was done manually, the slips sorted on the lexicographer's desk and the manuscripts prepared in handwriting.

In 2002 the work was reorganised and moved to a digital platform, making the editing process a lot more efficient. Increased funding allowed the project to employ more editors, and the work gained speed. During the period 2005-2013 7 volumes were published, with the last volume to be finished in 2014. However, the volumes produced before 2002 (volumes 1-4 and roughly half of volume 5) remain only partly digitised and show a number of discrepancies compared to the latter volumes. This has to do with changes in editorial practices that were implemented along the way. The digitisation of the volumes produced before 2002 and the revision of the contents of these volumes to bring them up to date are essential tasks that must be undertaken after the completion of the last volume. This will ensure that NO is a homogeneous dictionary that meets the scholarly standards of the age of electronic corpora and can be updated continuously in the future. Only when the entry database covers the whole alphabet can it be used for other purposes (e.g. the extraction of semantic structures to form the basis of a Nynorsk word net, the extraction of subsets of entries for new, thematic dictionaries etc.). In addition, revisions of the entry database itself can then be organised thematically, and not necessarily alphabetically.

Since 2012 an online version of NO has been available, but this version only contains the material from the letter *i* onwards. A complete online version is dependent on the complete digitisation of the early material and adaptation of this material to the database system used.

The reorganisation into the time-limited project NO 2014 also led to a change in profile for the dictionary. During the whole history of the dictionary, there has been a strict emphasis on constructing a



scholarly work that meets scientific demands. However, the editorial profile of the earliest volumes was that of a scientific paradigm still concerned with nation-building. The dictionary was part of the work to document and elaborate on Nynorsk as a cultural object and to further standardise this language, which was still in its formative stage. In the modern project organisation, the emphasis is on editorial practice as descriptive research work. This results in the inclusion of entries that were earlier not considered part of Nynorsk proper, but which have entered Nynorsk during the last 50 years. To take one example, during the work on the entries starting with the Norwegian privative prefix *u-* a number of instances were discovered where the ‘positive’ counterparts of these words from the earliest parts of the alphabet were not covered (words like *bekvem* ‘comfortable’, *bekymra* ‘worried’, *bemann* ‘manned’ etc.). These are loan words from Danish and German that were earlier only used in Bokmål, but they are now also part of modern Nynorsk and should thus be included.

## 2 The microstructure of *Norsk Ordbok*

The microstructure of NO entries is fairly similar to that found in other comprehensive scholarly monolingual dictionaries, such as the OED, the *Dictionary of the Danish Language* (ODS) and the *Swedish Academy Dictionary* (SAOB). Each headword is followed by a section containing information on early lexicographical sources listing the word and etymology, as well as pronunciation (mainly for borrowed words) and alternative written forms of the word. This section also provides attested dialect forms of the word with geographical indications. Only dialect forms that do not follow automatically from general and well-known rules of sound correspondences in Norwegian dialects are included. The introductory section is the part of the NO entries which has seen the most variation and change during the project period. In the early volumes there was a certain degree of experimenting both with the order of the information given here and the structuring of this information. The digital platform used from 2002 onwards ensures stringency, but the variation found in the introductory section in the early volumes presents big challenges when it comes to digitisation.

The part of the entry following the introductory section is fairly straightforward, with potentially three explicit levels of senses, each sense customarily followed by literary sources and/or geographical indications, as well as examples of usage. In the early volumes, multi-word expressions are treated largely on a par with ordinary examples. Starting from the letter *i*, such expressions have been edited as sublemmas, appearing in boldface.

### 3 Challenges involved in the digitisation and revision of the early volumes

The goal of the revision project is to bring volumes 1–5 up to the same standards and give the entries in these volumes the same structure as that found in volumes 6–12. The contents of volumes 1–5 must be evaluated in view of the present editorial policies and revised on all levels where necessary in order to reflect these policies. This involves restructuring, adding information and also (particularly in volumes 3–4) removing some information. The result will be a homogeneous product reflecting the Nynorsk of the 21<sup>st</sup> century as well as the history of this written standard and the diversity of the Norwegian dialects.

There are several possible ways of digitising the oldest volumes of the dictionary. One solution could be OCR-scanning. This process was chosen for the first online version of the *Swedish Academy Dictionary* (SAOB) in 1997, but the result was considered unsatisfying and also turned out to be very expensive (Mattisson 2012). SAOB is currently going through a second re-digitisation process. This time the printed text is punched and stored in digital files in China. When this part of the process is finished, the SAOB editorial staff themselves will process the files by hand into valid XML. A similar process was chosen by the Society for Danish Language and Literature when they digitised their 28-volume *Dictionary of the Danish Language* (ODS) in 2005 (cf. ODS FBTS). The solution chosen for the ODS and for the second digitisation of the SAOB seems to be a good choice for older dictionaries where all the text is produced as typed manuscripts to feed a print version. The situation for NO is not quite similar to these works. Firstly, the dictionary has been produced on a digital platform from the letter *i* onwards. When the work on the 1<sup>st</sup> edition finishes in 2014, approximately 2/3 of the dictionary entries will be digital entries feeding both the online dictionary and the printed version. Secondly, the punching part of the digitisation process is already done for the oldest volumes of NO. In order to make an online version which covers the whole dictionary, and in order to complete the dictionary database, the only fully satisfactory solution for our dictionary will therefore be to integrate the digitised text from the oldest volumes into the already existing entry structure of the digital dictionary.

The current state for volumes 1–5 of Norsk Ordbok is that the two first volumes were punched and proofread in 2001–02. The manuscripts for volumes 3–4 and the part of volume 5 that covers the letter *h* were produced in simple word processing programmes, and supplied with tags either during the editing process or afterwards. The original text for the oldest volumes of NO thus existed as digital manuscripts as early as in 2002. In 2005, the Norsk Ordbok 2014 project organisation made a pilot study on the integration of this digital text into the modern database system. The adaptation of the texts into the new and stringent database format proved too difficult and too time-consuming for the time-limited project organisation, and was therefore put on hold.

The entries from volumes 1 and 2 are integrated in the database system of NO 2014, but only in an incomplete version. The text is not in line with the current quality when it comes to consistency, and it does not give a complete coverage of older source material. Volumes 3 and 4 are partly integrated in

the database, but a lot of the text is not fitted into the correct fields, and the huge amount of dialect data and information on etymology is lacking altogether. The part of volume 5 covering entries starting with the letter *h* is not integrated in the database at all.

### 3.1 Why digitisation *and* revision?

Why is it so important to do the digitisation and the revision in one integrated operation? As mentioned above, the project organisation made a pilot study in 2005 to see if it would be possible to load the text of the oldest volumes into the modern editorial database. The pilot revealed that a lot of work has to be done to make the old text fit into the strict categories of the new editorial system, and that work inevitably also involves revision. One way of presenting the whole dictionary digitally without performing this integrated process of digitisation *and* revision would be to publish the oldest volumes as searchable PDFs on the Internet. This would be very unsatisfactory for several reasons: low user-friendliness, no possibility to perform searches across the base, lack of access to multi-word expressions in the earliest volumes, lack of possibility to do thematically based revisions and use the dictionary contents for other purposes etc.

Producing a digital dictionary which is identical to the printed version of NO is not the best solution in the view of the project organisation. Instead, we want to fit the entries from volumes 1–5 into the modern editorial database format. Preserving the contents of the oldest volumes in detail would force us to extend the existing database structures in order to adapt it to the structure and the idiosyncrasies of the old entries. Our goal is instead to modernise and standardise these entries and adapt them structurally to the modern online dictionary format.

### 3.2 Structural changes related to the digitisation

The first four and a half volumes of the dictionary were produced manually. The entries of these volumes are of a high quality for their time, but they often have a very tiered structure (Atkins & Rundell 2008:249) and from time to time include entry-specific structuring of data. This practice is possible and probably inevitable when the manuscripts are produced by hand, but it meets problems with the introduction of a digital production platform.

In 2002, the senior editing staff of the dictionary did a huge job extracting an ideal entry structure from the early volumes. This was used for setting up the electronic editing schema of the modern, digitised dictionary. The entry structure at the macro level (entry status, flat vs tiered structure, content selection etc., cf. Atkins (2008: 36ff)) was created on the basis of what was conceived to be the best practice of the old volumes, but this still leaves a lot of information that will not fit into the categories of the schema, and that will need to be given elsewhere in the entries or, if deemed superfluous, deleted. The planning of the entry structure at the macro level is much in line with the process of dictionary planning described by Atkins (2008), but for a dictionary project that has already published five

volumes the options when setting up the macro structure are not open in the way they are when planning new dictionary projects (see also Cantell & Sandström 2012: 166f).

Another task associated with the digitisation of the material is the electronic linking of words in definitions, etymologies and elsewhere, as well as the linking of the first part of compounds to the correct basic word. Such links are an integral part of the structure in the latter volumes, and must be added also in the early material. This linking also requires that the structure of the older volumes is possible to adapt into the new data base system.

### 3.3 Structural changes related to the revision

Several structural changes must be performed in the old material in order for it to meet the requirements of NO's present editorial practices; a few examples will be mentioned here. As stated above, multi-word expressions were in the early volumes treated more or less on a par with ordinary examples, while starting from the letter *i* they have been edited as sublemmas in boldface. In order to attain a unitary structure throughout the dictionary, multi-word expressions in the early volumes must be identified and changed into sublemmas. A case in point is the phrase *bita i graset* ≈ 'bite the dust' (literally 'bite in the grass'), seen in figure 1. The phrase appears as an example under sense 1a in the entry **bita**, but clearly deserves the status of sublemma in a revised version the entry.

I **bita** (*i*) v **bit**, **beit**, **biti** [VAgd16-,L,L 280,A, R3; målf òg ft pl *bito* (Hall), *betò* (Va); fp òg *bite*, *bele*; gno *bita* (*bit*, *beit*, *bitum*, *bitit*); grunntyd 'kløyva'; tyd 6 vel av eng. *beat* 'slå']. **1) a)** [...]

// *bita i graset*.

1) (eigl om hest som stoggar og tek seg ei grastugge når han møter folk; overf om folk) stogga og tala til ein som ein råkar (Rog): *han Ola for forbi meg utan å bita i graset*. 2) *bita i bakken*; *tapa: dei lyt bita i graset som hev minste makti* (RomsOrdt)

Figure 1: Part of the entry **bita** with the multi-word expression *bita i graset*.

The structure of senses will – particularly in longer entries – need to be made flatter, more transparent and thus easier to navigate. The fact that the editors have access to a much larger body of linguistic data today (including a ~100 mill. word corpus) than when the early volumes were produced has contributed to less tiered sense structures in the latter volumes, and this will necessarily also be the case for the early material after revision.

There are a lot of structural features where the early volumes differ from today's editorial practice, and where structural revisions along the lines of the present editorial guidelines are required. One example concerns the use of usage labels; certain labels are no longer in use, such as *lbr* (*lite brukande* ≈

‘should be used with caution’), which is connected with a certain puristic inclination in the early years of the dictionary. Figure 2 shows two entries with this label from volume 1:

**behandla** v -a [ty] lbr. **1**) stella med etter ein viss metodisk framgangsmåte for å nå eit visst resultat; handsama; ha føre: *behandla ein for feber* (Ra.F) / *behandla ei sak*. **2**) fara fram (slik el slik) mot, fara åt (slik el slik) med: *eg hugsa og korleis han behandla Husmanns-Knut* (HolmS 57). **behandling** f lbr, det å behandla; stell, handsaming; førehaving; medferd.

**Figure 2: The entries *behandla* ‘treat’ and *behandling* ‘treatment’ are equipped with the label *lbr*, although they are widespread in modern Nynorsk.**

Another example concerns the labels *zool* (zoology) and *bot.* (botany), which were earlier used for all definitions covering names of animals and plants respectively, but are today restricted to official terms, while e.g. local names for plants lack the label *bot.*, but are electronically linked to the official term.

### 3.4 Revision of the lemma list

Faced with the task of producing a definite number of volumes on the basis of a certain amount of data, the project NO 2014 has developed effective methods for determining which lemmas should be included and how much space each entry should occupy (Grønvik 2006). The existing lemma list in volumes 1-5 must be revised using the present criteria for inclusion in NO and taking into account the material we have at our disposal today, which is a lot larger than when the first volumes were edited and includes a corpus dominated by 21<sup>st</sup> century newspaper texts. Neologisms and words that were previously not represented or poorly represented in the material must be included, together with lemmas of German or Danish provenance that were left out for puristic reasons but are used in modern Nynorsk (cf. section 1). In other cases lemmas that were originally included must be excluded – especially in volumes 3-4, where the inclusion criteria were clearly more liberal than today. Thus one can fairly frequently find entries that are based on hapax legomena (figure 3) or exclusively on occurrences in bilingual dictionaries (figure 4). These entries do not qualify for inclusion in the dictionary according to the present editorial guidelines.

**franske-flokk** m flokk av franskmenn (Maul.I I,64).  
**fred-stor** adj poet., sj, fredfull: *\*i fredstore ævelengdi* (Hovd.SS 109).

**Figure 3: Entries in volume 3 based on hapax legomena.**

**fransk-etar** m [ett ty *franzosenfresser*] person som ottast el hatar franskmennene (Ra.E u *gallophobe*; Ra.F u *gallophobe*; Vo.Ty u *franzosenfresser*).

**fransking** f gallsisme, franskbragd (S.D u *gallicisme*; Vo.Ty u *gallizismus*).

Figure 4: Entries in volume 3 based exclusively on occurrences in bilingual dictionaries.

The oldest entries of the dictionary are not more than some 70 years old. This means that the diachronic dimension of the work itself is less challenging than for dictionaries with a production period that stretches over more than one century. Still, the oldest parts of NO show that some entry revision is needed. New entries have to be added, some old entries should be removed altogether and a lot of existing entries need revision due to broader and sounder empirical evidence, language change or both.

### 3.5 Revision of the dialect data given in the entries

The dialect material at the editors' disposal is substantially larger today than 80 or even 30 years ago. The geographical indications regarding special dialect forms and dialectal uses of words and word senses can thus be supplemented, in many cases possibly justifying the use of larger areas instead of single counties (the county is the smallest unit used for geographical references in NO). At the same time, the geographical indications in some of the early volumes reflect a more liberal practice than the one followed today, and they must be checked to make sure that the dictionary reflects the actual dialect material at our disposal.

The method of presenting dialect forms has changed somewhat during the history of NO; in particular, volumes 3–4 present such forms in greater phonetic detail and with more parallel forms than both the earlier and the latter volumes (cf. figure 5). Here the revision must imply a certain degree of simplification, following methods established in 2002 and later.

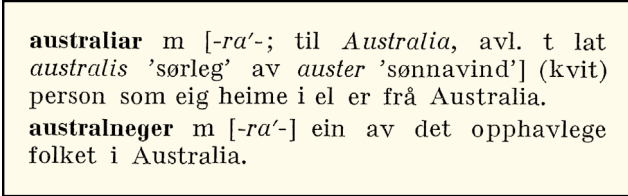
**fredag** m [Fyresdal1698 *frædajen*, VTel1821 118 *frædajæn*, A, R; målf *fre'dag* (Vanl, sjeldnare *fre'da*), *fræ'dag* (mest sjells, sjeldnare *fræ'da* Furnes), *fræ'dæ* (Vestf), *fræ'dæu* (ØvRendal), *fred'da(g)* el *fræd'da(g)* (Agd sumst, Sokndal, Stav), *frei(d)dag* (JrR), *frad'da* o l (Trysil, Tynset), *fre'dag* (Elsfjord, Gimsøy, Borge i Lof, Bjarkøy); gno *frjådagr*, mno *fræigjadagr* svar. t gno *\*friggjardagr* 'Friggs dag' etter lat *dies Veneris* 'Venus' dag', T]

**mån-dag** m [Fyresdal1698 *måndajen*, Stav1698, VTel-1821 118 og 121 *mådajen*, C, A, R, R2; skr òg *mandag* (ØklandGA362, Furs.ENE44, Hovl.SV63, I.LindSD 131); målf òg *man-* (vanl), *mån-* (heller vanl), *mon(n)-* (Torsnes, Vågå R, Uvdal, ATel R2, SAukra, Rindal), *må-* (Vinje i Tel R, Mo i Tel R), *mårn-* (Alvdal, Grong); norr. *månadagr* 'månedag', ett lat *dies lunae*, frå gr]

Figure 5: The introductory section of the entry *fredag* 'Friday' from volume 3 and that of *måndag* 'Monday' from volume 8. Note the differences in the notation of dialect forms (introduced by 'målf' and 'målf òg' respectively).

### 3.6 Revision of definitions

In the majority of the entries the actual wording of the definitions will hardly require a lot of revision. Still, since the publication of volume 1 in 1966 the language has certainly undergone quite a few changes on the level of semantics and pragmatics – changes that must necessarily lead to adjustments in a number of definitions. Obvious examples are words that were earlier used neutrally, but later developed derogatory connotations and often have become obsolete altogether. In figure 6, the definition of *australiar* ‘Australian’ is ‘(white) person belonging in or coming from Australia’; the first word in the definition should definitely be deleted. The second headword, *australneger*, is no longer in use due to its derogatory character. As NO is a descriptive dictionary which documents actual (historical) usage, the entry should be preserved, but the definition must be updated to reflect the stylistic properties of the words and the fact that it is obsolete. The modern neutral term *aboriginar* ‘Aborigine, Native Australian’, which is not found in volume 1, must also of course be added.



**australiar** m [-ra'-; til *Australia*, avl. t lat *australis* 'sørleg' av *auster* 'sønnavind'] (kvit) person som eig heime i el er frå Australia.  
**australneger** m [-ra'-] ein av det opphavlege folket i Australia.

**Figure 6: The entries *australiar* ‘Australian’ and *australneger* ‘Australian negro’ from volume 1.**

On a more general level, there is a tendency in some of the early material towards focusing on the particular rather than the general and to posit separate word senses where the present practice would prefer lumping rather than splitting. Thus especially in longer entries there will be a need for revising or rewriting definitions. Entries that lack a definition altogether but meet the criteria for inclusion in the dictionary must of course be provided with a definition.

Integrating new source material and meeting the requirements of a modern scholarly dictionary

The new source material – including corpus data – must be integrated at all levels of the early volumes of the dictionary. This will be reflected in the addition of new entries (cf. 3.4), the creation of new senses in existing entries, the introduction of new examples, especially the addition of more recent examples (sometimes due to reasons of space and clarity replacing some of the existing examples) as well as in new, updated geographical indications (cf. 3.5).

It is essential to ensure that the early volumes meet the requirements of a modern scholarly dictionary. This implies, firstly, that every entry must be linked to its source material and, secondly, that all entries and all word senses must have a documented source material behind them and contain at least one source reference at the level of definition and/or example. In the electronic version of the dictionary the links between entries and source material will be made explicit, enabling users to verify the information given and potentially falsify it.



**dyvels-** el **divels-klo** f [målf *divels-* (Masfjorden); ett ty *teufelsklaue*, ndl *duivelsklauw*, eng. *devils claw* eigl 'djevetsklo'] sjøm., fag., (mest i pl) eit par jarnkrokar i sams stropp til å huka i bjelkar o l som skal heisast; slike krokar i enden av eit seglskaut til å festa i storseglet på ein større båt med sprisegl.

**Figure 7: The definition in the entry *dyvelsklo* 'devil's claw (a kind of split hook)' from volume 2 lacks source references. One or more references must be added, or the entry must be excluded from the dictionary.**

## 4 Planning and implementing the revision project

As part of the planning it must be decided to what degree the dictionary entries should be rewritten. A plausible strategy is to assume that smaller entries – which constitute the majority – require only revision, while at least a part of the larger entries (especially large verbs and function words) will benefit from being re-edited. This re-editing must be performed with the editor at all times keeping a keen eye on the existing entry and making sure that all essential information that is given there and can be verified is transferred to the new version.

In the modern NO 2014 organisation, all the relevant source material is digitised and stored in a structured relational database system. This makes it possible to quantify relative space for each entry and to estimate the work load for the staff as a group and for each single editor. The experience from the last 12 years of project work shows that this way of working gives a high degree of prediction when it comes to how much time and money are needed to perform the whole operation of revising and digitising the oldest parts of the dictionary.

The whole of the source material behind the earliest volumes is included in the dictionary database system, and this provides a very sound way of estimating the work load for doing the integrated digitisation and revision work. For the whole bulk of 112,500 lemmas it is possible to make fairly accurate estimates that also take into account that some entries will be revised, while others will gain from a full rewriting. Based on experience with producing the last seven volumes of the dictionary over a period of 12 years, feeding both a printed publication (each volume includes 800 pages of entries) and an Internet version, the NO 2014 organisation estimates that the digitisation and revision of the first volumes will be possible with a staff of 16 editors working full time over a period of five years. This is approximately 45 % of the amount of work that was put into volumes 6–12.



## 5 Funding

The revision project will have a total cost of some 70 million NOK (approx. 8.5 million EUR). The production of NO has so far been funded by the University of Oslo and the Norwegian Ministry of Culture in a joint agreement, but this funding ends in 2014. Language infrastructure, including dictionaries, is cost-intensive and involves huge amounts of manual work. Norway is a relatively small language community, and the commercial potential of the basic language infrastructure resources for Norwegian is quite low. This means that in order to reach the central goals on the field of Norwegian language policy, the building up of basic language resources needs public funding.

A NO dictionary database covering the whole alphabet span will not only offer the public a comprehensive description of spoken Norwegian and written Nynorsk. The full dictionary database will also be an important component in future Norwegian language infrastructure and language technology. In this perspective, public funding of the digital integration of volumes 1–5 of Norsk Ordbok in the dictionary database would hopefully be within reach. In 2013 the Language Council of Norway set up a policy document for dictionaries and other basic lexical resources for the Norwegian languages, including Sami and the official minority languages of Norway. This policy document states the importance of a complete and updated online version of NO, and also states that this needs public funding (LCN 2013-08 and LCN 2014-03).

## 6 Conclusions

A lot of work is needed to bring the oldest volumes of NO up to the same digital standard as the rest of the dictionary. During the 80 years that have passed since the dictionary work started, the language itself, linguistic theory and preferred publishing platform have all changed. These changes have in turn led to changes in lexicographical practice. For a scholarly dictionary to be scientifically sound and relevant to the dictionary users, it is necessary to revise and upgrade its contents. For the dictionary database to become complete, it is not an option to choose only digitisation, or doing the process in two separate steps.

## 7 References

- Aasen, I. (1850). *Ordbog over det norske Folkesprog*. Christiania: Carl C. Werner & Comp.
- Aasen, I. (1873). *Norsk Ordbog med dansk Forklaring*. Christiania: P.T. Mallings Boghandel.
- Atkins, B.T.S. (2008). Theoretical Lexicography and its Relation to Dictionary-Making. In: T. Fontenelle (ed.) *Practical Lexicography. A Reader*. Oxford: Oxford University Press, pp. 31-50.
- Atkins, B.T.S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

- Cantell, I. & Sandström, C. (2012), Hur blir en traditionell, tryckt ordbok en webbordbok? In: B. Eaker, L. Larsson, A. Mattisson (eds.) *Nordiska studier i lexikografi 11*. Rapport från Konferensen om lexikografi i Norden Lund 24-27 maj 2011, pp. 157-168.
- Grønvik, O. (2006) Verknader av digitalisering på materialvurdering, redaksjonell metode og opplæring. In: *Nordiske Studier i Leksikografi*, 8, pp. 129-142.
- Hellevik, A. (1966) Til fyrste bandet. In: NO volume 1, pp. XV-XVI.
- LCN 2013-08 = En samlet ordbokpolitikk etter 2014 (letter from the Language Council of Norway to the Ministry of Culture). Accessed at: <http://bit.ly/1i06QPN> [08/04/2014].
- LCN 2014-03 = Norsk ordbokpolitikk (memorandum from the Language Council of Norway to the Ministry of Culture). Accessed at: <http://bit.ly/OBoRYU> [08/04/2014].
- NO = (1966-) *Norsk Ordbok. Ordbok over det norske folkemålet og det nynorske skriftmålet*. Oslo: Det Norske Samlaget. Web edition accessed at: <http://no2014.uio.no> [02/04/2014].
- ODS = (1918-2005) *Ordbog over det danske Sprog*. København: Gyldendal. Web edition accessed at: <http://ordnet.dk/ods> [02/04/2014].
- ODS FBTS = Fra bog til skærm. Accessed at: <http://ordnet.dk/ods/fakta-om-ods/fra-bog-til-skerm> [08/04/2014].
- OED = *Oxford English Dictionary*. Accessed at: <http://www.oed.com> [02/04/2014].
- SAOB = (1898-) *Ordbok över svenska språket*. Lund: Svenska Akademien.
- Skard, S. (1932) *Norsk Ordbok. Historie - plan - arbeidsskipnad*. Oslo: Det Norske Samlaget.

# A Dictionary Guide for Web Users

Valeria Caruso, Anna De Meo  
University of Naples 'L'Orientale'  
vcaruso@unior.it, ademeo@unior.it

## Abstract

The continuously growing number of specialised lexicographical resources on the Web calls into question the users' ability to solve their information needs autonomously. Neither terminological data-banks, nor dictionary aggregators actually represent valuable solutions to these needs (Lew 2011), and in order to guide users towards useful resources, a database was created, which collects evaluation forms of free internet specialised dictionaries and allows users to carry out customised searches on the basis of their subject field expertise (laymen, semi-experts, experts), the desired language, and the kind of support they need (basically with communicative problems or in acquiring new knowledge).

Using a specific evaluation system, the tool displays the best resources available for the desired parameters, assessing dictionaries on the basis of an evaluation scale and explicit guidelines that prevent contradictory responses, such as dictionaries that are simultaneously suited for laymen and experts.

The paper illustrates the current development of the tool, with special reference to its evaluation system, as well as its possible future improvements.

**Keywords:** Dictionary Guides; Dictionary evaluation; Online Dictionaries; Specialised lexicography

## 1 Information overload on the Web

Though the Web is the most used source of information, too much data are offered to the Internet surfers, causing what has been called "information death" (Tarp 2010: 41). Search engines are too generic to be of any assistance to users with these tasks, and metalexicographical resources have started to appear. The quickest searches are offered by dictionary aggregators (i. e. *OneLook*) and mesh-ups (i. e. *Your Dictionary*) which show definitions taken from different vocabularies on one page, a system that doesn't seem to be completely effective, because terminology archives - and hence the number of definitions provided - are either too small to cope with users' needs, or too big to solve the problem of an effective and efficient access to information (Heid 2011).

From this point of view, the World Wide Web poses stimulating metalexicographical issues, some of which will be outlined here while presenting a new lexicographical tool for guided searches on the Web, namely a rated inventory of free specialised dictionaries, managed through a relational database which allows users to carry out multiparametric searches.

## 1.1 Information accessibility

The issue of knowledge accessibility led to the creation of dictionaries, since:

(t)he truly unique thing about dictionaries is not the various types of data they employ in covering the information needs of users (...). Such data can generally be incorporated into other types of book and text as well. The truly unique thing is the way in which this data is made accessible so users can quickly and easily find the exact data they need. (Tarp 2008: 101)

Nowadays lexicographers focus their work on the customization of dictionaries for their users, and different approaches have been proposed in order to achieve this aim. One in particular seems to be useful not only for writing vocabularies, but also for their critical evaluation, since it offers a synthetic procedure to define the parameters a dictionary must have in order to fulfill its desired functions. Therefore the theory has been named *lexicographical function theory*, and was formulated by Sven Tarp (2008; also Tarp 2009, 2010) as a result of long metalexicographical reflections and debates carried out by the research group of the Aarhus University in Denmark (Nielsen 1994; Geeb 1998; Bergenholz & Tarp 1995). According to this theory, dictionary functions must be identified on the basis of the kind of users, as well as the situations in which the vocabulary is employed, therefore the compilers must think about the specific context in which the need for vocabulary consultation arises (Tarp, 2008: 81). For example, dictionaries may be used in many different situations, such as by students proofreading their homework, or by professional editors working on books to be published, or even in the less common situation of young people reading religious books, in such a case the dictionary “should only explain the meaning of a word or of phrase and noting more” (Bergenholz, 2012: 245). Therefore the more specific the target is, the easier it is to tailor the dictionary to the users’ desired functions. As a consequence, the traditional general language dictionaries (or *polyfunctional dictionaries*), offering aid for different kinds of tasks without a specific tailoring of the information they provide are judged as inefficient, since:

(they) are in many cases so overloaded that this causes information stress and in the worst case may even cause the search to be abandoned if the user cannot find the needle in the haystack (Bergenholz, 2012: 251).

The alternative model proposed is the *monofunctional* electronic vocabulary, extracted from lexical databases using search forms that allow users to tailor the entry to their needs. For example, if the dictionary must supply assistance for text production in an L2, the database will provide a dictionary article displaying grammar information, “synonyms, collocations and examples” (Bergenholz, 2012: 253). Conversely, if the user must understand a text, this information is probably inadequate and certainly not of the outmost important.

Lastly, by fixing explicit parameters that guide good practices of dictionary writing, the theoretical framework of the *lexicographical functions* proves to be suited for the opposite task too, namely dictio-

nary evaluations, which can be undertaken not only in general review terms (see Nielsen, 2009, 2013), but also in a more lexicographical direction, employing the same principles as orienteering parameters among the existing lexicographical resources.

Using these observations as a starting point, a database has been created. The resource, accessible at the *Web Linguistic Resources* (WLR) site, collects free specialized Internet dictionaries which are often more valuable for their unrestricted access than for their overall quality, since the Internet compilers have little or no lexicographical expertise at all. The usability of the majority of these dictionaries is therefore dependent on guides and filters that prevent users from wasting their time and being given inefficient information, in this way they might become quick reference tools for web surfers.

The archived dictionaries were collected during two extensive research sessions in 2010 for the sector of oenology and medicine in different languages: Italian, English and French. A similar intensive exploration was carried out in 2013 for Economics dictionaries of the English language, whilst other sporadic additions gave the database more resources from different specialised sectors on the basis of more occasional findings. A more systematic analysis and upgrade of the inventoried resources is planned to be carried out before the definitive version of the tool is released, since it is currently available only as a pivotal 'beta' version.

## 2 Dictionaries on the Web: the features to be rated

Instead of providing users with multiple definitions on one page, and leaving them with the task of selecting data, the WLR database offers a rated inventory of dictionaries which help users to find the best resources available for free on the Web.

Moreover, the adaptation of the lexicographical function theory parameters to critical principles of analysis in order to rate and filter dictionaries also fulfills the proposal of Nielsen (2009; 2013) to judge dictionaries on lexicographical principles that are generally applicable in order to make dictionary reviews an integral part of the academic field of lexicography.

The rated inventory of the WLR site is based on an evaluation form (fig. 4 below), managed by a relational database that allows multiparametric searches.

The 53 fields in the form (see table 1) correspond to the possible features of a dictionary, and address all the component parts of vocabularies, i. e. the overall organization and the host site, the medio-structure (Wiegand, 1996; Nielsen, 2003), and microstructure (Hausmann & Wiegand, 1989; Hartmann, 2001). The features were partly set in advance, and partly added - or modified - during the data collection, in order to portray adequately the characteristics of these atypical dictionaries - they are listed in table 1 according to the parts of dictionaries they belong to<sup>1</sup>.

---

1 See also Caruso [2011] and Caruso & Pellegrino [2012] for a more detailed description of the features considered.

Dictionary parts	Features and sub-features
General Organization and Host Site	Guide, Kind of Site: Amateur/ Blog/ Commercial/ Collective Resource/ Generalist/ Institutional/ Specialised, Learning Resources, Bibliographic Resources, Hyperlinks, User Feedback, Access: Browse / Search Engine / Advanced Search Engine, Entries: 0-49 / 50-100 / Over 100, General Organisation: Concepts / Words, Kind of Dictionary: Monolingual Dictionary/ Monolingual Word List/ Multilingual Dictionary/ Multilingual Word List/ Plurilingual Dictionary, Bidirectionality, Lemmata: Technical And Non-Technical Terms / Only Technical Terms;
Mediostructure	Cross-references, Related terms, Hypernyms & Hyponyms, Hypertexts;
Microstructure: Linguistic fields	Grammatical Category, Morphological Information, Syntactic Pattern, Phonetic Transcription, Pronunciation Notation, Stress Information, Audio Files, Syllabification, Frequency Of Use, Linguistic Variation, Technical Definitions, Translation Equivalences, Example Sentences, Quotations, Idioms, Collocations, Synonyms, Antonyms, Etymology;
Microstructure: Non Linguistic Fields	Definitions, Examples, Domain Field, Video Files, Pictures, Cultural Notes.

**Table 1: the listing of the dictionary features and sub-features assessed by the evaluation form of the *Web Linguistic Database*.**

The host site may be an important validation criteria of the dictionary quality, since it is to be expected that credited Institutions (universities, ministries, professional associations etc.) publish good lexical resources. In point of fact, ‘institution’ refers here to authoritative organizations within one field, and it has a more restricted use than in Fuertes-Olivera (2009), where the term refers generically to every dictionary not compiled ‘collectively’ by non-professional lexicographers (such as *Wiktionary*).

The overall organization, instead, comprises the dictionary type, whether a simple word list, a multilingual dictionary provided or not with bidirectionality (which is a separate field in the form), or a *plurilingual*, a new dictionary added to the list which is typical of the Internet, namely the dictionary within localized sites (Caruso 2011). These sites in fact are optimized for the market of different countries (Pym 2004), and therefore offer many language versions of their pages that are not interlinked with each other. Since one version is completely independent from the others, the many language dictionaries therein also have no direct connection. Therefore, the user must scan the entire word list and check for correspondences in the definitions in order to find any translation equivalences.

Moreover, Internet dictionaries may also offer special access facilities to users, such as advanced search engines (another field in the form). For example, the dictionary of the *Büro für angewandte Mineralogie* allows searches not only in the whole dictionary contents, but also in its classifying ontology: looking for *Elemente*, the listing provided by the engine will include also *Periodensystem der Elemente*, besides all the chemical elements in the dictionary (from *Antimony* to *Sulfur*), and the entries that contain the required word in their definitions.

During the data collection, special attention has also been paid to the mediostructure, or the cross-linking system, which is obviously a key component of electronic vocabularies. Accordingly, the evaluation form registers both *Cross-references* and *Related terms* (see table 1), only the former having direct hyperlinks to other entries, while *Hypernyms and hyponyms* signal semantic hierarchies that also function as internal references.

As for the microstructure, or the dictionary entry components, the evaluation form takes note of its linguistic and encyclopedic aspects, and accounts for specific fields that reveal the user-friendly character of these dictionaries, which generally offer non-technical definitions, and pronunciation notations rather than phonetic transcriptions.

### 3 The rating system

Since this lexicographical project does not aspire to the detailed dictionary reviews of Nielsen (2009, 2013), but to large scale qualitative estimations that filter dictionaries of poor quality or, at least, dictionaries not suited for a specific function, we limit the critical system to a few lexicographically relevant situations and only some types of users.

The most general situations of dictionary use are, according to Tarp, communicative and cognitive contexts in which someone needs to produce texts or know something - in the database we name them *Communication* and *Knowledge*. To these we add two others, which are more specific and are expected to be the most typical for web surfers: contexts in which someone needs to translate (*Translation* in the database) or learn something (*Learning*). Therefore our inventory is made up of three *lexicographical parameters*: three kinds of users, two general and two specific consultation situations (see fig. 1). The kind of user parameter is thus limited to laymen, experts, and semi-experts of one field, e. g. economy journalists who are not economists themselves (Bergenholtz & Kaufmann, 1997; Hartmann, 1989).

To the parameters, feature frequency (see fig. 1) has been added, in order to keep track of the features that are always present and those which occur only sometimes in one dictionary, since the majority of these lexicons lack any strict lexicographical organization, and offer unsystematic assistance to users.

Lexicographical parameters →	Users									General Situations						Specific situations						
	Layman			Semi-expert			Expert			Knowledge			Communication			Translation			Learning			
Feature frequency →	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	
Dictionary features ↓																						
General Organization and Host Site																						
Mediostructure																						
Microstructure																						
Linguistic fields																						
Non linguistic fields																						
Maximum score	24			24			24			30			30			25			25			

**Figure 1: Lexicographical parameters (*Users, General Situations, Specific Situations*), lexicographical profiles (*Layman, Semi-Expert, Expert, Knowledge, Communication, Translation, Learning*), and dictionary features (addressing the *General Organization, Mediostructure, Microstructure*) with their occurrence frequency (Yes, No, S.= Sometimes).**

On this basis, the features considered to be more relevant (Bothma & Tarp 2012) for one parameter receive 1 or 2 points score, conversely, negative scores (-1, -2) are given to those judged as contradictory. Thus the evaluation scale is made as follows:

- 2 points to the most relevant features
- 1 point to relevant features
- -1 to contradictory features
- -2 to the most contradictory features

The specifics of each lexicographical parameter determine what we call here a *lexicographical profile*, which is outlined by its characterizing features, as it is displayed in table 2 below.

Lexicographical profile	Features and scores
Layman	Institutional Site: Yes (2); Specialised Site: Yes (1); Technical and non-technical terms: Yes (2); Cross-references: Yes (2); Related terms: Yes (1); Hypernyms & Hyponyms: Yes (1); Pronunciation notation: Yes (1); Stress information: Yes (1); Audio files: Yes (2); Technical definitions: Yes (-2), No (2); Example Sentences: Yes (2), No (-2), Sometimes (1); Quotations: Yes (-2), Sometimes (-1); Synonyms: Yes (2), Sometimes (1); Antonyms: Yes (2), Definitions: Yes (2), Sometimes (1); Examples: Yes (2), Sometimes (1), Video files: Yes (2), Sometimes (1), Pictures: Yes (2), Sometimes (1).
Semi-Expert	Institutional Site: Yes (2); Specialised Site: Yes (1); Bibliographic resources: Yes (1) No (-1); Hyperlinks: Yes (1); Access: Advanced search engine: Yes (1); Entries: 0-49: Yes (-2); Entries: 50-100: Yes (-2); Technical and non-technical terms: Yes (1); Cross-references: Yes (1); Related terms: Yes (1); Hypernyms & Hyponyms: Yes (1); Phonetic transcription: Yes (1); Syllabification: Yes (1); Linguistic variation: Yes (2), Sometimes (1); Technical definitions: Yes (1) No (-1); Quotations: Yes (2); Idioms: Yes (2), Sometimes (1); Collocations: Yes (2), Sometimes (1); Etymology: Yes (1), No (-1); Definitions: Yes (1); Domain field: Yes (1); Pictures: Yes (1).
Expert	Institutional Site: Yes (2); Specialised Site: Yes (1); Bibliographic resources: Yes (2), No (-2); Hyperlinks: Yes (2); Access: Browse: Yes (-1); Entries: 0-49: Yes (-2); Entries: 50-100: Yes (-2); Hypertexts: Yes (1); Phonetic transcription: Yes (2), Sometimes (1); Syllabification: Yes (2), Sometimes (1); Linguistic variation: Yes (2), Sometimes (1); Technical definitions: Yes (2), No (-2), Sometimes (1); Quotations: Yes (2), No (-2), Sometimes (1); Idioms: Yes (2), Sometimes (1); Collocations: Yes (2), No (-2), Sometimes (1); Etymology: Yes (2), No (-1), Sometimes (1); Domain field: Yes (1).
Knowledge	Institutional Site: Yes (2); Specialised Site: Yes (1); Bibliographic resources: Yes (2); Hyperlinks: Yes (2); Cross-references: Yes (2); Related terms: Yes (2); Sometimes (1); Hypernyms & Hyponyms: Yes (2), Sometimes (1); Hypertexts: Yes (2); Quotations: Yes (2); Sometimes (1); Etymology: Yes (2), Sometimes (1); Definitions: Yes (2), Sometimes (1); Examples: Yes (2), Sometimes (1); Domain field: Yes (2), Sometimes (1); Video files: Yes (2), Sometimes (1); Pictures: Yes (2), Sometimes (1); Cultural notes: Yes (2).
Communication	Institutional Site: Yes (2); Specialised Site: Yes (1); Technical and non-technical terms: Yes (2); Grammatical category: Yes (2), Sometimes (1); Morphological information: Yes (2), Sometimes (1); Syntactic pattern: Yes (2), Sometimes (1); Phonetic transcription: Yes (2), Sometimes (1); Pronunciation notation: Yes (1); Stress information: Yes (1); Audio files: Yes (2), Sometimes (1); Syllabification: Yes (1); Frequency of use: Yes (1); Linguistic variation: Yes (2), Sometimes (1); Example Sentences: Yes (2), Sometimes (1); Idioms: Yes (2), Sometimes (1); Collocations: Yes (2), Sometimes (1); Synonyms: Yes (2), Sometimes (1); Antonyms: Yes (2), Sometimes (1).



Translation	Institutional Site: Yes (2); Specialised Site: Yes (1); Multilingual dictionary: Yes (2); Multilingual word list: Yes (1); Plurilingual dictionary: Yes (2); Bidirectionality: Yes (2); Technical and non-technical terms: Yes (2); Grammatical category: Yes (1); Morphological information: Yes (1); Syntactic pattern: Yes (2), Sometimes (1); Linguistic variation: Yes (2), Sometimes (1); Translation equivalences: Yes (2), No (-2), Sometimes (1); Idioms: Yes (2), Sometimes (1); Collocations: Yes (2), Sometimes (1); Cultural notes: Yes (2).
Learning	Institutional Site: Yes (2); Specialised Site: Yes (1); Learning resources: Yes (2), No (-2); Bibliographic resources: Yes (2); Hyperlinks: Yes (2); Monolingual dictionary: Yes (2); Multilingual dictionary: Yes (2); Related terms: Yes (1); Grammatical category: Yes (1); Morphological information: Yes (1); Syntactic pattern: Yes (1); Audio files: Yes (2), Sometimes (1); Example Sentences: Yes (2), Sometimes(1); Definitions: Yes (2), Sometimes(1); Examples: Yes (2), Sometimes(1); Video files: Yes (1); Pictures: Yes (2), Sometimes(1).
Lexicographical profile	Features and scores

**Table 2: Score assignment in the evaluation system of the *Web Linguistic Resources* database. Specific features receive different scores and outline the different *lexicographical profiles* considered.**

In addition, the scores were given the following basic guidelines:

- 1) profiles belonging to the same lexicographic parameter may reach the same maximum score;
- 2) complementary profiles don't share the same features;
- 3) similar profiles may share the same features.

According to the first rule, user profiles may reach 24 points maximum each, general situations 30, and more specific consultation situations 25 (see fig. 1 and fig. 2).

The second principle, however, prevents the database from giving contradictory responses, such as dictionaries suited for laymen and experts at the same time. Therefore, referring to figure 2 below, technical definitions are required in the vocabularies for experts (2 points), but not in those for layman (-2). The opposite is also true: if a dictionary doesn't have technical definitions, it is suited for laymen (2) but not for experts (-2). Similarly, example sentences are expected in dictionaries for laypeople, and quotations in those for experts.

Dictionary features	Layman			Semi-expert			Expert			Knowledge			Communication			Translation			Learning			
	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	Yes	No	S.	
Technical definitions	-2	2		1	-1		2	-2														
Example Sentences	2	-2	1										2		1					2		1
Quotations	-2		-1	2		1	2	-2	1	2		1										
Etymology				1	-1		2	-1	1	2		1										
Maximum rating	24			24			24			30			30			25			25			

**Figure 2: Score giving to features according to the different profiles (*Layman, Semy-Expert, Expert, Knowledge, Communication, Translation and Learning*) and their occurrence frequency (Yes, No, S.=Sometimes).**

On the contrary, the second guideline states that if the profiles are similar, they can share features and scores, such as a specialized host site and information on syntactic patterns for the translation and learning situations (see fig. 3).

Dictionary features	Translation			Learning		
	Yes	No	S.	Yes	No	S.
Institutional Site	2			2		
Specialised Site	1			1		
Multilingual dictionary	2			2		
Grammatical category	1			1		
Morphological information	1			1		
Syntactic pattern	2		1	1		
Audio files				2		1

**Figure 3: Features in common for the *Translation* and *Learning* profiles.**

The evaluation procedure adopted is thus purely proscriptive (Andersen & Nielsen 2009), and based on the careful distribution of scores among the profiles inventoried in order to fulfill the requirement of the guidelines stated above. This should guarantee a balanced critical assessment procedure, minimizing the possibility that some profiles are easier to fulfill because they require lower maximum scores. Consequently, even though the comparative methodology used for the distribution of grades among the different profiles is paramount and not dismissible (Caruso forthcoming), at least one test on real users has already been carried out in order to check the overall validity of the proposed evaluation system (Caruso & De Meo 2013). In this study, the higher scoring medicine dictionaries of the WLR database for the *Translation* profile were used by 39 university students in a controlled translation session, and despite the overall low-quality of these vocabularies, students who consulted them to overcome some of the main difficulties in the source text performed better than those who translated freely, without referring to any dictionary whatsoever.

Focusing on the post-consultation phase, this small study on real dictionary use is just a starting point for the examinations that may be carried out in order to validate the assessment procedure of the WLR system, and the features that have been chosen to outline each *lexicographical profile*.

## 4 How to search the database

The features and the lexicographical (or rating) profiles are the main search options of the Web Linguistic Resources database. Accessing the homonymous site, it is possible to search for the dictionary that is best suited to the user's needs. The available options are listed in the center of the page, where the dictionaries ratings are provided as a percentage, since the score gives evidence of the degree to which the dictionary corresponds to the desired profile. Figure 4, for example, shows the search for a dictionary of biology suited for a learning context. The sector "biology" is a subfield within the dictionary features, which are listed on the left, while in the upper right of the page users can choose the rating profile.

The other available search options are the translation languages, the language in which the dictionary is written (*Main Language*), but also other languages present in the entry list (*Languages Involved*), for example French terms in English wine dictionaries or Latin words in German law lexicons.

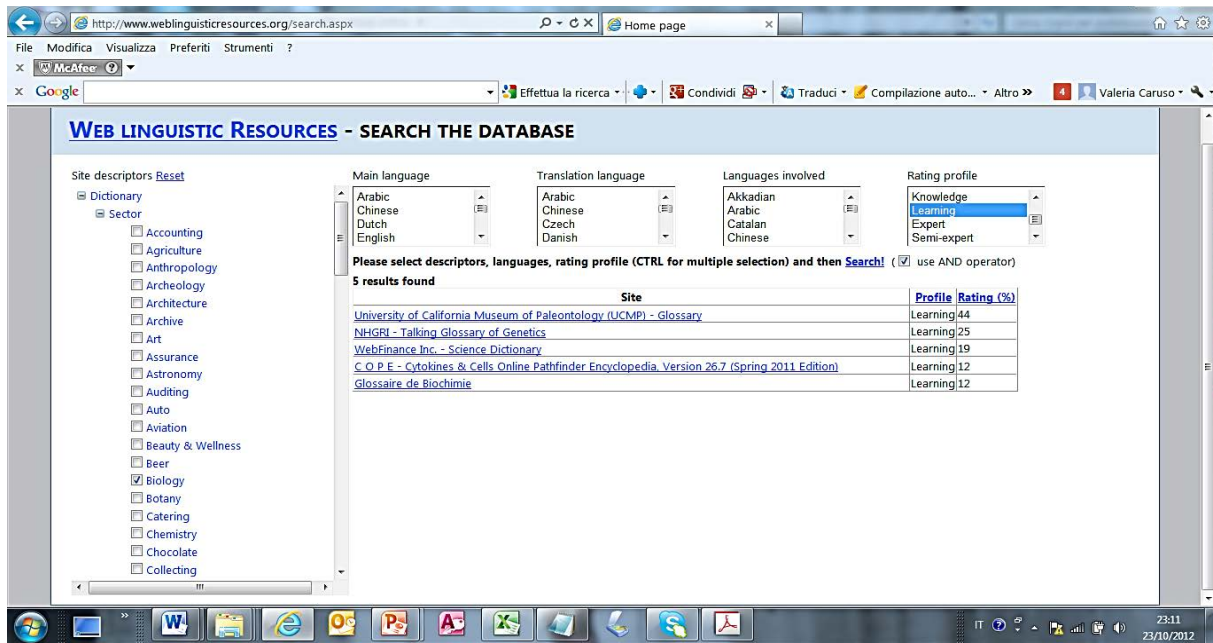


Figure 4: The search form of the *Web Linguistic Resources* database.

## 5 What remains to be done

At present the evaluation system filters dictionaries only on the basis of their features, according to explicit lexicographical parameters, but it doesn't provide any assurance about the reliability of contents, which nevertheless is one of the most urgent requirements for anyone browsing the Internet. Obviously it is impossible to vouch for the quality of every single piece of information provided by the web dictionaries or by any other dictionary. What is needed is to avoid resources that create problems for users instead of helping them. This is the case with the following explanations related to the enological term "extra dry":

Extra-Dry

Don't believe everything you read. What this really denotes is a sweet Champagne.

(Pacific Northwest Wine Company. Terminology and Descriptions)

extra dry

adj. Another step on the sweetness-level scale associated with Champagne. Starting on the low end with brut zéro, the scale ascends to brut nature, extra brut, and brut sauvage (all of which are bone-dry), then brut (dry), extra dry (a hint of sweetness), sec (slightly sweet), demi-sec (moderately

sweet), and *doux* (the sweetest of all). Why extra dry is sweeter than brut is a mystery to everyone but Francophiles. The only types of sparkling wine you're likely to see at the store are brut, extra dry, and demi-sec, of which brut is far and away the most popular. FYI, table wine that's slightly sweet is referred to as off-dry (*Wine Lovely - Glossary*).

In these examples, the discrepancy between the ordinary value of the adjective *dry* and its meaning in the *extra dry* specialised compound is particularly highlighted, and in the second definition the difference is also underlined using an indirect question: "Why extra dry is sweeter than brut is a mystery to everyone but Francophiles". However no answer is given.

One useful discriminatory criteria might be that of referring to dictionaries published by leading institutions of one field, but whilst browsing the Internet it is possible to collect examples of the lexicographical inexperience of experts responsible for dictionary writing. Firstly, if definitions are not compiled carefully, they can give bad explanations that eventually turn into information voids, this is the case with the entry *Chromosome* of the *Talking Glossary of Genetics*, published by the highly esteemed National Human Genome Research Institute. The definition says that: "Humans have 23 pairs of chromosomes(...), and one pair of sex chromosomes, X and Y", which is misleading, since XY is the chromosome pair of males, while women have XX, as is clearly explained in the voice for *Sex Chromosome*:

(...) Humans and most other mammals have two sex chromosomes, the X and the Y. Females have two X chromosomes in their cells, while males have both X and a Y chromosomes in their cells (...).

Secondly, sometimes the lack of any strict lexicographical organization prevents exhaustive meaning explanations. For example, the *University of California Museum of Paleontology* explains *Basement Rock* as follows:

basement rock -- n. The oldest rocks in a given area; a complex of metamorphic and igneous rocks that underlies the sedimentary deposits. Usually Precambrian or Paleozoic in age.

In fact, since no explicit label nor clear text divisions are provided, it is impossible to decide whether the first part of the definition "The oldest rocks in a given area;" is one possible meaning of "basement rock", or if "The oldest rocks in a given area;" is a synonym of the following part of the definition, particularly that which states: "Usually Precambrian or Paleozoic in age".

These brief examples give an idea of the kind of work that remains to be done, but not of the kind of solutions to be provided. In effect, after having established which features of the definition must be rated, two main evaluation options remain: one is to choose a pair of critical terms for each specialized field and analyze their definitions in every vocabulary, the other is to extract at random a fixed number of terms for each resource and provide statistically relevant assessments. Speaking in general

terms, the latter option is preferable, since the 'critical' terms of huge fields (e. g. medicine, economy etc.) are too numerous.

Therefore, the most suitable statistical evaluation model for the matter remains to be chosen, provided that the number of the rated definitions remains the same for every vocabulary, regardless of its entry number. Since the number of definitions considered doesn't change, it is necessary to provide each assessment of the dictionary entries with its variation coefficient, i. e. the precision index of the estimation made for the vocabulary considered. It is therefore unsurprising that small dictionaries will be rated more accurately than the big ones.

## 6 References

- Andersen, B., & Nielsen, S. (2009). Ten Key Issues in Lexicography for the Future. In H. Bergenholtz, S. Nielsen, S. Tarp (eds.) *Lexicography at a Crossroads. Dictionaries and Encyclopedias today, Lexicographical Tools tomorrow*. Bern etc.: Peter Lang, pp. 355-365.
- Bergenholtz, H. & Kaufmann, U. (1997). Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. In *Hermes*, 18, pp. 91-127.
- Bergenholtz, H. & Tarp, S., eds. (1995). *Manual of Specialised Lexicography*. Amsterdam, Philadelphia: John Benjamins.
- Bergenholtz, H. (2012). Concepts for Monofunctional Accounting Dictionaries. In *Terminology*, 18, 2, pp. 243-263.
- Bothma, T. J. D. & Tarp, S. (2012). Lexicography and the Relevance Criterion. In *Lexikos*, 22, pp. 86-108.
- Büro für angewandte Mineralogie*. Accessed at: <http://www.a-m.de/deutsch/inhalt.htm> [04/01/2014].
- Caruso, V. & De Meo, A. (2013). Comunicare i saperi sul Web: il caso dei dizionari specialistici. In C. Bosisio, C. Cavagnoli (eds.) *Comunicare le discipline attraverso le lingue: prospettive traduttiva, didattica, socioculturale. Proceedings of XII AItLA International Congress, Macerata, 23-24 febbraio 2012*. Perugia: Guerra.
- Caruso, V. & Pellegrino, E. (2012). Metadizionari digitali specialistici. In S. Ferreri (ed.) *Lessico e lessicologia. Atti del XLIV Congresso internazionale di studi della Società Italiana di Linguistica (SLI), Viterbo, 27-29 settembre, 2010*. Roma: Bulzoni, pp. 487-495.
- Caruso, V. (2011). Online specialised dictionaries: a critical survey. In I. Kosem, K. Kosem (eds.) *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011, Bled, 10-12 November*. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 66-75.
- Caruso, V. (Forthcoming). A guide for the quality assessment of dictionaries of economics. In P. Leroyer, S. Tarp (eds.) *Dictionaries of Economics in the 21st Century: The Challenges of Online Lexicography*. Berlin, Boston: De Gruyter.
- Fuertes-Olivera, P. A. (2009). The Function theory of lexicography and electronic dictionaries: Wiktionary as a Prototype of Collective Multiple-Language Internet Dictionary. In H. Bergenholtz, S. Nielsen, S. Tarp (eds.) *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographica Tools Tomorrow*. Bern etc.: Peter Lang, 99-134.
- Geeb, F. (1998). Semantische und enzyklopädische Informationen in Fachwörterbüchern. Eine Untersuchung zu fachinformativen Informationstypen mit besonderer Berücksichtigung wortgebundener Darstellungsformen. In *Hermes*, 21, pp. 205-216.
- Hartmann, R. R. K. (1989). Sociology of the Dictionary User: Hypotheses and Empirical Studies. In F. J. Hausmann, O., Reichmann, E. H. Wiegand, L. Zgusta, (eds.) *Wörterbücher/Dictionaries/Dictionnaires*.

- Ein internationales Handbuch zur Lexikographie/An International Encyclopedia of Lexicography/ Enciclopédie internationale de lexicographie. Berlin-New York: De Gruyter, vol. I, pp. 102-111.
- Hartmann, R. R. K., (2001). Teaching and Researching Lexicography. Applied Linguistics in Action. Harlow: Longman-Pearson Education.
- Hausmann, F. J. & Wiegand, H. E. (1989). Component Parts and Structures of General Monolingual Dictionaries: A Survey. In F. J. Hausmann, O., Reichmann, E. H. Wiegand, L. Zgusta, (eds.) Wörterbücher/ Dictionaries/Dictionnaires. Ein internationales Handbuch zur Lexikographie/An International Encyclopedia of Lexicography/Enciclopédie internationale de lexicographie. Berlin-New York: De Gruyter, vol. II, pp. 328-360.
- Heid, U. (2011). Electronic Dictionaries as Tools: Toward an Assessment of Usability. In P. A. Fuertes-Olivera, H. Bergenholtz (eds.) e-Lexicography. The Internet, Digital Initiatives and Lexicography. London, New York: Continuum, pp. 287-304.
- Nielsen, S. (1994) The Bilingual LSP Dictionary. Principles and Practice for Legal Language. Tübingen: Narr Francke Attempto Verlag.
- Nielsen, S. (2003). Mediostructures in Bilingual LSP Dictionaries. In R. R. K. Hartmann (ed.) Lexicography. Critical Concepts. Lexicography, Metalexigraphy and Reference Science, London: Routledge, vol. III, pp. 270-294.
- Nielsen, S. (2009) Reviewing Printed and Electronic Dictionaries: a Theoretical and Practical Framework. In S. Nielsen, S. Tarp (eds) Lexicography in the 21st Century. In honour of Henning Bergenholtz. Amsterdam, Philadelphia: John Benjamins, pp. 23-41.
- Nielsen, S. (2013). A General Framework for Reviewing Dictionaries. In O. Karpova, F. Kartashkova (eds.): Multi-disciplinary Lexicography: Traditions and Challenges of the XXIst Century. Cambridge: Cambridge Scholars Publishing: 145-157.
- OneLook*. Accessed at: <http://www.onelook.com> [04/01/2014].
- Pacific Northwest Wine Company. Terminology and Descriptions*. Accessed at: <http://pcnwwineco.com/terminology-and-descriptions> [04/01/2014].
- Pym, A. (2004). The moving text: localization, translation, and distribution. Amsterdam, Philadelphia: John Benjamins.
- Talking Glossary of Genetics*. Accessed at: <http://www.genome.gov/glossary/index.cfm> [04/01/2014].
- Tarp, S. (2008). Lexicography in the Borderland between Knowledge and Non-knowledge. General Lexicographical Theory with particular Focus on Learner's Lexicography. Lexicographica. Series Maior, Berlin-New York: De Gruyter.
- Tarp, S. (2009). Beyond Lexicography: New Visions and Challenges in the Information Age. In H. Bergenholtz, S. Nielsen, S., Tarp (eds.) Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow. Bern: Peter Lang, pp. 17-32.
- Tarp, S. (2010). Functions of Specialized Learners Dictionaries. In P. A. Fuertes-Olivera (ed.) Specialised dictionaries for learners. Lexicographica. Series Maior, Berlin-New York: De Gruyter, pp. 39-53.
- University of California Museum of Paleontology - Glossary*. Accessed at: <http://www.ucmp.berkeley.edu/glossary/glossary.html> [04/01/2014].
- Web Linguistic Resources*. Accessed at: [www.weblinguisticresources.org](http://www.weblinguisticresources.org) [04/01/2014].
- Wiegand, H. E. (1996). Textual Condensation In Printed Dictionaries: a Theoretical Draft. In *Lexikos*, 6, pp. 133-158.
- Wiktionary*. Accessed at: [https://en.wiktionary.org/wiki/Wiktionary:Main\\_Page](https://en.wiktionary.org/wiki/Wiktionary:Main_Page) [04/01/2014].
- Wine Lovely - Glossary*. Accessed at: <http://www.winelovely.com/index.asp?codice=43> [04/01/2014].
- Your Dictionary*. Accessed at: <http://www.yourdictionary.com> [04/01/2014].

## Acknowledgements

Special thanks are due to Gianluca Monti for providing the project with all the technical assistance needed.

# What a Multilingual Loanword Dictionary can be used for: Searching the *Dizionario di italianismi in francese, inglese, Tedesco* (DIFIT)

Matthias Heinz, Anne-Kathrin Gärtig  
Universität Salzburg  
matthias.heinz@sbg.ac.at, anne-kathrin.gaertig@sbg.ac.at

## Abstract

The paper outlines the structure and practical use of the *Dizionario di italianismi in francese, inglese, tedesco* (DIFIT), a dictionary registering Italian loanwords in three languages published in print and currently being prepared for digital reedition. This specialized lexicographical resource focuses on language contact resulting in synchronic and diachronic lexical transfer from the Italian language to French, English and German. Italian as a culturally extremely rich and diverse source for borrowings has extensively influenced other languages in different historic phases all along the past centuries. The DIFIT reflects this rich European heritage of borrowings in a maximal variety of lexical fields. Major methodological aspects in and prior to editing the dictionary included the delimitation and exploration of the source material and deciding on depth of historical coverage as well as defining the information programme of the microstructure. At the same time, limitations of the general lexicographical documentation on both Italian and the three recipient *languages presented particular challenges*. *The poster displays results of some first quantitative research on linguistic aspects of the borrowed elements in DIFIT and gives an outlook on the ongoing project of digitizing the lexicographical data by means of an online database documenting the linguistic impact of Italian as a global donor language (Osservatorio degli italianismi nel mondo/OIM, hosted by the Accademia della Crusca).*

**Keywords:** Multilingual Lexicography; Specialized Dictionaries; Online Lexicography; Language contact; History of the Italian/French/English/German Lexicon

*“Il faudrait encore, assurément, repérer, cartographier la diffusion de la langue italienne elle-même, cet élément insistant de toute culture européenne.” (Braudel 1989: 15)*

## 1 Introduction

This paper intends to give an overview of the structure and possible applications for linguistic research of the *Dizionario di italianismi in francese, inglese, tedesco* (DIFIT), published in print in 2008 by the Accademia della Crusca and currently being released online as part of a larger project (*Osservatorio degli italianismi nel mondo / OIM*). As a specialized lexicographic resource it focuses on language contact between Italian and three major European languages, French, English, German, whose lexical outcomes are presented in both synchronic and diachronic perspective.<sup>1</sup> The lexicographic metalanguage of the DIFIT is Italian, but the dictionary is in itself multilingual in so far as it registers lexical transfer from the donor language Italian to the three languages mentioned, the respective loan items following the lemmatization of their Italian source words (etyma).

Literary Italian (based on the Tuscan variety) along with many of its dialectal varieties stands out as a culturally extremely rich and diverse source for borrowings. While its extensive lexical influence on the other three European languages in different historic phases all along the past six centuries has been noted in a host of studies, the somewhat scattered lexicographical documentation is brought together for the first time in a reference work spanning more than one target language. By taking into account a variety of lexical fields, the DIFIT aims at reflecting the rich European heritage of borrowings which becomes evident in parallel loans and lexical interrelations within the French, English and German vocabularies.

A borrowing (It. *prestito*, G. *Entlehnung*) is seen here as the result of the imitation of a foreign linguistic pattern by a speech community (cf. Gusmani 1986, Pinnavaia 2001), meaning elements like the following:

- (1) it. *ciao* (→ present in Fr., Eng., G.)
- (2) it. *dolce far niente* (→ Eng., G.)
- (3) it. *-issimo* suffix (→ e.g. Fr. *Affairissimo*)
- (4) it. *eppur si muove!* → (Fr., Eng., in G. loan translation: *Und sie bewegt sich doch!*)

(DIFIT-OIM<sup>2</sup>: s.vv.)

The loan items present in the lexicon of a language can be single words like *ciao* (1), multiword expressions as (2) *dolce far niente*, or sometimes formatives as the Italianizing suffix with superlative meaning *-issimo* (cf. Fr. *affairissimo*, or G. *Transportissimo*, cf. Stammerjohann 2010)<sup>3</sup>, as well as phrases (pro-

1 For details of the macro- and microstructural makeup of the print dictionary cf. Heinz (2008).

2 Data from the DIFIT corpus available in digitized form via the OIM are labelled here as DIFIT-OIM. These are searchable at [www.italianismi.org](http://www.italianismi.org).

3 Cf. also examples in Austrian and Swiss German advertisements such as *Vielfaltissimo* (denominal derivation), *Perfektissimo* (deadjectival), *Verwöhnissimo* (deverbal) based on a product name consisting of a hybrid formation (*Caf-* 'coffee' + *-issimo*). Such productive formatives based on loan items are described as „induzierte“ by Gusmani (1986: 155).



verbal and idiomatic expressions) like „eppur si muove!“ (4) in French and English, rendered in German by the loan translation *Und sie bewegt sich doch!*<sup>4</sup>

At present the DIFIT lists 8951 Italianisms and 4660 Italian etyma (242 of which are without prior attestation elsewhere). Following a concise presentation of the dictionary and its digital counterpart, some examples of how the dictionary data can be used for linguistic analyses will be outlined in this paper.

## 2 Language contact and lexicographical documentation: DIFIT and OIM

From the metalexicographical point of view informing the classification proposed by Wiegand (2001), the DIFIT can be described as representative of the type of the „aktives polylaterales Sprachkontakt-wörterbuch“ (‘active, polylateral dictionary of language contact’); it is active insofar as its lemmatization sets out from the etyma and polylateral in its registering loanwords of one donor language in several recipient languages. Moreover, it is, as mentioned at the beginning, monolingual as to the metalanguage, Italian, while being multilingual as regards its set of recipient languages (French, English, German). The general scope of the dictionary is stated in the Introduction:

[I] presente *Dizionario* [...] vuole offrire più che una semplice somma dei dati finora raccolti e acclarati. Il suo **scopo** è più specifico, è quello di **mettere a confronto l’incidenza dell’italiano sul francese, l’inglese e il tedesco**, le tre lingue che sono al centro dello spazio europeo e sono a più stretto contatto tra loro, con l’intento di **ricostruire le trafele di penetrazione e la diversa sorte delle parole italiane** in questo circuito. (DIFIT: XI; emphasis added)

In summary, the intention of this work vis-à-vis specialized repertoires and studies centred on Italianisms present in single languages, on whose results the DIFIT draws, is to offer more than a simple listing of data yet collected and examined. Its intention is more specifically “mettere a confronto l’incidenza dell’italiano sul francese, l’inglese e il tedesco”, i.e. ‘compare the impact of Italian on French, English, and German’, three languages situated centrally in a European communicative space and in close contact among each other, in order to reconstruct the mechanisms of lexical interpenetration and the ‘variable fate of Italian words in this [historical and cultural] circuit’. By design it presupposes a certain depth of diachronic coverage both in the macrostructural preselection and in the microstructural organization of its entries.

---

4 Examples drawn from DIFIT-OIM.

The DIFIT is currently being expanded within the OIM project. The *Osservatorio degli italianismi nel mondo*, an international collaboration hosted by the Accademia della Crusca, has two main purposes:

- (a) making available and editing the data collected in the DIFIT in an online database;
- (b) extending the contact languages beyond the three major European languages French, English and German documented in the DIFIT, aiming at the lexicographical description of the impact of Italian as a source language e.g. in Spanish, Polish, Japanese etc.

Hence the goal of this collaborative effort is a linguistic ‘observatory’ for Italian loanwords in the world’s languages.

### 3 The contribution of a loanword dictionary to linguistic analyses

The OIM database<sup>5</sup> offers various search options. Besides a free search option (“ricerca libera”) complex searches in the list of etyma (“ricerca negli etimi”) are made possible by applying and combining the following criteria:

- Italian etyma (+ variants)
- grammatical categories (etymon)
- year of first attestation (etymon) with optional delimitation of time span
- archaic / obsolete / without lexicographical attestation
- domain of use / lexical field (etymon)
- dialectal origin (etymon)
- register
- recipient language (by year or range of years + one, two or all of the recipient languages)

The sources documenting the etyma can be displayed in a window with the full bibliographical reference by moving the cursor to the respective abbreviation. The “ricerca negli italianismi” (search Italianisms) option allows for complex search paths with the following set of criteria:

- Italianisms (+ variants)
- grammatical categories
- recipient language

---

5 The database was designed and implemented as part of the electronic resources of the Accademia della Crusca by Marco Biffi with the help of Giovanni Salucci and Maurizio Rago, while the task of populating the database by digitizing and adapting the DIFIT data for the new online user interface is the work of Gesine Seymer.

- year of first attestation with optional delimitation of time span
- archaic / obsolete / without lexicographical attestation
- domain of use / semantic field
- dialectal use
- register
- type of borrowing:
  - integral
  - partial (calque: formal / partial / semantic)
  - direct
  - indirect (with mediating language)
  - pseudo-borrowing<sup>6</sup>
  - borrowings whose (Italian) origin is doubtful or unclear (e.g. contradictory etymological information in authoritative sources)

Here too, sources for the Italianisms can be displayed with the full bibliographical reference by moving the cursor to the abbreviation.

In what follows, three exemplary fields of application are sketched out in order to show how the DIFIT data can be used for linguistic research, yielding some first results related to a number of research questions especially in contact linguistics.<sup>7</sup> These are namely the word formation types in the multiword lemmata identified in the DIFIT corpus, the typology of the borrowings and, as a further outlook, the visualization of the results of language contact in the OIMap project.

## 4 Language contact and word formation

Of the 4660 lemmata listed in the DIFIT, 4161 are single words, while 499 (more than 10%) are classified as *locuzioni*, multiword expressions including complex syntagmatic units or phrases as *con spirito* and lexicalized compounds like *pesce spada*. A semiautomatic count with manual classification of the formation types yields the following percentages:

---

6 Also *pseudo-loan* (It. *pseudo-prestito*, G. *Scheinentlehnung* alongside other terms), cf. Winter-Froemel (2011: 44-45), „Bildungen wie dt. *Picobello* [...], die aus einer anderen Sprache (hier dem Italienischen [...]) direkt entlehnt zu sein scheinen, die sich aber in der vermeintlichen AS [Ausgangssprache] in dieser Form als nicht existent erweisen.“ (45)

7 The relevance of studying loanwords for the enterprise of general linguistics and linguistic typology becomes evident when using resources like the World Loanword Database (WOLD, Haspelmath/Tadmor 2009), which analyzes lexical borrowability based on a restricted set of core vocabulary meanings (1460) in 41 (mainly non-Indo-European) recipient languages; Italian is present among the 369 donor languages, though only with a small number of loanwords (86 entries in the database, some having an identical source word).

Formation	Examples	%
N + Adj	<i>salto mortale, sinfonia concertante, opera comica</i>	33,3
Prep + N	<i>a battuta, a conto, con spirito</i>	29,7
Adj + N	<i>dolce vita, sacra conversazione, onorata società</i>	15,8
N + Prep + N	<i>giorno di respiro, lira da gamba, zuppa di pesce</i>	13,5
N + N	<i>pesce spada, pensione baby</i>	4,0
[N + N]	<i>acquamarina, acquatinta, autostrada</i>	3,6

**Table 1: Distribution and formation type of multiword expressions in DIFIT-OIM.**

In a further step these data (as well as the overall database) can be analyzed with regard to the single recipient languages in order to compare how processes of loan integration may display a language specific dynamics.

## 5 Typology of borrowings

Categorizing and quantifying the diverse types existing alongside classical loanwords (such as formal or semantic calques, rare pseudo-borrowings together with hybrid formations etc.) gives a general picture of the impact of formal and semantic processes on the diachronic evolution and synchronic stratification of the lexicon in the recipient languages.

**GENERALISSIMO** ◀ ▶

---

**GENERALISSIMO** s. m. mil. / armi stor. Sec. XVI. Un tempo, comandante supremo di un esercito. (DELI).

**F** **généralissime** (calco formale) s. m. mil. / armi 1600-12 (TLF).

**I** **I. generalissimo** s., pl. -os mil. / armi 1621 (OED).  
**II. generalissima** (pseudoprestito) s. f. mil. / armi 1643 (OED).

**T** **Generalissimus** (< it. × lat. -US) s. m., gen. -, pl. -mi e -musse mil. / armi 1674-78 (DuGW; Hechtenberg).

**Figure 1: Different loan outcomes in the recipient languages (DIFIT-OIM, s.v. ‘generalissimo’).**

As a lemma like *generalissimo* (Fig. 1) shows, the loan formations resulting from even a single etymon may be characterized by a variety of outcomes. Here French has a formal calque, English a pseudo-borrowing based on an Italian structural model, whereas the German *Generalissimus* is identified as a hybrid combining an Italian base with a Latin(ate) ending.

Typology of borrowings	%
Direct borrowing (loanword)	94,5
Formal calque (loan translation)	3,3
Partial calque	0,8
Semantic calque	1,3
Pseudo-Italianism	0,1

**Table 2: Distribution of various types of borrowings registered in DIFIT-OIM.**

Table 2 gives the percentages of different types of borrowings. Clearly direct borrowings (loanwords in a strict sense) are by far the most common type among the Italianisms registered in DIFIT, while the most frequent type of calque is the word for word rendering of a foreign model with the means of the recipient language known as loan translation (e.g. It. *monte di pietà* → Fr. *mont-de-piété*), followed by partial calque (It. *fare fiasco* → G. *Fiasko machen*) and semantic calque (It. *futurista* → Eng. *futurist*). Pseudo-Italianisms like Eng. *generalissima* or Fr./Eng./G. *tutti-frutti* (with various meanings) are extremely rare (0,1%) in the DIFIT documentation.

Furthermore, thematic indices based on the different sectors of the lexicon that have contributed Italianisms can be created with the OIM search engine. Thanks to functions permitting combined search criteria it becomes then possible to list the results by chronology and/or target language.<sup>8</sup>

## 6 Visualizing the areal distribution of Italianisms: OIMap

As a further outlook, we will briefly describe an instrument for visualizing the areal distribution (and dynamics) of Italianisms by means of a mapping tool currently being developed under the name of OIMap.<sup>9</sup> In fact, while lists of borrowed lexical items in a dictionary or database provide multi-faceted information in a very dense manner, new insights into the geographic distribution and the dynamic tendencies of contact relationships between languages can be gained through maps.

Drawing upon a free software tool for digital cartography (<http://batchgeo.com>), OIMap intends to show the geographic direction of language contact. Moreover, the dialectal origin of borrowed elements can be obtained (other than manually) by using refined search options. Thereby dialect varieties having contributed a relatively high number of loanwords to the lexicon of one, two or all three of the extant recipient languages can be singled out, and as in the case of Venetian (or, with a comparatively lower number of lexical items, Genoese) a clearer picture of the areal dynamics emerges. In Fig.

<sup>8</sup> The ,ups and downs' of Italian lexical influence in the three languages taken into account by DIFIT are illustrated by Stammerjohann/Seymer (2007: 51), with the availability of the OIM search engine the raw data for even more fine-grained analyses can be elicited as described in 3.3.

<sup>9</sup> OIMap is designed for the cartographic representation of language contact dynamics based on the DIFIT data (project situated at the University of Salzburg under the direction of M. Heinz).

3 an arrow indicates the main direction of borrowings where an absolute (Genoese: 7 out of 9) or relative (Venetian: 15 out of 45) majority of loan elements ‘migrates’ exclusively towards one language (French and German respectively).



**Figure 3: OIMap – Areal dynamics of Genoese and Venetian borrowings (FR = French, TED = German).**

As the OIM database is developed further and the number of recipient languages grows,<sup>10</sup> an interface with the cartographic tool OIMap becomes possible, resulting in a world map<sup>11</sup> of lexical ‘migration’ setting out from Italian and its varieties.

## 7 References

- Biffi, M. et al. (2014). *Osservatorio degli italianismi nel mondo*. Firenze: Accademia della Crusca ([www.italianismi.org](http://www.italianismi.org) [10/04/2014]).
- Braudel, F. (1989). *Le modèle italien*. Paris: Arthaud.
- DIFIT = Stammerjohann, H., Arcaini, E., Cartago, C., Galetto, P., Heinz, M., Mayer, M., Rovere, G., Seymer, G. (2008). *Dizionario di italianismi in francese, inglese, tedesco*. Firenze: Accademia della Crusca.
- Gusmani, R. (1986). *Saggi sull'interferenza linguistica. Seconda edizione accresciuta*. Firenze: Le Lettere.
- Haspelmath, M., Tadmor, U. (eds.) 2009. *World Loanword Database*. Leipzig: Max Planck Institute for Evolutionary Anthropology (<http://wold.livingsources.org> [10/04/2014]).

<sup>10</sup> For quite a number of languages electronic lists of Italianisms are at hand, though unpublished and in different formats, so integrating the existing material into the OIM database may soon result in substantial progress as for the documentation of Italian lexical influence.

<sup>11</sup> Such a map could provide more detailed cartographic and lexical information than the useful but rather plain one figuring on the WOLD website (Haspelmath/Tadmor 2009).

- Heinz, M. (2008). L'expérience du Dizionario di italianismi in francese, inglese, tedesco (DIFIT): objectifs, structure et aspects méthodologiques. In F. Pierno (ed.) *Aspects lexicographiques du contact entre les langues dans l'espace roman*. Strasbourg: Université Marc Bloch, pp. 165-180.
- Pinnavaia, L. (2001). *The Italian Borrowings in the Oxford English Dictionary: A lexicographical, linguistic and cultural analysis*. Roma: Bulzoni.
- Stammerjohann, H., Seymer, G. (2007). L'italiano in Europa: italianismi in francese, inglese e tedesco. In N. Maraschio (ed.) *Firenze e la lingua italiana fra Nazione e Europa*. Firenze: Accademia della Crusca, pp. 41-55.
- Stammerjohann, H. (2010). Italianismi. In *Enciclopedia dell'italiano*. Roma: Treccani ([http://www.treccani.it/enciclopedia/italianismi\\_\(Enciclopedia-dell'Italiano\)](http://www.treccani.it/enciclopedia/italianismi_(Enciclopedia-dell'Italiano))).
- Wiegand, H.-E. (2001) Sprachkontaktwörterbücher: Typen, Funktionen, Strukturen. In *Germanistische Linguistik* 161/162, pp. 115-224.
- Winter-Froemel, E. (2011). *Entlehnung in der Kommunikation und im Sprachwandel*. Berlin/New York: de Gruyter. <http://batchgeo.com> [10.04.2014]





# The Basic Estonian Dictionary: the first Monolingual L2 learner's Dictionary of Estonian

Jelena Kallas, Maria Tuulik, Margit Langemets  
Institute of the Estonian Language  
jelena.kallas@eki.ee, maria.tuulik@eki.ee, margit.langemets@eki.ee

## Abstract

This paper is a report on a lexicographical project completed by the Institute of the Estonian Language. The Basic Estonian Dictionary was published in print in March 2014, and the online version will be available by September 2014. The BED is aimed at learners of Estonian as a foreign language or as a second language at the elementary and intermediate levels (A2 to B1) according to the Common European Framework of Reference for Languages.

The dictionary contains about 5,000 headwords, including single items and multi-word lexical items. The BED provides lexicographical information on pronunciation, morphological information, definitions, word formation, government and collocation patterns, multi-word phrases, semantically related words and usage notes. In the online version, sound recordings (mp3 audio files) are provided. Morphological information was generated automatically. The most frequent government and collocation patterns were analysed and selected using the Sketch Engine corpus query system (Kilgarriff et al. 2004).

The BED contains approx. 400 illustrations, study pages and picture pages (e.g. related to animals). In the appendix, geographical names and grammar tables are included.

In addition, the Dictionary of the Estonian Sign Language (containing approx. 6,700 video recordings), based on the BED database, was published online in March 2014.

**Keywords:** L2 monolingual lexicography; active dictionary; Estonian

## 1 Purpose and structure of the dictionary

The Basic Estonian Dictionary (henceforth BED) is a monolingual active dictionary aimed at learners of Estonian as a foreign language or as a second language at the elementary and intermediate levels. The dictionary contains about 5,000 headwords, which were chosen on the basis of their frequency in the Estonian Reference Corpus<sup>1</sup>, with 250 million tokens as input. In addition, headwords that are necessary in everyday life, but might not be as frequent in corpora, were added, for example *pott* 'pot',

---

1 <http://www.cl.ut.ee/korpused/segakorpus/> [01/04/2014]

*pann* 'pan', *jahu* 'flour', and *kõhima* 'cough'. To get systemic content, some semantic classes (e.g. animals, plants and professions) were specially analysed.

Headword list includes not only single items, but also multi-word lexical items. Multi-word lexical items presented independently are multi-word verbs – particle verbs (verb + adverb particle, e.g. *alla kukkuma* 'fall down') and expression verbs (verb + noun/adjective phrase, e.g. *aru saama* 'understand') – and multi-word interjections (e.g. *tere õhtust* 'good afternoon'). The headword list of the BED was considered to be the controlled vocabulary list of the whole dictionary, other words were used in the entries. This was intentional so that users can look up unknown words in the dictionary.

Morphological information was generated automatically by using a morphological synthesizer for Estonian<sup>2</sup>. The BED as a learner's dictionary uses a comprehensive form-based presentation of data. For declinable words, grammatical cases in singular and plural, as well as the short form of the Illative, are presented explicitly. For verbs, the *-ma* and *-da* infinitives, and *he/she* forms, the past participle forms are given. However, after automatic generation there was a need for manual control of generated forms. Mostly this was necessary for the identification of homonymy. But forms were also deleted and added according to their frequency in corpora.

Information on pronunciation (palatalization, stress and syllabic quantity (in Estonian, a tripartite correlation of three syllabic quantities of stressed syllables exists)) is presented on the level of basic morphological forms of headwords. This is done by means of special palatalization, quantity and stress marks. Stress is shown only in cases where it is not on the first syllable, the normal stress pattern in Estonian.

Information on word formation is built into the micro-structure. Compounds with the headword as a second element (base word) are presented as references/links to their own entries, without additional information in the entry of the base word. Only transparent compounds, where the meaning of the base word has been preserved, were selected. All referenced compounds are presented as independent headwords as well.

Semantically related words (synonyms, antonyms and paronyms) of headwords are shown using the simplest possible metalanguage, e.g. *sama mis* 'same as' for synonyms and *vastand* 'opposite' for antonyms.

At the end of some entries, there are usage notes. Usage notes show differences between words and help to build vocabulary, e.g. polite phrases related to particular headwords are given and usages prone to error are pointed out.

The XML database of the Basic Estonian Dictionary is organized into several fields: lemma, pronunciation, inflectional information, definition, word formation, government, collocation, multi-word patterns, semantically related words and usage notes.

---

2 <http://www.eki.ee/keeletehnoloogia/projektid/morfana/> [01/04/2014]

## 1.1 Government and collocation patterns in the BED

As the BED is an active dictionary, the explicit presentation of syntagmatic relations (government and collocational patterns, also multi-word phrases) are of the utmost importance.

The most frequent government and collocation patterns were analysed and selected using the Sketch Engine corpus query system (Kilgarriff et al. 2004).

Estonian Sketch Grammar (Kallas 2013) is geared towards the specification of the Estonian Reference Corpus and it contains 85 rules (14 UNARY, four SYMMETRIC, 62 DUAL and five TRINARY grammatical relations). As a result, the system searches for 32 types of lexicogrammatical constructions.

For nouns, the system searches for modifying adjectives, participles, oblique-case substantives, adverbs, pronouns, prepositional phrases, non-finite verbs and (by identifying conjunctive words) subordinate clauses.

For adjectives, the system searches for modifying adjectives, adverbs, oblique-case substantives, prepositional phrases, non-finite verbs and (by identifying conjunctive words) subordinate clauses.

For adverbs, the system searches for modifying adverbs, oblique-case substantives, prepositional phrases and (by identifying conjunctive words) subordinate clauses.

For verbs, the system searches for substantives that function as subjects, objects and adverbials, and also for modifying adjectives, adverbs, prepositional phrases, non-finite verbs, gerundives and (by identifying conjunctive words) subordinate clauses.

Multi-word verbs, i.e. particle verbs (verb + adverb particle, e.g. *alla kukkuma* 'fall down'), expression verbs (verb + noun/adjective phrase, e.g. *aru saama* 'understand'), catenative verbs (verb + non-finite verb, e.g. *käima panema* 'start', lit. 'make [the engine] work'), and support verb constructions (e.g. *läbirääkimisi pidama* 'negotiate') are considered separately. Since adverbial particles are tagged in the corpus as regular adverbs, a list of adverbial particles was compiled. The system identifies the most frequent adverbial particles used with particular verbs. This feature has great value when lexicographers need to choose what kind of particle verbs should be presented in the dictionary. Secondly, it is possible to see components of expression verbs if the component concerned has the part-of-speech tag X. Other components of multi-word verbs are identified as objects, adverbials or modifying non-finite verbs.

In addition, constructions with the conjunctions *ja/või* 'and/or', and *kui/nagu* 'as' can be found for all content words. For nouns, the system also searches for predicatives (complements of the copula-like verb *olema* 'be').

Figure 1 shows the word sketch for the noun *diskussioon* ‘discussion’.

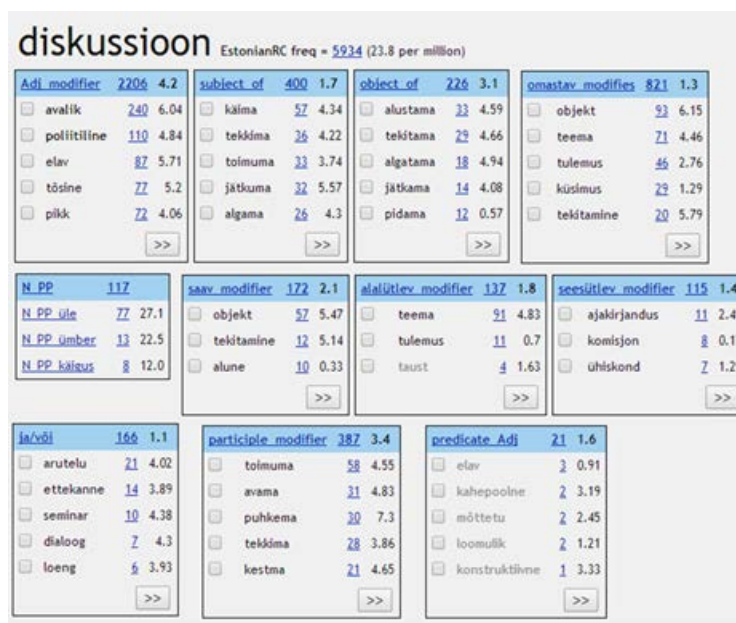


Figure 1: Word sketch of the noun *diskussioon* ‘discussion’.

The word sketch offers the lexicographer the most frequent collocates that occur as adjectival modifiers (e.g. *avalik* ‘public’, *poliitiline* ‘political’, *elav* ‘lively’, *tõsine* ‘serious’, *pikk* ‘long’ and *avatud* ‘open’), various oblique-case substantive modifiers (e.g. *diskussiooni objekt/teema/tulemus* ‘object/topic/result of discussion’) and in the ‘and/or’ relation to the node word (e.g. *diskussioon ja arutelu* ‘discussion and debate’).

Also identified are relations where the node word functions as subject (e.g. *diskussioon käib/tekib/jätkub* ‘discussion takes place/starts/continues’) and object (e.g. *diskussiooni alustama/algatama/jätkama/avama* ‘start/initiate/continue/open a discussion’).

The most frequent extracted patterns are mostly included in the entry of particular words and registered in the dictionary database.

In the BED database, the government pattern field contains data about the government pattern, together with attributes for the type of government (object, case, adposition, infinitive and conjunctive word government), as well as the position of the complements and complementation variability.

The collocation pattern field contains data about the collocation pattern, together with attributes for the type of collocation. Collocation patterns are described by means of categorical and functional-relational labels. There are 13 types of collocation in the BED database:

- N(S)+V noun (as grammatical subject) + verb: *päike paistab/tõuseb/loojub* ‘sun shines/rises/sets’;
- N(O)+V noun (as grammatical object) + verb: *arvutit sisse lülitama* ‘switch on the computer’;
- N(A)+V noun (as adverbial modifier) + verb: *kinnisvarasse investeerima* ‘invest in property’;
- Adj+V adjective + verb: *määravaks osutama* ‘prove decisive’;

- Adv+V adverb + verb: *kiiresti jooksmas* 'run fast';
- N+N noun + noun: *ekspertide hinnang/arvamus* 'assessment/opinion of experts';
- Adj+N adjective + noun: *hea/halb eeskuju* 'good/bad example';
- Num+N numeral + noun: *sada meetrit/kilo* 'hundred meters/kilograms';
- Adv+N adverb + noun: *kergesti süttiv* 'easily flammable';
- Adv+Adv adverb + adverb: *väga aeglaselt* 'very slowly';
- Prep+N preposition + noun: *enne/pärast jõule* 'before/after Christmas';
- N+Post noun + postposition: *interneti/raadio kaudu* 'on television/ radio'.

Collocations of the same type are divided into semantic sets and presented explicitly as separate bundles. Figure 2 shows the entry for *arve* 'invoice, account' in the printed version of the BED.

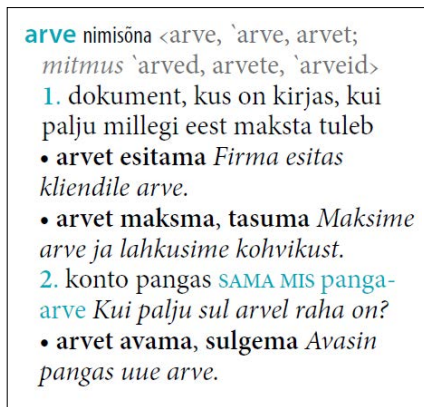


Figure 2: The BED entry for the noun *arve* 'invoice, account' in the printed version.

## 1.2 Extra materials

The BED also contains approx. 400 illustrations. These are single illustrations with legends, structural illustrations (particular objects are highlighted by means of arrows), functional illustrations (mostly for adpositions), scenic illustrations (mostly for phrasal verbs) and nomenclatory illustrations (see figure 3).



Figure 3: Nomenclatory illustration for the entry *maja* 'house'.

Besides illustrations, the dictionary has a centre section of 16 study pages (including instructions for producing numbers, time and dates, writing letters and emails, punctuation marks, common abbreviations, useful phrases) and 17 picture pages (e.g. on insects, animals, flowers and transportation). In the appendix, a list of countries, people and languages is given, as well as grammatical tables. Grammatical tables show how to decline and conjugate words, also they give guidance for producing all other word forms when moving on from the basic forms given in the dictionary.

## 2 The BED as an online-dictionary

The online version of the BED has some innovative features, which are implemented in the Estonian lexicography for the first time. Figure 4 illustrates the interface of the online version of BED.



Figure 4: Online BED entry for the noun *hiir* 'mouse'.

Pictures are aligned with particular word meanings. If the picture is topically related to one of the special picture pages provided in the dictionary as extra material (e.g. *hiir* 1. 'mouse' as related to *animals*), then these pictures are linked together. Otherwise it is possible to enlarge the picture.

Green musical note symbols indicate that there are sound recordings (mp3 audio files) linked to particular morphological forms. The audio files were pre-recorded.

The contents of the entire dictionary have been morphologically analysed. As a result, users can click on any word in a definition or example to find the entry for that word. And, vice versa, it is possible to type into the search box a word in any form (previous dictionaries allowed for searching only on the basis of lemma), and the lemma entry will be provided.

### 3 The BED as a basis for the Dictionary of the Estonian Sign Language

The BED database was used to compile the online Estonian Sign Language – Estonian Language dictionary<sup>3</sup>. Figure 5 illustrates the interface of the dictionary. There are approx. 6,700 video files. For every sign, the BED database contains information on the initial hand form, the location where the sign is articulated (face, lips, cheek, chest, neutral space etc.) and the movement with which the sign is formed. Based on these three parameters, it is possible – for the first time in Estonian lexicography – to search for a certain sign by choosing the hand form, the location, or the movement of the sign. This enables the deaf dictionary user to find the Estonian equivalent for a sign. The interface allows for searching in the opposite direction as well, making it possible for the non-deaf to learn Estonian Sign Language.

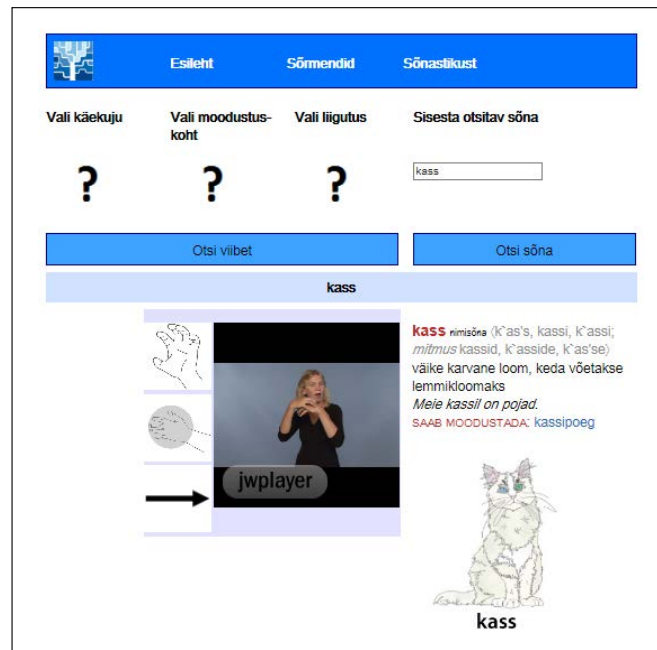


Figure 5: Online entry for the noun *kass* ‘cat’ in the Estonian Sign Language – Estonian Language online-dictionary.

### 4 The BED as a lexical resource in the Dictionary Writing System EELex

The BED was compiled in the web-based dictionary writing system EELex<sup>4</sup> (Jürviste et al. 2011). Nearly 50 dictionaries of different types (monolingual and bilingual, general and learner’s dictionaries, etc.)

3 <http://www.eki.ee/dict/viipekeel/> [01/04/2014]

4 <http://eelex.eki.ee/> [01/04/2014]

with a standard XML mark-up make EELex a multi-purpose lexicographical database. XML-based compilation allows for the generation of different outputs: for example, specialised dictionaries based on partial database output (Kallas, Langemets 2012). There are two options for the automatic generation of specialised dictionaries: reorganising the preview (and layout) of the existing dictionary articles, or generating a new dictionary database (i.e. to clone only a part of the source database).

The function of the article preview generator makes it possible to modify the preview, i.e. to set a character, text or line break between, in front of or after a specific element or group of elements, to show or hide specific elements in the article editing preview, to assign a condition for displaying a specific element (according to the value of the attribute or neighbouring elements) or to assign a hyperlink to an element. So, by specifying elements in the print preview, it is possible to get output consisting of only those elements that are specified by the user.

The same result may be achieved by the customization of the regular XML query. It is possible to select particular elements to be displayed instead of the whole content of the dictionary article. Table 1 shows a dictionary-like extract from the BED database consisting only of the following elements: lemma, collocational patterns and usage example.

<u>abielu</u>	abielu sõlmima	Noored sõlmisid abielu kirikus.
<u>abielu</u>	abielu lahutama	Mari ja Martin lahutavad abielu.
<u>abielu</u>	abielus olema	Kas ta on abielus või vallaline? Nad on juba 20 aastat abielus.
<u>aeg</u>	lähemal ajal viimasel ajal	Olen viimasel ajal kuidagi väsinud.
<u>ahi</u>	ahju kütma	Peremees kütab ahju.

**Table 1: An example of the collocations extracted from the BED database.**

In this way, it is possible to reuse the BED database in order to generate specialised dictionaries (e.g. a dictionary of government and collocations).

## 5 Conclusion

The XML-based compilation makes the BED database a useful lexical resource, which can be used in different ways for development materials meant for the teaching and learning of the Estonian language as a second or a foreign language. The dictionary is special in many ways. It is the first monolingual dictionary meant for learners of Estonian at the elementary and intermediate levels (previously there were bilingual dictionaries). Government and collocation patterns were analysed and selected using the Sketch Engine corpus query system. The online version allows learners to listen to the pronunciation of words. In addition, the morphological analysis implemented in the BED make it a very innovative and user-friendly dictionary.



The first online Estonian Sign Language – Estonian Language dictionary has also been compiled. This dictionary is unique in that it enables the deaf dictionary user to find the Estonian equivalent for a sign, and not only for a word.

In future, it may be possible to convert the dictionary web page into a language-learning portal by combining the dictionary with other resources (corpora, different specialised dictionaries etc.).

## 6 References

- Jürviste, M., Kallas, J., Langemets, M., Tuulik, M., Viks, Ü. (2011). Extending the functions of the EELEX dictionary writing system using the example of the Basic Estonian Dictionary. In I. Kozem, K. Kozem (eds.) *eLexicography in the 21st Century: New Applications for New Users*, Proceedings of eLex 2011, Bled, 10-12 November 2011. Ljubljana: Trojina, Institute for Applied Slovenian Studies, pp. 106-112.
- Kallas, J., Langemets, M. (2012). Automatic Generation of Specialized Dictionaries Using the Dictionary Writing System EELEX. In A. Tavast, K. Muischnek, M. Koit (eds.) *Human Language Technologies – The Baltic Perspective*, Proceedings of the Fifth International Conference Baltic HLT 2012. IOS Press, (Frontiers in Artificial Intelligence and Applications), pp. 103-110.
- Kallas, J. (2013). *Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias*. PhD thesis. Tallinn: Tallinna Ülikool.
- Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the XI Euralex International Congress*. Lorient: Université de Bretagne Sud, pp. 105-116.

### **Acknowledgements**

The project was supported by the National Programme for Estonian Language and Cultural Memory (2009–2013), as well as the National Programme for Estonian Language Technology (2011–2017).



# Die fremdsprachige Produktionssituation im Fokus eines onomasiologisch konzeptuell orientierten, zweisprachig-bilateralen Wörterbuches für das Sprachenpaar Deutsch - Spanisch: Theoretische und methodologische Grundlagen von DICONALE

Meike Meliss  
Universidad de Santiago de Compostela-Spanien  
meike.meliss@usc.es

## Abstract

Der Beitrag beschäftigt sich mit den verschiedenen Such-, Auffindungs- und Auswahlprozessen, die für die fremdsprachige Produktion notwendig sind und von DICONALE-*online*, einem onomasiologisch-konzeptuell ausgerichteten, zweisprachig-bilateral konzipierten Verbwörterbuch der spanischen und deutschen Gegenwartssprache, besonders berücksichtigt werden. Der Ausgangspunkt von DICONALE ist ein unbefriedigendes Informationsangebot in den bestehenden ein- und zweisprachigen Lernerwörterbüchern für den L2-*output* und bestätigt das Projektteam in der Notwendigkeit, ein neuartiges benutzer- und situationsdefiniertes *online*-Nachschlagewerk zu erstellen. Zwei Bezugsrahmen bilden die Grundlage für einen komplexen, konzeptuell und framegeleiteten Zugriffspfad, der dem Benutzer bei der Suche und Auswahl von Ausdrucksmöglichkeiten und der adäquaten Anwendung behilflich sein soll. Das Novum dieses Wörterbuchprojekts besteht hauptsächlich darin, eine onomasiologisch-konzeptuelle Perspektive für den fremdsprachigen Produktionsprozess nutzbar zu machen und mit einem semasiologischen Zugriff zu verbinden, durch den es möglich ist, die inter- und intralingualen Unterschiede zwischen den Lexemen eines lexikalisch-semantischen (Sub)Paradigmas hervorzuheben.

Ziel des Beitrages ist es daher, den Ausgangspunkt, sowie die theoretischen und methodologischen Grundlagen von DICONALE-*online* unter der speziellen Perspektive der Benutzer- und Situationsorientiertheit zur Diskussion zu stellen, die einzelnen Zugriffspfade für den Such- und Auffindungsprozess vorzustellen und das Angebot zur Auswahl und zum adäquaten Gebrauch aus inter- und intralingualer Perspektive zu präsentieren.

**Keywords:** Lernerlexikographie; Argumentstrukturgrammatik; kontrastive Linguistik

# 1 Einleitung

DICONALE-*online* ist ein Forschungsprojekt<sup>1</sup> zur Erstellung eines onomasiologisch-konzeptuell orientierten, zweisprachig-bilateralen Wörterbuches deutscher und spanischer Verben<sup>2</sup> im Kontrast, welches sich an fortgeschrittene Lernende des Deutschen bzw. Spanischen als Fremdsprache (DaF und Ele) richtet und besonders die fremdsprachige freie Produktion anvisiert. Es handelt sich somit um ein Wörterbuchprojekt, welches sowohl den Benutzerkreis als auch die Benutzersituation klar vordefiniert und damit sogleich deutlich einschränkt. Die Struktur des zukünftigen Wörterbuches, sowie die Auswahl des Informationsangebotes und der Zugang stehen in direkter Verbindung mit der vordefinierten Benutzersituation.

Trotz zahlreicher zweisprachiger Wörterbücher für das besagte Sprachenpaar in *print*- und *online*-Format besteht nach unserer Auffassung die Notwendigkeit, ein neuartiges, konzeptuell orientiertes *online*-Verbwörterbuch zu konzipieren, um damit gerade die Aspekte in den Mittelpunkt zu stellen, die in der herkömmlichen ein- und zweisprachigen Lernerlexikographie für das Sprachenpaar Deutsch-Spanisch bis jetzt zu kurz gekommen sind. Diese sind u.a. die konsequente Berücksichtigung des situationsbedingten Such- und Auffindungsprozesses eines geeigneten Lexems für den fremdsprachigen *output*, sowie eine benutzerorientierte Darbietung der Information für den Auswahlprozess und den anschließenden situationsgerechten Gebrauch in der jeweiligen Fremdsprache. Die sich daraus ergebenden notwendigen Suchsequenzen rechtfertigen eine primär onomasiologisch-konzeptuell orientierte Zugriffsstruktur. Diese Prämissen stellen uns vor die Herausforderung, einen Benutzerkreis, der prinzipiell an alphabetisch-semasiologisch konzipierte Wörterbücher gewöhnt ist, durch ein geeignetes Suchleitsystem zu den gewünschten Resultaten zu führen. Das Novum dieses Wörterbuchprojekts besteht daher hauptsächlich darin, eine onomasiologisch-konzeptuelle Perspektive für den fremdsprachigen Produktionsprozess nutzbar zu machen und mit einem semasiologischen Zugriff zu verbinden, durch den es möglich ist, die inter- und intralingualen Unterschiede zwischen den Lexemen eines lexikalisch-semantischen (Sub)Paradigmas hervorzuheben.

Ziel des Beitrages ist es daher, den Ausgangspunkt, sowie die theoretischen und methodologischen Grundlagen von DICONALE-*online* unter der speziellen Perspektive der Benutzer- und Situationsorientiertheit zu präsentieren (Kapitel 2). In Kapitel 3 sollen im Einzelnen die Zugriffspfade für den Such- und Auffindungsprozess vorgestellt, in Kapitel 4 das Angebot zur Auswahl und zum Gebrauch präsentiert und abschließend in Kapitel 5 ein kurzer Ausblick angeboten werden. Zur Veranschauli-

---

1 DICONALE (= *D*iccionario *c*onceptual del *a*lemán y del *e*spañol): Es handelt sich um ein von dem spanischen Ministerium gefördertes Forschungsprojekt (MINECO – FEDER: FFI2012-32658: 2013-2015), das außerdem in Verbindung mit dem lexikographischen Netzwerk RELEX (Xunta de Galicia: CN2012/290) entwickelt wird. Das Forschungsteam besteht aus Mitgliedern verschiedener spanischer, deutscher und portugiesischer Universitäten und Forschungseinrichtungen und wird von der Autorin dieses Beitrages an der Universidad de Santiago de Compostela (Spanien) geleitet.

2 Neben der Hauptlemmaliste werden in einer sekundären Lemmaliste auch deverbale Nomen, Adjektive und Adverbien und komplexe, mehrteilige Lexeme aufgenommen.

chung werden einige ausgewählte Beispiele aus dem Bereich der Verben der AUDITION herangezogen.

## 2 Ausgangspunkt und theoretisch-methodologische Grundlagen

Als Ausgangspunkt unserer Überlegungen stützen wir uns auf Untersuchungen, die aufzeigen konnten, dass die lexikographischen Ressourcen, die normalerweise im DaF- und Ele-Bereich zur Verfügung stehen, bis jetzt zu wenig Wert auf fremdsprachige Produktionssituationen für die fortgeschrittene Lernerebene (ab B2) gelegt haben. Dies gilt gleichermaßen für die zweisprachigen, als auch für die einsprachigen Lernerwörterbücher des besagten Sprachenpaars (Meliss 2013a, 2013b, 2014a, 2014b, 2014c).

Nach der Untersuchung der gängigsten zweisprachigen Großwörterbücher Spanisch-Deutsch (GWB-sp/dt)<sup>3</sup> in *print* und *online*-Format lässt sich zusammenfassen, dass in Verbindung mit der hier im Fokus stehenden fremdsprachigen Produktionssituation dem Benutzer zu wenig Information zu dem syntagmatischen Kombinationspotenzial der möglichen Entsprechungen in der fremdsprachigen Zielsprache angeboten und die semantisch orientierte Disambiguierung bei Entsprechungsvielfalt durch Angabe von paradigmatischen Sinnrelationen zu wenig für einen angebrachten Gebrauch genutzt werden (Fuentes Morán 1997, Haensch&Omeñaca <sup>2</sup>2004, Hausmann 1991, Meliss 2013a, 2013b, 2014a, 2014b, Model 2010).<sup>4</sup> Der Such- und Auswahlprozess zur geeigneten Benennung wird außerdem von der Muttersprache geleitet und führt daher in vielen Fällen nicht zu der zielsprachigen Ausdrucksvarietät (Meliss 2014c).

Einsprachige Lernerwörterbücher (LWB) für DaF<sup>5</sup> und Ele<sup>6</sup> weisen zwar in den meisten Fällen ein relativ hohes Informationsangebot bezüglich des Kombinationspotenzials der einzelnen Lexeme durch Angabe von Strukturformeln auf (Dentschewa 2006, Engelberg 2010; Meliss 2013b, 2014b) und bieten dem Benutzer somit nützliche Information zum korrekten Gebrauch.<sup>7</sup> Der klassische, alphabetisch orientierte Zugang und die semasiologische Zugriffsperspektive favorisieren hingegen nicht die Auffindung eines unbekanntes Lexems zur Benennung eines bestimmten Konzepts im fremdsprachigen

---

3 LHWB und LHWBe, LEO, Pons: Das Sprachenportal, SGIWBe (Slaby/Grossmann/Illig);

4 Dies steht im Einklang mit Untersuchungen zu zweisprachigen WB anderer Sprachen (Engelberg/Lemnitzner 42009, Herbst/Klotz 2003). Siehe dazu auch: Abel 2008.

5 LGWB-DaF (Götz et al.) & online-Version, GWB-DaF (Kemcke), PGWB-DaF und Pons-DaF-online, Duden-DaF, Wahrig-DaF und online-Version;

6 DS (Diccionario Salamanca), DA (Diccionario Alcalá) und online-Version, SM-Clave-online;

7 Dieses Informationsangebot ist besonders ausgeprägt in den DaF-Lernerwörterbüchern von Langenscheidt (LGWB-DaF: Götz et al. 32010) und Kempcke (1999). So ist z.B. die mikrostrukturelle Information zu den Verballemmata von Langenscheidt DaF von einer syntagmatisch orientierten Grundstrukturierung geprägt (Engelberg 2010: 116).

*output*-Prozess und zieht auch keine spezifische Hilfestellung zur Auswahl aus einer Vielfalt von bedeutungsähnlichen Ausdrucksmöglichkeiten in Betracht.<sup>8</sup>

Verschiedene andere lexikographische Ressourcen wie z.B. syntagmatische und paradigmatische Spezialwörterbücher<sup>9</sup> können zwar die genannten Informationslücken und Zugriffsblokaden für die fremdsprachige *output*-Situation teilweise beheben, stehen aber im DaF- und Ele-Bereich entweder nicht zur Verfügung, oder genießen im Falle von frei verfügbaren *online*-Ressourcen, zu denen der Benutzer durch externe Links der gängigen ein- und zweisprachigen Wörterbuchportale zwar fast automatisch gelangt<sup>10</sup>, nicht den erwünschten Bekanntheitsgrad (Meliss 2013b, 2014b). Die begrenzten Sprachkenntnisse und die mangelhafte lexikographische Vorbildung des hier anvisierten prototypischen Benutzers führt außerdem zu einer wenig optimierten Nutzung des inzwischen sehr breiten Informationsangebots bei gleichzeitiger Gefahr des „Sich Verirrens“ („lost in hyperspace“: Storrer 2010).

Das Forschungsprojekt DICONALE hat sich daher zum Ziel gesetzt, die unterschiedlichen Such-, Auffindungs- und Auswahlprozesse, die in der freien L2-Sprachproduktion durchlaufen werden müssen, konsequent zu berücksichtigen. Im Mittelpunkt der Überlegungen stehen daher folgende Problemkomplexe: (i) die Ausdrucks- bzw. Benennungssuche, (ii) die Ausdrucksauswahl aus der Vielfalt und (iii) der Gebrauch unter Berücksichtigung kontrastiv relevanter Divergenzen. Eine sich daraus ableitende onomasiologisch-konzeptuelle Zugriffsstruktur und ein modulares Informationsangebot bilden die Grundlagen für die Makro- und Mikrostruktur von DICONALE.

Dementsprechend setzt sich die MAKROSTRUKTUR in einer ersten Arbeitsphase aus 10 konzeptuellen Feldern<sup>11</sup> zusammen, die in weitere konzeptuelle Subfelder „zweiten und dritten Grades“ mittels einer immer feiner differenzierenden Konzeptualisierung gegliedert werden. Diese Subfelder bilden die Grundlage für die lexikalisch-semanticen (Sub)Paradigmen (SPLs), denen durch die auf dieser Stufe erfolgte Lesartdifferenzierung einzelne Lexeme in beiden Sprachen zugeordnet werden können. Das mehrstufige Beschreibungsmodell basiert auf unterschiedlichen lexikologischen Parametern die in 4 Modulen erfasst werden (Meliss & Sánchez Hernández 2014, González Ribao & Meliss 2014). Zur Lesartdisambiguierung wird besonderer Wert auf die Beschreibung der Bedeutung und bestimmter kombinatorischer Merkmale gelegt. Die Argumentstrukturbeschreibung zusammen mit entsprechender Information zu den morpho-syntaktischen, funktionalen und semantisch-kategoriellen Füllungen und Kollokatoren<sup>12</sup> stehen hier – neben den paradigmatischen Sinnrelationen – im Mittel-

---

8 Ausnahmen sind für das Deutsche der Teil 2 des Wörterbuches von Kempcke (1999) und für das Spanische der onomasiologisch-konzeptuell angelegte Teil 2 des zukünftigen „Diccionario de Coruña“ (Porto Dapena et al. 2008).

9 Insbesondere sind hier die Konstruktionswörterbücher (ValenzWB, KollokationsWB etc.) und die SynonymWB zu nennen.

10 So gelangt der Benutzer über das Pons-Sprachenportal zu einsprachigen Wörterbüchern der deutschen Sprache, wie z.B. DWDS und zu einigen Spezialwörterbüchern. Über CanooNet gelangt der etwas geschulte Benutzer über TheFreeDictionary zu der online-Version von Langenscheidts DaF-WB (LGWB-DaF-online). Allerdings muss festgehalten werden, dass kaum mit syntagmatisch-orientierten Ressourcen verlinkt wird.

11 Es werden z.Z. Felder der Wahrnehmung, Kommunikation, Zwischenmenschlichen Beziehung, Kognition, Transfer, Konsum, Fortbewegung und der Existenz untersucht.

12 Siehe dazu u.a.: Engelberg 2014a, Engelberg 2014b, Engelberg et al. 2012.

punkt der Module 2 und 3 (*Abbildung 1*).<sup>13</sup> Zu den empirischen Grundlagen in Verbindung mit der problematischen Erstellung von vergleichbaren Korpora für beide Sprachen soll auf die Studie von González Ribao (2014) verwiesen werden. Bezüglich des Formats haben uns jüngste Studien zur Benutzerforschung im ein- und zweisprachigen Kontext (Domínguez et al. 2013, Klosa et al. 2011)<sup>14</sup> in der Notwendigkeit bestätigt, ein lexicographisches Werk zu schaffen, welches über einen freien Internetzugang, ein schnell zugängliches, modular organisiertes, benutzerfreundliches und -adaptives, intern und extern verlinktes Informationsangebot offeriert<sup>15</sup>, welches unserem DaF- und Ele-Benutzerkreis auch sprachlich und metasprachlich entgegen kommt.

M 1	<p><b>Lesart übergreifende, feldrelevante allgemeine Information:</b></p> <ul style="list-style-type: none"> <li>▶ Ausdrucksseite: Wortart, Konjugationstyp, suprasegmentale Merkmale, morph. Aufbau (trennbar-untrennbar);</li> <li>▶ Inhaltsseite: Bedeutung; semantische Komponenten, Verbalcharakter: Aktionsart, Aspektualität;</li> <li>▶ Wortbildung: lesartübergreifend: feldrelevante Formen (Auswahl)</li> <li>▶ Szenarien – semantische Rollen</li> <li>▶ externe Links</li> </ul>
M 2	<p><b>Feldrelevante Lesarten (LAfr):</b></p> <ul style="list-style-type: none"> <li>▶ <b>Bedeutungserklärung</b> mit Hinweis auf paradigmatische <b>Sinnrelationen</b> und distinktive <b>Bedeutungsmerkmale</b>:</li> <li>▶ <b>Paradigmatische Sinnrelationen</b>: innerhalb &amp; außerhalb des Paradigmas</li> <li>▶ <b>Argumentstrukturmuster (ASTM) jeder Lesart</b>: ▶ Varianten</li> <li>▶ Register</li> <li>▶ Illustrative <b>Belegbeispiel</b></li> <li>▶ <b>Frequenz</b> jeder Lesart und Variante</li> </ul>
M 3	<p><b>Kombinatorik jeder Variante &amp; Entsprechungsangebot i.d. Kontaktsprache:</b></p> <p>Grundlage: <i>tertium comparationis</i>: Bedeutungsstruktur + ASTM</p> <ul style="list-style-type: none"> <li>▶ <b>Satzbauplan (SBP)</b></li> <li>▶ <b>Spezifizierung/Füllung der Argumente (A)</b>: syntaktisch-funktional + semantisch kategorial + Kollokatoren</li> <li>▶ <b>Frequenz</b></li> <li>▶ häufige Zirkunstanten, Belegbeispiele, Kommentare ...</li> </ul>
M 4	<p><b>Weitere gramm. Information:</b></p> <p><b>Gebrauch und Frequenz</b>: Passiv, passivische Ersatzformen, Modus, synt. Konversion (Auswahl: deverbale Nomen, Adjektive, Adverbien...) etc.</p>

**Abbildung 1: Die 4 Beschreibungsmodulare von DICONALE.**

### 3 Suchen und Finden

Der übliche und bekannteste Zugriff auf ein Wörterbuch erfolgt zwar aus einer semasiologisch-alphabetisch geleiteten Perspektive, für freie fremdsprachige Produktionszwecke ist diese Perspektive jedoch nur geeignet, wenn man das sprachliche Ausdrucksmittel schon kennt, bzw. schon ausgewählt hat und das WB nur noch zwecks Überprüfung bestimmter lexikologischer Parameter zur korrekten Anwendung konsultieren möchte. Wenn man aber noch kein sprachliches Ausdrucksmittel ausgewählt hat, weil man das Ausdrucksmittel für die Ausdrucksbedürfnisse gar nicht kennt, dann kann

13 Siehe dazu auch verschiedene sprachvergleichende Studien in Verbindung mit lexikalisch-semantischen Paradigmen in Engelberg et al. (eds.) (2014).

14 Im Rahmen von DICONALE ist eine breit angelegte Wörterbuchbenutzerumfrage entwickelt worden, die unter <https://www.usc.es/gl/proxectos/diconale/aleman/enquisa.html> [11.04.2014] abgerufen werden kann. Sie erfragt die Benutzergewohnheiten und Erwartungen im Bereich DaF und Ele an universitären und nicht universitären (Gymnasien, Sprachschulen, Volkshochschule etc.) Lehrinstitutionen in Spanien, Deutschland und Portugal. Die Umfrageauswertungen liegen Ende 2014 vor.

15 Siehe dazu auch Haß & Schmitz (2010), Klosa (ed.) (2008), Mann (2010), Müller Spitzer & Engelberg (2013), Storrer (2010) etc. und spezifisch zu dem Mehrwert der Internetlexikographie Engelberg & Lemnitzer 42009: 220 und Tarp 2012: 253).

man in einem einsprachig semasiologisch konzipierten WB nicht wirklich fündig werden (González Ribao & Proost 2014; Proost 2007). Daher soll die onomasiologisch-konzeptuelle Zugriffsstruktur<sup>16</sup> besonders für den fremdsprachigen Produktionskontext genutzt werden, wobei die konzeptuelle Referenz zusammen mit den jeweiligen verbalen Szenarien die zwei Hauptbezugsrahmen bilden und gleichzeitig das *tertium comparationis* für den Sprachvergleich stellen. Der erste Bezugsrahmen (BR1) bezieht sich auf die konzeptuellen Referenzen mit unterschiedlichem Spezifizierungsgrad, während der zweite Bezugsrahmen (BR2) sich an die Beschreibung der verbalen Szenarien annähert und die am verbalen Geschehen beteiligten Rollen beschreibt<sup>17</sup>. Beide Bezugsrahmen bilden die Grundlage des Zugriffspfades 1 mit den entsprechenden Spezifizierungen (1a-1c). Zusätzlich wird auch ein klassisch alphabetisch geleiteter Zugriffspfad 2 angeboten, der aber nicht zu der erwarteten semasiologischen Informationsperspektive führt, sondern die Benutzenden über eine Lemmaliste zu dem Zugriffspfad 1 zurückleitet. Die verschiedenen Zugriffspfade werden in *Abbildung 2* visualisiert und sollen im Folgenden genauer beschrieben werden. Als Beispiel dient uns das konzeptuelle Subfeld ersten Grades: AUDITIVE WAHRNEHMUNG (SFK<sup>1</sup>), welches dem Wahrnehmungsfeld untergeordnet ist (FK WAHRNEHMUNG). Der Benutzer greift auf das Wörterbuch aus einer konzeptuellen Suchperspektive zu, die durch anfängliche Benutzereinstellung in spanischer oder deutscher Sprache realisiert werden kann. Ein detailliertes Optionsleitsystem ermöglicht die Suche, Auffindung, Auswahl und den anschließenden Gebrauch der Ausdrucksform, die am besten dem kommunikativen Ausdrucksbedürfnis der jeweiligen Situation entspricht.

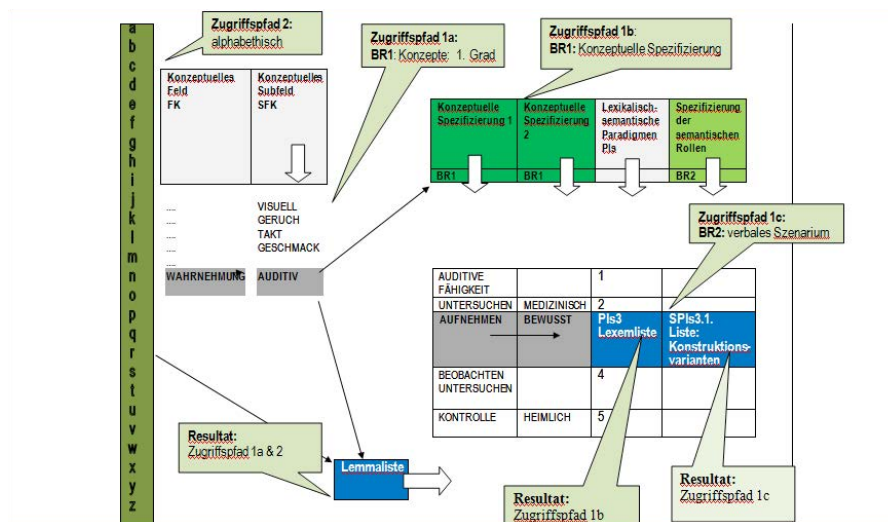


Abbildung 2: Unterschiedliche Zugriffspfade für DICONALE

16 Die zur Verfügung stehenden klassischen konzeptuell-onomasiologisch geordneten Wörterbücher und Nachschlagewerke wie Wehrle & Eggers und Dornseiff oder Casares für das Spanische sind wegen ihrer Komplexität für den hier anvisierten Benutzerkreis ungeeignet. Neuere Studien, wie z.B. zu den Kommunikationsverben (Harras et al.) mit online-Zugang (Proost) verfolgen einen sehr komplexen Bezugsrahmen, der für unsere Benutzersituation ebenfalls nicht geeignet ist.

17 Siehe dazu Studien zu FrameNet: Boas 2013, Boas & Dux 2013, FrameNet Spanish: Subirats 2009;



Der Zugriffspfad 1 führt den Benutzer mithilfe des entsprechenden Bezugsrahmens BR1 (WAHRNEHMUNG & AUDITION) von einer allgemeinen konzeptuellen Referenz zu einer Lemmaliste, in der neben den verbalen Lemmata in Simplexform und den entsprechenden affigierten Formen auch einige mehrteilige Lemmata und deverbale Formen aufgeführt werden (Abbildung 3).

Konzeptuelle Feld: WAHRNEHMUNG & AUDITION und mögliche Lexikalisierungen		Lemma	Spanisch	
<b>Deutsch</b>	<b>Link</b>	Simplexform	<b>oir, escuchar, auscultar ...</b>	
hören, horchen, lauschen ...		Affigierungen	...	
zu/an/ab/hin/mit ... –hören		Komplexe Strukturen	...	
sich an/um/herum/ver/auf ... –hören		Deverbale Ableitungen	el oído ..	
zu/an/ab/hin/mit/auf! ... –horchen			las escuchas ...	
sich (her)um ... –horchen				
belauschen				
vernehmen				
...				
die Ohren spitzen ...				
...				
der Hörer, Zuhörer, Lauscher ...				
die Anhörung ...				

Abbildung 3: Lemmalisten in beiden Sprachen zu dem BR1: WAHRNEHMUNG & AUDITION.

Diese Lemmalisten sind allerdings nur von Nutzen, wenn die angeführten internen Links den Benutzer zu den verschiedenen Lesarten führen und die Vernetzung innerhalb des entsprechenden konzeptuellen Feldes und den dazu gehörigen lexikalisch-semanticen Paradigmen abgerufen werden kann (Abbildung 4: Beispiel *abhören*).

Bezugsrahmen 1: konzeptuelle Spezifizierung					Bezugsrahmen 2: Szenarium und Rollen				
Lemma	Konzeptuelles Subfeld SFK <sub>1</sub>	Konzeptuelle Subfelder SFK <sub>2</sub>	Lexeme: Lesarten	Kompetenzbeispiel	Bedeutungsähnlichkeit	Lexikalisch-semantiche Paradigmen Pls	Lexikalisch-semantiche Subparadigmen SPis		
<i>abhören</i>	WAHRNEHMUNG & AUDITION	FÄHIGKEIT & AUDITIV	SFK1	-	...	-	Pls1		
		MEDIZINISCH & UNTERSUCHUNG	SFK2	<i>abhören1</i>	...	Med. untersuchen	Pls2		
		AUFNAHME & BEWUSST	SFK3	-	...	-	Pls3	A1 A2	A1 A3   A1 A3   ...
		BEOBACHTUNG / UNTERSUCHUNG	SFK4	<i>abhören2</i>	...	genau anhören / überprüfen	Pls4		
		KONTROLLE & HEIMLICH	SFK5	<i>abhören3</i>	...	kontrollieren / beobachten / bespitzeln	Pls5		

Abbildung 4: Lemma und Linkangebot zu Lesarten in Verbindung mit dem BR1 und BR2.

Auf dem Zugriffspfad 1 kann der Benutzer auch direkt – geleitet durch weitere konzeptuelle Spezifizierungen des BR1(1a-1c) - zu möglichen Benennungseinheiten gelangen, die mit Bezug auf das entsprechende lexikalisch-semantiche Paradigma in Lexemlisten mit Kompetenzbeispielen angeordnet werden. So wird z.B. der Benutzer durch den BR1 und eine Spezifizierung durch AUFNEHMEN & BEWUSST zu dem lexikalisch-semantiche Paradigma Pls3: „zuhören“ geführt. Die Ausdrucksmittel, die dem BR1 entsprechen, sind für das Deutsche: (*sich*) *anhören1*, *zuhören1*, *hören2*, *horchen1*, *lauschen1* und für das Spanische: *escuchar 2* y *oir2*. Eine Auswahl von Kompetenzbeispielen illustrieren die Be-

deutung durch den Kontext (*Abbildung 5*: Suchergebnisse bezüglich BR1). Verben wie Dt.: *abhören, abhören, belauschen* und Sp.: *auscultar* etc., werden in diesem lexikalisch-semantic Paradigma konsequenterweise nicht aufgeführt, da sie in keiner ihrer Lesarten zu dem besagten Paradigma gehören. Der Bezugsrahmen 2 ermöglicht die Spezifizierung des verbalen Szenariums durch die beteiligten Rollen und Muster. Die zum lexikalisch-semantic Paradigma Pls3 gehörigen Rollen sind u.a. der Hörer (R1), die Äußerung, die wahrgenommen wird (R2), ein Lebewesen, das als Geräuschquelle wahrgenommen wird (R3), eine nicht belebte Entität, die als Geräuschquelle wahrgenommen wird (R4) und das wahrgenommene Geräusch selbst (R5). Der BR2 stellt einen wichtigen Auswahlfaktor für den Benutzer dar, denn nicht alle Lexeme in einem lexikalisch-semantic Paradigma deuten auf dasselbe Szenarium hin. Aus der Kookkurrenzanalyse zu *lauschen* und *zuhören* (CCDB: Cyril) lässt sich z.B. schließen, dass *lauschen* häufiger in Verbindung mit dem Szenarium: “jemand (R1: Hörer) nimmt ein Geräusch wahr (R5: Klang, Gesang, Geräusch...)” auftritt, während *zuhören* andere Szenen vorzieht: (z.B.: R1: Hörer & R3: belebte Geräuschquelle: Leute ...). Je nach Beteiligung der Rollen liegt das eine oder andere Argumentstrukturmuster vor, welches in einem Menu mit den entsprechenden Vorgaben selektiert werden kann. Wenn der Benutzer Ausdrucksformen für “jemand hört aufmerksam zu, was jemand sagt” sucht, bezieht es sich auf ein ganz bestimmtes Szenarium, in dem die Rollen Hörer (R1) und Äußerung (R2) verbalisiert werden sollen. Die Argumentstruktur | jemand A1 V etwas A2 | bildet daher den BR2, durch den nur die Konstruktionsvarianten einer Lesart selektiert werden, die dieses Szenarium realisieren können. In unserer L2-*output*-Benutzersituation kann davon ausgegangen werden, dass der Benutzer genau und nur das Szenarium sucht, welches zu seinem Ausdruckswunsch passt. Nach der Auswahl des entsprechenden Musters wird dem Benutzer eine Liste von Konstruktionsvarianten eines lexikalisch-semantic Paradigmas zusammen mit Kompetenzbeispielen für die ausgewählte Sprache (Beispiele 1-5<sub>dt</sub>) angeboten, die beide Bezugsrahmen miteinander teilen (*Abbildung 5*).<sup>18</sup> Nach der Auswahl der einen oder anderen Konstruktionsvariante kann der Benutzer in weiteren Schritten detaillierte Information zu den 4 Beschreibungsmodulen erhalten (*Abbildung 1*).

- (1.1<sub>dt</sub>) Sie **hören sich** die Probleme **an**, die den Kindern auf den Nägeln brennen [...]. (R97/SEP.72386 Frankf. Rundschau, 15.09.1997, S. 4).
- (1.2<sub>dt</sub>) Vier Jahre hätten die Vorbereitungen gedauert, man habe sich in anderen Städten vergleichbare Ansagen **angehört** und sich jetzt für diese Lösung entschieden. (M13/JUL.01277 Mannh. Morgen, 04.07.2013, S. 20).
- (1.3<sub>dt</sub>) Auch die Mitarbeiterinnen im Bürgercenter werden aufatmen, denn sie mussten sich in den vergangenen drei Wochen so manche nicht immer freundlich vorgetragene Beschwerde **anhören**. (BRZ10/NOV.09775 Braunsch. Z., 19.11.2010).
- (2.1<sub>dt</sub>) Sie **hörte** die Aussagen der verängstigten Kinder. (RHZ05/JUN.16629 RZ, 15.06.2005).
- (2.2<sub>dt</sub>) Frei nach dem Grundsatz „Erst mal **hören**, **was** die Zeugen wissen“, verlegte sich das Muskelpaket und Vater von drei Kindern aufs Schweigen. (RHZ03/JUL.12417 RZ, 16.07.2003).

18 Zur Behandlung von Argumentstrukturen und Lesartdisambiguierung siehe u.a. Engelberg (2010) für das Deutsche und Porto Dapena et al. (2008) für das Spanische.

- (3.1<sub>dt</sub>) Man setzt sich, **horcht** der Feldpredigt und genießt die Sonne. (A00/JUN.38103 St. Galler Tagblatt, 02.06.2000)
- (3.2<sub>dt</sub>) Er **horchte** auf die Worte der Reisenden, die von Bord gingen, und wenn sie deutsch sprachen, redete er sie an. (P97/MAR.11400 Die Presse, 22.03.1997).
- (3.3<sub>dt</sub>) Vor kurzem lud der Tischtennisverein Züllig zur 15. Generalversammlung. Zahlreiche Mitglieder folgten der Einladung und **horchten** den erfreulichen Neuigkeiten. (A99/SEP.68281 St. Galler Tagblatt, 30.09.1999).
- 4.1<sub>dt</sub>) Jago, das Aas, **lauscht** hinter Säulenfluchten den honorigen Erklärungen Othellos vor dem Rat von Venedig. (UN93/JUN.01879 NN, 26.06.1993, S. 22).
- (4.2<sub>dt</sub>) Rund ein Drittel der Steinacher Ortsbürger **lauschten** den Worten des Präsidenten, als dieser auf das vergangene Jahr zurückblickte. (A13/APR.06663 St. Galler Tagblatt, 17.04.2013, S. 34).
- (4.3<sub>dt</sub>) Plätzchenduft schwebte durch die Schule in Wattenheim, denn dort entpuppten sich die Jungen und Mädchen nicht nur als gute Zuhörer, sondern auch als exzellente Plätzchenbäcker. Gemeinsam mit Lehrerin Ilka Peter hatten sich zehn Kinder aus den Klassen Zwei, Drei und Vier versammelt und **lauschten** auf die Erzählung „Weihnachten im Möwenweg“ von Kirstin Boie. (M08/DEZ.95513 Mannh. Morgen, 08.12.2008, S. 19).
- (5.1<sub>dt</sub>) Mit großem Interesse hatten die Pflegekräfte [...], dem Vortrag **zugehört** [...]. (M03/FEB.11701 Mannh. Morgen, 22.02.2003).
- (5.2<sub>dt</sub>) Helmut Mägdefrau, der stellvertretende Tiergartendirektor, hat der Debatte lange schweigend **zugehört** [...]. (NUZ13/JUL.01777 Nürnberger Zeitung, 20.07.2013, S. 11).
- (5.3<sub>dt</sub>) Er marschiert von Haustür zu Haustür und **hört zu**, was ihm die Leute erzählen. (RHZ96/AUG.03591 RZ, 07.08.1996).

Der Zugriffspfad 2 kann genutzt werden, wenn der Benutzer schon ein mögliches Lexem für seine Ausdrucksbedürfnisse kennt. Über eine alphabetisch angeordnete Leiste erhält er Zugriff zu dem gesuchten Lemma, und wird dann über die Zuordnung zu dem entsprechenden konzeptuellen Feld (*Abbildung 2*) und den feldrelevanten Lesarten zu dem einen oder anderen lexikalisch-semantischen Paradigma – entsprechend dem Zugriffspfad 1, geleitet (siehe *Abbildung 4: abhören*).



in Verbindung mit „Beschwerde / Beschimpfung“ scheint *anhören* (Beispiel 1.3<sub>dt</sub>) am adäquatesten zu sein (Abbildung 5). Nach Auswahl einer Konstruktionsvariante hat der Benutzer dann die Möglichkeit detaillierte einzellexematische Information zu den vier Beschreibungsmodulen (Abbildung 1) zu erhalten. An dieser Stelle des Such- und Auswahlprozesses angelangt, wird die onomasiologisch orientierte Ausgangsperspektive, für die zwei Zugangspfade angeboten wurden, mit einer semasiologischen Zugriffsstruktur verbunden (Blank & Koch 2003, Mingorance 1994, Proost 2007, Reichmann 1989) und bietet auf der Mikrostrukturebene die relevante Gebrauchsinformation an.

Daneben ist es aber auch möglich, die einzellexematische Information weiterhin im Kontrast zu den anderen Elementen des Teilparadigmas zu erhalten (Abbildung 6). In diesem Fall wird dem Benutzer z.B. deutlich, dass nicht alle Elemente des Paradigmas denselben Satzbauplan (SBP) aufweisen. Das zweite Argument erfährt unterschiedliche morphosyntaktische Realisierungsformen. Während *sich anhören1* eine Akkusativergänzung (Eakk) (Bsp. 1<sub>dt</sub>) und *zuhören1* (Bsp. 5.1<sub>dt</sub>, 5.2<sub>dt</sub>) eine Dativergänzung (Edat) regiert, ist bei *lauschen1* (Bsp. 4<sub>dt</sub>) und *hорchen1* (Bsp. 3<sub>dt</sub>) die Alternanz zwischen Edat/Eprp auffällig. Die Information zu den einzelnen SBP können zwar auch in Valenzwörterbüchern konsultiert werden, aber erst der Überblick der Vielfalt in einem Teilparadigma durch ein strukturiertes Informationsangebot ermöglicht dem L2-Benutzer eine bewusste Auswahl und Anwendung.

Ebenso ist der strukturierte Überblick der Information in Teilparadigmen für den Sprachenkontrast von höchstem Interesse. Bei der Selektion der möglichen Entsprechung *escuchar 2* zu allen Elementen des besagten Paradigmas erhält der Benutzer die Information zu dem spanischen Verb bezüglich der vier Beschreibungsmodule. Besonders auffällig sind Unterschiede in der morphosyntaktischen Realisierungsform. Dem dativisch realisierten A2 in *zuhören* entspricht z.B. ein direktes Objekt in *escuchar2* und *oir2* (Bsp. 1-2<sub>sp</sub>), während die Realisierungsmöglichkeit durch eine Präpositionalphrase in mehreren deutschen Lexemen des Subparadigmas in *escuchar2* nicht möglich ist. Eine weitere kontrastiv relevante Auffälligkeit in diesem Subparadigma ist die Beobachtung, dass sich die deutschen Verben teilweise durch distinktive Bedeutungsmerkmale unterscheiden lassen können, während das spanische Verb eine viel allgemeinere Bedeutung besitzt, und daher die kommunikative Notwendigkeit zu Spezifizierungen über adverbiale Zusätze erfolgen muss (1.4-1.5<sub>sp</sub>). Weitere aufschlussreiche Divergenzen zwischen beiden Sprachen und bezüglich aller konzeptueller Felder, die im Rahmen von DICONALE behandelt werden, sind in den unterschiedlichen satzförmigen Komplementrealisierung zu erwarten (2.3<sub>dt</sub>, 5.3<sub>dt</sub>, 1.3<sub>sp</sub>). Für den korrekten Gebrauch in der fremdsprachigen Produktionssituation sind derartige kontrastive Informationen von enormer Relevanz und sollten dem Benutzer klar vor Augen geführt werden.

- (1.1<sub>esp</sub>) En la tribuna de invitados **escucharon** el debate el secretario general de UGT [...] (El Mundo, 20/11/2002)
- (1.2<sub>esp</sub>) Allí, mientras **escuchaba** las noticias por televisión, se quedó impresionada cuando una locutora narraba con frialdad el siguiente suceso [...]: (El Diario Vasco, 31/01/2001)
- (1.3<sub>sp</sub>) Fue en tono de broma, pero también hay que **escuchar** lo que dicen los demás“ (El País, 02/06/1989)

- (1.4<sub>sp</sub>) La comitiva **escuchó** atentamente las explicaciones de Fernando Checa, director del museo, y de José Antonio Fernández Ordoñez, presidente del patronato. (El País, 18/11/1997)
- (1.5<sub>sp</sub>) Horacio **escucha** con atención mi relato. (La Razón, 20/12/2001)
- (2.1<sub>esp</sub>) Prácticamente a la misma hora, el juez de instrucción de París, Gilles Boulouque, que se encarga de los atentados de 1986, y un tribunal islámico, **oían** las declaraciones de los dos personajes. (El País, 01/12/1987)

Pls3 Paradigma: „zuhören“ ([Höruse], [Höruse], [Höruse])		Modul 2	Modul 3 → argument- strukturmarker: [A1 A2] A1: HÖRER A2: GEHORTE AUßERUNG	Modul 4	Ent- sprechungen		
(Sich) <b>anhören</b> Modul 1	Bedeutung	[A1 A2]	Freq.	A1	A2	Passiv: werden	<b>escuchar / oír</b>
	semantisch- distinktive Merkmale (-genau)	Jemand (A1) hört (sich) etwas (A2) an		(-hum)	(-frie)	Wb: die Anhörung ...	
	Synonymie Anonymie Hyperonymie ...	SBP <‘s a>		s	a		
				OR:	OR:		
	Belege						
	Kommentar						
(Lauschen) <b>lauschen</b> Modul 1	Bedeutung	[A1 A2]	Freq.	A1	A2	Passiv: werden	<b>escuchar</b>
	semantisch- distinktive Merkmale (-konzentriert)	Jemand (A1) lauscht etwas / auf etwas (A2)		(-hum)	(-frie)	Wb: der Lauscher ...	
		SBP <‘s d/ prp>		s	d/ prp/au fa		
				OR:	OR:		
	Belege						
	Kommentar						

Link : mit entsprechender Information zu den 4 Modulen und spezifischer Hinweis auf kontrastive Besonderheiten

Abbildung 6: Tabellarische Übersicht eines SPIs: Informationsangebot für Module 1-4 zur Auswahl (Teilinformation).

## 5 Ausblick

In dem Beitrag wurden die verschiedenen Zugriffspfade zu der für fremdsprachige Produktionszwecke relevanten Information vorgestellt und der Versuch unternommen, die Perspektive eines Wörterbuchbenutzers im fremdsprachigen Produktionsprozess zu verfolgen, um gemäß seiner Bedürfnisse den Such-, Auswahl- und Anwendungsprozess zu gestalten. Im Laufe der Ausführungen ist deutlich geworden, dass das Verfolgen einer onomasiologisch-konzeptuell angeordneten Informationsdarbietung im zweisprachigen Kontext ein komplexes Unterfangen darstellt. Das hier vorgestellte Projekt soll als Prototyp eines neuartigen Konsultationswerkes verstanden werden, welches durch einen primär konzeptuell und szenenorientierten Bezugsrahmen den Zugang anbietet und durch ein komplexes Such- und Selektionsverfahren den Benutzer zu einer Reihe von bedeutungsähnlichen Lexemen und Konstruktionsvarianten führt, aus denen nach verschiedenen Kriterien für den kontextadäquaten Gebrauch ausgewählt werden muss. Gebrauchsrelevante einzelsprachige und kontrastive Information sollen eine korrekte Anwendung in der jeweiligen L2 ermöglichen. Übersichtliche, tabellari-



sche inter- und intralinguale Gegenüberstellungen der Elemente eines Paradigmas sollen dem Benutzer bei der bewussten Auswahl im Falle der Ausdrucksvarietät behilflich sein. Obwohl gerade in den letzten Jahren immer mehr lexikographische Ressourcen, vor allem mit *online*-Zugang, entwickelt wurden, steht bis jetzt noch ein umfassendes, verlagsunabhängiges und benutzergerechtes lexikographisches Informationsangebot für den fremdsprachigen Produktionsprozess in den Bereichen DaF und Ele aus. DICONALE hat sich zum Ziel gesetzt, diesem Desideratum einen Schritt näher zu kommen.

## 6 Literatur

### 6.1 Wörterbücher und andere Ressourcen

- CanooNet: Portal: Deutsche Wörterbücher und Grammatik. <http://www.canoo.net/> [11.04.2014].
- Casares, J. (1942/<sup>3</sup>2001): Diccionario ideológico de la lengua española. Barcelona.
- CCDB: Cyril Belica: Kookkurrenzdatenbank V3.3. Eine korpuslinguistische Denk- und Experimentierplattform © 2001 ff., Institut für Deutsche Sprache, Mannheim. <http://corpora.ids-mannheim.de/ccdb/> [11.04.2014]
- DA=Diccionario de Alcalá: Alvar Ezquerro, M. (dir.): Diccionario para la enseñanza de la lengua española. Español para extranjeros, Barcelona, Vox y Universidad de Alcalá, (1995/<sup>2</sup>2000). Online-Version über <http://www.diccionarios.com/> [11.04.2014]
- Dornseiff, Franz/Wiegand, Herbert E./Quasthoff, Uwe (<sup>3</sup>2004). Der deutsche Wortschatz nach Sachgruppen. 8. Neubearbeitete Fassung. Berlin. (print+elektronisch).
- DS=Diccionario Salamanca: Gutiérrez Cuadrado, J. (dir.): Diccionario Salamanca de la lengua española, Madrid, Santillana y Universidad de Salamanca, 1996/2007.
- Duden-Portal: online <http://www.duden.de/> [11.04.2014]
- DWB-DaF: Duden (<sup>2</sup>2010): Deutsch als Fremdsprache – Standardwörterbuch. Mannheim.
- DWDS: Digitales Wörterbuch der deutschen Sprache. <http://www.dwds.de/> [11.04.2014]
- GWB-DaF: Kempcke, G.: Wörterbuch Deutsch als Fremdsprache. Berlin/NewYork: de Gruyter, 1999.
- LEO: zweisprachiges Wörterbuchportal. <http://www.leo.org/> [11-04.2014]
- LGWB-DaF: Götz, D., Haensch, G. & Wellmann, H.: Langenscheidts Großwörterbuch Deutsch als Fremdsprache. Neubearbeitung. Berlin, München: Langenscheidt, <sup>3</sup>2010. Online-Zugang über: TheFreeDictionary <http://de.thefreedictionary.com> [11-04.2014]
- LHWP: Langenscheidts Handwörterbuch Spanisch <sup>12</sup>1982: Teil 1: Spanisch - Deutsch (LHWP-SD), Teil 2: Deutsch - Spanisch (LHWP-DS), Berlin, München: Langenscheidt.
- LHWBe: Langenscheidts Handwörterbuch Spanisch 2006. Spanisch - Deutsch (LHWBe-SD) / Deutsch - Spanisch (LHWBe-DS). Berlin, München: Langenscheidt. Elektronische Fassung.
- PGWB-DaF: Pons Großwörterbuch DaF (2004): Stuttgart: Pons.
- Pons: Das Sprachenportal. <http://de.pons.eu/> [11.04.2014]
- SGIWBe: Slaby, R. J. & Grossmann, R. & Illig, C. (<sup>3</sup>2003): Wörterbuch der spanischen und deutschen Sprache. Spanisch - Deutsch (SGIe-SD), Deutsch - Spanisch (SGIe-DS). Wiesbaden: Brandstetter Verlag. + Elektronische Fassung.
- SM-Clave: SM-portal: Diccionario de uso del español actual. <http://clave.smdiccionarios.com/app.php> [11-04.2014]
- SM-Diccionario: Maldonado, C. (dir.): Diccionario de español para extranjeros. Madrid: SM, 2002.

The free Dictionary: <http://de.thefreedictionary.com/> [11.04.2014]

Wehrle, H. & Eggers, H. (1961/<sup>17</sup>1993): *Deutscher Wortschatz: ein Wegweiser zum treffenden Ausdruck*. Stuttgart.

WGWB-DaF: Wahrig (2008): *Großwörterbuch Deutsch als Fremdsprache*. Berlin. Online-Zugang über Wissens-Portal: <http://www.wissen.de> [11.04.2014]

Wortschatz Universität Leipzig: Portal <http://wortschatz.uni-leipzig.de/> [11.04.2014]

## 6.2 Fachliteratur

Abel, A. (2008). *ELDIT* (Elektronisches Lernerwörterbuch Deutsch-Italienisch) und *ellexiko* im Vergleich. In Klosa, A. (ed.) 1/2008. 175-189. <http://pub.ids-mannheim.de/laufend/opal/pdf/opal2008-1.pdf>. [11.04.2014]

Blank, A. & Koch, P. (eds.) (2003). *Kognitive romanische Onomasiologie und Semasiologie*. Tübingen: Niemeyer.

Boas, H. (2013). Wieviel Wissen steckt in Wörterbüchern? Eine Frame-semantische Perspektive. In *Zeitschrift für Angewandte Linguistik* 57, 75-97.

Boas, H. C. & Ryan Dux (2013). Semantic frames for foreign-language education: Towards a German frame-based dictionary. In *Veritas On-Line* 1/2013, 81-100.

Dentschewa, E. (2006). DaF-Wörterbücher im Vergleich: Ein Plädoyer für „Strukturformeln“. In Dimova, Ana et al. (eds.): *Zweisprachige Lexikographie und Deutsch als Fremdsprache*. Hildesheim: Olms, 113-128.

Domínguez Vázquez, M<sup>a</sup> J. et al. (2013). Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen. In Domínguez Vázquez, M<sup>a</sup> J. (ed.), 135-172.

Domínguez Vázquez, M. (ed.) (2013). *Trends in der deutsch-spanischen Lexikographie*. Frankfurt: P. Lang Edition.

Domínguez Vázquez, M<sup>a</sup> J., Gómez Guinovart, X. & Valcárcel Riveiro, C. (eds.). *Lexicografía románica. Aproximaciones a la lexicografía moderna y contrastiva*. (Coord.: Sánchez Palomino, M<sup>a</sup>. D. & Domínguez Vázquez, M<sup>a</sup> J. (vol. 2). Berlin: de Gruyter (im Druck).

Engel, U. (2004). *Deutsche Grammatik*. – Neubearbeitung. München: iudicium.

Engelberg, S. & Lemnitzer, L. (2001), (\*2009). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Engelberg, St. (2010). Die lexikographische Behandlung von Argumentstrukturvarianten in Valenz- und Lernerwörterbüchern. In Fischer, K., Fobbe, E. & Schierholz, St. (eds.), *Valenz und Deutsch als Fremdsprache*, Frankfurt a. M.: P. Lang, 113-141.

Engelberg, St. (2014a). The argument structure of psych-verbs: A quantitative corpus study on cognitive entrenchment. In Boas, H. & Ziem, A. (eds.): *Constructional approaches to argument structure in German*. Boston, Berlin: De Gruyter Mouton. (im Druck).

Engelberg, St. (2014b). Gespaltene Stimulus-Argumente bei Psych-Verben. Quantitative Verteilungsdaten als Indikator für die Dynamik sprachlichen Wissens über Argumentstrukturen. In: Engelberg, St. et al. (eds.): *Argumentstruktur – Valenz – Konstruktionen*. Tübingen: Narr. (im Druck).

Engelberg, St., Kopleinig, A., Proost, K., Winkler, E. (2012). Argument structure and text genre: cross-corpus evaluation of the distributional characteristics of argument structure realizations. In *Lexicographica* 28, 13-48.

Engelberg, St., Meliss, M., Probst, K. & Winkler, E. (eds.) (2014). *Argumentstruktur – Valenz – Konstruktionen*. Tübingen: Narr. (im Druck)

Fuentes Morán, M<sup>a</sup> T. (1997). Gramática en la lexicografía bilingüe. Morfología y sintaxis en diccionarios español-alemán desde el punto de vista del germanohablante. Tübingen: Niemeyer.

González Ribao, V. & Meliss, M. (2014). Vorschläge zur Ausarbeitung eines onomasiologisch-konzeptuell orientierten Produktionswörterbuches im zweisprachigen Lernerkontext: Deutsch-Spanisch. In Calañas Contente, J.A. et al. (eds.). *Wörterbücher des Deutschen*. Frankfurt a. M.: P. Lang (Reihe: Studien zur Linguistik des Deutschen – Spanische Akzente) (im Druck).



- González Ribao, V. & Proost, K. (2014). El campo léxico al servicio de la lexicografía: Un análisis contrastivo en torno a algunos subcampos de los verbos de comunicación en alemán y español. In Domínguez Vázquez, M<sup>a</sup> J. et al. (eds.): *Lexicografía*. (Coord.: Sánchez Palomino, M<sup>a</sup>. D. & Domínguez Vázquez, M<sup>a</sup> J.). (vol. 2). (im Druck).
- Haensch, G. & Omeñaca, C. (1997, <sup>2</sup>2004). (coords.): *Los diccionarios del español en el siglo XXI*. Salamanca: Ediciones Universidad.
- Haß, U. (ed.) (2005). Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. Berlin: de Gruyter.
- Haß, U. & Schmitz, U. (2010). Lexikographie im Internet 2010 – Einleitung. In Gouws, R. H. et al. (eds.). *Lexicographica. Internationales Jahrbuch für Lexikographie*. 26. Berlin: de Gruyter, 1-18.
- Hausmann, F. J. (1991). Die zweisprachige Lexikographie Spanisch-Deutsch, Deutsch-Spanisch. In Steger H. & Wiegand, H.E. (eds.). *Wörterbücher: Ein Internationales Handbuch zur Lexikographie*. Berlin/NewYork: de Gruyter. 2987-2991.
- Herbst, Th. & Klotz, M. (2003). *Lexikografie*. Paderborn: Schöningh.
- Harras, G., Winkler, E. et al. (2004): *Handwörterbuch deutscher Kommunikationsverben*. Teil 1: Wörterbuch. Berlin.
- Harras, Gisela, Proost, K. & Winkler, E. (2007). *Handbuch deutscher Kommunikationsverben*. Teil 2: Lexikalische Strukturen. Berlin. Und: Proost, K.: [Online-Nachschlagewerk Kommunikationsverben](#):
- Kemmer, K. (2010). Onlinewörterbücher in der Wörterbuchkritik. Ein Evaluationsraster mit 39 Beurteilungskriterien. In *Online publizierte Arbeiten zur Linguistik*. OPAL2/2010, 1-33. <http://pub.ids-mannheim.de/laufend/opal/pdf/opal2010-2.pdf>. [11.04.2014]
- Klosa, A. (ed.) (2008). *Lexikographische Portale im Internet*. OPAL-Sonderheft 1/2008, . [<http://pub.ids-mannheim.de/laufend/opal/pdf/opal2008-1.pdf>]. [11.04.2014]
- Klosa, A., Koplenig, A. & Töpel, A. (2011). Benutzerwünsche und Meinungen zu einer optimierten Wörterbuchpräsentation – Ergebnisse einer Onlinebefragung zu „elexiko“. In *Online publizierte Arbeiten zur Linguistik: OPAL 3/2011*. Mannheim: Institut für Deutsche Sprache. [11-04.2014]
- Mann, M. (2010). Internet-Wörterbücher am Ende der „Nulljahre“. In Hass, U. & Schmitz, U. (eds.): Thematic Part: Lexikographie im Internet 2010. *Lexicographica. Internationales Jahrbuch für Lexikographie*, 26/2010. Berlin, 19-45.
- Martín Mingorance, L. (1994). La lexicografía onomasiológica. In: Hernández, H. & Mederos, H. (Coord.). *Aspectos de lexicografía contemporánea*. Barcelona: Biblograf, 15-27.
- Meliss, M. (2013a). Das zweisprachige Wörterbuch im bilateralen deutsch-spanischen Kontext. Alte und neue Wege. In Domínguez Vázquez, M<sup>a</sup> J. (ed.), 61-87.
- Meliss, M. (2013b). Online-Lexikographie im DaF-Bereich: Eine erste kritische Annäherung: Bestandsaufnahme – Nutzen – Perspektiven. In *Real Revista de Estudos Alemães*, 4, 176-199. [http://real.fl.ul.pt/textos.page/pag/2](http://real.fl.ul.pt/textos/page/pag/2). [11.04.2014]
- Meliss, M. (2014a). (Vor)überlegungen zu einem zweisprachigen Produktionslernerwörterbuch für das Sprachenpaar DaF und ELE. In Reimann, D. (ed.). *Kontrastive Linguistik und Fremdsprachendidaktik Iberoromanisch – Deutsch*. Studien zu Morphosyntax, nonverbaler Kommunikation, Mediensprache, Lexikographie und Mehrsprachigkeitsdidaktik (Spanisch/Portugiesisch/Deutsch) (Reihe: Romanistische Fremdsprachenforschung und Unterrichtsentwicklung). Tübingen: Narr.
- Meliss, M. (2014b). Das verbale Kombinationspotenzial in einsprachigen DaF-Lernerwörterbüchern: Kritische Bestandsaufnahme – Neue Anforderungen. In *ZDaF* (im Druck).
- Meliss, M. (2014c). Propuestas para un diccionario conceptual bilingüe para ELE y DaF. In: Domínguez Vázquez, M<sup>a</sup> J. et al. (eds.): (Coord.: Sánchez Palomino, M<sup>a</sup>. D. & Domínguez Vázquez, M<sup>a</sup> J. (vol. 2). (im Druck).
- Meliss, M. & Sánchez Hernández, P. (2014). Theoretical and methodological foundations of the DICONALE project: a conceptual dictionary of German and Spanish. In Silvestre João Paulo et al. (eds.): *Dicionários que não existem*. Lissabon.

- Model, B. (2010). Syntagmatik im zweisprachigen Wörterbuch. Berlin: de Gruyter.
- Müller-Spitzer, C. & Engelberg, St. (2013). Dictionary Portals. In Rufus H. Gouws et al. (eds.): *Dictionaries. An International Encyclopedia of Lexicography*. Supplementary volume: Recent developments with special focus on computational lexicography. Berlin, New York: de Gruyter (im Druck).
- Porto Dapena, J. A., Conde Noguero, E., Córdoba Rodríguez, F. & Muriano Rodríguez, M<sup>a</sup> M. (2008): Presentación del diccionario “Coruña” de la lengua española actual. En: Bernal, E. & DeCesaris (eds.): *Proceedings of the XIII Euralex International Congress*. Barcelona: Documenta Universitaria, Série Activitats, 20.753-762.
- Proost, K. (2007). *Conceptual structure in lexical items: The lexicalisation of communication concepts in English, German and Dutch*. Pragmatics & Beyond New Series; 168. Amsterdam/Philadelphia: Benjamins.
- Reichmann, O. (1989). Das onomasiologische Wörterbuch: Ein Überblick. In Sterger H. & Wiegand, H.E. (eds.). *Wörterbücher: Ein Internationales Handbuch zur Lexikographie*. Berlin/New York: de Gruyter, 1057-1067.
- Storrer, A. (2010). Deutsche Internet-Wörterbücher: Ein Überblick. In Gouws, R. H. et al. (ed.). *Lexicographica. Internationales Jahrbuch für Lexikographie* 26. Berlin: de Gruyter, 154-164.
- Subirats, C. (2009). Spanish Framenet: A frame-semantic analysis of the Spanish lexicon. In Boas, H. (ed.). *Multilingual FrameNets in Computational Lexicography. Methods and Applications*. Berlin/New York: Mouton de Gruyter, 135-162.
- Tarp, S. (2012): Online dictionaries: today and tomorrow. In Heid, U. (ed.): Thematic Part: Corpora and Lexicography. *Lexicographica (International Annual for Lexicography)* 28/2012. Berlin, 253-267.

# Graph-Based Representation of Borrowing Chains in a Web Portal for Loanword Dictionaries

Peter Meyer  
Institut für Deutsche Sprache, Mannheim  
meyer@ids-mannheim.de

## Abstract

This paper reports on an ongoing lexicographical project that investigates Polish loanwords from German that were further borrowed into the East Slavic languages Russian, Ukrainian, and Belorussian. The results will be published as three separate dictionaries in the *Lehnwortportal Deutsch*, a freely available web portal for loanword dictionaries having German as their common source language. On the database level, the portal models lexicographical data as a cross-resource directed acyclic graph of relations between individual words, including German ‘metalemmata’ as normalized representations of diasystemic variants of German etyma. Amongst other things, this technology makes it possible to use the web portal as an ‘inverted loanword dictionary’ to find loanwords in different languages borrowed from the same German etymon. The different possible pathways of German loanwords that went through Polish into the East Slavic languages can be represented directly as paths in the graph. A dedicated in-house dictionary editing software system assists lexicographers in producing and keeping track of these paths even in complex cases where, e.g., only a derivative of a German loanword in Polish has been borrowed into Russian. The paper concludes with some remarks on the particularities of the dictionary/portal access structure needed for presenting and searching borrowing chains.

**Keywords:** online dictionary; graph databases; loanwords

## 1 Introduction

### 1.1 The *Lehnwortportal Deutsch*

The *Lehnwortportal Deutsch* ([lwp.ids-mannheim.de](http://lwp.ids-mannheim.de)) is a freely accessible online lexical information system developed at the Institute for German Language (IDS) that has been designed to provide unified access to a large number of both existing and newly produced XML-based dictionaries of German loanwords in other languages.<sup>1</sup> The modular architecture of the portal allows for easy integration of new resources of possibly very heterogeneous structure; each portal dictionary may have its own XML schema, as long as the underlying lexicographical information of the different constituent parts of an

---

1 The web portal in its present form has been developed in a project funded by the Federal Government Commissioner for Culture and the Media upon a Decision of the German Bundestag.

entry are unambiguously and explicitly encoded and separated in the markup, analogous to what is called the ‘lexical view’ in the TEI.dictionaries module, cf. Burnard, Bauman 2007, section 9.5 (online at <http://en.guidelines.tei-c.org/html/DI.html#DIMVLV> [04/11/2014]).

Apart from conventional access to the individual dictionaries, the portal offers complex cross-dictionary search functionality; in particular, it can be used as an ‘inverted loanword dictionary’ (Engelberg 2010) to trace the way of German words into different recipient languages, comparable to the manually compiled dictionary of Dutch loanwords in the world’s languages by van der Sijs (2010). As any German etymon may appear in a variety of orthographical, phonetic/phonological and diasystemic variants in different entries within and across loanword dictionaries, these different forms are mapped in manual lexicographical work to etymologically corresponding ‘normalized’ word forms, wherever possible contemporary Standard German words. This is accomplished at the IDS with the help of an in-house software tool during the integration of a loanword dictionary into the web portal. These normalized entries, henceforth *metalemmata*, are used as headwords of the inverted loanword dictionary.

## 1.2 Graph-based Data Modeling

The XML-based representation of entries in the individual component dictionaries mainly serves as input for XSLT transformations that produce a fairly conventional, dictionary-specific HTML-based online presentation of the entries. For advanced search functionalities, however, a relational database is used that represents lexicographical information as a cross-resource network (a *directed acyclic graph*) of relations between words that are, as we say throughout this paper, ‘recorded’ in the individual dictionaries. These recorded words include *metalemmata*, etyma and loanwords alongside their variant forms, derivatives etc. Interactive visualizations of parts of this graph are available online; figure 1 below shows the subgraph for the German *metalemma Gestalt* ‘shape’. Differently colored discs correspond to words recorded in different dictionaries (vertices/nodes in the graph); different kinds of relations (arcs/directed edges) are symbolized by different types of arrays between two discs. In the example we see that the ‘normalized’ contemporary German *metalemma Gestalt* ‘corresponds to’ [=dark solid arrow] a New High German etymon *Gestalt* and a Middle High German etymon *gestalt* as recorded in the portal’s Polish loanword dictionary (color tag: dark blue) and to a Middle High German / Bavarian etymon *g·stalt* as recorded in the Slovene dictionary (color tag: green). We further see that, e.g., the etymon *gestalt* ‘has been borrowed into’ Polish [=grey solid arrow] as *hsztalt* which ‘has a variant phonetic’ form *hstalt* [=black long-dashed arrow] and from which (amongst other words) the verb *hsztalcíc* ‘has been derived’ [=dashed-dotted arrow]. The relationships between words recorded in the same loanword dictionary entry are programmatically extracted from the XML source of the entry on a per-dictionary basis, making use of the fact that different kinds of relations correspond to different structural configurations in the entry document that depend on the XML schema of the dictionary and can be described using XPath expressions. Every word in the graph has a set of attributes (diasystemic, grammatical, semantic information) obtained by encoding the appropriate pieces of in-



## 2 Modeling and Editing Borrowing Chains in the Portal's Graph Database

### 2.1 Data Modeling Aspects: Borrowing Chains as Paths in the Graph

Borrowing chains are the premier *raison d'être* for the graph-based data modeling in the *Lehnwortportal*. Figure 2 (below) shows, if only in a highly schematic fashion, the sample case of German *Drucker* 'printer' that has entered the East Slavic languages mostly through Polish. The different pathways obviously form a small directed graph that can be added more or less directly as a new subgraph to the portal data graph. The dashed arrows indicate less likely borrowing pathways; correspondingly, edges in the portal graph may be assigned weights to indicate likelihood of a borrowing relation and to calculate rankings of search results. Note that there are three different paths in this subgraph all leading from the German etymon *Drucker* to the Russian loanword *drukar'*.

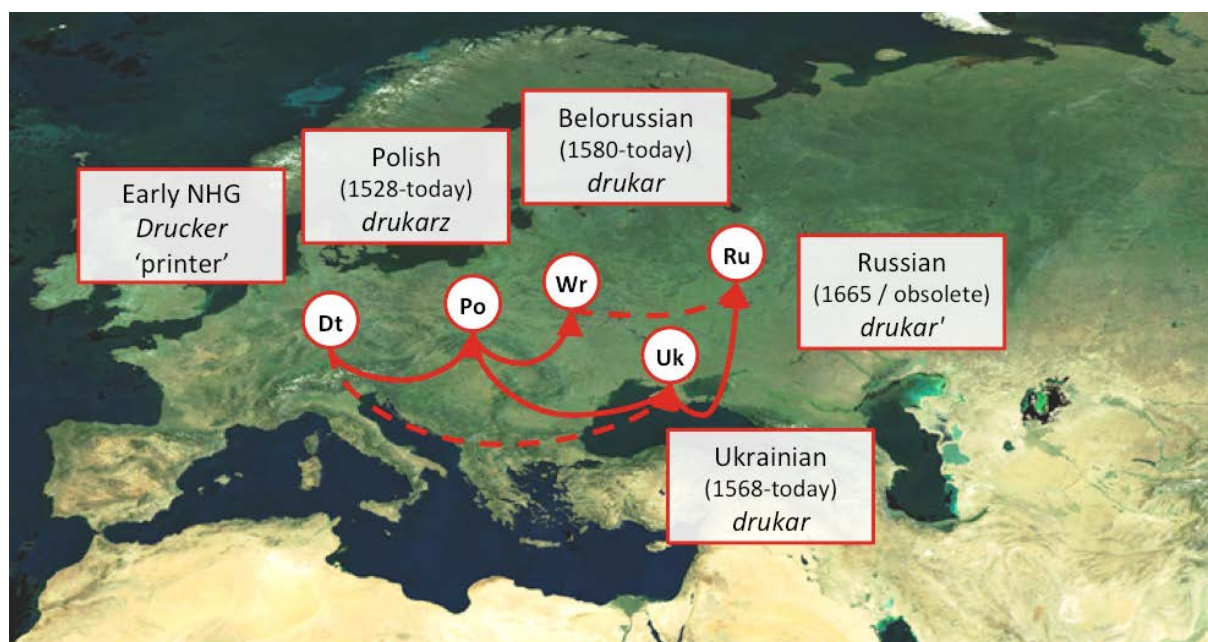


Figure 2: Possible borrowing paths of German *Drucker* 'printer' into the East Slavic languages.

### 2.2 Some Technical Aspects of the Lexicographical Process

The graph data layer atop the XML resources of individual loanword dictionaries requires a highly specialized lexicographical process of its own (cf. Meyer, to appear). The graph is not a self-contained resource; instead, it must be constructed anew from the individual dictionaries and portal-specific cross-reference information as soon as any of the portal's data sources changes. Tracing borrowing chains complicates the picture considerably. For the project presented here, a complex in-house desktop dictionary editing software is being developed which allows lexicographers to collaboratively



compile and edit excerpts *using the lemmata (and other recorded words and meaning definitions) of the portal's Polish loanword dictionary* (de Vincenz, Hentschel 2010) *as a common frame of reference*. Figure 3 (below) shows a screenshot of a preliminary version of the editor used for excerpting. The working lexicographer selects a Polish loanword from (de Vincenz, Hentschel 2010) such as *browar* 'brewer; brewery' from Middle High German *brouwer* 'brewer' (1). A preview of this entry is displayed in the main window (2). All hitherto produced excerpts of existing dictionary entries on East Slavic borrowings from the selected Polish word are listed as a tree structure in the editor (3); for each such entry, the tree shows all recorded (phonetic and diasystemic) variants, meanings, derivatives (with their own range of variants and meanings) and competing near-synonyms that have been input so far. Clicking on a tree item (here, on the variant *provar* of the entry *brovar* in the multivolume Belorussian Historical Dictionary *Historyčny sloŭnik belaruskaj movy*) opens an input panel (4) for all pertinent lexicographical information, including an arbitrary number of records and quotations. A preview of the current state of the whole excerpt is also available (5). There is a separate input panel, not shown in figure 3, for editing cross-dictionary information on the possibly multiple borrowing paths within the East Slavic languages. Often Polish loanwords from German have formed compounds and derivatives; it is well possible that only one of these derived forms, but not the 'original' loanword, has been passed on into an East Slavic language. The editing software also offers convenient input options for such cases. There are additional tools for compiling the entries of the three new East Slavic loanword dictionaries from the excerpts.

There are many reasons why an off-the-shelf software solution would not have been suitable for the lexicographical tasks of the project. To begin with, it would have been next to impossible to customize a commercial dictionary editing application in order to incorporate cross-referencing functionality to an existing dictionary. In this particular case, cross-references are needed not only to whole entries of the Polish dictionary (de Vincenz, Hentschel 2010), but also to derivatives and compounds recorded in these entries, and, most important, to the different word senses given in the entries since they will serve as a *tertium comparationis* for word sense distinctions in the East Slavic loanwords. It would have been possible to customize a professional XML editor by implementing some kind of cross-referencing plugin. However, there is another layer in the editing process that cannot easily be managed in XML: After compiling excerpts of entries on a German loanword in, say, Ukrainian, in a number of Ukrainian loanword dictionaries, these excerpts have to be merged in a rather complex way to produce a new entry in the Ukrainian loanword dictionary of the portal. The excerpted loanword dictionaries (which may or may not cover different periods of the language) will have differing lemmatizations, list different variants of the word, use incompatible word sense distinctions and so on. On the other hand, there will usually be a lot of duplicate information. As a consequence, the amalgamation process of creating entries in the three new East Slavic portal dictionaries is far from trivial; doing this by cutting XML fragments from the excerpts and pasting them into the XML structure of the newly created entries would be an excessively tedious, error-prone and confusing task, even more so since word sense distinctions in parallel entries in the three dictionaries should be made in as uniform a

manner as possible, based on the distinctions in the entries of (de Vincenz, Hentschel 2010). It is not a realistic goal to develop software tools for these tasks as simple XML editor plugins; even the very idea of using XML as the basic frame of reference is problematic in such a complex cross-resource editing context.

In fact, the software developed at the IDS is not directly XML-based, but uses a straightforward object-oriented data model for both the excerpts and the newly produced entries. This greatly simplifies the underlying cross-referencing and the implementation of tools for merging and validating lexicographical data from a large number of resources. The software produces XML serializations of the data that can be used both to construct HTML views of the data and to define the directed graph of the portal.

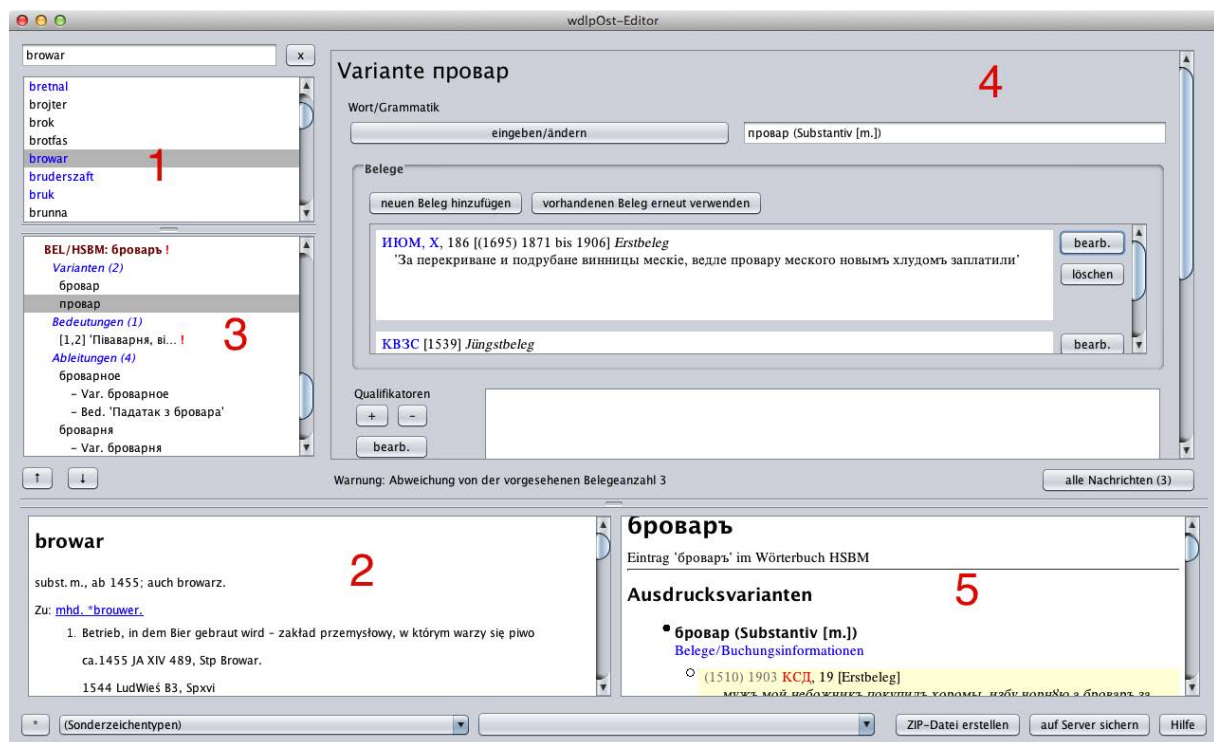


Figure 3: Screenshot: Preliminary user interface of the dictionary writing software.

## 2.3 General Lexicographical Issues Concerning Borrowing Chains

As explained above, the project presented in this paper strives to compile new loanword dictionaries that directly reference a Polish loanword dictionary already integrated into the portal. As the *Lehnwortportal* has been designed with a focus on leveraging existing resources, we expect to see other cases in the future where information from multiple existing loanword dictionaries is combined to reconstruct borrowing chains. Here, the data graph is an ideal means of abstracting from micro- and mediostuctural idiosyncrasies of the dictionaries involved. Dutch may serve as a good example since



it has served as a ‘hub’ that mediated German lexis into many languages, in particular those of colonized countries. Thus, we might try to combine information on Dutch loans from German – as represented in a traditional loanword dictionary (e.g., van der Sijs 2005) – with information on Dutch loanwords in other languages – as given in (van der Sijs 2010) – to (re)construct borrowing chains from German via Dutch into other languages. In these cases, an ‘intermediate’ Dutch loan corresponds to *two* connected vertices in the graph as it appears in two independent lexicographical resources – both as a loanword from German and as an etymon for Dutch loans in other languages. The etymological identification of these two ‘instances’ of the intermediate loanword is part of the lexicographical process to be carried out for the portal. There are many technical and lexicographical issues arising in borderline cases, e.g., when borrowing chains are directly specified in a loanword dictionary entry (such as “from German *Drucker* via Polish *drukarz*”). Here, multiple cross-resource etymological identifications with words in other resources might become necessary, even with the possibility of conflicting information, e.g. if another loanword dictionary of the portal gives a different German etymology for an intermediate loan.

### **3 Access Structures for Borrowing Chains in the Portal**

#### **3.1 Online Entry Presentation**

The information on borrowing chains is, in many cases, not present within the confines of a single loanword dictionary entry, but is instead distributed between different resources. The online presentation of individual dictionary entries should nevertheless make this information visible in all of the individual dictionary entries involved. At present, loanword dictionary entries and entries of the ‘inverted loanword dictionary’ of German metalemmata systematically cross-reference each other in the *Lehnwortportal*; in the case of ‘indirect’ loanwords from German, the web application will use the data graph to add another layer of information, viz. on the ‘intermediate’ or ‘terminal’ loanwords, to the existing entries in the Polish and East Slavic dictionaries. These additions only concern the presentation layer; the underlying entries remain unaltered. A special feature of the presently compiled East Slavic dictionaries will be the presence of cross-dictionary commentaries (including schematic visualizations of borrowing pathways) on all entries that refer to the same Polish loanword, since often German loanwords in an East Slavic language could also have been borrowed directly from German or were mediated by another East Slavic language (cf. figure 2 above).

#### **3.2 Advanced Search Options**

It is highly desirable that portal users can include search criteria concerning borrowing chains into their advanced queries such that, e.g., German loanwords in language X that were possibly mediated

through language Y can be found. The *Lehnwortportal* offers fairly advanced and granular search options (cf. Meyer 2013a) that allow the inclusion of complex criteria concerning both German etyma (including metalemmata) and loanwords. With the inclusion of the Slavic dictionaries, search criteria will also be attributable to ‘intermediate’ loans in a borrowing chain. Search results will be ranked according to the weight of the edges of the graph path. Borrowing paths will be specifiable through a planned extension of the declarative domain-specific query language that is currently available for advanced portal users (figure 4); cf. (Wood 2012) for a general overview on graph database query languages and (Meyer 2013a) for more information on the portal’s query language. Even a graphical search through an interactive visual query language for graph databases is conceivable (cf. Blau et al. 2002) and would allow users to literally draw the borrowing paths they are looking for.

### Suche im Portal-Wortnetzwerk

Für Fachleute besteht auf dieser Seite die Möglichkeit, mit einer speziellen →**Abfragesprache** gezielt nach Konstellationen im →wörterbuchübergreifenden Wortnetzwerk (Graphen) des Portals suchen. Die Verweise rechts neben dem nachstehenden Eingabefeld öffnen verschiedene Eingabehilfen.

suche etymon herkunftswort.  
suche lehnwort entlehnung.  
suche lehnwort ableitung.

herkunftswort ist vogaenger zu entlehnung.  
ableitung ist derivat zu entlehnung.

die bedeutung von herkunftswort enthaelt 'Geld'.  
(entlehnung ist verb oder entlehnung ist substantiv).  
nicht(die sprache von entlehnung ist 'Teschener Polnisch').  
nicht(ableitung ist substantiv).

**Einfügen von Suchkriterien**

- Knotendeklarationen
- Eigenschaften von Knoten
- Relationen zwischen Knoten
- weitere Suchbedingungen

Vollständiges Beispiel einfügen

Abfrage ausführen
Abfragefeld leeren

### Suchergebnisse

Wort 'herkunftswort'	Wort 'entlehnung'	Wort 'ableitung'
<span style="color: green;">█</span> gesuoch <i>Etymon in 'žuh'</i>	<span style="color: green;">█</span> žuh <i>Lehnwort in 'žuh'</i>	<span style="color: green;">█</span> žuhati <i>Ableitung in 'žuh'</i>
<span style="color: blue;">█</span> Rechnung <i>Etymon in 'rachunek'</i>	<span style="color: blue;">█</span> rachunek <i>Lehnwort in 'rachunek'</i>	<span style="color: blue;">█</span> rachunkowy <i>Ableitung in 'rachunek'</i>
<span style="color: blue;">█</span> Rüge <i>Etymon in 'rug'</i>	<span style="color: blue;">█</span> rug <i>Lehnwort in 'rug'</i>	<span style="color: blue;">█</span> rugowy <i>Ableitung in 'rug'</i>

**Figure 4: Screenshot: Example search using the portal’s graph query language (<http://lwp.ids-mannheim.de/search/prof>).**

For illustration purposes, here is a rough preview of how a simple query involving borrowing chains will look like in the portal’s query language. Note that the original query language has a German-like context-free grammar; here, we present a corresponding English-like version. The query reads: “Find all Ukrainian or Belorussian words (including variants, derivatives etc.) in the database that have

been borrowed through Polish from a German noun ending in *ung*.” The query uses graph-theoretical terms where appropriate; thus, any loanword borrowed from German is represented by a node in the portal’s directed graph that is a *descendant* of the node corresponding to the German etymon. The results of the query are ordered triples (germanWord, polishWord, eastSlavicWord) of those words recorded in entries of the portal’s dictionaries that comply with all of the constraints specified in the query.

find etymon germanWord.

find loanword polishWord.

find loanword eastSlavicWord.

the language of germanWord is German.

the language of polishWord is Polish.

(the language of eastSlavicWord is Ukrainian OR the language of eastSlavicWord is Belorussian).

germanWord is a noun.

germanWord ends in ‘ung’.

polishWord is descendant of germanWord.

eastSlavicWord is descendant of polishWord.

## 4 References

- Blau, H., Immerman, N. & Jensen, D. (2002). *A Visual Language for Querying and Updating Graphs*. Technical Report 2002-037. University of Massachusetts, Amherst.
- Burnard, L., Bauman, S. (eds.) (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Charlottesville, Virginia: TEI Consortium. Online: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html> [04/11/2014].
- de Vincenz, A., Hentschel, G. (2010). *Wörterbuch der deutschen Lehnwörter in der polnischen Schrift- und Standardsprache. Von den Anfängen des polnischen Schrifttums bis in die Mitte des 20. Jahrhunderts.* (= *Studia slavica Oldenburgensia*, vol. 20). Oldenburg: BIS-Verlag. Online edition: <http://diglib.bis.uni-oldenburg.de/bis-verlag/wdlp> [04/11/2014].
- Engelberg, S. (2010). An inverted loanword dictionary of German loanwords in the languages of the South Pacific. In A. Dykstra, T. Schoonheim (eds.) *Proceedings of the XIV EURALEX International Congress (Leeuwarden, 6-10 July 2010)*. Ljouwert (Leeuwarden): Fryske Akademy, pp. 639-647.
- Meyer, P. (to appear). Von XML zum DAG: Der lexikographische Prozess bei der Erstellung eines graphenbasierten Wörterbuchportals. In F. Mollica, M. Nied, M.J. Domínguez Vazquez (eds.) *Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik*. (to appear in: *Lexicographica*, Series Maior).
- Meyer, P. (2013a). Advanced graph-based searches in an Internet dictionary portal. In I. Kosem, J. Kallas, P. Gantar, P. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 488-502. Accessed at: [http://eki.ee/elex2013/proceedings/eLex2013\\_34\\_Meyer.pdf](http://eki.ee/elex2013/proceedings/eLex2013_34_Meyer.pdf) [04/11/2014].

- Meyer, P. (2013b). Ein Internetportal für deutsche Lehnwörter in slavischen Sprachen. Zugriffsstrukturen und Datenrepräsentation. In S. Kempgen, N. Franz, M. Jakiša, M. Wingender (eds.) *Deutsche Beiträge zum 15. Internationalen Slavistenkongress, Minsk 2013*. München: Otto Sagner, pp. 233-242. (=Die Welt der Slaven. Sammelbände, vol. 50).
- Meyer, P., Engelberg, S. (2011). Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen. In H. Hedeland, Th. Schmidt, K. Wörner (eds.) *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*. Hamburg: Universität Hamburg, pp. 169-174 (=Arbeiten zur Mehrsprachigkeit/Working Papers in Multilingualism, Series B, No. 96).
- van der Sijs, N. (2005). *Groot leenwoordenboek*. Utrecht: Van Dale Lexicografie.
- van der Sijs, N. (2010). *Nederlandse woorden wereldwijd*. Den Haag: SDU Uitgever.
- Wood, P.T. (2012). Query Languages for Graph Databases. In *SIGMOD RECORD*, 41(1) (March), pp. 50-60.

# At the Beginning of a Compilation of a New Monolingual Dictionary of Czech (A Report on a New Lexicographic Project)

Pavla Kochová, Zdeňka Opavská, Martina Holcová Habrová  
Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i.  
kochova@ujc.cas.cz, opavska@ujc.cas.cz, holcova@ujc.cas.cz

## Abstract

The aim of the article is to present the new lexicographic project that is being implemented at the Institute of the Czech Language of the Academy of Sciences of the CR, v. v. i. Since 2012, its Department of Contemporary Lexicology and Lexicography has worked on the creation of a new medium-sized monolingual dictionary of Czech with the working title *Akademický slovník současné češtiny* (The Academic Dictionary of Contemporary Czech). With its size and method of treatment, the dictionary ranks among academic dictionaries, i.e. dictionaries with an elaborated, standardised and structured explanation of the meaning of lexical units, with an adequately rich exemplification documenting the typical use of lexical units, with a sufficiently elaborated description of the basic semantic relations, mainly synonymy and antonymy, with the appropriate description of the grammatical properties of lexical units and with usage labels (a description of the stylistic, temporal, spatial, frequency and pragmatic markedness) of lexical units.

**Keywords:** lexicography; monolingual dictionary; macrostructure of a dictionary; microstructure of an entry

## 1 Introduction

*Akademický slovník současné češtiny* (hereinafter as the ASSČ) builds on the tradition of the general monolingual dictionaries of Czech that emerged at the Institute for the Czech Language in the 20th century.<sup>1</sup> More than forty years have passed since the publication of a dictionary of a larger size, i.e. the

---

1 This tradition has developed from the largest *Příruční slovník jazyka českého* (Reference Dictionary of the Czech Language; *PSJČ*, 1935–1957), through the medium-sized *Slovník spisovného jazyka českého* (Dictionary of the Standard Czech Language; *SSJČ*, 1960–1971) to the one-volume *Slovník spisovné češtiny pro školu a veřejnost* (Dictionary of Standard Czech for Schools and the Public; *SSČ*, 1st edition in 1978; 2nd, revised edition in 1994; 3rd, revised edition in 2003). The *PSJČ* is a scientific descriptive dictionary of a large size (ca 250,000 entries); it describes Czech vocabulary since 1880; it does not use run-on entries; examples are provided by quotations. The *SSJČ* is a medium-sized dictionary (192,908 entries); it captures the literary lexical standard of the time, but the range of the word list exceeds it (by including obsolete, infrequent words etc.); it describes contemporary Czech vocabulary (approximately from the 1930s, selectively from 1880); the *SSJČ* uses run-on entries; exemplification is mainly based on the minimal typical contexts. The *SSČ* is a smaller-size dictionary (2nd ed.: 45,366 entries) focusing on the widest range of users; it describes the central vocabulary of contemporary Czech (mainly from 1945) with an overlap to variously marked words; it has a normative character; exemplification is very limited and is based on the minimal typical contexts. On the history and characteristics of the modern Czech lexicography, see mainly the detailed study by Hladká (2007).

*Slovník spisovného jazyka českého* (since the publication of the first volume it has even been fifty years), which is very long considering vocabulary dynamics, linguistic methodology,<sup>2</sup> in terms of the platform for the creation of a dictionary as well as the medium for its publication.

## 2 The Basic Characteristics of the ASSČ

The ASSČ is a *medium-sized* dictionary,<sup>3</sup> with the expected number of 120–150 thousand lexical units. Its *aim* is to capture widespread contemporary Czech vocabulary used in public official and semi-official communication as well as in everyday (i.e. non-public, unofficial) communication. A natural part of the lexis described are terminological expressions, but not highly specialised terms. To a limited extent, the dictionary presents units utilised in professional and interest-group communication, namely if their use has been extended beyond their professional, interest milieu. Dialectal expressions have been included if they are common in a wider area and are used especially in oral communication or in literature. The expected *user* of the dictionary is a secondary-school educated native speaker; nevertheless, also those interested in Czech as a foreign language are marginally taken into account (since Czech is a language of a small nation, specialised monolingual dictionaries of a larger size for learners are not created). The dictionary being prepared will be continuously *published* on the Internet (on the website of the Institute of the Czech Language). After the work on the dictionary has been completed, it will be possible to publish the work as a whole in a book form.

## 3 The Dictionary Development Method: Selected Aspects

The essential *material basis* is the synchronic corpus of written texts SYN of the Institute of the Czech National Corpus of a size of 2.2 milliard words. Other material resources are the electronic archives of the company Newton Media, a. s. (the archives of both nationwide and regional printed periodicals and transcripts of current affairs television and radio programmes), the internet and the databases of the Institute of the Czech Language.<sup>4</sup>

---

2 In connection with lexicography and lexicology, it is mainly the creation and development of computational and corpus linguistics and corpus lexicography (language corpora, excerpt databases, electronic archives, special software tools, eg. for an analysis of collocations the Word Sketch Engine – Kilgarriff et al. 2004). Cf. Čermák, Blatná 1995; Čermák 2010.

3 It should be emphasised that the dictionary being developed is not a lexical database. The relation between a lexical database and a monolingual dictionary is understood in accordance with Hanks (2010: 581): “A lexical database is a fundamental background resource for use in the creation of many important linguistic artefacts – dictionaries, course books, computer programs for natural language processing among them. A great monolingual dictionary has a different function: it brings together speakers of a language, it has a socially integrative function, making explicit the basis of words and meanings and usage, which all uses of the language rely on.”

4 An excerpt database of neologisms (focused on new lexical phenomena), a database of specialised vocabulary, the Pralex – preparatory lexical database and the Modern Czech lexical archives created in 1911–1991.

## 4 The Macrostructure of the Dictionary

The word list of the ASSČ is built using a different lexicographic technique than before.<sup>5</sup> It draws on a set of three balanced corpora, SYN 2000, SYN 2005 and SYN 2010. The entries are selected from an automatically generated word list mainly based on the frequency criterion and the criterion of the commonness of their usage (i.e. only widespread lexical units are included; specialised terms, professional and slang expressions etc. are included only selectively). On the other hand, the word list has been expanded on the basis of word-formation relations (members of word-formation groups) and on the basis of co-hyponymic and other relations (members of lexical-semantic classes have been added).

Unlike in earlier dictionaries (SSJČ, SSČ), *derivatives* (relational adjectives, adverbs, names of properties), which used to be added to the lemmas as run-on entries, are listed as separate entries now. The new method (including the explanation of the meaning and exemplification) makes it possible to give an adequate lexicographic description, which however requires a detailed, often demanding analysis, cf. the explanation of the meaning in the entry for the relational adjectives (*badatel* n. “researcher” → *badatelský* adj. “vztahující se k badateli, k badatelství • složený z badatelů • určený pro badatele, pro badatelství” =pertaining to researchers, researching • consisting of researchers • intended for researchers, for researching), see figure 1.

**badatelský** příd.  
vztahující se k badateli, k badatelství • složený z badatelů • určený pro badatele, pro badatelství; syn. vědecký: intenzivní badatelská práce; badatelské projekty; moderní badatelské přístupy; badatelské zaujetí; badatelský tým; nastupující badatelská generace; poskytovat badatelské a knihovnické služby; špičkové badatelské pracoviště  
□ *badatelský list* tiskopis sloužící k evidenci údajů o badateli a jeho výpůjčkách z knihovny, archivu ap.

Figure 1: Entry *badatelský*.

Only some lexical types are treated as *run-on entries*. These include words derived by adding feminine suffixes (*herečka* ← *herec* “actress ← actor”), diminutives (*pejsek* ← *pes* “doggie ← dog”) and frequentative verbs (*balívat* ← *balit* “to pack”), where the semantic structure of the derivative does not differ from the lemma. Cf. the lemma *bouček* “small beech” treated as a run-on entry in the entry *buk* “beech” (see figure 5).

In the ASSČ, greater autonomy has been given also to *multi-word lexical units*. The treatment distinguishes between: phraseological units (*balit si kufry* “to pack one’s bags”) and non-phraseological units (terminological – *akciová společnost* “joint-stock company”; non-terminological – *bílá technika* “white goods”; multi-word grammatical expressions – *bez ohledu na* “regardless of” preposition etc.). In the dictionary, these are listed in an one-word lemma entry, but it is taken into account that they are

5 The word list of the PSJČ relied on a comprehensive and sophisticated excerption of 5 million excerpts. The word list of the SSJČ built on the word list of the preceding dictionary, i.e. PSJČ, and its own excerption. Similarly, the latest of these modern dictionaries, the one-volume SSČ, proceeded from the word list of the SSJČ and its own excerption.



independent formal-semantic lexical units; therefore, the meaning explanation and exemplification are provided for a large part of them (always for phrasemes; for non-phraseological units, the explanation is given where the meaning is not compositional). The independence of multi-word lexical units is indicated also by the method of their presentation in the entry (a highlighted multi-word lemma, labelling with special symbols), see figure 2 and figure 3.

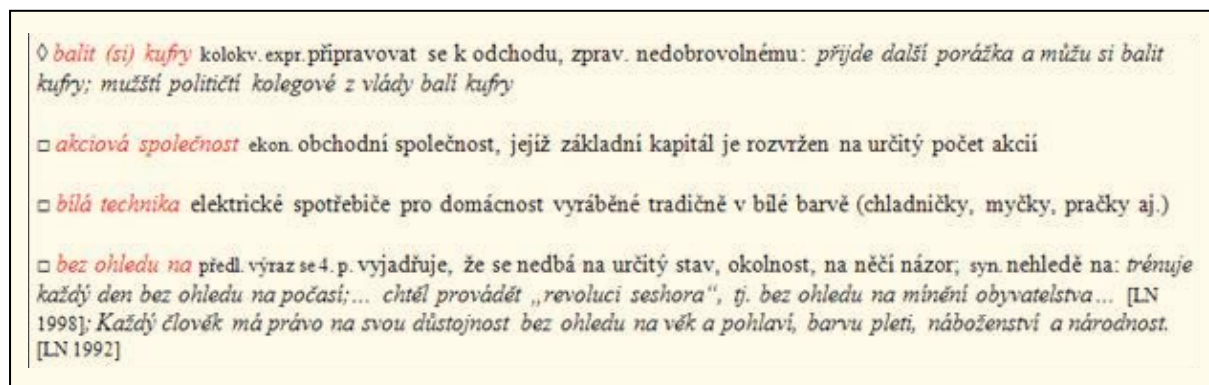


Figure 2: Multi-word lexical units *balit (si) kufry*, *akciová společnost*, *bílá technika* and *bez ohledu na*.

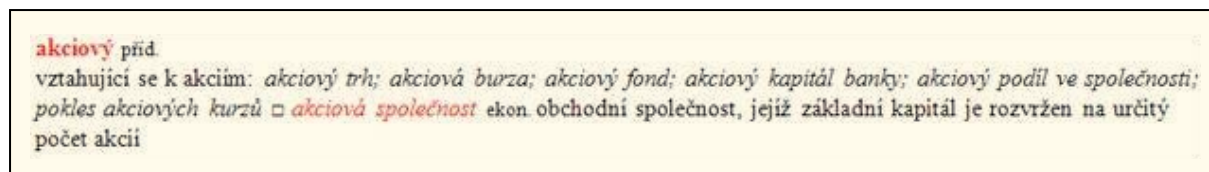


Figure 3: Multi-word lexical unit *akciová společnost* listed in the one-word entry *akciový*.

## 5 The Microstructure of an Entry

An entry in the ASSČ consists of the following parts: the lemma (including variant forms), information on homonyms, pronunciation, the etymology of the lexical unit, grammatical information (word class, morphology, valency), the usage label, the explanation of the meaning (including synonyms and antonyms), exemplification, notes (e.g. encyclopaedic information, further etymological information)<sup>6</sup> and cross-reference to (semantically, grammatically) related entries.

In the ASSČ, *grammatical information* (see figure 4) is treated more comprehensively than in previous monolingual dictionaries.<sup>7</sup> The morphological data in the ASSČ entries include mainly doublet forms and forms where the users may hesitate. The information on valency is systematically given for verbs

6 On the usage of notes, see e.g. the Oxford Dictionary of English (Soanes, Stevenson 2005), in Czech lexicography the neological dictionaries (Martincová et al. 1998, 2004).

7 In some respects, this transcends the genre of a general monolingual dictionary; on the other hand, it accommodates the users, who expect this type of information in a dictionary.



(both right and left valency), selectively also for nouns and adjectives. The valency information is semantically specified, if necessary, in the explanation of the meaning, or in the examples.

**bafat** (3. j. bafá, bafe, rozk. (ne)bafej!, čin. bafal, podst. jm. bafání) ned. expr.  
4. (kdo || ~) (zprav. o psu nebo jiné psovitě šelmě) vydávat jednotlivě vyřážené zvuky baf, haf; syn. štěkat: *pes bafal jako divý; Malý pokojový psík řafá podstatně vyšším hlasem, než jakým bafá mohutná doga.* [Týdeník Rozhlas 2010]

**Figure 4: Sense 4 of the lemma *bafat*.**

When giving the *lexical meaning* of the entries, the ASSČ proceeds from the basic concept of determining the species classification – genus proximum – and differential semantic elements – differentia specifica (bearing in mind that besides notional elements also pragmatic elements need to be described). A part of the lexical meaning, however, are also those semantic elements that cannot be considered as necessary distinctive features but which mirror the complex of information on the denoted extra-linguistic reality that the language users have on the level of common knowledge. To a certain extent, the explanation of the meaning may hence contain “encyclopaedic” data (especially those that are objectively reflected in the word-formation structure of a word, in set similes and other phrasemes and in semantically derived meanings, on the basis of a metaphor).<sup>8</sup> In order to eliminate circularity, the explanation by means of synonyms is limited to a minimum, only to some slang, expressive or dialectal words.

The *exemplification part* of an entry includes both typical examples illustrating typical usage and extended examples that show less common, unusual and sometimes even authorial use of the word (mainly in the case of less frequent words and those belonging to peripheral areas of vocabulary). In addition, the examples are to illustrate grammatical information (especially on valency) and demonstrate (semantic) collocability. The exemplification may further contain those connotations which are not included in the explanation of the meaning but which the user (proto)typically connects with the unit concerned.

8 Dolník (2012: 45); cf. Buzássyová, Jarošová (2006: 27–28).

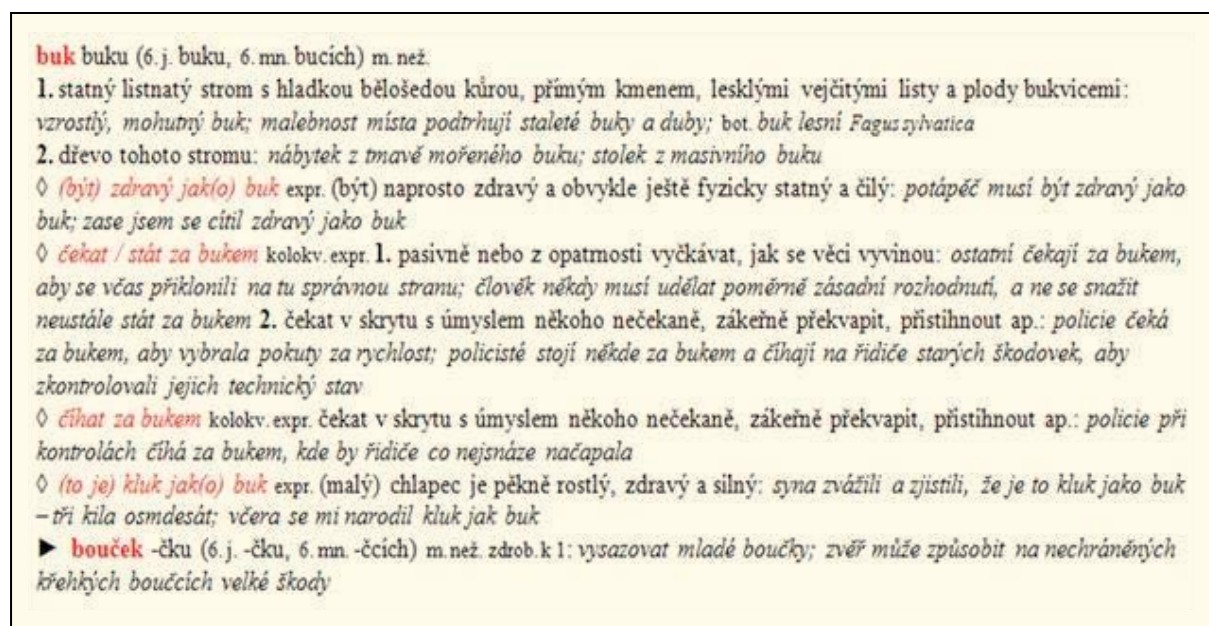


Figure 5: Entry *buk* “beech”.

## 6 Software

Unlike earlier monolingual dictionaries created in the Institute of the Czech Language, the ASSČ has been compiled since the very beginning by means of specialised lexicographic software for dictionary creation (DWS). After various possibilities were considered, it was decided that new, special software needs to be developed as the creation of the dictionary has its significant specifics and the programme must be flexible. The software development has received grant support from the Ministry of Culture of the CR within the National and Cultural Identity (NAKI) applied research and development programme. (For more details on the software, see Barbierik et al. 2013; Barbierik et al. 2014.)

## 7 Conclusion

In the creation of the ASSČ, we are seeking new paths for the resolution of the issues that lexicographers have always faced as well as those of the modern present. Although the preparation of a good monolingual dictionary is a Sisyphean task – “the pursuit of perfection in lexicography is doomed to constant failure” –,<sup>9</sup> it must be attempted. “A dictionary of the national language is one of the basic needs of an educated man.” (J. Jungmann, the preface to *Slovník česko-německý* (A Czech-German Dictionary)).

9 We borrowed the metaphor at the end from Hanks (2005: 254).

## 8 References

- Barbierik, K. et al. (2013). A New Path to a Modern Monolingual Dictionary of Contemporary Czech: the Structure of Data in the New Dictionary Writing System. In K. Gajdošová, A. Žáková (eds.), *Natural Language Processing, Corpus Linguistics, E-learning, Proceedings of the conference Slovko 2013, Bratislava, 13–15 November 2013*. Lüdenscheid: RAM-Verlag 2013, pp. 9–26.
- Barbierik, K. et al. (2014). Simple and Effective User Interface of Dictionary Writing System. In *Euralex 2014 Proceedings, Bolzano 15–19 July 2014*.
- Buzássyová, K., Jarošová, A. (eds.) (2006). *Slovník současného slovenského jazyka A–G*. (First edition.) Bratislava: Veda.
- Czech National Corpus – SYN2000*. Institute of the Czech National Corpus, Prague 2000. Accessed at: <http://www.korpus.cz> [07/04/2014].
- Czech National Corpus – SYN2005*. Institute of the Czech National Corpus, Prague 2005. Accessed at: <http://www.korpus.cz> [07/04/2014].
- Czech National Corpus – SYN2010*. Institute of the Czech National Corpus, Prague 2010. Accessed at: <http://www.korpus.cz> [07/04/2014].
- Czech National Corpus – SYN*. Institute of the Czech National Corpus, Prague. Accessed at: <http://www.korpus.cz> [07/04/2014].
- Čermák, F. (2010). Notes on Compiling a Corpus-Based Dictionary. In *Lexikos 20 (AFRILEX-reeks/series 20:2010)*, pp. 559–579.
- Čermák, F., Blatná, R. (eds.) (1995). *Manuál lexikografie*. Jinočany: H & H.
- Dolník, J. (2012). Lexikálna pragmatika. In K. Buzássyová, B. Chocholová & N. Janočková (eds.), *Slovo v slovníku. Aspekty lexikálnej sémantiky – gramatika – štylistika (pragmatika)*. Na počesť Alexandry Jarošovej. Bratislava: Veda, pp. 41–49.
- Hanks, P. (2005). Johnson and Modern Lexicography. In *International Journal of Lexicography*, 18(2), pp. 243–266.
- Hanks, P. (2010). Compiling a Monolingual Dictionary for Native Speakers. In *Lexikos 20 (AFRILEX-reeks/series 20:2010)*, pp. 580–598.
- Hladká, Z. (2007). Lexikografie. In J. Pleskalová, M. Krčmová et al. (eds.), *Kapitoly z dějin české jazykovědné bohemistiky*. Prague: Academia, pp. 164–198.
- Jungmann, J. (1835 (1834) – 1839). *Slovník česko-německý*. (5 vol.) Prague: Knížecí arcibiskupská knihtiskárna.
- Kilgariff, A. et al. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.), *Proceedings of the eleventh EURALEX International Congress EURALEX 2004 Lorient, France, July 6–10 2004*. Lorient: Université de Bretagne-Sud, pp. 105–116.
- Martincová, O. et al. 1998. *Nová slova v češtině. Slovník neologizmů 1*. Prague: Academia.
- Martincová, O. et al. 2004. *Nová slova v češtině. Slovník neologizmů 2*. Prague: Academia.
- Příruční slovník jazyka českého 1935–1957*. Prague: Státní pedagogické nakladatelství / SPN.
- Slovník spisovné češtiny pro školu a veřejnost* (1978). (Second, revised edition 1994; third, revised edition 2003.) Prague: Academia.
- Slovník spisovného jazyka českého* (1960–1971). (First edition.) Prague: Nakladatelství ČSAV.
- Soanes, C., Stevenson, A. (eds.) (2005). *Oxford Dictionary of English* (Second, revised edition.) Oxford: Oxford University Press.

### Acknowledgements

The article has been written within the grant project of the National and Cultural Identity (NAKI) applied research and development programme A New Path to a Modern Monolingual Dictionary of Contemporary Czech (DF13P01OVV011).



# Frame Semantics and Learner's Dictionaries: Frame Example Sections as a New Dictionary Feature

Carolin Ostermann  
Friedrich-Alexander Universität Erlangen-Nürnberg, Germany  
carolin.ostermann@fau.de

## Abstract

Frame semantics has so far been neglected or even been rejected in the context of EFL-lexicography, although lexicographic description within a frame semantics approach would have advantages for learners, e.g. the coherent presentation of several relevant lexical items at a time, as well as their conceptual connection, both of which would also further vocabulary acquisition. This proposal will detail how a frame semantics approach for the example section in English monolingual learner's dictionaries can contribute to the notion of cognitive lexicography, i.e. a lexicography that puts an emphasis on how users process language, which would in turn facilitate a user's understanding of an entry. For this purpose, so-called *frame example sections* were developed on agentive nouns (e.g. *bridegroom*, *plaintiff*); these are small coherent text passages that define and exemplify the noun in relation to its whole frame. The frame example sections mention related frame elements, collocating verbs and describe the typical scenario underlying a semantic frame, in order to promote decoding, i.e. understanding the meaning of lexical items, as well as encoding, i.e. learning words and finding related language material. The paper will be rounded off by presenting the results of a small user-study that was conducted on the frame example sections.

**Keywords:** frame semantics; learner's dictionaries; cognitive lexicography; user-study

## 1 Frame Semantics and Learner's Dictionaries

Frame semantics in Fillmore's terms (1982) has come to be a widely accepted notion of semantic description, and in relation to lexicography, it has inspired the FrameNet online project (cf. Fontenelle 2003). In traditional lexicography, however, the approach has been neglected so far and even deemed useless (Bublitz and Bednarek 2004: 50). This paper will, however, demonstrate how frame semantics can be used in English monolingual learner's dictionaries. The approach is part of the larger concept of cognitive lexicography (cf. Ostermann 2012 and fthc.), in which theories and semantic analyses of cognitive linguistics are used in common lexicographic practice in order to create dictionary features and entries which are more accessible to the dictionary user, since they use and describe language in the same way the users process it.

Frame semantics is a very useful tool for meaning description in lexicography: Fillmore and Atkins (1992, 1994, 2000) have demonstrated several times in how far a frame approach can help with distinguishing meanings of polysemous items (*'risk'*) and ensure a more realistic display of their different senses. This is one example of what Geeraerts (2007: 1168) refers to by stating generally that cognitive linguistics can enrich lexicography by a more realistic conception of semantic structure.

The feature proposed here aims at a more vivid exemplification of lexemes within the context of their frame, enabling the user to acquire new vocabulary from the frame and find important collocates for encoding, e.g. writing purposes. The feature replaces or complements example sentences in traditional dictionary entries as a so-called *frame example section (FE-section)*. In the following, the structure and composition of FE-sections will be outlined, illustrating how they fit into a dictionary entry while at the same time offering an onomasiological access to the dictionary's macro-structure. A few remarks on a user-study conducted will round the paper off.

## 2 Frame Example Sections

### 2.1 Theory and Structure

For the application of frame semantics to a traditional dictionary entry the example section has been selected. Example sentences are especially suitable for being replaced or supported by *FE-sections* since they do not carry the main burden of rendering meaning but complement the definition by showing the meaning in context and offering typical collocations (cf. Drysdale 1987: 218-222). Since the FE-section is a small coherent text passage on a lexical item and mentions the frame with its frame elements and most important collocations, it additionally allows the user to grasp the meaning better. Regarding its language, the style of FE-sections is natural and typical, informative and intelligible, as good examples are supposed to be (Atkins and Rundell 2008: 458). This generally follows Fillmore's demands (2003: 283) that we should define "not words but only families of words that jointly express one frame".

For the writing of FE-sections, a suitable lexeme has to be chosen; in addition to the lexical items treated, the frame semantic content for each FE-section has to be established. The relevant frame and its frame elements are determined by using elicitation techniques, i.e. simple questions such as 'who', 'where', 'what', 'which aim?'. Superordinate place and collocating verbs are determined, and information from FrameNet is taken into consideration if the frame also figures there. Authentic language material is also collected from the BNC web, especially for collocations and related lexical items. Rundell (1988: 135) observed here very early that "(...) any account in a learner's dictionary of the word *problem* should at the very least mention as significant collocates the verbs *pose* and (especially) *solve*" and this can be ensured by an analysis of authentic language material. The FE-sections are written with the help of this collective input. Once the text has been produced, various perspectives in ac-

cordance with the various frame elements are created in order to be able to enter the FE-section at all the lexemes in the dictionary that are part of the frame. Finally, the potential for a related frame is checked, i.e. synonyms, antonyms or related semantic fields. The figure below summarizes the process of writing FE-sections.

<p>SET-UP OF FRAME EXAMPLE SECTIONS:</p> <ol style="list-style-type: none"> <li>1. Choice of the lemma: person-denoting noun.</li> <li>2. Identification of the frame and frame elements.</li> <li>3. Collection of authentic language material from the BNC, esp. of collocations.</li> <li>4. Writing of the (main) frame example section with its annotations.</li> <li>5. Check for perspectives of the frame example section.</li> <li>6. Check for semantic ‘spin-offs’, i.e. related frames.</li> <li>7. Decision of places to enter in the dictionary (in line with perspectives).</li> </ol>
---

**Table 1: Set-up of frame example sections.**

## 2.2 The Set of Frame Example Sections

The FE-sections developed in this proposal centre on so-called person-denoting nouns. These are nouns that occupy an agentive slot in a frame, denoting a person and its habitual activities, and therefore provide a good perspective as a start, especially since they comprise actions and objects, as well as people or places that interact. The table below lists all the lexemes with their respective frame for which FE-sections have been produced. These 17 lexemes can also be divided into three groups: EVENT-frames (where something happens, usually starting with a preposition of time), ACTIVITY-frames (starting with *when* and introducing the setting of the frame), and PLACE-frames (taking place at typical locations).

<p><i>bridegroom</i> WEDDING ▪ <i>caretaker</i> BUILDING ▪ <i>conductor</i> ORCHESTRA ▪ <i>conductor</i> TRAIN ▪ <i>landlord</i> RENT ▪  <i>librarian</i> LIBRARY ▪ <i>mayor</i> CITY ▪ <i>midwife</i> BIRTH ▪ <i>pawnbroker</i> MONEY ▪ <i>plaintiff</i> COURT ▪ <i>striker</i> FOOTBALL                  ▪ <i>surgeon</i> OPERATION ▪ <i>suspect</i> POLICE ▪ <i>umpire</i> SPORT ▪ <i>undertaker</i> FUNERAL ▪ <i>usher</i> PERFORMANCE ▪  <i>waiter</i> RESTAURANT</p>
--

**Table 2: Person-denoting nouns and their frames.**

For the lexeme *bridegroom*, the FE-section is reproduced below with annotations: the WEDDING-frame is an event frame, i.e. one where something happens. *Bridegroom*, *bride*, *husband* and *wife* are frame-constitutive elements, i.e. those which are necessary to understand the frame, and are printed in small capitals. Frame-supportive elements, i.e. those which are optional for an understanding of the frame and rather expand it, here *priest/pastor*, *church*, *reception*, are underlined. Collocations (on the wedding day, to get married) are given a dotted underlining. The full annotations including sources of au-

thentic language material (here from BNC web and FrameNet) and perspectives for the FE-section on *bridegroom* can be found in the appendix, as well as the FE-sections for all the other items.

<b><i>bridegroom</i> - WEDDING</b>
On their WEDDING day, the BRIDE and the BRIDEGROOM get married and become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
<p>ANNOTATION:</p> <p>[On<sup>Coll</sup> their wedding day]<sup>Event</sup>, the bride<sup>Partner1/WhoColl</sup> and the bridegroom<sup>Partner 2/Who</sup> get married<sup>CollActivity</sup> and become<sup>[change relationship] / Goal</sup> husband and wife<sup>Partners</sup>. A priest or pastor in church<sup>where</sup> traditionally marries them with family and friends present. Afterwards, there often is a wedding reception<sup>Coll</sup>.</p>

Table 3: FE-section for *bridegroom*.

### 2.3 A Cognitive Macro-structure

Many of the person-denoting nouns are rather rare items (cf. *pawnbroker*, *plaintiff*, *usher*) or only supposedly transparent lexical items (*caretaker*, *landlord*), which makes them very interesting in a language-learning perspective. It would be ideal if, once these items are looked up, they would become attached to a user’s mental lexicon, possibly via familiar material and links within their frame. This is also the reason why the various perspectives of FE-sections are written and the FE-sections should be entered repeatedly at the entries of all their participating lemmata.

In this way, all the person-denoting nouns also contribute to a macrostructure that exhibits more links between single items than traditional dictionaries do, and one that also allows for onomasiological access. Every FE-section spans a small net over the macro-structure with its single frame elements; all FE-sections together span an even larger net since they often share elements or deal with polysemy (cf. the two FE-sections for *conductor*).

This is also in accordance with Geeraerts’ assumption (2007: 1169) that “Cognitive Linguistics may also suggest ways of dealing with the links between the senses of lexical items that go beyond common practice”. If we suppose that the FE-section is – whether incorporated within the dictionary entry or in a box nearby – clearly delimited regarding its layout (e.g. use of colours, etc.), it almost automatically leads the user to related entries, especially since the same information of one frame can be found in all the places of the frame in the dictionary. In an electronic, online or CD-ROM-version of a dictionary, this could even be achieved more effectively by hyperlinking. FE-sections therefore also fulfill lexicographically the function of signposts (the capital print of the frame as a meaning indication via synonym, cf. DeCesaris 2012) and of component-internal implicit cross-references (cf. Svensén 2009: 388 and 391), in which many entries of one frame, but also across frames are linked.



### 3 A User-Study on Frame Example Sections

#### 3.1 Methodology

In order to determine the usefulness of FE-sections, a small-scale user-study was conducted with 50 university students of English. The hypothesis was that in a two-part production-oriented primed vocabulary task, the group of students in the target group ( $n^t=25$ ) who received dictionary entries of the respective lexemes, complemented by FE-sections, would perform better than those in the control group ( $n^c=25$ ) who worked with traditional dictionary entries only.

The participants received in the first part of the experiment a randomised reading booklet with the LDOCE5-dictionary entries of 12 of the above-mentioned lexemes as a prime (two groups of six items: *caretaker, midwife, pawnbroker, plaintiff, umpire, usher* and *conductor<sup>1</sup>, conductor<sup>2</sup>, landlord, striker, surgeon, undertaker*). The participants in the target group worked with reading-booklets in which the dictionary entries were complemented by the FE-sections; the participants in the control group received dictionary entries complemented by reading material on the lexemes taken from the BNC, so that both groups had the same amount of reading material to master. On each page of the booklet, they found one entry and were supposed to read it carefully within ca. 25 seconds, turning the page only when being told to do so and not going backwards. This session was devised as primed input for the second part of the experiment, which followed after a break of approximately 45 minutes. In this second part, the test subjects received a worksheet on the 12 person-denoting nouns, on which they were supposed to give a German translation of each noun, define it in their own words and tick off in a list whether they had known the word before.

It must be noted generally, however, that the hypothesis could not be verified, since the experiment yielded inconclusive, statistically non-significant results.

#### 3.2 Results and Discussion

Regarding the knowledge of the test items, it can be concluded that the test was conducted in a homogeneous group with approximately the same level of knowledge of all the items across the participants. The items from the first group, such as *pawnbroker, plaintiff, usher* and *umpire*, were rated very low and were fairly unknown, whereas the items from the latter six received higher ratings.

For the results of the translation task (reproduced in the chart below), the scores of correct translations were counted for each item in both groups and compared; the significance of difference was checked with the help of the  $\chi^2$ -test. The numbers of correct translations are approximately equal for all items, with the exceptions of *pawnbroker* and *landlord*, which proved to be statistically significant ( $\chi^2= 2.01, p<0.20$  and  $\chi^2= 3.57, p<0.10$ ). It should be noted, however, that many students seemed to have had problems in coming up with a good translation, since a certain number of the participants suggested e.g. German 'Torschütze' instead of 'Torjäger' for *striker* and did not even seem to be aware of the

semantic difference. Therefore, demanding a German translation might not have been the best measure, as it yielded problems of its own, even when the concepts behind the lexical items were apparently understood, since in many cases, correct paraphrases were given.

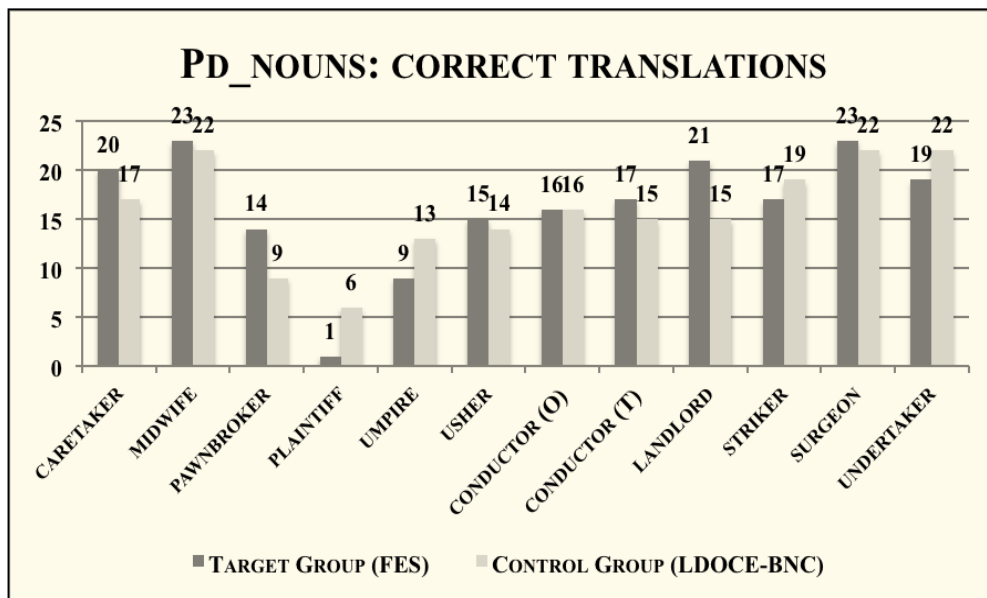


Figure 1: Results of translation task

In order to evaluate the results of the paraphrasing task, a point system was devised. Points were assigned in the participants' paraphrases to a correct paraphrase in general, to the frame mentioned and to all the frame elements reproduced. Generally, the participants in the target group scored higher for each item and in general (36.22 points on average compared to 29.34 points for the participants in the control group), and their paraphrases were also longer (14.97 words on average compared to 11.48 words in the control group). It has to be admitted, however, that there is a certain correlation between the amount of input and output, which, on the other hand, admits the conclusion that more input in the form of FE-sections is indeed beneficial. It should be noted that a learning effect (i.e. when people indicated that they had not known the word before but gave a correct definition) could be achieved more often in the target group and that the number of paraphrases given compared to the number of correct paraphrases given was equal in more instances in the target group. The non-transparent item *landlord* ('Vermieter' in German, but its parts often translated literally as 'Landherr' / 'Lehensherr') caused fewer misunderstandings among the participants in the target group, the effect of which can also be attributed to the cognitive FE-sections. Therefore, the FE-sections did score an effect, even if it was small and statistically not significant.

Overall, it can be concluded that the complexity of the task probably made it difficult to measure the effect that FE-sections can have. The reading time might not have been sufficient for vocabulary acquisition, especially since "lexical acquisition is not immediate" (Béjoint 1988: 145), and vocabulary items will not get a real foothold in one's mental lexicon through decoding alone (Atkins and Rundell

2008: 410). The more difficult the items were (e.g. *plaintiff* compared to the simpler *midwife*), the poorer the results were, or the more blanks could be found on the worksheets; only single instances of a better performance with one item or another, or cases of real vocabulary acquisition in the target group could be ascertained. Possibly, the wealth of information in the FE-sections also hindered immediate acquisition with difficult items. These effects could be elucidated in another test condition or in a longer testing phase with repeated tasks or dictionary training of the participants.

## 4 Conclusion

All in all, it can be concluded that FE-sections are a new approach for EFL-lexicography which would probably work best in an individual look-up situation. Although no superior results over traditional dictionary entries could be proven statistically, the benefits still come into play, and this is one step on the way to a more cognitive and more onomasiological dictionary of encyclopaedic nature.

## 5 References

- Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Béjoint, H. (1988) Psycholinguistic evidence and the use of dictionaries by L2 learners. In: M. Snell-Hornby (ed.). *ZüriLEX '86 Proceedings*. Tübingen: Francke. pp. 139-148.
- Bublitz, W. and M. Bednarek. (2005) Nur im begrenzten Rahmen. *Frames im Wörterbuch*. In: T. Herbst, G. Lorenz, B. Mittmann, and M. Schnell (eds.) *Lexikographie, ihre Basis- und Nachbarwissenschaften. (Englische) Wörterbücher zwischen "common sense" und angewandter Theorie*. Tübingen: Niemeyer. pp. 35-52.
- DeCesaris, J. (2012) On the Nature of Signposts. In: *Proceedings of the 15th EURALEX International Congress*. Universitetet i Oslo. pp. 532-540.
- Drysdale, P.D. (1987) The role of examples in a learner's dictionary. In: A. P. Cowie (ed.). *The dictionary and the language learner: papers from the EURALEX seminar at the University of Leeds, 1-3 April 1985*. Tübingen: Niemeyer. pp. 213-223.
- Fillmore, Ch. J. (1975) An Alternative to Checklist Theories of Meaning. In: C. Cogen, H. Thompson, G. Thurgood, K. Whistler and J. Wright (eds). *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society. pp. 123-131.
- 2003. Double-Decker Definitions: The Role of Frames in Meaning Explanation. In: *Sign Language Studies* 3(3): pp. 263-295.
- Fillmore, Ch. J. and B.T.S. Atkins (1992) Toward a Frame-Based Lexicon: The Semantics of RISK and its Neighbors. In: A. Lehrer and E. Feder Kittay (eds.). *Frames, Fields and Contrasts. New Essays in Semantic and Lexical Organization*. Hillsdale, New Jersey / London: Erlbaum. pp. 75-102.
- 1994. Starting where the dictionaries stop: The challenge of Corpus Lexicography. In: B.T.S. Atkins and A. Zampolli (eds.). *Computational Approaches to the Lexicon*. Oxford: Oxford Univ. Press. pp. 349-393.
- 2000. Describing Polysemy: The Case of 'Crawl'. In: Y. Ravin and C. Leacock (eds.). *Polysemy. Theoretical and Computational Approaches*. Oxford: Oxford University Press. pp. 91-110.
- Fontenelle, Th. (ed.) (2003) Special issue on FrameNet and frame semantics. In: *International Journal of Lexicography* 16(3).

- FrameNet*. Accessed at: <https://framenet.icsi.berkeley.edu> [10/04/2014].
- Geeraerts, (2007) *Lexicography*. In: D. Geeraerts and H. Cuyckens (eds.). *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press. pp. 1160-1174.
- LDOCE5 = *Longman Dictionary of Contemporary English*. 5<sup>th</sup> ed. 2009. Ed. director Michael Mayor. Harlow: Pearson – Longman.
- Ostermann, C. 2012. Cognitive Lexicography of Emotion Terms. In: *Proceedings of the 15th EURALEX International Congress*. Universitetet i Oslo. pp. 493-501.
- fthc. Cognitive Lexicography. In: S. Niemeier and C. Juchem-Grundmann, with D. Schönefeld (eds.). *Dictionaries of Linguistics and Communication Science (WSK) online, Vol 14: Cognitive Grammar*. Berlin: Mouton de Gruyter. [online. 1 page].
- Rundell, M. (1988) Changing the rules: Why the monolingual learner's dictionary should move away from the native-speaker tradition. In: M. Snell-Hornby (ed.). *ZüriLEX `86 Proceedings*. Tübingen: Francke. pp. 127-137.
- Svensén, B. (2009) *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.
- The BNC web*. Accessed at: <http://bncweb.lancs.ac.uk> [10/04/2014].

**Appendix 1:** An Annotated Example for *bridegroom*

<b>1. Lemma:</b> <i>bridegroom</i>	
<b>2. Frame:</b> WEDDING	
2. Frame elements	bride, bridegroom, husband, wife, church, priest / pastor ⇒ superordinate place: church ⇒ collocating verb: marry ⇒ kind of frame: EVENT
2.a Elicitation techniques	⇒ who, where, activity, goal?
2.b FrameNet Frame FEs from FN FrameNet definition	Forming_relationships ⇒ Partner 1, Partner 2, Partners; Epistemic stance ⇒ Partner 1 interacts with Partner 2 (also collectively called Partners) to change their social relationship.
<b>3. Authentic language material</b>	
Collocations from BNC	to get married; (on their) wedding day, wedding reception, bride
<b>4. Frame example section</b>	
On their WEDDING day, the BRIDE and the BRIDEGROOM get married and become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.	
ANNOTATION	
[On <sup>Coll</sup> their wedding day] <sup>Event</sup> , the bride <sup>Partner1/WhoColl</sup> and the bridegroom <sup>Partner 2/Who</sup> get married <sup>CollActivity</sup> and become <sup>[change relationship] / Goal</sup> husband and wife <sup>Partners</sup> . A priest or pastor in church <sup>where</sup> traditionally marries them with family and friends present. Afterwards, there often is a wedding reception <sup>Coll</sup> .	
<b>5. Different perspectives</b>	
BRIDE	On their WEDDING day, the BRIDE gets married to her BRIDEGROOM and they become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
GROOM	On their WEDDING day, the BRIDEGROOM gets married to his BRIDE and they become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
WIFE	On their WEDDING day, the BRIDE and the BRIDEGROOM got married and became HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
HUSBAND	On their WEDDING day, the BRIDE and the BRIDEGROOM got married and became HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
<b>6. Semantic spin-off</b>	
antonym	divorce
<b>7. Place(s) in the dictionary</b>	
wedding ■ bridegroom ■ bride ■ husband ■ wife	

**Table 4: Full annotation for *bridegroom*.**

## Appendix 2: The set of frame example sections

bridegroom WEDDING	On their WEDDING day, the BRIDE and the BRIDEGROOM get married and become HUSBAND and WIFE. A priest or pastor in church traditionally marries them with family and friends present. Afterwards, there often is a wedding reception.
caretaker BUILDING	In a public BUILDING, e.g. a school, a CARETAKER (or also JANITOR) is the person who looks after the BUILDING. S/he takes care of the BUILDING's maintenance and makes sure that everything is in order, that broken things are repaired or that rules are obeyed. The CARETAKER usually has his or her own OFFICE in the BUILDING where s/he can be found.
conductor TRAIN	In a TRAIN, a CONDUCTOR (or also GUARD) is responsible for checking and collecting or also selling the PASSENGERS' TICKETS; s/he furthermore is in charge of the train, making sure everything is in order or answering the passengers' questions. Conductors also travel on BUSES where they collect the fare.
ORCHESTRA	When an ORCHESTRA or CHOIR performs, either as a rehearsal or in front of an AUDIENCE, a CONDUCTOR stands in front on a podium and conducts, i.e. directs the MUSICIANS' PERFORMANCE with a baton (small thin stick). The MUSICIANS follow the CONDUCTOR'S movements so that all play in a coordinated way and the PERFORMANCE sounds good.
landlord RENT	When you RENT a PLACE TO LIVE, i.e. an apartment/flat or a house, you pay MONEY, the RENT, to the LANDLORD who owns the building and lets you live there. You are then the TENANT and a formal contract, the lease, guarantees your rights as a TENANT.
librarian LIBRARY	In a LIBRARY, a LIBRARIAN is the person who is in charge of running the institution, i.e. lending BOOKS or other materials to LIBRARY users. People can read the BOOKS there or they can borrow them. Schools and universities usually have their own libraries and their use is often free of charge.
mayor CITY	In a CITY or TOWN, the MAYOR is the head of the local GOVERNMENT. S/He is elected directly by the citizens and resides in a city or town hall. S/he fulfils official duties and functions and makes decisions in local politics.
midwife BIRTH	When a pregnant WOMAN goes into labour and is about to give BIRTH to a BABY, she usually goes to hospital. There, she gets help from a MIDWIFE, who is a nurse helping women to get through labour pains and who also takes care of the MOTHERS and their BABIES before and after birth.
pawnbroker MONEY	When you are in urgent need of MONEY, but cannot or do not want to borrow money from a bank, you may turn to a PAWNBROKER in a PAWNSHOP. S/he will lend you money in exchange for valuable OBJECTS, e.g. jewellery or electronic devices. If you cannot pay back the MONEY after a certain while, the pawnbroker will sell what you have PLEDGED.
plaintiff COURT	In COURT, a PLAINTIFF brings a CASE against another person, the defendant. The PLAINTIFF is usually supported by a LAYWER (in Britain a solicitor in the lower courts of law) to fight the case successfully, and the judge or a jury has to decide on the verdict.
striker FOOTBALL	In a FOOTBALL MATCH, the STRIKER is the PLAYER whose main task on the PITCH it is to score a GOAL and help his team to win, which the other team's PLAYERS and especially the goalkeeper try to prevent.
surgeon OPERATION	During an OPERATION, a SURGEON is the doctor who cures and rescues PATIENTS by performing surgery, i.e. by operating on patients in a HOSPITAL in an OPERATING THEATRE with nurses and other doctors assisting. Patients who undergo surgery are seriously ill and usually stay in hospital to recover.
suspect POLICE	When the POLICE think that a person took part in a CRIME, they arrest this person, who is a SUSPECT. After the ARREST, the SUSPECT is taken into custody at the POLICE STATION for a police interview / an interrogation.
umpire SPORT	During a SPORTS COMPETITION in an arena, an UMPIRE is the person who makes sure that RULES are obeyed. There is an UMPIRE present in e.g. baseball, tennis, cricket, hockey, or athletics COMPETITIONS; s/he also calls the score, decides on penalties, starts races, or reports irregularities to chief UMPIRES (depending on the discipline).
undertaker FUNERAL	After somebody's DEATH, a FUNERAL is held at a CEMETERY. AN UNDERTAKER or FUNERAL DIRECTOR prepares the deceased person's burial or cremation and arranges the FUNERAL service, so that people can attend the ceremony and mourn the loss of the deceased.
usher PERFORMANCE / EVENT	When people go to see a public PERFORMANCE OR EVENT, e.g. in a theatre, a cinema, a concert hall, or a sports stadium, they show their TICKETS to an USHER (or USHERETTE) who shows them their SEATS or even guides them there. Often, the USHER also keeps order during a show.
waiter RESTAURANT	In a RESTAURANT, people sit at TABLES and eat a MEAL for which they have to pay the bill at the end. A WAITER or WAITRESS brings customers the MENU first and later serves the food they ordered.

Table 5: The Set of Frame Example Sections.

# Translating Action Verbs using a Dictionary of Images: the IMAGACT Ontology

Alessandro Panunzi<sup>°</sup>, Irene De Felice<sup>\*</sup>, Lorenzo Gregori<sup>°</sup>, Stefano Jacoviello<sup>†</sup>, Monica Monachini<sup>\*</sup>, Massimo Moneglia<sup>°</sup>, Valeria Quochi<sup>\*</sup>, Irene Russo<sup>\*</sup>

<sup>°</sup>University of Florence, <sup>\*</sup>ILC CNR (Pisa), <sup>†</sup>University of Siena

alessandro.panunzi@unifi.it, irene.defelice@ilc.cnr.it, lorenzo.gregori@unifi.it,  
stefano.jacoviello@gmail.com, monica.monachini@ilc.cnr.it, moneglia@unifi.it,  
valeria.quochi@ilc.cnr.it, irene.russo@ilc.cnr.it

## Abstract

Action verbs have many meanings, covering actions in different ontological types. Moreover, each language categorizes action in its own way. One verb can refer to many different actions and one action can be identified by more than one verb. The range of variations within and across languages is largely unknown, causing trouble in all translation tasks. IMAGACT is a corpus-based ontology of action concepts, derived from English and Italian spontaneous speech corpora, which makes use of the universal language of images to identify the different action types extended by verbs referring to action in English, Italian, Chinese and Spanish. This paper presents the IMAGACT search interface and the various kinds of linguistic information the user can derive from it. IMAGACT makes explicit the variation of meaning of action verbs within one language and allows comparisons of verb variations within and across languages. Because the action concepts are represented with videos, extension into new languages beyond those presently implemented in IMAGACT is done using competence-based judgments by mother-tongue informants, without intense lexicographic work involving underdetermined semantic descriptions.

**Keywords:** Action verbs; Image ontology; Multilingual dictionary; Computer-aided translation

## 1 Introduction

In all language modalities, action verbs bear the basic information that should be understood in order to make sense of a sentence. Moreover when we communicate, we have to refer to actions very often. Native speakers do not have a problem finding the right verb for a specific action in their own language. However, in a foreign language, they often have difficulty choosing the appropriate verb. The reason is that the more common action verbs, in their own meaning, refer to many different actions: in this sense, they are “general” verbs. Moreover, each language categorizes actions in its own way. These facts imply that there are not one-to-one translation relationships between different general verbs in different languages (Majid et al. 2007; Kopecka & Narasimhan 2012). If we take the English verb *to*

*cross*, for instance, we could argue that it can refer to at least two different action types, as in the sentences:

- (1) John crosses the street
- (2) John crosses his arms

On the contrary, in Italian we must use two different verbs to translate the previous sentences, namely *attraversare* (for *crossing the street*) and *incrociare* (for *crossing arms*):

- (3) Gianni attraversa la strada
- (4) Gianni incrocia le braccia

The problem is a significant one because reference to action is very frequent in ordinary spoken communication (Moneglia & Panunzi 2007) and specifically high-frequency verbs can each refer to many different action types (Moneglia in press).

The IMAGACT project has now delivered a corpus-based language ontology covering the set of actions most frequently referred to in everyday language. Using English and Italian spoken corpora, we have identified 1010 distinct action concepts and visually represented them by means of prototypical scenes, either animated (3D) or filmed (Moneglia et al. 2012; Frontini et al. 2012). The cross-linguistic correspondences to action concepts of 521 Italian verbs and 550 English verbs (i.e., the verbal lexicon most likely to be used when referring to action) are stored in a database. The action concepts in IMAGACT have already been extended to Chinese and Spanish (included in the first IMAGACT release). Perhaps more importantly, the action concepts can be easily identified by speakers of any language, since they are represented in an ontology of animated and filmed scenes.

This paper presents the IMAGACT online interface and how queries are made to the database. The user can search in IMAGACT in three main ways: a) as a bilingual dictionary, based on concept selection; b) through explicit comparison of the range of actions that can in principle be referred to by two lexical entries, of the same language or of different languages; c) through the direct selection of an action concept in the gallery of prototypic scenes, independently of the language of the user. In the last section, the paper also introduces an initiative aimed at the extension of the IMAGACT database to other languages.

## 2 Dictionary

If the user wonders how an English action verb translates into Italian or into another target language (Spanish and Chinese in the IMAGACT first release), IMAGACT can be used as a multilingual dictionary of images. Figure 1 shows the thumbnail images of the main types of actions identified by the English verb *to cross*.



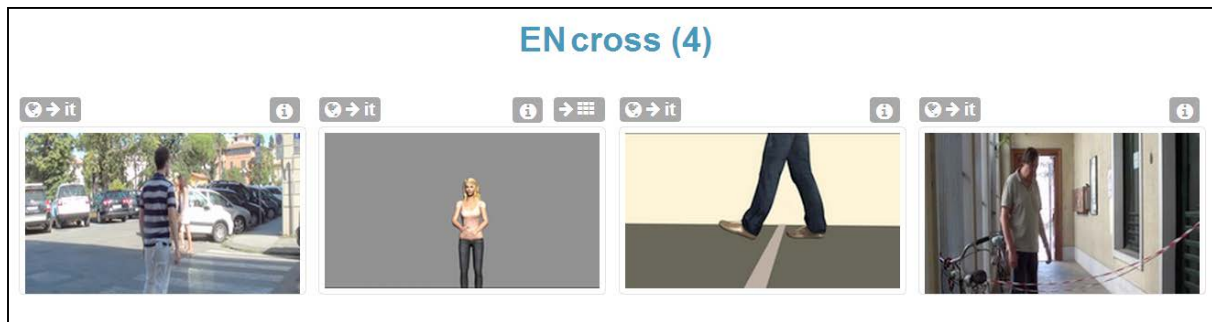


Figure 1: The variation of *to cross* across action types.

Looking at the various action types this verb expresses, the user can:

- select the action type he is interested in
- look at the animation to clarify the referred action
- see how this action is identified in the target language

IMAGACT returns one main verb and an additional set of verbs which equally identify this specific type of action. For each scene, which represents a distinct action type, Italian gives different translation for this verb, as shown in Figure 2.



Figure 2: The cross-linguistic relation of verb(s) to action types.

### 3 Comparison

The user can compare verbs that in principle should translate between each other from two different languages. Searching with this function, the system illustrates the set of action types in which both verbs can be respectively applied. The result of such a search for *to cross* and *attraversare* (see Figure 3) supports the intuition that the two verbs can translate to each other, at least with respect to some of the action types they can refer to. At the same time, however, the system shows which actions can be indicated by one verb but not by the other, and *vice versa*. As a consequence, the difference between the

Italian verb *attraversare* and the English verb *to cross* becomes explicit. The Italian user will learn that, in English, *to cross* cannot be applied to the types on the right column in Figure 3. In this case, he can go directly to the English translation of verb *attraversare*, as shown in Figure 4: for these two actions he has to use, respectively, *to traverse / to pass* and *to stab / to pierce*.

Comparison between two verbs can also be requested within the same language, to allow the user exploring more deeply the differences in meaning between the lexical entries suggested by the system. For instance, an English user can learn that both the verbs *passare* and *attraversare* can be applied to the action type illustrated by the scene on the left column, second row, in Figure 4. The user may wonder what the difference is between the two Italian lemmas suggested by the system. So, he can then compare the two verbs (of the same target language), clarifying the differences between their referential properties (Figure 5).

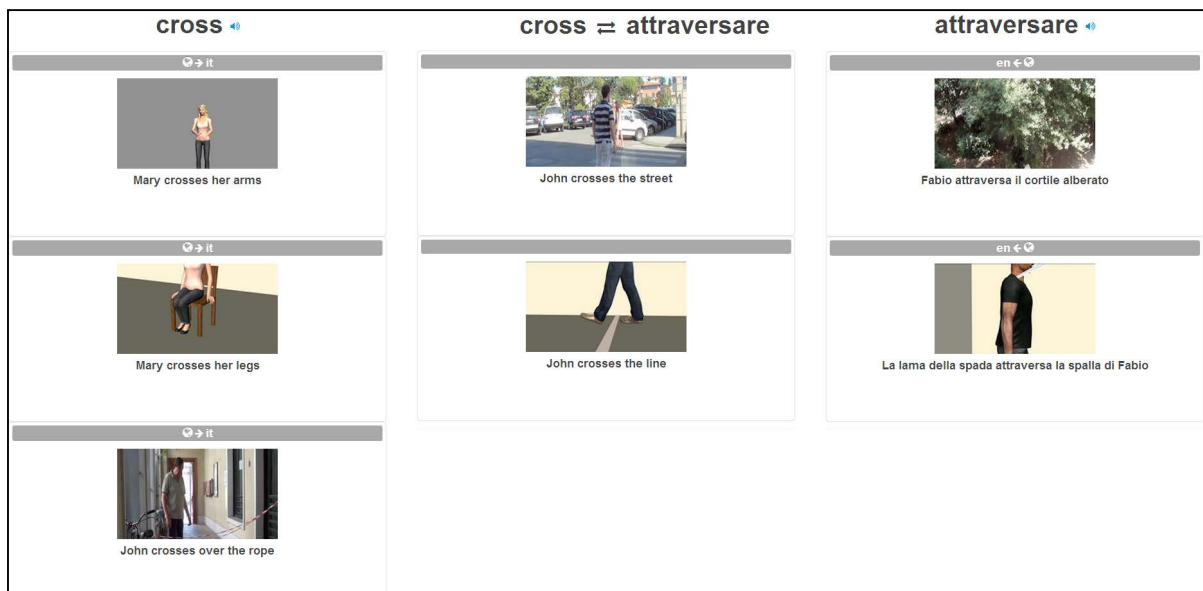


Figure 3: Comparison of *turn* vs. *girare* (results of the query interface with graphic adaptations).



Figure 4: From comparison to linguistic categorization.



Figure 5: Intra-linguistic comparison.

## 4 Gallery

If the language of the user is not represented in IMAGACT, he can use the system directly as a gallery of scenes. This may be of special interest to users who speak minority languages.

The system works through the selection of one “meta-category” of action among the ones proposed by the interface. Such meta-categories are represented by a series of 3D animations, which are continuously played in loop, as the thumbnails in Figure 6 suggest.

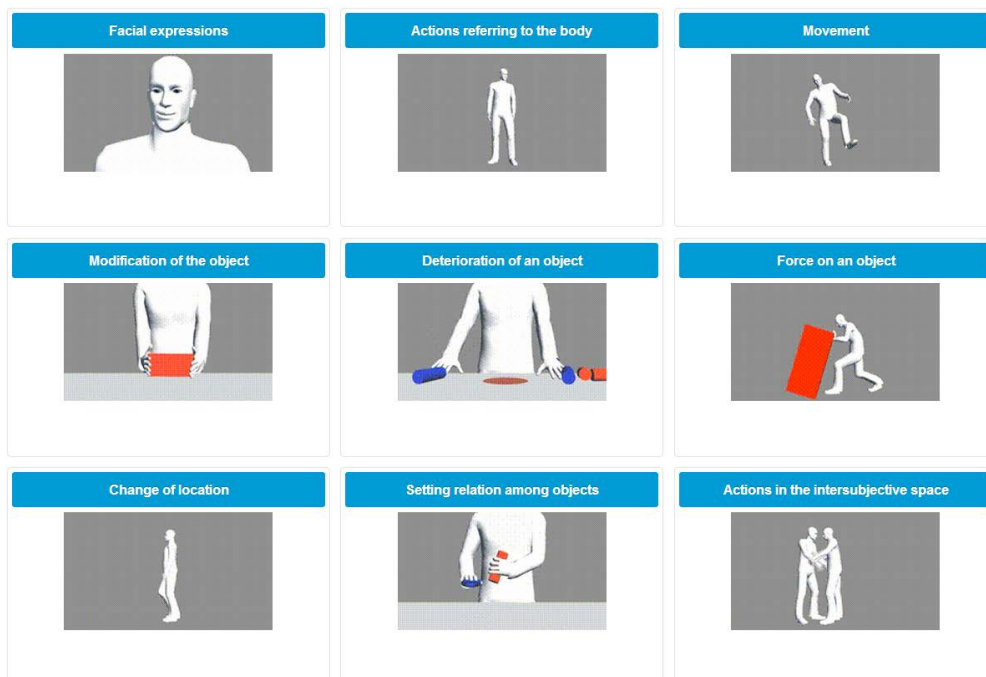


Figure 6: Representation of action meta-categories through avatars.

The numerous actions covered by IMAGACT are gathered into 9 macro-classes, which have high relevance in human categorization of action. Meta-categories are ordered according to criteria which take into account the informative focus of the action, as reported in Table 1.

Perspective centered on the Actor	Perspective centered on the Actor-Theme relation	Perspective centered on the Theme-Destination relation
Actions referring to facial expression	Modifications of the object	Change of location of the object
Actions referring to the body	Deterioration of the object	Setting relations among objects
Movement in space	Forces on the object	Actions in inter-subjective space

**Table 1: Criteria for meta-categories.**

The user can figure out what kind of action these stand for by looking at the abstract representation heading each class, and of course through a quick look at the actions gathered under each one. The user identifies the action he is interested in independently of the word he has for that action in his own language; after choosing the action via its visual representation, he is able to reach its linguistic categorization in the required target language. From this point of view, the IMAGACT gallery reverses the ordering of the dictionary: it goes from concepts to language instead of from language to concepts.

Once the user has understood the meaning of the action groups, it will be easier to search for the specific action he is interested in. He will click on one scene in the gallery headed by one category and get the linguistic categorization of the concept in one of the possible target languages in IMAGACT. For instance, Figure 7 is what the system returns when asked for the Chinese verb for the action corresponding to the verb *to cross* under the category *Actions referring to the body* (i.e., *crossing the arms*).



**Figure 7: From gallery to linguistic categorization (Chinese).**

## 5 Extending the dictionary

Because IMAGACT’s direct representation of actions through scenes can be interpreted independently of language, the system allows the mapping of lexicons from different languages onto the same cross-linguistic ontology. On this basis, it is possible to ask mother-tongue informants which verb(s) in their languages should be applied to each scene, thus extending the ontology to any language (IMAGACT4ALL).

In the simplified interface for the Competence Based Extension of the IMAGACT database to other languages (called *CBE light*), the set of action concepts represented by the IMAGACT prototypic scenes is assumed as a fixed-reference universe, and the work starts directly from such scenes.

An informant receives the action types as input. Figure 8 shows the interface the informant would use for processing one action type and how this has been done in the case of Chinese. The interface presents the informant with the scene prototype and the matching English and Italian verbs derived from corpus analysis. The informant assesses the action represented in the video and provides the verb or verbs in his language that can be used to refer to that specific action.

Lemmas are annotated in its citation form, as it is commonly reported in dictionaries, in the box corresponding to his language. For each lemma he then writes in the caption box a simple sentence in the present tense, third-person singular, filling all the arguments of the verb that properly describes the action. This sentence will be used as the caption of the scene in the language of the informant.

Both the verb and the caption should be written in the current writing system of the language of the informant. If this system does not use Latin characters, the informant also provides the verb and its caption in Latin characters, as can be seen for Chinese.

Corpus verbs	Type	Lang.	Caption	
fold	PRO		Mary folds her arms	
cross	INST		Mary crosses her arms	
incrociare	INST		Marta incrocia le braccia	



Assigned verbs					
Verb	Transliteration	Rejected	Lang.	Caption	Transliterated caption
交叉	jiāo chā	<input type="checkbox"/>		李娜在胸前交叉双臂	lǐ nà zài xiōng qián jiāo chā shuāng bì
抱	bào	<input type="checkbox"/>		李娜把双臂抱在胸前	lǐ nà bǎ shuāng bì bào zài xiōng qián

**Figure 8: Simplified Competence Based Extension (CBE light).**

Given that verbs with different meanings can identify the same action, the informant is asked to find multiple lemmas allowed by his language for each action. However, simply viewing one short clip may be not sufficient to elicit all the alternatives. The infrastructure provides one simple means to stimulate the thinking of the informant. More specifically, corpus-based annotation generated English and Italian alternatives that fit with the represented scene. These verbs will function as sugge-

sions for figuring out alternatives in the language of the informant. Therefore, after the first lemma has been determined, the annotator is requested to judge whether or not the alternatives suggested have translations in his language, translations that can be used in referring to the event in question. If so, he will report a new verbal lemma and a new caption by adding a line to his language options. The work of the informant must be supervised by a mother-tongue expert linguist before the language is mapped onto the IMAGACT database. More specifically, an annotation can be rejected by the supervisor during revision if considered inappropriate. Spanish and Chinese have already been implemented through IMAGACT4ALL, and various initiatives are currently being pursued to extend the database to a number of different languages.

## 6 References

- Frontini, F., De Felice, I., Khan, F., Russo, I., Monachini, M., Gagliardi, G., Panunzi, A. (2012). Verb Interpretation for Basic Action Types: Annotation, Ontology Induction and Creation of Prototypical Scenes. In M. Zock, R. Rapp (eds) *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, Mumbai, 15 December 2012. Red Hook (NY): Curran Associates, pp. 69-80.
- IMAGACT. Accessed at: <http://www.imagact.it> [04/04/2014].
- Kopecka, A., Narasimhan, B. (2012). *Events of Putting and Taking, A Cross-linguistic perspective*. Amsterdam/Philadelphia: John Benjamins.
- Majid, A., Bowerman, M., van Staden, M., Boster, J. S. (2007). The semantic categories of cutting and breaking events: A crosslinguistic perspective. In *Cognitive Linguistics*, 18(2), pp. 133-152.
- Moneglia, M. (in press). The semantic variation of action verbs in multilingual spontaneous speech corpora. To appear in T. Raso, H. Mello (eds) *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins.
- Moneglia, M., Monachini, M., Calabrese, O., Panunzi, A., Frontini, F., Gagliardi, G., Russo, I. (2012). The IMAGACT Cross-linguistic Ontology of Action. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis (eds) *Proceedings of Eighth Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, 23-25 May 2012. Paris: ELRA, pp. 2606-2613.
- Moneglia M., Panunzi A. 2007. Action Predicates and the Ontology of Action across Spoken Language Corpora. The Basic Issue of the SEMACT Project. In M. Alcántara Plá, Th. Declerck (eds) *Proceeding of the International Workshop on the Semantic Representation of Spoken Language, SRSL7, at Conferencia de la Asociación Española de Inteligencia Artificial, CAEPIA 2007, Salamanca, 12-16 November 2007*. Salamanca: Universidad de Salamanca, pp. 51-58.

### Acknowledgements

The IMAGACT project is funded by the PAR/FAS program of the Tuscan region in Italy. Further research on the IMAGACT database, including this paper, is being accomplished with the contribution of the MODELACT project (2013-2016), funded by the national programme Futuro in Ricerca.

# Degrees of Synonymity as the Basis of a Network for German Communication Verbs in the Online Reference Work *Kommunikationsverben* in OWID

Kristel Proost, Carolin Müller-Spitzer  
Institut für Deutsche Sprache, Mannheim  
proost@ids-mannheim.de, mueller-spitzer@ids-mannheim.de

## Abstract

This contribution presents the procedure used in the *Handbuch deutscher Kommunikationsverben* and in its online version *Kommunikationsverben* in the lexicographical internet portal OWID to divide sets of semantically similar communication verbs into ever smaller sets of ever closer synonyms. *Kommunikationsverben* describes the meaning of communication verbs on two levels: a lexical level, represented in the dictionary entries and by sets of lexical features, and a conceptual level, represented by different types of situations referred to by specific types of verbs. The procedure starts at the conceptual level of meaning where verbs used to refer to the same specific situation type are grouped together. At the lexical level of meaning, the sets of verbs obtained from the first step are successively divided into smaller sets on the basis of the criteria of (i) identity of lexical meaning, (ii) identity of lexical features, and (iii) identity of contexts of usage. The stepwise procedure applied is shown to result in the creation of a semantic network for communication verbs.

**Keywords:** communication verbs; speech act verbs; synonymity; synonymy; conceptual field; semantic network; access structures; advanced search options

## 1 Kommunikationsverben in OWID

This contribution deals with the synonymy relations of German communication verbs and the way in which they were used to create a semantic network for these verbs in the *Handbuch deutscher Kommunikationsverben* (cf. Harras et al. 2004, Harras/Proost/Winkler 2007) and in its online version *Kommunikationsverben*, which has recently been integrated into the lexicographical internet portal OWID ('Online Wortschatz- und Informationssystem Deutsch') of the Institut für Deutsche Sprache. In both the print and the online reference work, the meaning of German communication verbs is described on two levels: a conceptual and a lexical level. The distinction between these two levels of meaning derives from two-levels-semantic (cf. Bierwisch & Lang 1989, Bierwisch & Schreuder 1992, Lang 1994). On the conceptual level, communication verbs are described as referring to different types of situations in which a speaker utters something to a hearer. The situation types referred to by communication verbs are represented as consisting of several components relating to the attitudes of the speaker,

to properties of the propositional content of the speaker's utterance and to specific aspects of the situation in question. Verbs used to refer to the same specific situation type constitute a "paradigm" or conceptual field. On the lexical level, communication verbs belonging to the same field are differentiated with respect to their lexical meaning and their lexical features. Verbs which are identical with respect to their lexical meaning are subsumed under the same lemma and hence appear in the same dictionary entry. Of a set of verbs having the same lexical meaning, only one is lemmatised – usually the one which is least specific with respect to its contexts of usage – while the others are listed as synonyms of the verb lemmatised. *Kommunikationsverben* contains about 800 verbs, 241 of which are lemmatised and appear with an entry of their own. All other verbs are listed as synonyms of the verbs lemmatised. On the whole, *Kommunikationsverben* lists 170 fields of German communication verbs.

In *Kommunikationsverben* in OWID, the conceptual and the lexical level of the meaning of communication verbs have each been implemented in different types of access structures (cf. Müller-Spitzer & Proost 2013). Particularly, the online version provides some advanced search options allowing the user (i) to combine components of situations to "create" many different situation types and find the verbs matching them, and (ii) to search for verbs sharing a smaller or larger number of lexical features.

Since the conceptual and the lexical level of the meaning of communication verbs are each associated with different degrees of semantic specification, verbs grouped together on each of these two levels are synonymous to different degrees. In this contribution, we will show that the notion of the graded nature of synonymy may be used to divide sets of semantically similar communication verbs into ever smaller sets of increasingly closer synonyms, a procedure which ultimately results in the creation of a semantic network for communication verbs. By providing the two advanced search options, not available in the print version, the online version facilitates the user's access to the different degrees of similarity in meaning among synonymous communication verbs, thereby enhancing the structure of *Kommunikationsverben* as a semantic network.

## 2 Synonymy as a Graded Feature

Synonymy is a relation of similarity or identity of meaning among the senses of different lexical items (cf. Cruse 1986: 267; Cruse 2002: 486). Since similarity of meaning is a matter of degree, different types of synonymy relations have been distinguished, depending on the degree of similarity of the senses of the lexical items compared. Absolute synonymy involves complete identity of meaning and forms one end-point on the scale of synonymy (cf. Cruse 1986: 268). All other types of synonymy proposed encompass not only similarity of meaning, but also some degree of semantic difference between the senses of two or more lexical items. Difference in meaning is involved, for example, in the relation between propositional synonyms (e.g. *begin-commence*) and that between plesionyms or near-synonyms (e.g. *giggle-chuckle*), the difference between these two types of synonym being that substitution of one item by the other yields sentences with equivalent truth-conditions in the case of the



former but not in that of the latter (cf. Cruse 1986: 270-289; Cruse 2002: 489-490). On the scale of synonymy, propositional synonymy occupies a position in between that of absolute synonymy and that of plesionymy. The latter shades off into non-synonymous difference of meaning, which constitutes the zero-point on the scale of synonymy (cf. Cruse 1986: 268).

### 3 The Meaning of Communication Verbs

#### 3.1 Communication Verbs

Communication verbs are verbs used to refer to different types of situations in which a speaker (henceforth: S) utters something to a hearer (henceforth: H). In the default case, the speaker's utterance also contains a proposition (henceforth: P). Some but not all of these verbs lexicalise combinations of speaker attitudes such as the speaker's propositional attitude, i.e. the attitude of the speaker towards the proposition of his/her utterance, the speaker's intention and the speaker's presuppositions. This smaller set of communication verbs is called "speech act verbs" (cf. Proost 2006: 65; 2007: 8-9). Examples of German speech act verbs include *behaupten* ('assert'), *mitteilen* ('inform'), *lügen* ('lie'), *auffordern* ('demand'), *versprechen* ('promise'), *loben* ('praise'), *kritisieren* ('criticise'), *schimpfen* ('scold'), and *klagen* ('complain'). Examples of German communication verbs which are not part of the narrower set of speech act verbs in the sense outlined above are *sagen* ('to say'), *sprechen* ('to speak'), *brüllen* ('to scream'), *unterbrechen* ('to interrupt'), and *faxen* ('to fax'). *Kommunikationsverben* focuses on speech act verbs.

#### 3.2 Representing the Meaning of Communication Verbs

##### 3.2.1 The Conceptual Level of the Meaning of Communication Verbs

All situations referred to by communication verbs are characterised by the presence of four features or situational roles: a speaker, a hearer, a set of speaker attitudes, and an utterance (mostly) containing a proposition. Since these four elements are part of any situation referred to by communication verbs, they constitute the unifying feature of the meaning of these verbs (cf. Verschueren, 1980: 51-57; 1985: 39-40; Wierzbicka, 1987: 18; Harras et al. 2004: Introduction; Proost, 2006: 651). The type of situation referred to by all speech act verbs is therefore called the 'general resource situation type'.

Two of the roles of the general resource situation type, the role of the speaker attitudes and that of the utterance, may be specified in different ways. The role of the speaker attitudes may be specified as consisting of the speaker's attitude to the proposition of his/her utterance, the speaker's intention, and the speaker's presuppositions. The speaker's propositional attitude may be further specified as S's taking P to be true, S's knowing P, S's wanting P, S's evaluating P positively or negatively, and so on. Specifications of the speaker's intention include S's intention to make H recognise S's propositional attitude (for example, to make H recognise that S knows P or takes P to be true) or to get him/her to do

something. The speaker’s presuppositions may concern an attitude of H (whether H takes something to be true, whether he/she knows something), the interests of S and H concerning P (whether P is in the interest of S or in the interest of H), or properties of P (for example, whether P is the case). The role of the utterance is specified by properties of the propositional content. These include the event type of P (whether P is an action, event, or state of affairs), the temporal reference of P (whether P precedes, coincides with, or follows the time of S’s uttering P) and, in the case that P is an action, the agent of P (S, H, S & H, and so on).

Different combinations of specifications of the different types of speaker attitudes and of the properties of the propositional content constitute special resource situation types. These are referred to by distinct types of verbs. For example, verbs like *mitteilen* (‘inform’), *lügen* (‘lie’) and *loben* (‘praise’) and their synonyms are used to refer to the situation types characterised by the specifications listed in Tables 1-3:

<b>Special Resource Situation Type: Representatives. Information. mitteilen</b>	
Propositional Content (P)	
Event Type	not specified
Temporal Reference	not specified
Agent	not specified
Speaker Attitudes	
Propositional Attitude	S knows: P
Intention	S wants: H know: P
Presuppositions	H does not know: P

**Table 1: Situation type referred to by *mitteilen* (‘inform’), *informieren* (‘inform’), *instruieren* (‘advise’) and *unterrichten* (‘advise’).**

<b>Special Resource Situation Type: Representatives.Assertives.lügen</b>	
Propositional Content (P)	
Event Type	not specified
Temporal Reference	not specified
Agent	not specified
Speaker Attitudes	
Propositional Attitude	S does not take to be true: P
Intention	S wants: H recognise: S takes to be true: P
Presuppositions	H does not know: P

**Table 2: Situation type referred to by lügen ('lie'), schwindeln and flunkern (both 'fib') and their prefixed forms anlügen ('lie to sb. '), belügen ('lie to sb. '), erlügen ('lie about sth. '), rumlügen ('tell lies'), vorlügen ('lie to sb. about sth. '), anflunkern ('fib to sb. '), rumflunkern ('tell fibs'), vorflunkern ('fib to sb. about sth. '), anschwindeln ('fib to sb. '), beschwindeln ('fib to sb. '), rumschwindeln ('tell fibs'), vorschwindeln ('fib to sb. about sth. ').**

<b>Special Resource Situation Type: Expressives.evaluative.positive.loben</b>	
Propositional Content (P)	
Event Type	action
Temporal Reference	past
Agent	H of 3 <sup>rd</sup> person
Speaker Attitudes	
Propositional Attitude	S considers: P good
Intention	S wants: H recognise: S considers: P good
Presuppositions	P is the case

**Table 3: Situation type referred to by loben ('praise'), huldigen ('pay tribute to'), ehren ('honour'), würdigen ('acknowledge') and honorieren ('appreciate').**

The combinations of the specifications of the speaker attitudes and of the properties of the propositional content lexicalised by verbs like *mitteilen*, *lügen* and *loben*, respectively, may also be conceived of as the concepts lexicalised by these verbs. Thus, *mitteilen* lexicalises the concept of a verbal action performed by a speaker who knows P and assumes that H does not know P with the intention that H know P, P being an action, event or state of affairs preceding, co-occurring with or following the time of S's utterance. The information in Table 2 captures the idea that verbs like *lügen* express the concept of a verbal action whereby a speaker who does not take P to be true and assumes that H does not know P intends the hearer to recognise that he/she – i.e. the speaker – takes P to be true. The verb *loben* lexicalises the concept of a verbal action performed by a speaker who evaluates P, a past action by H or a 3<sup>rd</sup> person, positively and intends the hearer to recognise this.

Verbs which are used to refer to the same special resource situation type constitute a “paradigm” or conceptual field. With respect to the examples in Tables 1-3, this means that the sets {*mitteilen, informieren, instruieren, unterrichten*}, {*lügen, anlügen, belügen, erlügen, rumlügen, vorlügen, flunkern, anflunkern, schwindeln, anschwindeln, beschwindeln, rumflunkern*} and {*loben, huldigen, ehren, würdigen, honorieren*} each represent a conceptual field.

### 3.2.2 Methods Used to Describe the Conceptual Level of the Meaning of Communication Verbs

Following a procedure proposed by Baumgärtner (1977: 260-264), the different specifications of the role of the speaker attitudes and the role of the utterance as well as the lower-level specifications of each of these are obtained from a comparison of sentences containing speech act verbs. The well-formedness of some of these and the ill-formedness of others show which elements are relevant to the meaning of the verbs they contain. For example, a comparison of the sentences in (1) and (2) shows that *to order* lexicalises the values ‘future’, ‘action’ and ‘hearer’ for the specifications of the temporal reference, the event type and the agent of P, respectively, while *to promise* lexicalises the values ‘future’, ‘action’ and ‘speaker’, respectively, for these specifications:

- (1) a. I order youi to PROi leave the room.  
 b. \*I order youi to PROi have left the room.  
 c. \*I order youi for mej to PROj leave the room.
- (2) a. Ii promise you to PROi leave the room.  
 b. \*Ii promise you to PROi have left the room.  
 c. \*Ii promise youj to PROj leave the room.

The introspective analysis exemplified in (1)-(2) has shown that the higher-level specifications of the speaker’s propositional content, the speaker’s intention, the speaker’s presuppositions and the propositional content are essential aspects of the meaning of speech act verbs. These four aspects correspond to five of the seven components of illocutionary force which Searle & Vanderveken (1985: 12-20) and Vanderveken (1990: 103-136) have argued to determine the conditions under which a particular type of speech act is both successful and non-defective. Particularly, the aspect of the speaker’s propositional attitude corresponds to the component of the sincerity conditions, the aspect of the speaker’s intention to the component of the illocutionary point, the aspect of the speaker’s presuppositions to the components ‘mode of achievement of the illocutionary point’ and ‘preparatory conditions’, and the aspect of the propositional content to the component of the propositional content conditions (cf. Harras 2001: 26-31, Proost 2006: 654-655).

While the higher-level specifications of the speaker’s propositional attitude, the speaker’s intention, the speaker’s presuppositions and the propositional content are obtained from the type of analysis exemplified in (1)-(2), the lower-level specifications of each of these are calculated systematically, i.e. irrespective of any existing lexicalisations. For example, the specification ‘temporal reference of P’ is

assumed to have the specifications 'Past', 'Present' and 'Future', the specification of the event type of P the specifications 'action', 'state' and 'event', and so on. The question of which values are lexicalised by a particular verb was decided on the basis of examples from the Mannheim German Reference Corpus DeReKo ("Deutsches Referenzkorpus"). Methodological issues are dealt with in detail in the introductions to both volumes of the *Handbuch deutscher Kommunikationsverben* (cf. Harras et al. 2004, Harras 2007), which are also available in the online version.

### 3.2.3 The Lexical Level of the Meaning of Communication Verbs

Verbs which belong to the same conceptual field but differ from each other with respect to their lexical meaning appear with an entry of their own. In the *Handbuch deutscher Kommunikationsverben* and its online version in OWID, lexical meanings were differentiated on the basis of examples from the IDS-corpora of written German. All other verbs are listed as synonyms of the verbs lemmatised. With respect to the *lügen*-field, this means that *lügen* ('lie') and *schwindeln* ('fib') each appear with a separate entry. These verbs differ from each other in that *schwindeln* but not *lügen* expresses an evaluation by a discourse situation speaker, i.e. a speaker who uses this verb to comment on the utterance of the resource situation speaker. Particularly, a speaker who uses the verb *schwindeln* to refer to the resource situation speaker's act of lying thereby indicates that he/she does not consider S's act of lying to have serious consequences for H. In the *Handbuch deutscher Kommunikationsverben* and its online version *Kommunikationsverben* in OWID, this difference in the lexical meaning of *lügen* and *schwindeln* is reflected by the meaning paraphrases of these verbs in their respective entries. Since the evaluation expressed by *schwindeln* is an evaluation by a discourse situation speaker, it is not an element of the resource situation referred to by this verb. Hence, within the framework of *Kommunikationsverben*, it is not part of the conceptual component of its meaning. Rather, it is an essential part of the lexical meaning of this verb. With respect to the *lügen*-field, this means that this contains two lemmatised verbs, *lügen* and *schwindeln*. The verb *flunkern* is subsumed under the lemmatised verb *schwindeln*, because it has the same lexical meaning.

Verbs which belong to the same field *and* have the same lexical meaning are differentiated with respect to the following lexical features: (i) expression of thematic roles, (ii) syntactic realisation of the thematic roles, (iii) passivisation, (iv) resultativity, (v) evaluation by a discourse situation speaker (a speaker describing the speech act performed by the reference situation speaker), (vi) polysemy, (vii) performativity (the possibility for a verb to be used performatively), and (viii) stylistic markedness. Each member of a field is characterised as possessing or lacking each of these features as exemplified for the verb *lügen* und its prefixed forms *anlügen*, *belügen*, *rumlügen* and *vorlügen* by the screenshot in Figure 1:

Lexikalische Merkmale								
Verben	Merkmale							
	Seman- tische Rollen	Argu- ment Struktur	Passiv	Resulta- tivität	Bewer- tung	Poly- semie	Performa- tivität	stilistische Markiert- heit
lügen	H (block)		+	-	-	-	-	-
	P (block)							
anlügen	H (obl)	NP<Akk>	+	-	-	-	-	-
	P (block)							
belügen	H (obl)	NP<Akk>	+	-	-	-	-	-
	P (block)							
erlügen	H (block)		+	-	-	-	-	-
	P (obl)	NP<Akk>						
rumlügen	H (block)		+	-	-	-	-	+
	P (block)							
vorlügen	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE Inf						

**Fig. 1: Lexical Features of lügen ('lie') and its prefixed forms. “block” (“blocked”) means that the thematic role in question either cannot be realised at all or can be realised only by a prepositional phrase headed by vor ('before') or by an adpositional phrase headed by gegenüber ('in front of'); “obl” (obligatory) means that the thematic role in question must be realised.**

In addition to being differentiated with respect to their lexical features, verbs with the same lexical meaning may be distinguished with respect to their typical contexts of usage. Information on the range of contexts the non-lemmatised verbs may occur with is provided in the section *Kommentar* ('commentary') in the dictionary entry of the corresponding lemmatised verb. *Schwindeln* and *flunkern*, for example, are identical with respect to the specific type of situation they are used to refer to and regarding their lexical meaning but differ with respect to the contexts in which they are typically being used. Particularly, *flunkern* is used more frequently than *schwindeln* when reference is made to situations involving children telling lies, as illustrated in (1):

(1) Fast jeder fünfte Schüler (19 Prozent) verschweigt seinen Eltern schlechte Noten. 32 Prozent der Kids flunkern, wenn es allgemein um das Thema Schule geht. [Frankfurter Rundschau, 03.02.1999] [‘Almost every fifth pupil (19 Percent) keeps quiet to his parents about bad marks. 32 percent of the kids fib when the topic school is dealt with in general.’]

Because of this restriction on the range of contexts in which it is typically used, *flunkern* is not lemmatised and hence does not appear with an entry of its own. Rather, it is mentioned as a synonym of the verb *schwindeln*, which is less restricted than *flunkern* with respect to its contexts of usage and is therefore lemmatised and appears with an entry of its own.

## 4 Criteria for the Synonymy of Communication Verbs

As shown in the previous section, both the *Handbuch deutscher Kommunikationsverben* and its online version *Kommunikationsverben* in OWID describe communication verbs on different levels of analytical

detail. Communication verbs may be grouped together on each of these levels. Depending on the analytical level on which they are grouped together, communication verbs may be regarded as being synonymous in either a broader or a narrower sense. As an illustration of how the different criteria apply, they will be explained with respect to the verbs of the *lügen*-field, which has been introduced in the previous section. Additional examples will be discussed in section 5.

#### 4.1 The Criterion for Synonymy in the Broader Sense: Membership in the Same Field

On the lowest level of specification, verbs which are used to refer to the same special resource situation type and hence constitute a field in the sense outlined in section 3.2.1 may be regarded as synonyms in a broader sense. The corresponding criterion for synonymy on this level is membership within the same conceptual field. This means that all of the verbs mentioned underneath Tables 1-3 are synonyms in a broader sense. The degree of synonymy relating these lexical items is low. Membership within the same conceptual field is the minimum requirement for communication verbs to be considered synonyms at all. All other criteria for synonymy concern the lexical level of meaning and/or restrictions of usage. Verbs grouped together by these criteria are synonyms in a narrower sense.

#### 4.2 Criteria for Synonymy in the Narrower Sense

##### 4.2.1 Identity of Lexical Meaning

The first criterion relating to the lexical level is identity of lexical meaning. When applied to the verbs of the *lügen*-field, this criterion groups together *schwindeln* and *flunkern* as synonyms, distinguishing them from *lügen* by virtue of the fact that they both express an evaluation by a discourse situation speaker not part of the meaning of the latter verb (see section 3.2.3).

##### 4.2.2 Number of Shared Lexical Features

The degree of synonymy of communication verbs is additionally determined by the number of their shared lexical features. For example, *anlügen* and *belügen* are identical regarding all lexical features including their argument structure properties (see Figure 1). By contrast, *anlügen* and *belügen* on the one hand and *lügen* on the other differ in their argument structure properties while being identical with respect to all other lexical features. Specifically, *lügen* blocks the realisation of the roles of H and P while *anlügen* and *belügen* both obligatorily realise the role of the hearer as an NP in the accusative case and block the realisation of the role of P (see Figure 1). Due to these differences in the argument structure properties of *anlügen* and *belügen* on the one hand and *lügen* on the other, the degree of synonymy between the former two verbs is higher than that between either of them and *lügen*.

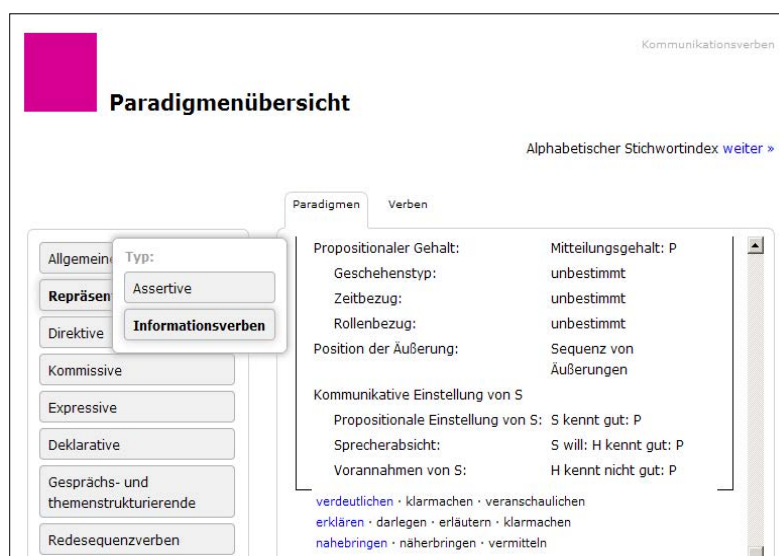
### 4.2.3 Substitutability *salva veritate*

The verbs *schwindeln* and *flunkern* are identical with respect to (i) the specific type of situation they are used to refer to, (ii) their lexical meaning, and (iii) their lexical features. As discussed in section they differ merely with respect to the contexts in which they are typically used: *flunkern* is used more frequently than *schwindeln* when reference is made to situations involving children telling lies as illustrated by example (1). Since substitution of *flunkern* by *schwindeln* in (1) does not yield a sentence with different truth-conditions, *flunkern* and *schwindeln* are substitutable *salva veritate*. Substitutability *salva veritate* is commonly regarded as an essential condition for propositional or cognitive synonymy. For a more detailed discussion of this particular type of synonymy, see Harras (2007b: 329-365).

## 5 Some Applications

### 5.1 Example I: Representatives of the Type ‘verdeutlichen’ (‘explain’)

Different degrees of synonymity may also be observed among the verbs of the field containing the lemmatised verbs *verdeutlichen* (‘explain’), *erklären* (‘explain’) and *nahebringen* (‘bring sth. home to sb.’). These verbs and the synonyms of each of these all express the concept of a verbal action whereby a speaker who knows something (P: a past, present or future action, event or state of affairs) well and assumes that H does not have sufficient knowledge of P makes several utterances to make H know P well. In *Kommunikationsverben* in OWID, information about special resource situation types is represented in the section *Paradigmenübersicht* (‘overview of paradigm’). The screenshot in Figure 2 represents the special resource situation type referred to by *verdeutlichen*, *erklären* and *nahebringen* and the synonyms of each of them:



**Fig. 2: Situation type referred to by *verdeutlichen* (‘explain’), *erklären* (‘explain’) and *nahebringen* (‘bring sth. home to sb.’) and their synonyms.**



Since the verbs *verdeutlichen* (‘explain’), *klarmachen* (‘make sth. clear to sb.’), *veranschaulichen* (‘illustrate’), *erklären* (‘explain’), *darlegen* (‘explain’), *erläutern* (‘explain’), *nahebringen* (‘bring sth. home to sb.’), *näherbringen* (‘bring sth. home to sb.’) and *vermitteln* (‘pass on knowledge’) are all used to refer to the same special resource situation type, they are synonyms in a broader sense.

On the lexical level of meaning, *verdeutlichen*, *erklären* and *nahebringen* are differentiated on the basis of their lexical meaning as follows:

- *verdeutlichen*: ‘to make sth. more comprehensible; to explain the crucial aspects of an issue or problem to sb. in order to make that person understand this issue or problem well’. Since *klarmachen* and *veranschaulichen* have the same lexical meaning as *verdeutlichen*, these three verbs are synonyms in a narrower sense.
- *erklären*: ‘to represent difficult and/or complex facts exactly and comprehensibly to sb. in order to make that person understand them well’. The verbs *darlegen*, *erläutern* and *klarmachen* are listed as having the same lexical meaning. *Erklären*, *darlegen*, *erläutern* and *klarmachen* may therefore be regarded as synonyms in a narrower sense.
- *nahebringen*: ‘to make sb. familiar with sth., usually with knowledge concerning a specific field, in order to arouse that person’s interest’. Since *näherbringen* and *vermitteln* are listed as having the same lexical meaning, these two verbs and *nahebringen* may be considered synonyms in a narrower sense.

On a more detailed level of analysis, verbs which are synonymous in as far as they have the same lexical meaning may be further differentiated by their lexical features. As indicated by Figures 3 and 4, *verdeutlichen*, *klarmachen* and *veranschaulichen* on the one hand and *nahebringen*, *näherbringen* and *vermitteln* on the other are completely identical with respect to all of their lexical features, including the syntactic realisation of their arguments:

Lexikalische Merkmale								
Verben	Merkmale							
	Semantische Rollen	Argument Struktur	Passiv	Resultativität	Bewertung	Polysemie	Performativität	stilistische Markiertheit
<b>verdeutlichen</b>	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
<b>klarmachen</b>	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
<b>veranschaulichen</b>	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						

Fig. 3: Lexical features of *verdeutlichen*(‘explain’), *klarmachen* (‘make sth. clear to sb.’) and *veranschaulichen* (‘illustrate’).

Verben	Merkmale							
	Seman- tische Rollen	Argu- ment Struktur	Passiv	Resulta- tivität	Bewer- tung	Poly- semie	Performa- tivität	stilistische Markiert- heit
<b>nahebringen</b>	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
<b>näherbringen</b>	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
<b>vermitteln</b>	H (obl)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						

**Fig. 4: Lexical features of nahebringen, näherbringen (both ‘bring sth. home to sb.’) and vermitteln (‘pass on knowledge’).**

Since all of the verbs mentioned in Figures 3 and 4 are identical with respect to all of their lexical features, they are very close synonyms.

Within the set {*erklären, darlegen, erläutern, klarmachen*}, *darlegen, erläutern* and *klarmachen* are also identical with respect to all of their lexical features. Like the verbs of the two other sets mentioned above, these three verbs are therefore synonymous to a very high degree. Since *erklären* differs from each of the three other verbs in that it is polysemous while the others are not, the degree of synonymity between this verb and each of the other three verbs is lower than among the other three verbs. Figure 5 lists the lexical features of *erklären* and its synonyms *darlegen, erläutern* and *klarmachen*:

Verben	Merkmale							
	Seman- tische Rollen	Argu- ment Struktur	Passiv	Resulta- tivität	Bewer- tung	Poly- semie	Performa- tivität	stilistische Markiert- heit
<b>erklären</b>	H (fak)	NP<Dat>	+	-	-	+	-	-
	P (obl)	NP<Akk> SE						
<b>darlegen</b>	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
<b>erläutern</b>	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						
<b>klarmachen</b>	H (fak)	NP<Dat>	+	-	-	-	-	-
	P (obl)	NP<Akk> SE						

**Fig. 5: Lexical features of erklären (‘explain’), darlegen (‘explain’), erläutern (‘explain’) and klarmachen (‘make sth. clear to sb.’)**

## 5.2 Example II: Expressives of the Type ‘klagen’ (‘complain’).

None of the verbs which are part of the field represented in Figures 2-5 share any special contextual restrictions on the basis of which they may be claimed to be even closer synonyms. Synonymy relations of this kind may be observed from a comparison of the contextual restrictions associated with the use of the verbs *klagen* (‘complain’), *jammern* (‘moan’) and *lamentieren* (‘lament’). These verbs are used to refer to situations in which a speaker who feels sorrow because of something (P: a past action, event or state of affairs) makes one or more utterances with the intention that the hearer recognize that he/she, i.e. the speaker, feels sorrow because of P. The situation referred to by *klagen* and its synonyms is represented by the screenshot in Figure 6:

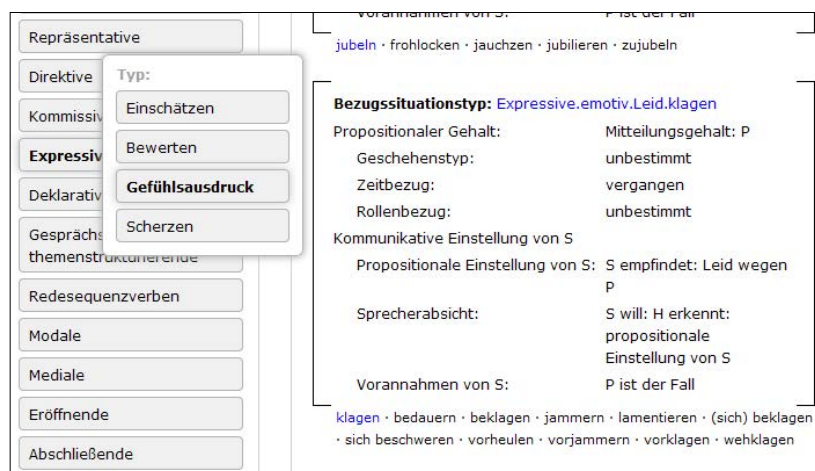


Fig. 6: Situation referred to by *klagen* (‘complain’) and its synonyms.

Zooming in, for the sake of brevity, on the verbs *klagen*, *jammern* and *lamentieren*, these verbs differ regarding the contexts in which they are typically used. Though *klagen* and *lamentieren* may be used in most of the contexts in which *klagen* is used, *klagen* is used more frequently in combination with expressions designating diseases:

- (2) Seltener wird über Kopfschmerzen geklagt. [‘People rarely complain about a headache.’]
- (3) ?Seltener wird über Kopfschmerzen gejammert. [‘People rarely moan about a headache.’]
- (4) ?Seltener wird über Kopfschmerzen lamentiert. [‘People rarely lament about a headache.’]

To the extent that *jammern* and *lamentieren* are less restricted with respect to their typical contexts of usage than *klagen*, the degree of synonymy between them is higher than that between either of them and *klagen*. As indicated by Table 7, the degree of synonymy between *jammern* and *lamentieren* is also higher than that between either of them and *klagen* by virtue of the fact that the former two verbs are identical with respect to all lexical features while *klagen* differs from *jammern* and *lamentieren* in that it is polysemous, which the latter two verbs are not:

Lexikalische Merkmale								
Verben	Merkmale							
	Seman- tische Rollen	Argu- ment- Struktur	Passiv	Resulta- tivität	Bewer- tung	Poly- semie	Performa- tivität	stilistische Markiert- heit
klagen	H (block)							
	P (fak)	PP SE PPKorrSE	+	-	-	+	-	-
bedauern	H (block)							
	P (obl)	NP<Akk> SE NPKorrSE	+	-	-	+	+	-
beklagen	H (block)							
	P (obl)	NP<Akk> SE NPKorrSE	+	-	-	+	-	-
jammern	H (block)							
	P (fak)	PP SE PPKorrSE	+	-	-	-	-	-
lamentieren	H (block)							
	P (fak)	PP SE PPKorrSE	+	-	-	-	-	-
(sich) beklagen	H (fak)	PP						

Fig. 7: Lexical features of klagen ('complain'), jammern ('moan') and lamentieren ('lament').

## 6 Conclusion: The Creation of a semantic network

The procedure whereby sets of verbs used to refer to the same special resource situation type are divided in a stepwise fashion into ever smaller sets of ever closer synonyms ultimately results in the creation of a semantic network for communication verbs. Fig. 8 represents the section of this network comprising representatives of the type 'verdeutlichen' ('explain'):

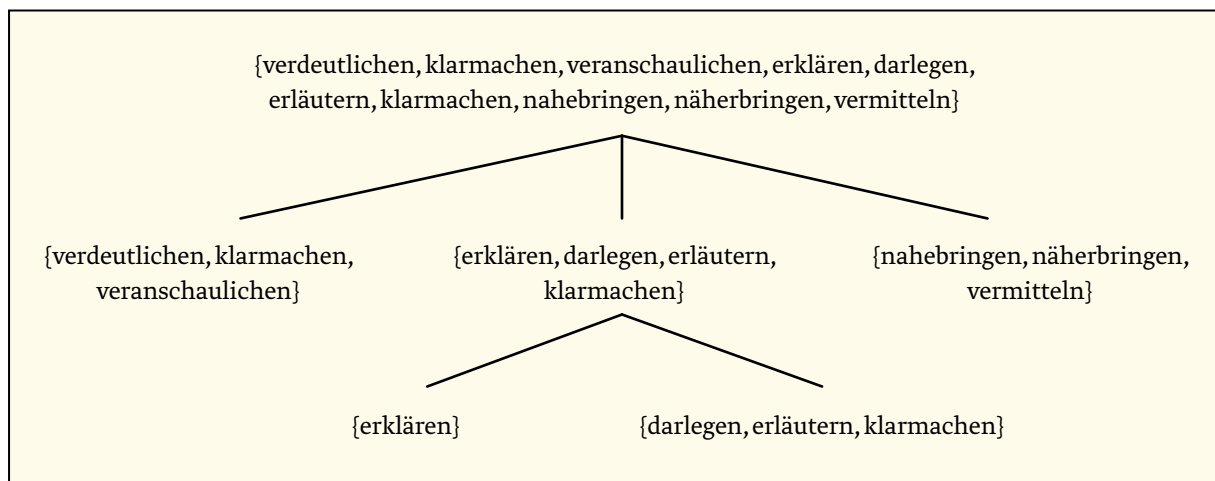


Fig. 8: Section of the network for communication verbs comprising representatives of the type 'verdeutlichen' ('explain').

Searching for verbs with varying degrees of synonymy is significantly facilitated by the online version, which provides two advanced search options allowing the user to automatically search for verbs sharing a smaller or larger number of conceptual and/or lexical features by selecting them from an input mask.

## 7 References

- Barwise, J. & Perry, J. (1983). *Situations and Attitudes*. Cambridge/MA: The MIT Press.
- Baumgärtner, K. (1977). Lexikalische Systeme möglicher Performative. In *Zeitschrift für Germanistische Linguistik*, 5, pp. 257-276.
- Bierwisch, M. & Lang, E. (1989). Somewhat Longer – Much Deeper – Further and Further: Epilogue to the Dimensional Adjective Project. In M. Bierwisch & E. Lang (eds.) *Dimensional Adjectives: Grammatical Structure and Conceptual Interpretation*. Springer Series in Language and Communication; 26. Berlin: Springer, pp. 471-514.
- Bierwisch, M. & Schreuder, R. (1992). From Concepts to Lexical Items. In *Cognition*, 42, pp. 23-46.
- Cruse, D. A. (1986). *Lexical Semantics*. Cambridge: CUP.
- Cruse, D. A. (2002). Paradigmatic relations of inclusion and identity III: Synonymy. In D. A. Cruse, F. Hundsnurscher, M. Job, P. R. Lutzeier (eds.) *Lexikologie-Lexicology: Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen – An international handbook on the nature and structure of words and vocabularies*. Vol. 2. Berlin/New York: Walter de Gruyter, pp. 485-497.
- DeReKo - Das Deutsche Referenzkorpus. <http://www1.ids-mannheim.de/kl/projekte/korpora/>
- Harras, G. (2001). Performativität, Sprechakte und Sprechaktverben. In G. Harras (ed.) *Kommunikationsverben: Konzeptuelle Ordnung und Semantische Repräsentation*. Studien zur deutschen Sprache; 24. Tübingen: Narr, pp. 11-32.
- Harras, G. (2007a). Lexikalische Strukturen der Repräsentative. In G. Harras, K. Proost, E. Winkler *Handbuch deutscher Kommunikationsverben. Teil II: Lexikalische Strukturen*. Schriften des Instituts für Deutsche Sprache; 10.2. Berlin/New York: de Gruyter, pp. 73-124.
- Harras, G. (2007b). Partielle und totale Synonymie von Sprechaktverben. In G. Harras, K. Proost, E. Winkler *Handbuch deutscher Kommunikationsverben. Teil II: Lexikalische Strukturen*. Schriften des Instituts für Deutsche Sprache; 10.2. Berlin/New York: de Gruyter, pp. 329-365.
- Harras, G., Winkler, E., Erb, S. & Proost, K. (2004): *Handbuch deutscher Kommunikationsverben. Teil I: Wörterbuch*. Schriften des Instituts für Deutsche Sprache; 10.1. Berlin/New York: de Gruyter.
- Harras, G., Proost, K. & Winkler, E. (2007): *Handbuch deutscher Kommunikationsverben. Teil II: Lexikalische Strukturen*. Schriften des Instituts für Deutsche Sprache; 10.2. Berlin/New York: de Gruyter.
- Kommunikationsverben*. Electronic version of the *Handbuch deutscher Kommunikationsverben*. Adaptation for the online version by Kristel Proost. Accessed at: <http://www.owid.de/wb/komvb/start.html>. [06/04/2014].
- Proost, K. (2006). Speech Act Verbs. In K. Brown (ed. -in-chief) *Encyclopedia of Language & Linguistics*. 2<sup>nd</sup> ed. Vol XI. Oxford: Elsevier, pp. 651-656.
- Proost, K. (2007). Conceptual Structure in Lexical Items: The Lexicalisation of Communication Concepts in English, German and Dutch. *Pragmatics & Beyond New Series*; 168. Amsterdam/Philadelphia: Benjamins.
- Lang, E. (1994). Semantische vs. konzeptuelle Struktur: Unterschneidung und Überschneidung. In M. Schwarz (ed.) *Kognitive Semantik: Ergebnisse, Probleme, Perspektiven*. Tübinger Beiträge zur Linguistik; 395. Tübingen: Narr, pp. 9-40.
- Müller-Spitzer, C., Proost, K. (2013): Kommunikationsverben in OWID: An Online Reference Work with Advanced Access Structures. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 296-309.
- Searle, J. R. & Vanderveken, D. (1985). *Foundations of Illocutionary Logic*. Cambridge, UK: Cambridge University Press.

- Vanderveken, D. (1990). *Meaning and Speech Acts I: Principles of Language Use*. Cambridge, UK: Cambridge University Press.
- Verschueren, J. (1980). *On Speech Act Verbs*. Amsterdam: John Benjamins.
- Verschueren, J. (1985). What people say they do with words: Prolegomena to an empirical-conceptual approach to linguistic action. Norwood, NJ: ALEX.
- Wierzbicka, A. (1987). *English speech act verbs: A semantic dictionary*. Sydney: Academic Press.

# Job-hunting in Italy: Building a glossary of “English-inspired” job titles

Virginia Pulcini, Angela Andreani<sup>1</sup>\*  
Università degli Studi di Torino  
virginia.pulcini@unito.it, angela.andreani@unito.it

## Abstract

This paper reports on a study of “English-inspired” job titles retrieved from a specialized corpus of job advertisements posted on Italian web pages. This corpus was created using the WebBootCat tool in the Sketch Engine, following the methodology described by Baroni and Bernardini (2004) and Baroni et al. (2006). The aim is to build a glossary of English job titles to be published online as a tool for prospective job applicants. Checking their status in English and Italian dictionaries, we will establish whether the titles collected are current English terms, false Anglicisms, or “English-inspired” creations. The preliminary findings consist of a list of 30 job titles which are analyzed in terms of form and meaning, and grouped into categories depending on whether an Italian equivalent is available or not. The corpus of job postings is used to analyze the lexical profile of job titles, their meaning and/or possible covert manipulative intent. In fact, data shows that some English job titles may be preferred to Italian equivalents to attribute greater status to the actual job designation and description. Moreover, some job titles are characterized by complex pre-modification which may confuse the ultimate users, i.e. job hunters themselves.

**Keywords:** job title; Anglicism; Anglicization

## 1 Introduction and research aims

Owing to the process of internationalization and globalization of business and trade, the job market is one of the many areas in which the influence of the English language is quite strong. A growing number of multinational companies have adopted English as a company language and most of them use English as a lingua franca on a regular basis for business communication. An emblematic case is the recent transformation of the historic Turin-based FIAT car company into a multinational through the merger with the American Chrysler and its adoption of a new “non-Italian” name – FCA – which stands for Fiat-Chrysler Automobiles. By the same token, small and medium-sized enterprises, even though operating domestically but aspiring to expand beyond national borders, also find it advisable to take on an international profile by using English for branding and product advertising. Today a

---

1 \* Both authors are responsible for the overall planning of this research. V. Pulcini drafted sections 1, 2, 3, and 4. A. Andreani drafted sections 3.1 (3.1.1, 3.1.2, 3.1.3), 3.2, 3.3 (3.3.1, 3.3.2).

good level of competence in English is an indispensable asset to hold a high-level job in the world of business. A working knowledge of spoken and written English is normally requested also for lower level occupations such as technical and clerical jobs, as emerges from advertisements in the national and international job market.

The key role of English in professional settings has greatly enhanced its desirability as a foreign language to learn. As a result, today there are more learners of English and competent non-native speakers of English than ever before, and the vocabularies of many languages have adopted a large stock of English words and terms, especially in specialized domains (Furiassi et al. 2012). The use of English is dictated primarily by practical reasons but also because non-native speakers have a favourable perception of it. As pointed out by Pulcini:

What is crucial in favouring the adoption of English loanwords are speakers' positive attitudes towards Anglicisms [...]. For better or for worse, the English language enjoys status and prestige, and Anglicisms are perceived by most speakers as modern, dynamic, fashionable and are thought to convey a higher level of competence and professionalism. (Pulcini et al. 2012: 16)

This study focuses on the influence of English on the designation and description of job titles in Italy, which appears to be a widespread and growing phenomenon in all non-English-speaking countries. Previous research carried out by Van Meurs et al. (2006; 2011) on job advertisements in the Netherlands has highlighted and described by means of quantitative data several particular aspects of the presence of English in Dutch job postings. For example, the use of English is greater in adverts posted by multinational companies and organizations than by domestic ones. English is more pervasive in ads for higher-level and academic jobs rather than medium-level vacancies and more frequent in specific domains such as transport, storage, communications and commerce as compared to, for example, the financial sector. Van Meurs et al. (2014) have also studied the perception of English loanwords with respect to Dutch equivalents in job advertisements, showing that English and Dutch terms have different associative meanings in the minds of the users. Another study carried out by Taavitsainen and Pahta (2003) points out that English is mandatory for recruitment in Finnish companies, and many Nordic companies have chosen English as their official language, abandoning their domestic names in favour of "English-inspired" ones in order to favour internationalization and, at the same time, sound young, modern and trendy. As for job postings, sometimes they are written entirely in English, and the use of English is quite frequent in vacancies for Scandinavian and Finnish companies, as well as for Swiss ones, as pointed out by Watts (2002). Taavitsainen and Pahta mention a recent campaign at the University of Helsinki against the use of "this odd form of business jargon", arguing that English job titles "blur the job description and unnecessarily mystify functions in the business world." (Taavitsainen & Pahta 2003: 8)



The research study illustrated in this paper is part of a wider project focussed on the influence of the English language in Italy<sup>2</sup>, including its impact on the world of business. The focus is on the use of English or “English-inspired” job titles retrieved from a corpus of job advertisements posted on Italian web pages. The aim of this research is to build a glossary of English-looking job titles to be published online as a tool for job hunters in Italy. Using dictionaries and corpora in order to observe the lexical profile of these job titles, we will try and establish which of these are current Anglicisms, false Anglicisms, or “English-inspired” creations. We will argue that some terms are rather opaque to the Italian user and their adoption is motivated by the intention to give “higher status” to a particular job or to camouflage its real nature and thus confuse or deceive the prospective applicant.

## 2 Methodology

The collection of English-looking job titles began with a preliminary survey of the websites of some Italian online job finding agencies<sup>3</sup> and of the websites of the Italian branch of some multinational human resource consulting companies.<sup>4</sup> On the websites, the user can select, among other options, a professional category (*categoria professionale* or *funzione aziendale*, e.g. retail, HR, banking), an industry sector (*settore*), and, in some instances, a specific role or job position (*mansione* or *funzione aziendale*, e.g. receptionist). The dropdown menus often include, alongside Italian ones, professional categories and functions already in English, which formed our preliminary list of English job titles.<sup>5</sup>

This was then expanded by querying a domain-specific corpus of Italian job advertisements, which we built using the WebBootCat tool in the Sketch Engine (Kilgarriff et al. 2004). Drawing on the methodology described in Baroni and Bernardini (2004) and Baroni et al. (2006), we selected a number of seedwords from among the most frequent terms and phrases in job postings: *annunci di lavoro; offerte di lavoro; si offre; si propone; si richiede; annuncio; lavoro; azienda; contratto; candidato; settore; profilo; esperienza; competenze*.<sup>6</sup> The corpus was then compiled using the TreeTagger for Italian (Baroni’s model) and opened in the Sketch Engine to compare it with the itTenTen10 corpus and extract further key terms to be used as seeds. The procedure was iterated twice, and then repeated at approximately three

---

2 The project is “The English language in Italy: linguistic, educational and professional challenges”, promoted by the University of Turin in conjunction with the *Compagnia di San Paolo* (2013-2015) and coordinated by Virginia Pulcini. [www.englishinitaly.wordpress.com](http://www.englishinitaly.wordpress.com)

3 Accessed at <http://www.adhr.it>; <http://www.alispa.it>; <http://www.carrieraefuturo.com>; <http://www.euroin-terim.it>; <http://www.gigroup.it>; <http://www.humangest.it>; <http://www.obiettivolavoro.it>; <http://www.orienta.net>; <http://it.quanta.com>; <http://www.umana.it/it-IT/home-page> [13/10/2013]

4 Accessed at <http://www.adecco.it>; <http://www.manpower.it>; <http://www.randstad.it>; <http://www.synergie-italia.it> [13/10/2013]

5 Data entry, development engineer, hostess, order entry, promoter, receptionist, telemarketer, visual merchandiser, web designer.

6 The total number of seeds, 14, was set following Baroni and Bernardini: “For well-defined specialized domains, a small list of seeds (in the 5-to-15 range) is typically sufficient” (2004: 1314). The additional parameters (tuple size, minimal and maximal file size, max URLs per query, etc.) were set according to the default settings of the WebBootCat in the Sketch Engine.

weeks' distance, obtaining a final corpus of 241,021 tokens.<sup>7</sup> The corpus was queried to retrieve additional English or English-looking job titles in context.

### 3 Preliminary Findings

The preliminary findings consist of a list of 30 job titles which are analyzed in terms of form, meaning and Italian equivalents in English and Italian general and specialized dictionaries and in our corpus. The English dictionaries considered are the *Collins English Dictionary* online (CED) and the *Cambridge Business English Dictionary* online (CBED); the Italian dictionaries are *Zingarelli 2014* (ZING) and the bilingual encyclopaedic dictionary *Economics&Business* (Picchi 2011, henceforth E&B).

Table 1 shows the attestation of the terms in the reference dictionaries. The cells highlighted in grey indicate Anglicisms with a current Italian equivalent. Items in italics are dictionary headwords that slightly diverge in form from our listed titles though they retain the same expected meaning.

As several terms were not recorded in dictionaries, we also browsed through the specialised glossary of job types published by the UK job finding website *Prospects*,<sup>8</sup> and through the International Standard Classification of Occupations elaborated by the International Labour Organization (ISCO08).<sup>9</sup> Also available online are the *Classificazione delle Professioni* (Classification of Occupations) produced by the Italian National Institute for Statistics (ISTAT CP2011) and the ISCO-ISTAT table of correspondences, the *Raccordo* ISCO08-CP2011, issued by the same Institute.<sup>10</sup> In order to account for the currency of Anglicisms in Italian we referred to the online historical archives of the Italian newspapers *La Stampa* (1867-2000) and *la Repubblica* (1984-present).<sup>11</sup>

---

7 In order to increase visibility, a single job advertisement is normally posted on several websites; therefore, queries run within a short time span from one another will tend to retrieve many duplicates.

8 [www.prospects.ac.uk](http://www.prospects.ac.uk)

9 [www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm](http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm)

10 <http://www.istat.it/it/archivio/18132>

11 <http://www.archiviolaStampa.it/>; <http://ricerca.repubblica.it/>

	<b>CED</b>	<b>CBED</b>	<b>ZING</b>	<b>E&amp;B</b>
accountant	accountant	accountant		accountant
area manager	area manager		area manager	area manager
baby sitter	baby-sitter	babysitter	baby-sitter	
barman	barman	barman	barman	
beauty sales agent	agent	agent	agent	sales agent
data entry		(other meaning)	(other meaning)	
deejay	deejay	deejay	deejay	
development engineer	engineer		mod+engineer	mod+engineer
electrical practical instructor	instructor	instructor		
export area manager	export manager			export manager
(financial) controller	(financial) controller	(financial) controller	controller	controller
first article inspector	inspector	inspector		
hostess	hostess	hostess	hostess	
instrument practical instructor	instructor	instructor		
mystery shopper	mystery shopper	mystery shopper		
order entry				
(junior) programmer	programmer	programmer		
project manager	project manager	project manager	project manager	project manager
promoter	promoter	promoter	promoter	promoter
receptionist (junior)	receptionist	receptionist	receptionist	
retail sales manager		retail manager		
runner	runner		(other meaning)	
sales account		(other meaning)		(other meaning)
sales manager	sales manager	sales manager	sales manager	sales manager
shop assistant	shop assistant	shop assistant		shop assistant
store manager				store manager
store specialist				
telemarketer	telemarketer	telemarketer		telemarketer
visual merchandiser	merchandiser	merchandiser	merchandiser	merchandiser
web designer	web designer	web designer		

**Table 1: Job titles in CED, CBED, ZING and E&B.**

Formally, a small group of job titles are polymorphemic one-word items, characterized by endings such as -er, -ist, -or, -ant, -man, that typically realize the {noun agent} morphemic function, i.e. denote the agent of the action indicated by the root element (e.g. promote ® promoter). In fact, most job titles are typically complex words, either solid compounds, like *barman* or 2- or 3-word compounds,

characterized by the modifier+head structure, in which the head element indicates the job function and the left-hand modifying element functions as classifier, i.e. it indicates a sub-class of the head element (e.g. sales manager = a person in charge of a company's sales activities and its sales force, CBED). This word-formation mechanism can trigger even more complex items if further classification of duties or skills need to be specified (e.g. beauty sales agent). However, since Italian is a language that typically modifies on the right of the head element, in complex job titles the order of the elements may be changed, as in *project manager junior* (instead of *junior project manager*): “Stiamo ricercando un *project manager junior* per gestione progetti C++/C#.”

In the following paragraphs, we present a sample of the analysis carried out on the start-list of job titles of our glossary, distinguishing between: a) Anglicisms which coexist with Italian equivalents; b) Anglicisms which do not have Italian equivalents; c) English (inspired) job titles which are not recorded in the selected dictionaries or are recorded with another meaning (false Anglicisms).

### 3.1 Anglicisms with Italian equivalents

The English job titles with a current Italian equivalent are *accountant*, *area manager*, *baby sitter*, *barman*, (*financial*) *controller*, *programmer*, *project manager*, *sales manager*, *shop assistant* and *telemarketer*. The presence of these terms in Italian and English monolingual and bilingual dictionaries online makes it theoretically viable for job hunters in Italy to check the meaning of unknown or unfamiliar terms.

#### 3.1.1 Area manager and sales manager

The head *manager* of these compounds is recorded as part of the core Italian lexicon in ZING (ultimately from It. *maneggiare* according to the *Oxford English Dictionary*). The first attestation of this Anglicism in Italian is 1895. It coexists alongside Italian *direttore* and *dirigente*, which are recorded as equivalents in E&B. This is a highly productive loanword in Italian, which functions as the head of numerous occupational titles.<sup>12</sup>

E&B defines *area manager* as the title used especially by American businesses and organizations to denote the person responsible for the sales force and for the marketing and distribution of products within a specific geographical area. The Italian equivalent proposed is *direttore di zona*. The ZING definition is consistent with E&B, but the recorded equivalent is the Italian *capoarea*. There are no occurrences of the Italian *direttore di zona* in our corpus, which contains instead 9 occurrences of the Anglicism and 3 occurrences of *capoarea*. The examples below show the use of this Anglicism in context:

- (1) Per piccola e solida azienda di prodotti per l'edilizia ricerchiamo 1 *Area Manager* Italia. La figura si interfacerà con la proprietà per seguire e consolidare i clienti acquisiti e espandere il pacchetto clienti. Viaggerà spesso e farà da referente per la rete agenti su tutto il territorio nazionale.

---

12 Other titles recorded in the selected Italian dictionaries are account manager, office manager, risk manager, and property manager.

- (2) La risorsa sarà inserita come *Area Manager* per il Mercato Italia e si occuperà del contatto e gestione degli agenti e dei clienti; della ricerca e sviluppo di nuovi contatti; studi di settore; redazione di offerte commerciali e partecipazione a fiere di settore.
- (3) La posizione, che riporta all'*Area Manager* della zona di competenza, ha la funzione di presidiare dei punti vendita specializzati del canale di riferimento, garantendo il raggiungimento degli obiettivi qualitativi e quantitativi stabiliti.

Examples (1) and (2) are extracts of job ads for *area managers*, which provide a general description of the tasks required by the position, such as maintaining contacts with existing customers and acquiring new ones, liaising with other agents and representatives on the assigned territory, creating proposal documents and representing the company at trade exhibitions. Example (3) is the extract of a job vacancy for a *commercial hostess*, whose tasks will include “reporting to the area manager of the assigned (geographical) area.”

In the Standard Classification of Occupations “managers responsible for specialized functions within a specific geographic area” are clearly distinguished from “managing directors and chief executives.” (ISCO08: 15). The Italian *direttore o dirigente di dipartimento* are provided in the EN-IT table as the standard equivalents of area manager-level occupational titles (irrespective of the department and specialization, e.g. sales or HR).<sup>13</sup>

With the exception of one occurrence of HR area manager, all instances of the Anglicism in our corpus refer to sales department area managers. The second compound analysed here is, in fact, *sales manager*, translated by both E&B and ZING as *direttore delle vendite*. E&B also records *direttore commerciale*. While this Anglicism is found 5 times in our corpus, *direttore commerciale* and *direttore vendite* also occur 7 times each. These job titles are shown in context in the examples below:

- (4) Nell’ambito del potenziamento dell’organico della filiale Svizzera in Ticino di multinazionale americana in costante crescita ricerchiamo *Sales Manager / Sales Account* da inserire all’interno della nostra struttura. Dimestichezza ed interesse per la tecnologia. Requisiti richiesti: -residenza a 25-30 km dal confine svizzero -età compresa tra 25 e 35 anni -esperienza di vendita o simile in servizi business to business di 2 anni -conoscenza lingua inglese.
- (5) In un’ottica di potenziamento della rete commerciale, il Gruppo ricerca nuove risorse per il ruolo di *Sales Manager*, da inserire all’interno della filiale di Milano. La funzione prevede lo sviluppo del portafoglio clienti corporate [...].
- (6) Importante gruppo tedesco, attivo nella commercializzazione di materiale elettrico e sistemi di fissaggio per il settore fotovoltaico, in un’ottica di forte sviluppo, ricerca un/una *sales manager*.

Example (4) seems to suggest that *sales manager* and *sales account* might be treated as equivalent roles; a more detailed discussion of this pair is provided in section 3.3 below. All job advertisements point to

---

13 At a higher level of a company structure we find *direttore generale, imprenditore, dirigente e amministratore* as the standard Italian equivalents to Chief executives or managing directors.

the commercial development of the company through the expansion of its customer base as one of the key responsibilities of the position.

### 3.1.2 Accountant and (financial) controller

(7) *Accountant / Financial controller* </p> <p> Veneto, Veneto / Permanenti </p><p> Per nostra azienda cliente, realtà multinazionale, ricerchiamo un *Accountant / Financial Controller* per la loro sede in Pennsylvania (USA). Il candidato si dovrà occupare di tutta la gestione contabile, fiscale, tesoreria, crediti, liquidità.

Example (7) shows the only occurrence of the Anglicism *accountant*, which, in our corpus, has been superseded by its Italian equivalent *contabile*, with 14 occurrences. The choice of the Anglicism might, in fact, depend on the type of company advertising the vacancy, e.g. the American branch of a multinational corporation. *Accountant* appears to be regarded as a synonym for *financial controller* in the posting, although the two denote different level positions in English: “an executive who is the head of a company’s finance or accounts department” the former, and “a person or company whose job is preparing the financial records of people, companies, or organizations” the latter (CBED). The CBED lists *controller* and *comptroller* as alternative forms of this compound. These are also recorded in ZING, which marks them as business terms, and E&B, in which *controller* and Italian *controllore della gestione* are recorded as current equivalents. There are no occurrences of *controllore della gestione* in our corpus, which contains instead 1 occurrence of *controller* alongside the Italian *responsabile amministrativo*:

(8) *Controller* Filiale Svizzera. Dinamico gruppo metalmeccanico italiano ci ha incaricato di selezionare un responsabile amministrativo-*controller* per la sede svizzera di un’azienda.

In fact, the lack of further information in the job advertisement makes it difficult to ascertain whether the position advertised in example (8) is exactly the same as the one described in example (7).

### 3.1.3 Baby-sitter, barman, telemarketer and programmer

The Anglicism *baby-sitter*, borrowed in the mid-20<sup>th</sup> century (1950 according to ZING), occurs 20 times in the corpus vs. 2 occurrences of its Italian equivalent *tata*. No occurrences are found for the other Italian equivalent *bambinaia*, which has registered a steady decline in use since the second half of the 20<sup>th</sup> century (*La Stampa*):

(9) [...] ricerca urgentemente una *babysitter* per attività didattiche e ludiche con bimbo di 6 anni. Si richiede: -Esperienza pregressa nella mansione -Preferibile titolo di studio ed esperienze in pedagogia -Ottima conoscenza lingua inglese o madrelingua inglese

*Barman* and *barista* are both attested in our corpus, where the false Anglicism *barlady* is also found for the feminine form instead of English *barmaid*, or the gender-neutral form *bartender*.<sup>14</sup>

(10) [...] ricerca per noto locale del fossanese un/a *barman /barlady* con esperienza documentabile nella mansione per inserimento con contratto di somministrazione.

14 Furiassi records the false Anglicism *barwoman* (2010: 110-11, 145).

The Italian equivalent *barista* may denote both the person serving drinks in a bar and a bar owner (ZING). The polysemy of the Italian term might obscure the intended meaning of such a job vacancy as the one offered in example (11), where the job requirements: “experience of at least 5 years in the management of bar activities”, are generic, and could be read as experience in serving and dealing with customers as well as experience in the actual management of a bar:

(11) Agenzia per il lavoro [...] ricerca per importante bar un barista / *barman*. Requisiti richiesti: esperienza di almeno 5 anni nella gestione delle attività di bar.

*Telemarketer* is rare in Italian, quoted in the *la Repubblica* daily newspaper 3 times from 2003 but recorded in E&B and translated as *televenditore*. The term *telemarketer* appears in the menu of one of the job finding agencies considered. In our corpus it appears as *telemarketing* preceded by the Italian nouns *operatori*, *operatrici*, *risorse* or *addetti* (equivalent to the English worker/s, workforce) and produces the hybrid compounds *operatrici telemarketing*, *risorse di telemarketing* and *addetto telemarketing*:

(12) Ricerchiamo Operatori *telemarketing* per fissaggio appuntamenti telefonici per conto di Consulente certificato Telecom/Tim. Il lavoro potrà essere svolto da casa.

Finally, an anomalous case in this first group is represented by the Anglicism *programmer*. *Programmatore*, derived from the Italian verb *programmare*, is well established in Italian, occurring in our corpus 57 times vs. a single instance of *programmer*. In fact, as shown in example (13), the job posting in which *programmer* occurs features an unusually high frequency of Anglicisms, italicized below:

(13) Job Title: Stage - Junior Programmer Job ID: 143142 Location: Milano Organization: Siemens S.p.A. Mode of Employment: Stage, Full time. Per il nostro ufficio *Energy Automation Solution Operation* del settore *Infrastructures & Cities* di Siemens Italia, nella sede di Milano (Vipiteno) cerchiamo un *Junior Programmer*. Scopo formativo dello stage è l'affiancamento al nostro personale che si occupa dello sviluppo di sistemi informatici per la gestione di reti di pubblica utilità con l'obiettivo di acquisire la conoscenza per sviluppare applicazioni relative ai sistemi e alle soluzioni progettate nella divisione *Smart Grid*.

### 3.2 Anglicisms with no Italian equivalent

This group includes *deejay*, *hostess*, *mystery shopper*, *promoter*, *receptionist*, *runner* and *web designer*. *Deejay* is a well-established loan, first attested in Italian in 1987. More recent is the Anglicism *mystery shopper* (CED= “a person who is employed, often by the owners, to visit shops, hotels, etc, incognito, and assess the quality of the service offered”), not recorded in Italian dictionaries but quoted in the *la Repubblica* newspaper (single instance in 1994, then occasionally from 2001 both in its English spelling and with <y> graphically adapted to <i>). Alongside the job details, the advertisement in the corpus also provides a definition of this Anglicism:

(14) Cerchiamo urgentemente una *mystery shopper* per veloce lavoro nel mese di settembre [...] Il *mystery shopper* è il cosiddetto cliente misterioso ossia una persona che fingendosi cliente effettua una visita presso un punto vendita.



Also *web designer*, though unrecorded in Italian dictionaries, appears to be quite transparent in meaning for the Italian user, as clearly denoting “someone whose job is to design websites” (CBED). In some advertisements it appears that the role is treated as an equivalent to the Italian *grafico* (graphic designer) omitting the website-specific function of the role, as in the following example:

(15) Si richiede esperienza consolidata nella mansione di grafico e/o *web designer*. Il candidato ideale è in possesso di conoscenza approfondita dei principali applicativi di grafica (Flash, Illustrator, Coreldraw, DreamWeaver, ecc).

The meanings and lexical profile of the remaining titles are more complex. Beginning with *receptionist*, the examples retrieved from the corpus indicate that the term might also be used as an equivalent of the Italian *centralinista* (telephone operator), or even of *telemarketer*:

(16) Per azienda nel settore moda ricerchiamo centralinista/ *receptionist* che abbia già maturato esperienza di almeno 2 anni presso aziende strutturate e modernamente organizzate.

(17) Centro Fitness vicino a Padova, cerca una *Receptionist* con mansioni di vendita interna abbonamenti, gestione clienti acquisiti, utilizzo del telefono (*telemarketing* in e out).

Another interesting example in this second group is *runner*, defined in the CED as “a messenger for a bank or brokerage firm” (meaning 2) or “a person who operates, manages, or controls something” (meaning 7). The synonyms proposed are “messenger, courier, errand boy, dispatch bearer”, thus describing an unskilled, entry-level position and a possible equivalent of the Italian *fattorino*, *addetto allo spostamento merci*.<sup>15</sup> The only occurrence in the corpus describes this position as follows:

(18) Per azienda moda lusso ricerchiamo 1 *runner*. Si richiede esperienza all’interno di negozi di moda e abbigliamento in qualità di venditore e di magazziniere. La risorsa si occuperà del ricevimento merce e preparazione dei prodotti per la vendita e supporterà in caso di necessità i colleghi venditori.

Thus, to some extent, the job description might be deceptive: *runner* is translated as *venditore e magazziniere* (=salesperson and runner, note that *salesperson* occurs first), although the tasks are the reception and preparation of products (*ricevimento merce e preparazione prodotti*) in support, if necessary, of the actual salespersons. The analysis shows that the Anglicism might in fact find a current Italian equivalent in *fattorino*, and thus typically an unskilled job, although the employer – a luxury-fashion house (*azienda moda lusso*) – requires specific experience in fashion and clothing (*si richiede esperienza*). It should also be pointed out that ZING records the Anglicism with different meanings: a) a person who runs, Italian *corridore* and b) strip of linen placed across a table, which has no current Italian equivalent.

Even more complex are the lexical profiles of the Italian *promoter* and *hostess*. The wordsketch of *promoter* indicates that this term might be used to refer to marketing positions, as an equivalent to *salesperson*, sometimes preceded by “sales”, as in “[...] offre la posizione di *Sales Promoter* e un percorso di

15 cf. *OED*: “A person employed to perform various (generally menial or unskilled) tasks, typically involving moving from place to place. Also more generally: an assistant.” (meaning 2d). See also ISCO-08: 543 and 564 “Transport and storage labourers”.



crescita professionale” and as an equivalent to *telemarketer*, as in “Azienda settore Telecomunicazioni seleziona operatori / *promoter* telefonici da casa per servizio di promozione e vendita abbonamenti”. The term *hostess*, borrowed in 1948, is used in Italian to denote a) women flight attendants or b) conference assistants. In the corpus, however, the meanings found referred to either conference assistants and nightclub hostesses, as illustrated in example (19):

(19) Ragazza *hostess* per lavoro di figurante di sala night club [...] Cerchiamo *hostess* da assumere con regolare contratto per il lavoro di figurante di sala per eleganti ed esclusivi night club di alto livello.

### 3.3 “English-inspired” job titles

The remaining job titles are complex job titles characterized by a modifier+head structure, in which the head element, generally recorded in English dictionaries, indicates the job function. None of these compounds is recorded in Italian or English dictionaries, although they might indeed sound plausible or acceptable both in form and in meaning, especially considering that they are sometimes accompanied by a description of duties and functions in job advertisements. This third group also features the false Anglicisms *data entry*, *order entry* and *sales account*, which will be treated separately in section 3.3.2.

#### 3.3.1 Complex job titles

This group includes the following titles: *beauty sales agent*, *development engineer*, *electrical practical instructor*, *export area manager*, *first article inspector*, *instrument practical instructor*, *retail sales manager*, *store manager*, *store specialist* and *visual merchandiser*.

The Anglicism *engineer* makes a particularly interesting candidate for our analysis. This is in fact a highly productive head for new job titles, as shown in both Italian and English dictionaries which record a wide variety of occupational titles like *safety engineer*, *civil engineer* and *mechanical engineer*. While Italian *ingegnere* denotes professionals with a University degree, the English *engineer* might also refer to a technician with specialist competence, but not necessarily a graduate, that the Italian would translate as *tecnico*. In the job market, this is a crucial difference with respect to the salary offered to the prospective candidate, and to the perceived prestige of the position. Corpus analysis might help to profile this term and its usage in a larger corpus of job postings.

Our corpus features one posting for a *testing engineer*. In fact, the search for an entry level position is clarified in the actual job description: the job title is repeated in the first lines of the advertisement, preceded by the adjective *giovane* (young), and the job requirements include details pertaining to experience (“even limited”) and level of education (“degree in engineering or other technical qualification”).

(20) Requisiti: - laurea in ingegneria meccanica o altro titolo di studio di formazione tecnica. - esperienza progressa, anche minima, nella mansione maturata all'interno di aziende operanti nell'automotive, su motori a benzina con competenze in particolare su Fuel Injection.

*Beauty sales agent* might be perceived as more economical and effective than its longer Italian equivalent *agente di vendita (di prodotti) di bellezza*. In fact, even in the job description the Italian standard equivalent of *sales agent*, *agente di vendita* is avoided and replaced by the euphemism *animatrice commerciale* (commercial performer, entertainer or catalyst). Yet, this is the only occurrence of *sales agent* in our corpus, which tends to prefer *agente di vendita* and to specify by means of the job description the business sector of the position advertised (supplies for coffee makers, real estate business, telephone market, electricity, etc.).

*Electrical practical instructor* and *instrument practical instructor* are plausible English compounds considering "(practical) instructor" as the head of the compound and "electrical" and "instrument" as modifiers. The current Italian equivalent for the head of the Anglicism would be *istruttore* or *formatore*, and denote an instructor for electrical technicians in the one example, and an instructor for instrument technicians in the other ("Vocational education teachers" in ISCO08: 112). Perhaps a strategic function underlies the creation of the "English-inspired" titles for the job posting, which advertises positions for an international training centre which will require fluency in English:

(21) Stiamo cercando un *Instrument Practical Instructor* per il Training Center ECU di Cortemaggiore. Il nostro cliente è l'Eni Corporate University. Sarà un On the Job Training Indirizzato a Instrument Technician iracheni, quindi il corso sarà tenuto in inglese.

(22) Stiamo cercando un *Electrical Practical Instructor* per il Training Center ECU di Cortemaggiore. Il nostro cliente è l'Eni Corporate University. Sarà un On the Job Training Indirizzato a Electrical Technician iracheni, quindi il corso sarà tenuto in inglese.

### 3.3.2 False Anglicisms

As a job title, the compound *sales account* seems to be an innovation typical of the Italian job market. The examples in the corpus refer to such activities as fostering business to business commercial transactions and expanding the customer base of a company, as example (23) shows:

(23) Si cerca un *sales account* con esperienza nel settore e predisposizione alle attività commerciali per inserimento in importante azienda che opera nella progettazione e realizzazione di prodotti e macchine speciali. La risorsa, rispondendo al responsabile commerciale si occuperà di attività consulenziale tecnica pre e post vendita e di implementazione del portafoglio clienti

*Sales account* may be considered as equivalent to *sales manager*, as already pointed out with reference to example (4) above. In English, *sales account* indicates "a record of the total cash or credit sales for a particular period" or "a customer that a business sells its products to" (CBED). It is not present in the CED. In fact, *sales manager*, "a person in charge of a company's sales activities and its sales force" (CBED), and *account manager*, "someone employed by a company to be responsible for one or more of its customers, especially someone in the banking or advertising industry" (CBED), are the best candi-

dates as quasi-equivalents to Italian *sales account*, which might in fact be an ellipsis of *sales account manager*, or denote a lower – i.e. not managerial – professional level. This third group also includes such false Anglicisms as *data entry* and *order entry* used in the websites of job finding agencies (cf. footnote 4) to refer to the agent rather than to the activity. While *order entry* is not recorded in the selected dictionaries, *data entry* is recorded in ZING to denote the activity, and not as an agent noun. Examples of the use of either *data* or *order entry* as agents were not retrieved in our corpus – which contains instead one occurrence of the correct usage of the Anglicism in the hybrid compound *impiegata data entry* (“data entry clerk”) – though an advanced Google search for the terms in Italian pages published in Italy can easily produce results like “Per importante società multinazionale ricerchiamo un *data entry* con pregressa esperienza nel ruolo da inserire con contratto di somministrazione”, “agenzia per il lavoro ricerca per azienda cliente un *data entry*”, “Per importante azienda cliente ricerchiamo un *order entry*. La risorsa si occuperà dell’inserimento degli ordini esteri.”

## 4 Conclusion

The Anglicisation of the job market gives the opportunity to linguists to observe language change and lexical innovation and reflect on the underlying mechanisms that trigger the introduction of new job titles. As has emerged from the present corpus-driven research, there is a growing habit of using Anglicisms or English-looking coinages to refer to functions or positions in Italian job postings. As a phenomenon of lexical innovation, the adoption of loanwords is motivated by the need to fill a lexical gap in the recipient language, but, especially in the case of Anglicisms, the main reason is to comply with international terminology in global business, and to express modernity and professionalism.

Our start list contains a few instances of “necessary” Anglicisms, i.e. *deejay*, *hostess*, *mystery shopper*, *promoter*, *receptionist*, *runner* and *web designer*. For these terms there are no competing Italian equivalents. Their success in the recipient language can be ascribed to several characteristics, such as brevity and conciseness for *deejay*, modernity for *web designer*, *promoter* and *baby-sitter* (taking over the old-fashioned *bambinaia* and *balia*, or the childish *tata*). When a domestic equivalent exists, the preference for English is dictated primarily by pragmatic and stylistic reasons, since English terms better answer the need for monoreferentiality and conciseness (e.g. *beauty sales agent* / *agente di vendita di prodotti di bellezza*). However, the coexistence of a foreign term along with a native equivalent can be regarded as a case of multiple terminology (*controller/controllore della gestione*), which violates the terminological principle according to which a term identifies a single concept (Pulcini 2012). As the job market develops giving rise to new jobs or professional profiles, a new term may in fact describe different duties as in the case of *receptionist*, whose tasks consist not only in answering to incoming calls (It. *centralinista*) but to attend to a wider range of services, including telemarketing. Finally, multinational companies may opt for an English job title to comply with the established international profile of the company, as in the case of *accountant/financial controller*, which is advertised by a company based in

Pennsylvania, USA. An example of a term which has been successfully assimilated into Italian and also displays great productivity is *manager*. Although many equivalent terms exist to identify different levels of managerial statuses (*direttore, dirigente, etc.*), *manager* seems to be an “all-purpose” term, lending itself to a variety of pre-modifications to indicate the management area involved (e.g. sales manager, area manager). We may add that *manager* is a long-standing and very productive Anglicism in Italian, ultimately a re-borrowing from Italian *maneggiare*, which is the source of the English term.

On the other hand, several advertised jobs may indeed be deceptive for job seekers. The very productive term *engineer*, for example, which resembles Italian *ingegnere* because of the common classical source, may refer to a technician with specialist competence and not necessarily to a professional with a degree in engineering. The former meaning may slowly be filtering into Italian as well, to attribute greater prestige to the actual job designation.

Among the terms discussed in this paper, some may be deceptive for the prospective applicant for different reasons. For instance, the term *hostess* has extended its meaning from air hostess, which has been replaced by the gender-neutral *assistente di volo* in Italian, to other jobs for which a female assistant or attendant is sought, e.g. in the meeting and event industry. In our data, however, many job positions also referred to nightclub hostesses. For the term *runner*, instead, both job designation and description were obscure, referring to functions as salespersons or storage labourers. Finally, the terms that we labelled as false Anglicisms were possibly derived from the ellipsis of multi-word compounds, e.g. *data entry* for *data entry clerk*.

In conclusion, the adoption of English and “English-inspired” job titles within the context of the Italian job market is a growing phenomenon, partly dictated by the need to name new occupations but especially to comply with the Anglicization of the job market and specialized terminology. Therefore, in this research we aimed to provide the theoretical framework on which to ground the compilation of a glossary of English (or English-looking) job titles – and their potentially misleading nature – to be made publicly available online as a dedicated tool for prospective job hunters in Italy.

## 5 References

- Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*. Lisbon: ELDA, pp. 1313-1316.
- Baroni, M., Kilgarriff, A., Pomikalek, J., Rychly, P. (2006). WebBootCaT: Instant domain-specific corpora to support human translators. *Proceedings of EAMT-2006*, pp. 247-252.
- Cambridge Business English Dictionary Online*. Accessed at: <http://dictionary.cambridge.org/dictionary/business-english/> [07/04/2014]
- Classificazione delle Professioni CP2011*. Accessed at <http://www.istat.it/it/archivio/18132> [26/03/2014]
- Collins English Dictionary Online*. Accessed at: <http://www.collinsdictionary.com/dictionary/english> [07/04/2014]
- Furiassi, C. (2010). *False Anglicisms in Italian*. Monza: Polimetrica.

- Furiassi, C., Pulcini, V., Rodríguez González, F. (eds.) (2012). *The Anglicization of European Lexis*. Amsterdam: John Benjamins.
- International Standard Classification of Occupations. Group definitions*. Accessed at [www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm](http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm) [26/03/2014]
- Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In *Proceedings EURALEX 2004*. Lorient: Franc, pp 105-116. Accessed at: <http://sketchengine.co.uk> [13/10/2013]
- la Repubblica*. Accessed at <http://ricerca.repubblica.it/> [02/04/2014]
- La Stampa*. Accessed at <http://www.archiviolaStampa.it/> [02/04/2014]
- Oxford English Dictionary Online*. Accessed at: <http://www.oed.com> [02/04/2014]
- Picchi, F. (2011) *Economics & Business. Dizionario enciclopedico economico e commerciale inglese-italiano con glossario italiano-inglese*. Bologna: Zanichelli. Accessed at: <http://dizionarioonline.zanichelli.it/dizionariOnline/#economics> [07/04/2014]
- Prospects. Types of Jobs*. Accessed at [http://www.prospects.ac.uk/types\\_of\\_jobs.htm](http://www.prospects.ac.uk/types_of_jobs.htm) [07/04/2014]
- Pulcini, V. (2012). Register variation in tourism terminology. In R. Facchinetti (ed.) *A Cultural Journey through the English Lexicon*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp.109-132.
- Pulcini V., Furiassi C., Rodríguez González, F. (2012). The lexical influence of English on European languages: From words to phraseology. In C. Furiassi, V. Pulcini & F. Rodríguez González (eds.) *The Anglicization of European Lexis*. Amsterdam: John Benjamins, pp. 1-24.
- Raccordo ISCO08-CP2011*. Accessed at <http://www.istat.it/it/archivio/18132> [26/03/2014]
- Taavitsainen, I., Pahta, P. (2003). English in Finland: globalisation, language awareness and question of identity. In *English Today*, 4, pp.3-15.
- Van Meurs, F., Korzillius, H., den Hollander, A. (2006). The use of English in job advertisements on the Dutch job site Monsterboard.nl and factors on which it depends. In *ESP across cultures*, 3, pp. 103-123.
- Van Meurs, F., Planken, B., Gerritsen, M., Korzillius, H. (2011). Reasons given by Dutch makers of job ads for placing all-English, partly English or all-Dutch job advertisements in Dutch newspapers: An interview-based study. In C. Degano, G. Garzone (eds.) *Discursive practices and textual realizations in organizational communication: Product and process, frontstage and backstage*. Trezzano sul Naviglio: Arcipelago Edizioni, pp. 53-57.
- Van Meurs, F., Hornikx, J., Bossenbroek, G. (2014). English loanwords and their counterparts in Dutch job advertisements: an experimental study in association overlap. In E. Zenner, G. Kristiansen (eds.) *New Perspectives in Lexical Borrowing*. Boston/Berlin: De Gruyter Mouton, pp. 171-190
- Watts, R. J. (2002). English in Swiss job adverts: A Bourdieuan perspective. In A. Fischer, G. Tottie, H. M. Lehmann (eds.) *Text Types and Corpora*. Tübingen: Gunter Narr Verlag, pp. 103-122.
- Zingarelli, N. (2013) *loZingarelli2014*, Bologna, Zanichelli. Accessed at: <http://dizionarioonline.zanichelli.it/dizionariOnline/#zingarelli> [07/04/2014]

## Acknowledgements

We wish to acknowledge the financial contribution of the *Compagnia di San Paolo* (Torino) for the project “The English language in Italy: linguistic, educational and professional challenges” of the University of Torino (2013-2015).



# A Small Dictionary of Life under Communist Totalitarian Rule (Czechoslovakia 1948-1989)

Věra Schmiedtová

The Institute of the Czech national corpus, Charles University in Prague

vera.schmiedtova@ff.cuni.cz

## Abstract

The intention is to preserve the vocabulary of the period for the younger generation; also to remind the older generation of vocabulary that they used to encounter, but are gradually forgetting. The dictionary is specific in that it is made up of two types of vocabulary – the language of communist propaganda and the spoken language emerging from how people reacted to the pressure of propaganda, often including popular humour. The first type of vocabulary has been collated through Corpus of Totalitarianism, for the second type a corpus-based source does not exist (it is the language as spoken, which it was possible to collate through quotes from fiction, journalistic writings and from the author's own observations. This language has been checked in contemporary written corpuses, on some occasions it is to be found in the Corpus of Totalitarianism or on the internet).

**Keywords:** dictionary of communist propaganda; czech language

## 1 Historical context

The Czechoslovak Republic underwent huge political changes in 1989. The period of communist totalitarian rule ended (1948-1989) and the country returned to democracy. This paper attempts to show how language changed with the change of political discourse. We have to bear in mind that totalitarian language changed its character over time. In this sense, three main historical periods can be identified: The fifties: major ideological pressures dominate Czech society. The focus is on building a new (socialist, communist) society and defining the conflict between the system and its real and alleged opponents. The focus is on the future and the prevailing tone is one of enthusiasm. Young people and children are targeted to represent these values. In some speakers, a full identification between their identity and the ideology of the system - and thus its language - can be observed. The sixties: This period is one of sobering up. Language reflects two main themes: (1) the attempt to escape from the restraints of the communist regime (socialism with a human face), (2) the end of all hope for political change after the Prague Spring, starting with the Soviet occupation of the former Czechoslovakia on August 21, 1968. The seventies and the eighties: The time of disillusionment and so-called normalization. Typical for speakers is not to identify themselves with their language.

## 2 Vocabulary collected

### 2.1 “Language of the rulers” – language of propaganda

This was gathered on the basis of the Corpus of Totalitarianism

“Totalitarian Corpus”

This is composed from journalistic texts. It includes three samples of Rudé právo (Red Justice), the daily newspaper of the Communist Party of Czechoslovakia, which reflected the ideological standpoints of the communist government:

The period 1948-1989 can be divided into three periods

#### 2.1.1 The 1950s (1952, a total of 926 texts, from 16.6 to 31.12.1952)

Examples of vocabulary

##### **Building a new order**

*agitátoři a propagandisté*

*agitators and propagandists*

*v agitačních střediscích*

*at “agitation centres”*

*v rudých koutcích pomocí agitek*

*in red cells helped by propaganda songs*

*stěngazety a desky cti*

*“wall newspapers”, “lists of honour”*

*agitace tlampači*

*agitation through loudspeakers.*

*komunisté, nestráníci*

*communists, non-party members .*

*reakcionáři, vykořisťovatelé, kulaci, fabrikanti*

*reactionaries, exploiters, “kulaks”, factor owners*

*podvracení republiky, velezradu, vlastizrada*

*subverting the republic, treason,*

*psovi psí smrt*

*“a dog’s death for a dog”*

##### **Atmosphere of the time**

*průvody*

*marches*

*manifestace s transparenty a s mávátky*

*demonstrations, banners, flags*

*alegorické vozy*

*floats*

*Sovětský svaz, náš vzor*

*The Soviet Union, our model*

*akademik Lysenko*

*The academic Lysenko*

*Mičurin, generalissimus Stalin*

*Micurin, Generalissimo Stalin*

*stachanovci*

*Stakhanovites*

*Zlobinova metoda*

*the Zlobin Method*

*etc.*

##### **Agriculture**

*kolektivizace*

*collectivization*



<i>združstevňování vesnice</i>	<i>“cooperativizing” a village</i>
<i>scelování pozemků</i>	<i>rationalizing parcels of land</i>
<i>rozorávání mezí</i>	<i>ploughing in the gaps between fields</i>
<i>etc.</i>	

### **Industry**

<i>havíři/horníci, úderníci, novátoři a vynálezci</i>	<i>miners/colliers, “shock-workers”, innovators, inventors</i>
<i>budování socialismu</i>	<i>building socialism</i>
<i>pětiletky</i>	<i>five-year plans</i>
<i>stavby socialismu</i>	<i>socialist constructions</i>
<i>stavby mládeže</i>	<i>youth constructions</i>
<i>závazky</i>	<i>commitments</i>
<i>zlepšovateľského hnutí</i>	<i>the “innovation movement”</i>
<i>etc.</i>	

### **2.1.2 1960s (1969, a total of 1038 texts, from 1.4 to 31.7.1969) Prague Spring**

Examples of vocabulary

<i>reformátoři</i>	<i>reformers</i>
<i>socialismus s lidskou tváří</i>	<i>socialism with a human face</i>
<i>ekonomická reforma, obrodný proces</i>	<i>economic reform, process of renewal</i>
<i>demokratizace, pluralita</i>	<i>democratization, plurality</i>
<i>deformace</i>	<i>deformation</i>
<i>dogmatik, konzervativec, kolaboranti</i>	<i>dogmatic, conservative, collaborators</i>
<i>bratrská pomoc, internacionální pomoc</i>	<i>fraternal support, international help</i>
<i>etc.</i>	

### **2.1.3 1970s and 1980s (1977, a total of 800 texts, from 3.1 to 31.3.1977)**

#### **Continuation of the period of “normalization”**

Examples of vocabulary

<i>normalizace</i>	<i>normalization</i>
<i>konsolidace</i>	<i>consolidation</i>
<i>agent, banda, bdělost, diverze</i>	<i>agent, band, vigilance, diversion</i>
<i>oportunist, područí, reakcionář</i>	<i>opportunist, bondage, reactionary</i>
<i>spiklenecká banda</i>	<i>conspiratorial gang</i>
<i>američtí váleční paliči</i>	<i>American warmongers</i>
<i>dřít kůži s těl dělníků</i>	<i>tearing the skin from the workers’ backs</i>
<i>grandiózní stavba socialismu</i>	<i>a grandiose socialist construction</i>

<i>krvavý pes Tito</i>	<i>that bloodstained dog, Tito</i>
<i>šťastné zítřky</i>	<i>a happy future</i>
<i>zahnívající kapitalismus</i>	<i>decaying capitalism</i>
<i>kontrarevoluce, krizové období</i>	<i>counter-revolution, period of crisis</i>
<i>prověrky</i>	<i>screening/vetting</i>
<i>výměna členských legitimací</i>	<i>renewal of party membership cards</i>
<i>pomýlený</i>	<i>misguided</i>
<i>vyloučení nebo vyškrtnutí ze strany</i>	<i>expelled or deleted from the party</i>
<i>zdravé jádro</i>	<i>the healthy core</i>
<i>exponent pravice</i>	<i>right-winger</i>
<i>souhlasit/ nebo nesouhlasit se vstupem</i>	<i>expressing agreement/disagreement with the intervention by</i>
<i>(spřátelených) vojsk</i>	<i>(friendly) troops</i>
<i>Chartra 77</i>	<i>officially described as a pamphlet</i>
<i>signatáři</i>	<i>signatories</i>
<i>samozvanci, zaprodanci rozvratníci</i>	<i>pretenders/usurpers, traitors, disruptive elements</i>
<i>samizdat</i>	<i>samizdat</i>
<i>edice Petlice</i>	<i>"Petlice" edition</i>
<i>pokrývač</i>	<i>"roofer"</i>
<i>jít do stoupy</i>	<i>"to be sent to the shredder"</i>
<i>trezorový film</i>	<i>"a film to be kept in the safe"</i>

Together with scans of 91 propaganda publications of varying lengths.

## 2.2 "Language of the ruled" – material has been gathered from

*Extracts from literary sources – novels etc. Personal experience – existing only in spoken form, these are expressions used among people who felt they could trust each other*

Language of the "ruled" – unofficial language

### 2.2.1 1950s

<i>kdy se to (v)obrádí</i>	<i>when will it turn round</i>
<i>kdy to rupne/ praskne</i>	<i>when will it crack/burst</i>
<i>je načichlej</i>	<i>he's "impregnated"</i>
<i>kopečkář, utýct (za kopečky)</i>	<i>runaway, running away "over the hills"</i>
<i>partajník, fabrika, fárplán</i>	<i>party man, factory, plan</i>
<i>1960s, 1970s and 1980s</i>	
<i>pravý džíny</i>	<i>real jeans</i>

## 2.2.2 1970s and 80s

<i>byl odejít ze strany</i>	<i>to be made to leave the party</i>
<i>Husákovo ticho</i>	<i>Husák's silence</i>
<i>Vokovická Sorbona</i>	<i>The Vokovice Sorbonne</i>
<i>RSDr. (ironicky Rodné cision of the Party"</i>	<i>(an academic title, referred to ironically as "Doctor by the De- nebo Rozhodnutím strany doktor) strany doktor</i>
<i>rychlokvaška</i>	<i>upstart, fast-track expert</i>

## 2.2.3 Used throughout the communist period

<i>aparátčik</i>	<i>apparatchik</i>
<i>papaláš</i>	<i>bigwig</i>
<i>Dederon (dederonský), dederon</i>	<i>slang for someone from East Germany</i>

## 3 Description of A small dictionary of life under communist totalitarian rule (Czechoslovakia 1948-1989)<sup>1</sup>

Includes more than 1,400 entries, drawn from a number of fields.

Includes:

1. Language of propaganda – drawn from the “Totalitarian Corpus”
2. Everyday language, capturing how people respond to propaganda, gathered through extracts from texts, through surveys, on the basis of personal experience and knowledge;
  - a) Language which captures the life of the time, through surveys and on the basis of personal experience and knowledge;
  - b) The entries also include very specific uses of language (e.g. the language of the secret police, of dissidents, prisoners.) It only includes words that came into common parlance

### 3.1 Some types of entries in A small dictionary of life under communist totalitarian rule: encyclopedia-type entries

The entry is made up of an encyclopedia-type explanation, taken from an example of the word used in context and stating the source of the example.

**Action Kulak** was the code name for a secret police operation between 1951-1954, under which awkward peasant families were forced to move and their property confiscated, they were tried on false

---

1 Schmiedtová (2012) *Malý slovník reálií komunistické totality*, Nakladatelství Lidové noviny, Praha ISBN 978-80-7422-192-7

pretences, imprisoned and discriminated: *Exactly fifty years have gone by since the beginning of Action Kulak, which the communist regime directed against peasants throughout the country in 1952* / internet.

**ensorship** /occurring only in texts from the 60s and 70s/ a central pillar of the regime; its discontinuation was one of the main prerequisites of the Prague Spring; it functioned under the auspices of the Federal Press and Information Department (FÚTI), up to 1968 under the Press Monitoring dept.: *It was far worse previously, when real censorship was exercised in newspapers and periodicals, cleverly managed and concealed as “journalistic solidarity”; following the badly organized, politically ill-prepared and ill-considered cancellation of censorship, the press came under the control and decisive influence of rightwing, opportunistic groups* / Corpus of Totalitarianism

**agitation centre** these were centres established by the Communist Party in villages, town districts and later in workplaces. Political agitation was carried out here, party education, information was published on noticeboards, instant messages and notices were put together and radio broadcasts were prepared, which were broadcast to people living nearby or to people at the workplace: *Under the principles approved by the secretariat of the Central Committee of the Communist Party agitation centres have been established in various places and socialist organizations* / Corpus of Totalitarianism

### 3.2 Some examples of words and phrases typical for the dictionary

**agent** /occurring predominantly in the 50s/ = **diversionist, spy** the high occurrence is the result of a phobia, seeking out people perceived as trying to subvert the new regime; people working for enemy intelligence organizations, trying to damage the communist order: *an agent of the American intelligence service; agents of American imperialism; agents of western imperialists; an agent of the bourgeoisie and an enemy of the Communist Party; CIA agents; with the help of a treacherous gang of agents* / Corpus of Totalitarianism

**gang** /the word occurs frequently primarily in the 50s/: *What was this **gang of conspirators** Slanský and his accomplices aiming at?; Slanský and his **criminal gang**; **a gang of Tito supporters**; smashing the **treacherous and marauding gang** of Clementis and co.* / Corpus of Totalitarianism

**not one grain should go to waste!** a popular slogan, primarily during the period of collectivization; the slogan also came to be parodied: *so that there will be enough bread in our republic, so that not a single grain of our rich harvest goes to waste* / Corpus of Totalitarianism

**facing the masses** a communist slogan: *Each communist is committed to the words of comrade Gottwald “Facing the masses”; during the continuous work to win the masses for the political work of the party – fulfilling the principle “facing the masses”* / Corpus of Totalitarianism

### 3.3 Words which reflect the real life of the time

**“androš”** /the word does not occur a single time in the Corpus of Totalitarianism/ **1** independent musical style, underground: *The only real underground Czech music is that of the Plastics and DG 307* /internet **2** an underground musician: *Brabenec’s journey from the underground to exile is a clear example of how the regime dealt with those it couldn’t control* /internet **3** a person with the outward appearance and lifestyle of the musical underground (long hair, shapeless sweater, scruffy jeans, avoiding regular work, hanging around in pubs, a kind of Czech “hippy”): *it’s true that for many years I haven’t given a shit about your average citizen, I’m more interested in non-average citizens – I mean guys with long hair, hippies, underground people [androše] or punks* /SYN [The word derives from the English word “underground”]

**bon** /the word does not occur a single time in the Corpus of Totalitarianism/ a token which could be obtained in exchange for hard western currency, and through which it was possible to buy goods in “Tuzex” shops. These were special shops where primarily western goods were sold. People without access to western currency could only buy these tokens on the black market from illegal currency traders. Officially one token was worth one Czechoslovak crown. On the black market in the 1980s the price for a token was around five crowns: *a whole hierarchy of illegal traders came into being, through whom even “ordinary” citizens could obtain tokens.* /SYN

## 4 Software used

We use Bonito (created by Pavel Rychlý, 2004) and TchwaneLex TLex Suit, version 7.1.0.726.

## 5 Conclusion

As we would expect under any political system, the language of totalitarianism in the former Czechoslovakia works within the semantic structure of Czech. However, it uses this structure for propaganda purposes, so words from the usual vocabulary are often abused to propaganda ends. The language is aggressive and monotonous, it frequently repeats certain associations, phrases and slogans. To certain words it adds its own evaluating positive or negative gloss. For example, the word *American* always has a negative semantic connotation, even though it is referring to a geographical concept; the word *Soviet* is always positive. Totalitarianism often abuses, to its ideological ends, words with a positive semantic connotation. It creates new meanings for words by expanding their polysemy, for example *western* = *capitalist*. It is fond of certain semantic connections, such as *building a better future*; *the struggle against enemies of the new order*; “*democratization of culture and education*”, which is a veiled re-

ference to censorship in these fields. With the aim of concealment it often uses euphemisms (*struggle for liberation*). This language is not creative, it draws from automatized components of the language. It often uses set phrases. To this day users often apply these phrases as ironic quotes, referring to the period.

The various tools of propaganda – techniques of persuasion, brainwashing, euphemisms – separate people into those who are with us and those who are against us, into the good and the bad, words take on new meanings, which have a political sense, linguistic stereotypes are used, which are repeated again and again, the propaganda works on the emotions, it is directed at ordinary people, which it perceives as a mass and a collective group, it tries to build its legitimacy on science, it speaks out strongly against the church.

The language of the ruled is spoken language, reacting to the pressure of propaganda. It is highly creative. It often parodies official language, it very often uses humour (e.g. *the "Vokovice Sorbonne", "to be made to leave the party of your own free will"*). It also captures the atmosphere of the time, which was influenced by the way the communist regime functioned (e.g. *real jeans, Tuzex token, Lenon Wall*). "

## 6 References

- Čermák, F., Cvrček, V., Schmiedtová, V. (2010) *Slovník kmunistické totality*. Prague: Lidové noviny.
- Schmiedtová, V. (2006) What did the totalitarian language in the former socialist Czechoslovakia look like? *The First Conference of The Slavic Linguistics Society* - <http://www.indiana.edu/~sls2006/page2/page2.htm>
- Schmiedtová, V. (2007) Totalitní jazyk v bývalém Československu. Koncept slova práce. In: *Totalitarismus 3, sborník z konference, katedra antropologie, FF ZU, s.110 – 116*
- Schmiedtová, V. (2008) Hodnotící prostředky v totalitním jazyce 1948-1989 v bývalém Československu. In: *Totalitarismus 4, sborník z konference, katedra antropologie FF ZU. Plzeň, s. 186 – 196*
- Schmiedtová, V. (2011) Die Sprache der Propaganda in der Tschechoslowakei 1948-1989 In: *Brücken, Germanistisches Jahrbuch Tschechien-Slowakei*, Nakladatelství Lidové noviny ISBN 1803-456X, s. 93-115
- Rychlý, P.: Bonito - graphical user interface to the system Manatee, version 1.80
- Кобрин, К. (2012) *Vita Sovietica. Неакадемический словарь-инвентарь советской цивилизации*. Издательство Август, Пермь Россия
- Mokijenko, V. M., Nikitina T. G. (1998) *Tol'kovyj slovar jazyka Sovdep'ii, Sankt Peterburgskij gosudarstvennyj universitet Possija*
- Mokijenko, V. M. (2003) *Novaja russkaja frazeologija*, Opole, Polsko

### Aknowledgements

This study was written within the Programme for the Development of Fields of Study at Charles University, No. P11 Czech national corpus

# A Frequency Dictionary of Dutch

Carole Tiberius, Tanneke Schoonheim, Adam Kilgarriff  
Institute of Dutch Lexicology, Lexical Computing Ltd  
{carole.tiberius,tanneke.schoonheim}@inl.nl, adam@lexmasterclass.com

## Abstract

In this paper, we present a corpus-based frequency dictionary of Dutch containing the 5000 most frequent words of Dutch. The dictionary has been published at the beginning of 2014 as part of the Routledge Frequency Dictionaries series, a well-established series with titles available for 11 languages at the time of writing. Novel in the Dutch frequency dictionary is that genre has been foregrounded. The dictionary does not contain one single frequency list, but multiple lists are presented, of which four are genre specific covering fiction, newspaper, spoken and web. Throughout the dictionary there are also thematically organised lists featuring the top words from a variety of key topics such as animals, food and other areas of daily and cultural life. Words specific to Dutch in Belgium are also included. The dictionary is based on a 290-million-word corpus which includes both written and spoken material from a wide range of sources.

**Keywords:** frequency dictionary; genre; Dutch.

## 1 Introduction

The *Frequency Dictionary of Dutch* provides the 5000 most frequently used words in contemporary Dutch and is specifically targeted at the beginning and intermediate language learner. This is not the first and only frequency list for Dutch, but there was certainly a need for an update. The best-known reference for word frequencies in Dutch is *Woordfrequenties in geschreven en gesproken Nederlands* by P.C. Uit den Boogaart from 1975. Another much used resource, the CELEX database, is more recent (the second release dates from 1996), but it is still over 15 years old and not widely distributed amongst language learners. The current frequency dictionary is contemporary. It is based on a large corpus of Dutch, spanning the past forty years and concentrating on the last twenty.

In Section 2, we briefly summarise the methodology used to compile the frequency dictionary. Section 3 presents the dictionary and discusses a number of issues we encountered while compiling the dictionary and how we have dealt with them. Section 4 concludes the paper.

## 2 Methodology

The dictionary is based on a 290 million word corpus of contemporary Dutch divided between four genres: fiction, newspaper, spoken and web.<sup>1</sup> This corpus is the result of a compilation of existing Dutch corpus material (Corpus Spoken Dutch (CGN), fiction from INL corpora and newspaper and web material from the SoNaR corpus).

A central problem in preparing frequency lists on the basis of corpora is the ‘*whelks*’ problem: if there is a text about *whelks* (a variety of mollusc) then the word *whelk* will probably occur many times in this text but not in the other texts of the corpus. If all occurrences of the word *whelk* are given equal weight, the resulting word frequency list will be skewed as this one text about *whelks* will push up the count of this otherwise rare word. To deal with this problem, we used a fixed-sample-size corpus (cf. the Brown corpus). We first truncated very long texts at 40,000 words, so that we did not have too many samples from any single text, and then we simply concatenated all the texts of each genre and cut into samples of 2000-words each.

Once the corpus had been assembled, it was lemmatised and tagged using the Frog software (van den Bosch et al. 2007).

Some manual checks were carried out (see Section 2.1), and then we calculated, for each genre, for each word, what proportion of samples it occurred in and normalised these figures to give percentages.<sup>2</sup> We then defined an algorithm for determining which words go into which list(s). See Kilgarriff and Tiberius (2013) for a detailed description. As some words occur in more than one of the four genre lists (e.g. *aankomst*<sub>fiction(885) | newspapers(499)</sub> ‘arrival’ occurs in fiction and newspaper), the sum is slightly higher than 5000. The words are distributed across the lists, as follows:

LIST	Core	Fiction	Newspaper	Spoken	Web	General
<b>WORDS</b>	943	1084	1129	155	523	2004
<b>CORPUS SIZE (millions of words)</b>		23	167	9	91	

**Table 1: Number of words in the different lists and subcorpora.**

### 2.1 Manual checking and correction

While the Frog tagging and lemmatisation software is good, it does produce occasional unwanted results. For instance, inflected forms were sometimes analysed as separate lemmas. This occurred in particular with singular and plural forms of certain nouns (e.g. *belasting* ‘tax.SG’ and *belastingen* ‘tax.PL’, *maand* ‘month.SG’ and *maanden* ‘month.PL’), as well as with masculine and feminine forms of cer-

1 We use genre in the general sense referring to broad text types.

2 Thus frequency in the dictionary is always the percentage of documents that a word occurs in.



tain nouns (e.g. *advocaat* ‘laywer.MASC’ and *advocate* ‘laywer.FEM’) and with diminutive forms (e.g. *lied* ‘song.SG’ and *liedje* ‘song.DIM’). Inflected forms of some pronouns, adjectives and verbs also produced double lemmas (e.g. *elk, elke* ‘each’; *raar, rare* ‘strange’ and *herkend, herkennen* ‘to recognise’). We have corrected the most frequent and evident errors manually, producing a list of lemmas of which the frequencies had to be counted together. In addition, we decided to count abbreviations together with their corresponding full forms, e.g. *kilometer, km*. This was also a manual task.

One of the characteristics of Dutch is that it is possible to separate parts of compound verbs like *uitleggen* ‘to explain’, *vasthouden* ‘to hold’ etc. in the sentence allowing others parts of the sentence to occur in between them (e.g. *hij legt het probleem duidelijk uit* ‘he explains the problem clearly’ and *hij hield het meisje stevig vast* ‘he held on to the girl firmly’). Automatic recognition of such separable verbs is error-prone and there were many instances where they were tagged as separate lemmas. Unwanted particles resulting from these split separable verbs have been manually filtered out of the resulting lemma list.

### 3 The dictionary

The main part of the dictionary is formed by the six frequency lists. These are:

- **Core:** words occurring with high frequency in all four genres
- **Fiction:** high-frequency fiction words
- **Newspaper:** high-frequency newspaper words
- **Spoken:** high-frequency words in spoken Dutch
- **Web:** high-frequency words on the Dutch web
- **General:** the next band of words which have high frequencies across at least three of the genres.

The words in the lists are sorted by frequency. In the core and the general list sorting is done on the basis of the overall frequency of the words in all four genres. In the genre-specific lists, the ordering is based on the frequency within that genre, rather than the overall frequency. Each entry in these lists contains the headword, its part of speech, an English translation of the commonest sense, and an example sentence showing how the word is used as is illustrated below:

**core(509) televisie noun de(f) television**

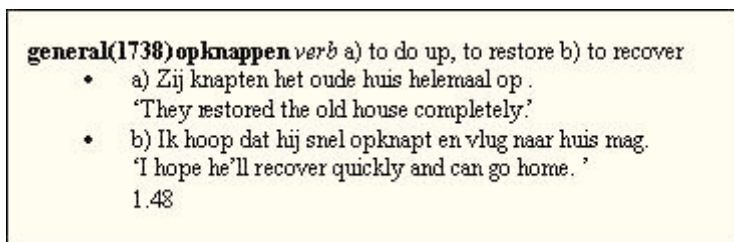
- Hij zette de televisie aan om naar het nieuws te kijken.

‘He switched the television on to watch the news.’

16.55

This entry shows that word number 509 in the rank order list for the core vocabulary is the noun *televisie* ‘television’. It is a feminine noun, which takes the article *de* in Dutch and has an overall frequency of 16.55 per 100 documents. The example sentence is taken from the corpus and shows the word as much as possible in a representative natural context.<sup>3</sup>

Normally, only an example of the commonest sense is given. If however a word has two meanings which are both equally common, two example sentences are given so both meanings can be illustrated.



In addition to the six frequency lists, the dictionary contains:

- an alphabetically-sorted index;
- an index of the commonest words by part of speech (nouns, verbs, adjectives, adverbs, prepositions, conjunctions and interjections).

Furthermore, there are boxes throughout the book which contain smaller lists of thematically related words, e.g. body, food, materials or grammatical information, e.g. paradigms of auxiliary verbs or lists of pronouns.

In the remainder of this section, we discuss a number of difficulties that we encountered whilst compiling the dictionary and how we solved them.

### 3.1 Example sentences

For each entry in the dictionary, an example sentence is given. The example sentences were supplied semi-automatically using the GDEX tool (Kilgarriff et al. 2008) from the Sketch Engine. GDEX (*Good Dictionary Examples*) is a tool which automatically sorts the sentences in a concordance according to how likely they are to be good dictionary examples. That is, the best examples are sorted to the top of the list and they are the ones the lexicographer sees first. GDEX was designed for English, so the heuristics that are used are specific to English or they were set with a particular group of users in mind. The tool had not been used on a large scale for Dutch before this project.

---

3 Examples are not translated in the dictionary, but a translation has been added here for clarity.

For the frequency dictionary GDEX automatically provided six candidate sentences from the corpus (or from the relevant subcorpus for the genre lists) for each headword which were put in an EXCEL spreadsheet. From these six examples, the best one was chosen manually, marking it with a Y.

	Zij was dan iemand net als zichzelf en niet als de mooie dames op de televisie.
	Hij was iets kleiner dan ik had gedacht op grond van zijn optreden op de televisie.
	De televisie staat op sneeuw.
	Op een avond heb ik haar betrappt terwijl ze huilend voor de televisie zat.
Y	Hij zette de televisie aan om naar het nieuws te kijken.
	Ik ga soms pontificaal voor de televisie staan als ik iets wil zeggen.

**Figure 1: Automatically generated example sentences for the noun televisie ‘television’.**

This worked surprisingly well considering that the tool has not been customised to Dutch. In many cases we shortened or simplified the original corpus sentences to make them more suitable for the language learner. For instance, referential pronouns and personal names have been replaced by personal pronouns.

If none of the automatically selected example sentences were good enough, an alternative example was selected and prepared after examining more corpus examples. This applied to words, like the noun *gek*, which also occur frequently as part of a phrase (i.e. *voor de gek houden* ‘to pull someone’s leg’, *voor gek staan* ‘to look like a fool’) or as another part of speech (i.e. the adjective *gek*).

	Montaigne schrijft ergens dat hij niet weet wie wie voor de gek houdt als hij met zijn kat speelt.
	Of gekken als geheime agenten.
	Die bol draait als een gek in de rondte en slaat zonder onderscheid van alles bij je bewustzijn naar binnen.
	Ze staat hier voor gek.
	Zij acht aan artiesten als aan gekken die elk ogenblik gevaarlijk konden worden.
	Maar het is gek dat je bij die dingen nooit denkt dat het ook zo dicht bij je gebeuren kan.

**Figure 2: Automatically generated example sentences for the noun gek ‘fool, idiot’.**

### 3.2 Translations

The dictionary contains the 5000 most frequent words of Dutch. For each, an English translation of the commonest sense is given. High frequency words are often polysemous and it has not always been straightforward to determine what the commonest meaning of a word is or whether there are different meanings which are all equally common. An example is the verb *optreden*<sub>core(727)</sub> which has

been translated as ‘to appear’, but can also mean ‘to perform’. As the corpus is not sense-tagged, this is a grey area and decisions on what the commonest meaning is have been made after manually inspecting the corpus data and relying on other resources (ANW, Van Dale).

There are also cases where a different translation is more appropriate depending on the genre in which the word is used. For instance, the verb *besturen*<sub>newspaper(681) | web(429)</sub> has been translated as ‘to govern’ in the newspaper list and as ‘to drive’ in the web list.

As the dictionary is targeted at language learners we have tried to assure as much as possible that the translations used belong to the core vocabulary of English. This has not always been possible. We have had long discussions about the appropriate translation for *wijf*<sub>fiction(552)</sub> in English. In unmarked cases it can be translated as ‘woman’. For the marked case we have ultimately settled for the word ‘broad’ which is neither core, nor general vocabulary (Van Dale marks it as American-English), but seems to express the Dutch connotations of the word best. Another example of a problematic translation was the opposite *gelovig*<sub>general(1893)</sub> – *ongelovig*<sub>fiction(889) | web(512)</sub> which we have translated as ‘faithful’ and ‘faithless’ in the thematic box of opposites. The adjective *gelovig* is mostly used in a religious context, whereas the opposite *ongelovig* also has a broader sense namely of expressing disbelief which seems to be more common in fiction texts.

Translation of specific terms related to local politics such as *gemeente*<sub>newspaper(16) | web(7)</sub> ‘municipality’, *schepen*<sub>newspaper(153)</sub> ‘local councillor’ also proved difficult, because these do not match exactly the words that look like their English counterparts.

### 3.3 Syntactic category

As a rule of thumb, we used the part of speech assigned by the Frog tagging and lemmatisation software in the dictionary. However, there are a few cases where we have diverted from this strategy. This is the case for the adverbial use of adjectives, where the adverbial use of the word is considered secondary to the adjectival use. In the dictionary, these words have been labelled as adjectives, even if the adverbial use was more common in the corpus. An example sentence of both uses is given as is illustrated in the entry for *absoluut*:

core(589) *absoluut* *adj* absolute

- Twintig juni is de absolute deadline.  
‘Twenty June is the absolute deadline.’
- (adv) Ik was het absoluut niet met haar eens.  
‘I absolutely did not agree with her.’

14.19

Note that the English translation for the adjectival and adverbial use are not the same.

There were also a few lemmas where it was difficult to assign a single and consistent part of speech, for example, the lemmas *meer* ‘more’, *meest* ‘most’ and *minder* ‘less’, *minst* ‘least’. Existing resources for

Dutch (e.g. WNT, Van Dale, the official Dutch spelling guide *Woordenlijst Nederlandse Taal*) do not agree here on the part of speech, indicating the words as adverbs, adjectives and numerals in various combinations (see Table 2):

Lemma	WNT	Woordenlijst	Van Dale GW	Van Dale Hedendaags
meer	adv;num	adj	adv;num	adv;num
meest	adj;adv;num	adj	adj;adv	adv

**Table 2: Comparison of the part of speech attributed to meer and meest.**

The most frequent use of these words in the corpus appeared to be as an indication of a certain amount. This use is considered to be typical for numerals and so in the *Frequency Dictionary of Dutch* these words are labelled as numerals.

### 3.4 Other cases

In some cases, an entry headword has been given a subentry. This has been done with headwords which are known to cause spelling errors, such as *ten minste* and *tenminste*. Both lemmas exist in Dutch, but they have a different meaning. The word *tenminste* means ‘at least’ while *ten minste* written in two separate words means ‘with a minimum of’

as is illustrated by the two example sentences in the entry below:

<p><b>core(821) tenminste</b> <i>adv</i> at least</p> <ul style="list-style-type: none"> <li>Dat klinkt mij tenminste heel bekend in de oren. ‘That sounds at least very familiar to me.’</li> </ul> <p><b>ten minste:</b> Je moet ten minste tien punten hebben om verder te mogen. ‘You need to have a minimum of 10 points to go through.’</p> <p>9.00</p>
---

It is very likely that in the corpus the appropriate form has not always been used in the appropriate context and thus counts will be skewed anyway. Our approach has been to list them in a combined entry.

Subentries have also been used for multi word expressions as for example in the case of the noun *beslag*<sub>general(444)</sub> which means both ‘batter’ and ‘fittings’, but also occurs as part of the phrase *in beslag nemen* ‘to confiscate’.

Reflexive verbs have been marked by including the reflexive pronoun *zich* behind the verb entry. For example *beklagen (zich)*<sub>fiction(1070) | newspapers(847)</sub>. The verb *beklagen* is not obligatory reflexive. In a sentence like *Zij kijkt hem vol medelijden aan en beklaagt hem* it means ‘to pity’. When used with the reflexive pro-

noun *zich*, the verb *beklagen* means ‘to complain’: *Hij beklagde zich erover dat hij zijn kantoor niet kon bereiken*. ‘He complained that he could not reach his office.’ An example of each is included in the dictionary.

## 4 Conclusion

In this paper we have discussed the *Frequency Dictionary of Dutch* that has just appeared as part of the Routledge Frequency Dictionary series. It provides first and foremost a valuable resource for learners of Dutch, but it is fascinating for anyone interested in the Dutch language. The web material (never used before in a Dutch frequency dictionary) appears to be an interesting mixture of informational language like the language used in the newspaper genre, and a written form of spoken language, as used in blogs and discussion groups. Newspaper material shows a focus on economy and sports, whereas the material taken from fiction tends to be rather conservative.

Besides this, it is material which provides lots of new research questions for (socio-)linguists and lexicographers. As van Oostendorp (2014) points out, while reading the dictionary, you can’t help wondering why *schouders* ‘shoulders’ are so popular in fiction and why *januari* ‘January’ is the most frequently mentioned month on the web followed by *juni* ‘June’, *mei* ‘May’, *december* ‘December’, *oktober* ‘October’ and *maart* ‘March’. The frequency dictionary itself does not provide the answers, but these are intriguing observations about our use of the Dutch language.

## 5 References

- Bosch, van den A., Busser, G.J., Daelemans, W., and Canisius, S. (2007). An efficient memory-based morpho-syntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium. 99-114.
- Kilgarriff, A. and Milos Husák, Katy McAdam, Michael Rundell, Pavel Rychlý. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In: Elisenda Berndal and Janet De Cesaris (eds), *Proceedings of the XIII EURALEX International Congress*, Barcelona, Spain. 561-569.
- Kilgarriff, A. and C. Tiberius (2013). Genre in a Frequency Dictionary. In: Andrew Hardie and Robbie Love (eds.) *Corpus Linguistics 2013 Abstract Book*. Lancaster. UCREL, 142-144.
- Oostendorp, van M. (2014). Het karakollenprobleem. Accessed at: <http://nederl.blogspot.nl/2014/03/het-karakollenprobleem.html>. [04/04/2014].
- Uit den Boogaart, P.C. (1975). *Woordfrequenties: in Geschreven en Gesproken Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

### Resources:

- Algemeen Nederlands Woordenboek (ANW)* Accessed at: <http://anw.inl.nl/> [20/08/2013]
- The CELEX Lexical Database (1995), R.H. Baayen, R. Piepenbrock and L. Gulikers, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Corpus Gesproken Nederlands (CGN), (2004), Nederlandse Taalunie, Den Haag.  
INL-Corpora, Instituut voor Nederlandse Lexicologie, Leiden. Accessed at <http://chn.inl.nl> [01/02/2014]  
SoNaR (2010), Nederlandse Taalunie, Den Haag.  
*Woordenboek der Nederlandsche Taal (WNT)* Accessed at: <http://gtb.inl.nl/> [04/04/2014]  
Woordenlijst Nederlandse Taal, Nederlandse Taalunie, Den Haag. Accessed at: <http://woordenlijst.org/>  
[04/04/2014]  
Van Dale *Groot Woordenboek van de Nederlandse Taal*, Van Dale Lexicografie, Utrecht. Online version  
[04/04/2014]  
Van Dale *Groot Woordenboek Hedendaags Nederlands*, Van Dale Lexicografie, Utrecht. Online version  
[04/04/2014]

### **Acknowledgements**

We gratefully acknowledge our colleagues at INL who have helped us in the process of compiling the frequency dictionary. We also like to thank the Sketch Engine team for their support and Karen Newton and James McAllister for their help with the English.





# The Corpus of the Croatian Church Slavonic Texts and the Current State of Affairs Concerning the Dictionary of the Croatian Redaction of Church Slavonic Compiling

Vida Vukoja  
Old Church Slavonic Institute, Zagreb  
vidal@stin.hr

## Abstract

Croatian Church Slavonic is a literary, bookish idiom used in Croatia from the XI/XII until the XVII c. based on Old Church Slavonic, an idiom created by Sts. Cyril and Methodius, and shaped by the Croatian vernacular.

The Croatian Church Slavonic corpus consists of the texts excerpted from 11 breviaries, 4 missals, 3 psalters, 3 rituals, 15 miscellanies. It also incorporates all the 26 fragments dated from the period up to and including the XIII c. and several auxiliary sources. The corpus was created over a period of thirty years (from 1959 to the early 1990s) at the Old Church Slavonic Institute in Zagreb. It is a historical, referential, representative paper card-file of excerpts. It is also a parallel corpus, as it contains Latin and Greek parallel texts, those that were identified as closest to the actual source texts for the translational Croatian Church Slavonic texts.

The *Dictionary of the Croatian Redaction of Church Slavonic* has been compiled on the basis of the Croatian Church Slavonic corpus. The fascicles of the *Dictionary* have been published since 1991. So far, 1 (1991)–19 (2012), with the dictionary articles A–ŽRBTVĀ (according to the Old Cyrillic alphabet) have been printed.

**Keywords:** corpus of the Croatian Church Slavonic; Dictionary of the Croatian Redaction of Church Slavonic; (Paleo)slavistic lexicography

## 1 Basic Information on the Croatian Church Slavonic Language (acr. CCS)

The Cyrillo-Methodian landmark mission, with its far-reaching influence on Medieval European culture and history, felt on a broad scale even today, commenced in 863, when the Thessalonian brothers arrived in Great Morava. Cyril and Methodius created a new language, now known as Old Church Slavonic (acr. OCS), for the purpose of translating Greek biblical and liturgical texts, literary texts as well as those concerning administration and law. That language did not match any particular Slavic vernacular, as it was constructed to potentially serve all the Slavic peoples ready to be evangelized in the

Slavic language (i.e. OCS), and willing to embrace literacy in a newly-invented script for its notification – Glagolitza or the Glagolitic script, created probably by Cyril.

After the end of the mission, a certain number of Cyril and Methodius' disciples arrived in the territory populated by the Croats. Under the influence of the Croatian vernacular, a new Church Slavonic language system came to be. That literary, bookish idiom used in Croatia from the XI/XII until the XVII c. is known as the Croatian Church Slavonic. It had a privileged status throughout the Croatian Middle Ages within the Croatian diasystem,<sup>1</sup> characterized by the Croatian/CCS diglossia<sup>2</sup>, as it was a liturgical language, whose usage regularly marked a high literary style.

Two things should be mentioned in order to signal the importance of CCS. The first one considers European culture and history, and the even broader context of the Catholic Church history. CCS was the only close-vernacular idiom which gained and retained the explicit permission of the Pope to be used for liturgical purposes (besides Latin, Greek and Hebrew). Therefore, ahead of the decision of the Second Vatican Council allowing Catholic liturgy in vernacular, a CCS mass was served in the Roman St. Peter's basilica. The second thing to mention considers Croatian literacy. Namely, CCS is the first Croatian literary language, used from the end of XI c. until 1561. Its significance is reflected in the fact that it is the language of the *Baška tablet* (Cro. Bašćanska ploča, dated in 1100), one of the first and most important Croatian written monuments, but also the language of the first Croatian incunabula, *Missale Romanum Glagolitice*, which is the first missal in Europe not published in the Latin script and Latin language. Six out of nine Croatian incunabula were printed in CCS. The preserved Croatian texts in the Cyrillic script date from later periods, as is the case with the texts in Latin script.

## **2 The Corpus of the Croatian Church Slavonic texts (abbr. the CCS Corpus)**

### **2.1 Basic Information on the Corpus of the Croatian Church Slavonic texts (abbr. the CCS Corpus)**

Here, the term “corpus” is understood as a language material, a cluster of texts, purposefully collected to testify choices and combinations of choices made by users of a particular language (Sinclair 2003:167; cf. Svensén 2009: 43). The CCS corpus is primarily prepared for the compiling of the *Dictionary of the Croatian Redaction of Church Slavonic* (acr. DCRCS), but due to its features, it serves as a prime source for various linguistic investigations and other types of research.

---

1 For the meaning of the term diasystem v. Weinreich (1954), cf. Brozović (1970 [1967]): 14. For the CCS and Croatian vernacular and literary idioms as constituents of one common diasystem v. Katičić (1992).

2 For the character of the Croatian/CCS diglossia v. Mihaljević (2010).

The history of the CCS corpus started with the Fourth International Slavistic Congress, which took place in Moscow in 1958<sup>3</sup>, where the suggestion to compile a thesaurus of the Church Slavonic language was embraced by the leading (paleo)slavists. The planned thesaurus was supposed to incorporate all the national Church Slavonic versions (or redactions): Bulgarian, Macedonian, Czech, Russian, Romanian, Bosnian, Serbian. It is important to mention that during that very time, the first fascicles of the landmark *Slovník jazyka staroslověnskeho* (abbr. *Slovník*; v. Štefanić 1962; Nazor 1991: I), based on the corpus of the canonical OCS texts, started to appear. The form and principles applied to the corpus prepared for the *Slovník* compiling, and the *Slovník* compiling itself, decisively influenced the CCS corpus and the DCRCS compiling respectively. The decision of the Croatian lexicographers to follow their Czech colleagues was not due to any lack of ingenuity, but was incited by the long-sighted priorities to shape the CCS corpus as similarly as possible to the corpus for *Slovník* compiling and to shape the DCRCS as similarly as possible to the *Slovník*. The aim of those decisions was to enhance the possibilities for comparative research ultimately aimed at attaining the structural knowledge of: (a) particular national versions of Church Slavonic, (b) the (Old) Church Slavonic language system, taken apart from any (national) vernacular's influence.

For that reason, the arrival of the reputable and experienced Czech paleoslavistic lexicographer, František Václav Mareš, one of the principle collaborators at the compiling of *Slovník*, was very much welcomed, as was his cooperation with the then-young Croatian paleoslavists, in the task of his laying-out of the main principles for the CCS corpus and the compiling of the DCRCS drafted in Mareš (2007[1962]), as well as setting up the long-lasting work of the DCRCS compiling. It took about thirty years to complete the CCS corpus (from 1959 until the beginning of the 1990s). Despite its apparent excessive duration, it was actually the expected length for such a demanding task.

## 2.2 The Constituents, Card-files and Features of the CCS Corpus

The CCS corpus consists of selected CCS sources, manuscripts and incunabula dating from (XI)/XII to mid-XVI c., with the priority given to earlier and integral versions of particular texts. Its constituents are as follows: 11 breviaries, 4 missals, 3 psalters (1 with Psalter commentary), 3 rituals, selected texts from 15 Croatian Glagolitic miscellanies, all the fragments dating from the period up to and including the XIII c. (altogether 26 pieces), auxiliary sources are excerpted in the cases of lexicographically interesting lemma occurrences: another 2 missals and 2 breviaries.<sup>4</sup> A vast range of text genres found their place within the CCS corpus: liturgical texts (including biblical passages), biblical and apocryphal texts, sermons and homilies, moral and didactical texts, legal texts, legends and visions, hagiographies, disputations and other literary texts.

---

3 For the information on the prehistory of the CCS corpus v. Nazor (2008).

4 For the exhaustive list of the sources, constituents of the CCS corpus v. Nazor (1991).

The exact number of the CCS corpus tokens is not known, but it should be somewhere between 1 400 000 and 2 100 000. The excerpts are of various lengths and multiplied so that they can be organized within two card-files (an example of excerpt cards can be seen in the Figure 1).

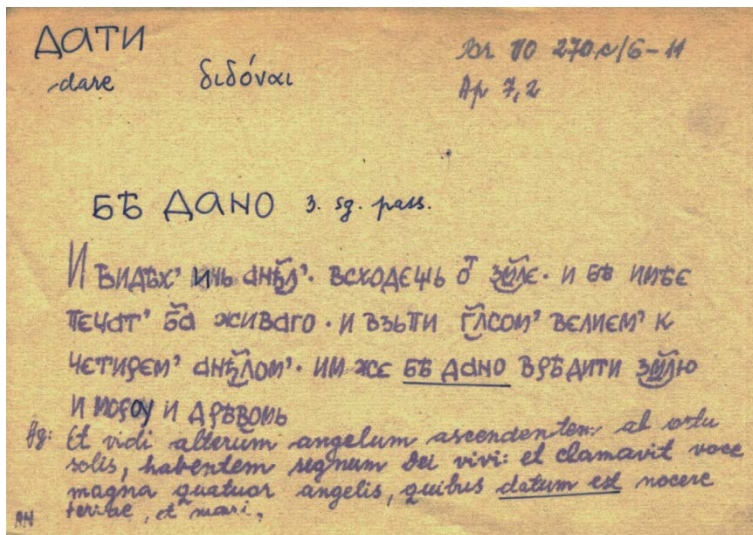


Figure 1: An excerpt card (used in the sources card-file and in the azbuka card-file).

Along the excerpt cards run the so-called parallel cards with the variations of the lexical constituents of the excerpt as found in other sources (an example of a parallel card is given in Figure 2.).

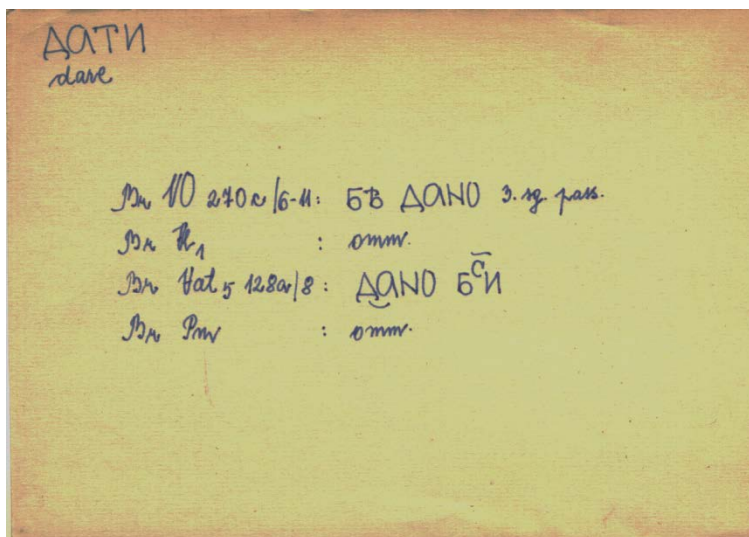


Figure 2: A parallel card (used in the sources card-file and in the azbuka card-file).

The first card-file is established according to the sources (informally, the sources card-file) and it contains approximately 420 000 cards. The second card-file is established according to the azbuka sequence of the lemmas (informally, the azbuka card-file), with more than 400 000 cards. The azbuka

card-file contains fewer cards than the sources card-file because not every token is taken into consideration for the compiling of the entry of the DCRCS, and consequently, its card doesn't appear in the latter card-file.

There are three auxiliary card-files. The first one to be mentioned is the card-file of the CCS lemmas (systematized according to the Old Cyrillic azbuka) with approximately 18 100 cards (of which approximately 8500 nouns, 5400 verbs, 2500 adjectives; an example of a card is shown in the Figure 3).

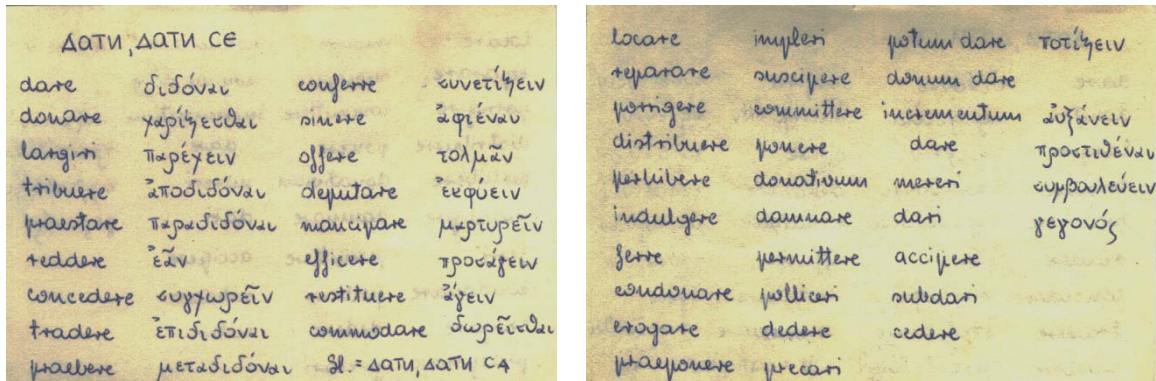


Figure 3: A card from the card-file of CCS lemmas (recto and verso).

The second one is the card-file of the Greek parallels of the CCS lemmas (systematized according to the Greek alphabet) with approximately 60 000 cards (two examples are given in the Figure 4).

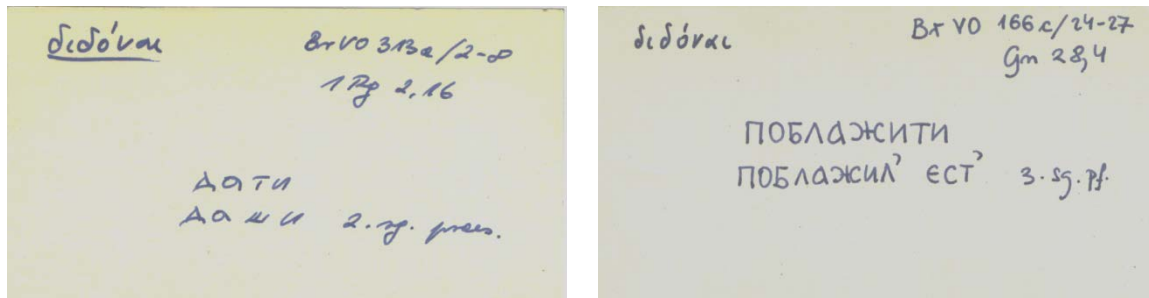


Figure 4: Two cards from the card-file of Greek parallels.

Also, there is the card-file of the Latin parallels of the CCS lemmas (systematized according to the Latin alphabet) with approximately 200 000 cards (two examples can be seen in the Figure 5).



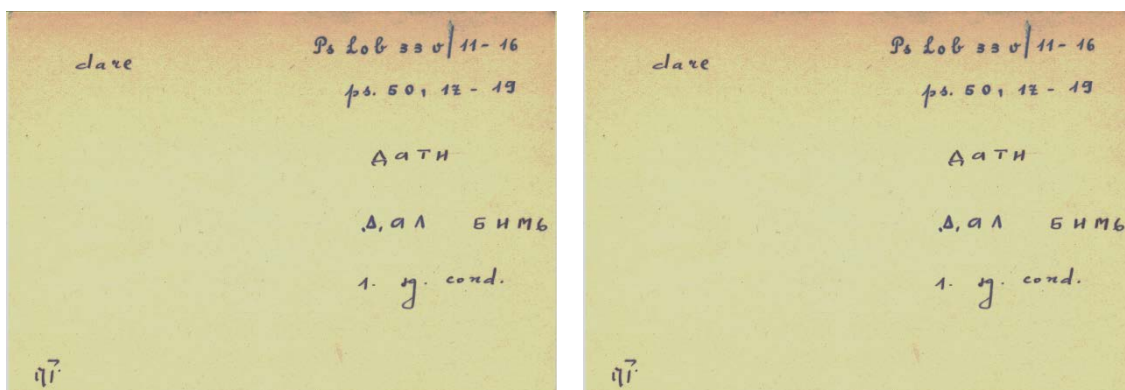


Figure 5: Two cards from the card-file of Latin parallels.

The main features of the CCS corpus are shaped by the principles articulated by Mareš (2007[1962]), and they can be listed as follows:

- (1) The CCS corpus is *historical*: it contains texts originating from (XI)/XII to XVI c.
- (2) The corpus is made with the aspiration to enable the production of a *thesaurus*, i.e. a dictionary containing the entire vocabulary range confirmed in the CCS texts.
- (3) It is *referential*, general; i.e. it contains virtually all of the confirmed CCS lexemes in a number of instances and types of context which represent the situation in the entire (preserved) corpus of the CCS texts. Also, it contains various versions of the same lemma in terms of its graphic features (such as manners of noting /b/, /ъ/, /m/, Latin or Cyrillic script initials in otherwise Glagolitic texts etc.) as well as its phonological, morphological, lexical, textological features. Therefore, the corpus meets the requirements for various types of research, be it linguistic or non-linguistic, such as textological, literary, liturgical, historical etc.
- (4) The corpus is *representative*. It contains texts of different genres, mirroring in quantity the real share of particular genres in the entirety of CCS texts.
- (5) In form, the corpus is *paper card-file* of excerpts, where excerpts are used for practical reasons, as expected considering the period it was set up. Still, it cannot be perceived as a mere collection of citations, as the sequences of the card-files contain integral texts, and even entire hundreds-of-pages long manuscripts, just as it is expected in the case of a proper corpus (and unlike expected for a collection of citation).
- (6) The corpus *constituents* are *not strictly divided*. Despite being fundamentally systematized on the basis of the source type (missals, breviaries, psalters, rituals, fragments, texts selected from miscellanies), the corpus is not subdivided into compartments. But, if important versions of a particular text appear in different groups of sources, the differences are noted on a string of parallel cards.
- (7) It has characteristics of *static* and *dynamic* corpora. A static corpus (also, *sample corpus*) is a structured collection of a limited number of examples of the lemmatized lexemes, which, once introduced in the collection, cannot be excluded from the corpus. And the CCS corpus is such a collec-

tion in its azbuka card-file. A dynamic corpus (also, *monitor corpus*) consists of entire texts and it is open to receiving a newly-found text or confirmation of lemmas shown to be desirable concerning the referential and representative features of the corpus, and the CCS corpus is open to the incorporation of such newly-found texts.

- (8) The corpus contains *only written texts*, with a great majority of texts being manuscripts and a limited number of incunabula. As it concerns medieval language material, no spoken texts are expected.
- (9) Besides the above reason, the CCS corpus can contain written texts only, as CCS itself is a *bookish* idiom (not spoken), *non-organic*, and *not a native language* of anyone.
- (10) Texts for the corpus are *excerpted directly from the original sources or photocopies of the original sources* (not from any secondary editions of the selected texts).
- (11) The corpus is *parallel*, not monolingual (on parallel corpora v. Svensén 2009:55-57; also Borin 2002). Whenever an excerpted CCS text is actually a translation from Greek or Latin (and most often this indeed is the case), if the parallel Greek or Latin text has been determined, the cards contain that parallel text below the CCS texts, as strictly aligned as possible.
- (12) The corpus is formed in such a manner that the *basic version of the text is determined and differentiated* from the secondary versions of the same text found in various sources.
- (13) The *critical, phototypical and facsimile editions* of the texts included in the corpus are *published*, the most ambitious undertakings being the capital editions of Hrvoje's Missal (Grabar et al. 1973) and the Second Novi Breviary (Pantelić & Nazor 1977).
- (14) Approximately *18,100 lemmas* are *expected* in the dictionary once it is finished.
- (15) Without exception, the texts incorporated in the CCS corpus are written in the *Glagolitic script*, but on the card-files of the CCS corpus they appear transliterated in the Old Cyrillic script, according to the decision made at the Fourth International Slavistic Congress. The principles of the transliteration from Glagolitic into Old Cyrillic script are appropriated from Jagić (the table of Glagolitic and Old Cyrillic correspondents can be found in Jagić 1879: XXXVII; cf. Bratulić 1981: 145-146). The lemmas of the DCRCS are also written in the Old Cyrillic script. Of course, Greek parallel texts are written in the Greek alphabet and Latin parallel texts are written in the Latin script.
- (16) The *tokens* of the CCS corpus, appearing in the sources card-file, are *grammatically tagged* (parsed) and inserted into the azbuka card-file in all the cases except those of extremely frequent types, such as the forms of the verb *biti* ('to be') or the noun *bog* ('god').
- (17) All the types are *lemmatized* in a *normalized form*. The principles of normalization are given in Grabar et al. 1991: VIII-XIII, XIX-XXV.

## 2.3 The European and Croatian Context of the CCS Corpus and the Compiling of the DCRCS

### 2.3.1 The Situation in the (European) Paleoslavistic Lexicography

Currently, after the completion of the landmark four-volumes OCS dictionary (i.e. *Slovník*), the Czech paleoslavistic lexicography has commenced five major lexicographic works based on several paleoslavistic (meta-)corpora: *Etymologický, Slovník V.*, revision of *Старославянский*<sup>5</sup>, *Srovnávací index k slovníkům zpracovávaným v rámci Komise pro církevněslovanské slovníky*<sup>6</sup>, *Řeckostaroslavěnsky*. Among these, special attention should be drawn to *Srovnávací*, as the CCS corpus is a component of its meta-corpus, and the DCRCS' lemmas are taken into consideration during the process of determining similarities and differences of various Church Slavonic idioms.

After publishing major Old and Middle Bulgarian<sup>7</sup> dictionaries (*Старославянский; Бончев 2002-2012*), Bulgarian paleoslavists took up the compiling of a voluminous digitized corpus *Компютърни и интерактивни средства за исторически езиковедски изследвания*<sup>8</sup>, which is to become a permanent basis for the comprehensive diachronic description of Bulgarian (including its Church Slavonic constituent) from X to XVIII c. Also, Bulgarian paleoslavistic production abounds with dictionaries and indices of particular texts and sources (e.g. *Тасева 2010*, esp. pp. 533-818; *Димитров 2010, 2013; Илиева 2013a, 2013b*). Macedonian paleoslavists are working on the dictionary of the Macedonian Church Slavonic, based on the corpus comprising documents from the XII to XVI c. (*Речник*). Once the corpus of the Serbian Church Slavonic texts have been created (and it is currently in preparation), the Serbian Church Slavonic dictionary compiling can be expected (*Српскословенски* serves as its introductory fascicle).

Important state-financed projects were recently run or are being run at present in non-Slavic countries, among which two will be mentioned: *SlaVaComp. COMPutergestützte Untersuchung von VARIabilität im KirchenSLAVischen* in Germany (Freiburg, 2013-)<sup>9</sup> and *Die kirchenslavische Übersetzung der Werke von Gregorios Palamas und Barlaam von Kalabrien* in Switzerland (Bern, 2010-2013)<sup>10</sup>.

### 2.3.2 The Situation in the Croatian Historical Lexicography

Currently, in Croatia, besides the DCRCS, there are three historical lexicographical projects in different phases of progress: *Dictionary of Croatian Kajkavian literary language* (Cro. *Rječnik hrvatskoga kajkav-*

---

5 Emilie Bláhová is the redactor of the revisited and supplemented edition of *Старославянский*.

6 The first volume should be published in foreseeable future.

7 Bulgarian slavists use the term "Middle Bulgarian" for the idiom other slavists usually name "Bulgarian Church Slavonic", and "Old Bulgarian" for the idiom the other slavists usually name "Old Church Slavonic".

8 The corpus has been created at the Софийски университет „Св. Климент Охридски“ under redaction of Dora Ivanova-Mirčeva. For more information on that corpus v. Totomanova (2012).

9 The project is led by J. Besters-Dilger and G. Schneider (University of Freiburg). More information can be found at the web-site: <http://www.slavacomp.uni-freiburg.de/> [15/10/2013].

10 The project was led by Y. Kakridis (University of Bern) and financed by *Schweizerischer Nationalfond*.



*koga književnog jezika*), *Dictionary of the Croatian literary language from the National Revival to I. G. Kovačić* (Cro. *Rječnik hrvatskoga književnoga jezika od preporoda do I. G. Kovačića*; compiling of which was resumed in 2008, after 18 years of hiatus), *Old Croatian dictionary* (Cro. *Rječnik starohrvatskoga jezika*, in the middle of the elaborate preparation of its corpus).

Here, the last of the three lexicographical projects is the most important one, because the corpus which is being prepared for the *Old Croatian dictionary* compiling shares two very important features with the CCS corpus: (a) both are referential which also makes the dictionaries based on them referential; (b) both corpora consist of integral texts found in manuscripts and incunabula originated from the same period, namely (XI)/XII—XV/XVI c. But, the fact that the two corpora and the two dictionaries are compatible is even more important than their aforementioned overlapping features. Indeed, the two corpora and the two dictionaries taken together cover two major components of the medieval Croatian diasystem: CCS (in the case of the DCRCS) and literary medieval Croatian idioms (in the case of the *Old Croatian dictionary*).<sup>11</sup> Thus, to once have both corpora and both dictionaries at disposal is critical for the trustworthy diachronic investigation of the Croatian language. It cannot be emphasized enough that it is virtually impossible to conclude the investigation of the medieval Croatian diasystem without both of the corpora and respective dictionaries published.

The research of the CCS component of the Croatian diasystem appears to be more demanding for a contemporary linguist than the research of the literary medieval Croatian idioms due to the fact that the CCS corpus of texts is much less familiar, even obscure, to the majority of linguists dealing with the diachrony of the Croatian language, which is not much of a surprise if the following is kept in mind:

- (1) it is written in an unusual script (i.e. Glagolitic);
- (2) it is written in an idiom (i.e. CCS) not entirely comprehensible if relying solely on the good command of the Croatian vernacular diasystem;
- (3) in numerous cases, the linguistic features of a given CCS text cannot be interpreted correctly if its Greek or Latin parallel text is unfamiliar to the investigator;
- (4) CCS documents are scattered not only across Croatia, but also across the world (including countries such as Russia, Austria, Slovenia, United Kingdom, USA and so forth);
- (5) in many cases, the CCS sources are at least partially damaged due to age and mal-manipulation, sometimes even to the point of being barely readable.

It can be added that the above-mentioned features of the CCS corpus give a clue to a range of competences CCS lexicographers have to have at their disposal, as well as to the indispensability of their contribution to the research of the overall diachrony of the Croatian language.

---

11 For the comparison of the two corpora v. Vukoja (2012).

### **3 The Contribution of the DCRCS Compiling to Croatian Studies and to the Historical Lexicography in General**

#### **3.1 The Contribution of the DCRCS Compiling to Croatian Studies**

Being referential and representative, besides its other features, the CCS corpus provides the most effective basis available for the relevant research of the CCS idiom. If conducted in a methodologically correct manner, the conclusions drawn from the analyses of the corpus can be taken as trustworthy for the whole of the CCS idiom.

In general, every research of the CCS has to take into consideration three layers of factors: (i) the inherited Old Slavonic state, with its characteristic Greek and Latin influences; (ii) the Latin influences that CCS acquired through the adaptation to the Western (Church) tradition, which should be taken separately from the Romance influences acquired through spoken language; (iii) the influences of Croatian vernacular idioms. At present, systematic research of CCS is conducted on the basis of the CCS corpus in several fields: grammar (phonology, morphology, syntax, semantics), lexicology, textology, translation theory and practice, research on the medieval conceptualization of feelings.

With the DCRCS, Croatian lexicography would vastly improve the starting point for all the research on the Croatian diachronic (medieval) diasystem, as it implies diglossia consisting of the Croatian vernacular and CCS. The compiling of the DCRCS adds up critically to another enterprise of the Croatian lexicography, namely the creation of the corpus for the *Old Croatian dictionary* and its compiling. Without the DCRCS, the work of the lexicographers engaged in the project of *Old Croatian dictionary* would be considerably more difficult, even hardly accomplishable in numerous cases, as the CCS features in prevalingly vernacular texts would easily remain unnoticed, which, in turn, would result with a poor diachronical description of the Croatian language.<sup>12</sup> The DCRCS and its corpus offer help to all the scholars who for various reasons need to understand CCS texts. Due to the vast range of the CCS text genres, such scholars may be hagiographers, liturgists, general historians, historians of specific fields (feelings, medicine etc.) ethnologists, and so forth.

#### **3.2 The Contribution of the DCRCS Compiling to the Historical Lexicography in General**

In the context of the (international) historical lexicography, the corpus for the DCRCS and the DCRCS itself are one of the pivots of the international (paleo)slavistic lexicography. At the moment, the results achieved within the management of the CCS corpus and the compiling of the DCRCS are the constitutive elements of the work on *Srovnávaci*, but the CCS corpus and the DCRCS are also at the disposal of all the lexicographers working on or with various Church Slavonic corpora. They are both also

---

12 Cf. e.g. the difficulties in determining the origin of certain features of language forms used by Bartol Kašić, v. Vrtič (2009:117-118.218-219.291-293).

valuable help for all the linguists engaged with the national Church Slavonic idioms (Bulgarian, Macedonian, Czech, Russian, Romanian, Bosnian, Serbian). As national Church Slavonic idioms have been rightly recognized to be integral components of their respective national standard languages, the CCS corpus and the DCRCS support, within their capacities, the research of various Slavic languages. The presentation of the CCS material in both formats, corpus and dictionary, is most-welcome because CCS in its original text, written in Glagolitic and Old Cyrillic scripts usually proves to be a tough nut for slavists, especially those not versed in dealing with the Slavonic idioms.

## 4 The DCRCS Dictionary Article

The DCRCS is compiled in accordance with the general lexicographical theoretical knowledge (Zgusta 1971, Sinclair 2003; Svensén 2009), fashioned in accordance with the practical knowledge of the most experienced paleoslavistic lexicographers, those from the Czech paleoslavistic lexicographical tradition (v. Mareš 2007[1962]), but finally formed so that it recognizes and appropriately displays a specific range of the CCS features in semantics, grammatical forms, textological and translational traits (Nazor 1991; Grabar et al. 1991; Vukoja 2012).

The methodology of the DCRCS compiling is based on the compiling methodology of the earlier Old Slavonic dictionary, *Slovník*, in order to achieve the formal lexicographical concordance needed to enable, help and enhance various paleoslavic research.<sup>13</sup> However, the particularities of the CCS material has asked for extensive adaptations, e.g. in the areas of normalisation, differentiation of semantic variants etc. By all means, the DCRCS is made according to the highest standards of the relevant historical dictionaries.

The head of the dictionary article is written in the Glagolitic and Old Cyrillic alphabets, but its body is in the Latin script, except for the Greek parallels of the lemma and pertaining citations. The body of the dictionary article contains Croatian and English translations of the CCS lemma, Greek and Latin parallel lexemes, as well as an encyclopaedic identification (in the Latin language), if the given lemma is an anthroponym, a toponym or a technical term (most often liturgical). Every recognised meaning is accompanied by CCS examples followed by Greek and Latin parallel phrases as well as CCS variants of the lemma, if existing. In choosing the examples, not only semantic variations are sought to be presented, but also a range of different contexts (and text genres) as well as the range of the lemma forms. Special attention is given to the phrasemes, the differences in spelling (words abbreviated under a tilde, shortened by suspension or by omission of the first syllable with which the preceding word ends) and other particularities of the lemma. At the end of the dictionary article, relevant synonyms are enlisted. Also, possible presence of the given lemma in *Slovník*, Miklosich (1862-1865) and ARj is indicated.

---

13 For the basic methodological principles of the DCRCS compiling v. Mareš (2007[1962]); also Grabar et al. (1991: VIII-XXX).

## **5 The Current State of Affairs Regarding the CCS Corpus and the DCRCS Compiling**

### **5.1 The Current State of the CCS Corpus Conversion into Digitized Form**

At present, two main paper card-files (the sources and the azbuka card-files) are scanned and preserved in the JPEG format (digitization editor: Marica Čunčić, software: Antonio Magdić, scanned by the Croatian State Archives, ArchivePRO). This way, the safety of the data is largely improved, but very little is done in terms of the digital manageability of the corpus (digital readability and searchability, practicality of the DCRCS' compiling process). For internal use, one of the DCRCS compilers edited the JPEG-formatted sources card-file in a more user-friendly manner, and the DCRCS' compilers are using that version in their daily work. Still, paper card-files are indispensable in the majority of research cases.

At the moment, a combination of factors (among which the lack of sufficient financial support is the most notable one) contributed to the decision of putting on hold the project of digitization, which should end only when the full digital readability and searchability of the CCS corpus is achieved. Another notable obstacle is that the cards in the card-files are written by more than two dozen different hands, which practically excludes all known Optical Character Recognition (acr. OCR) options, and which requires an exploration of the Intelligent Character Recognition (acr. ICR) possibilities. Still, the collaborators on the project of the compiling of the DCRCS, as well as the authorities of the Old Church Slavonic Institute are constantly looking for a solution that would make digitization viable.

### **5.2 The Current State of Affairs Concerning the DCRCS Compiling**

The fascicles of the DCRCS, one per year, each containing 64 pages, have been published since 1991. So far, 1 (1991)–19 (2012), with the dictionary articles A–ŽRBTV A (according to the Old Cyrillic alphabet) have been compiled. The first 10 fascicles are bound in Vol. 1. (DCRCS 2000), which has been peer-reviewed as an extraordinary lexicographical accomplishment on several occasions.<sup>14</sup> The fascicles 1–19 exist also in the PDF format, with a limited range of text-searching options. They are available on the Institute's intranet, and placed at the disposal of any interested researcher.

The fascicle 20 is only days away from publishing. Once it is printed, Vol. 2 of the DCRCS is to be bound. Despite their dedication, five lexicographers who are engaged in the compiling of the DCRCS are not able to produce the fascicles at a faster pace, but hopefully with additional collaborators the pace will be intensified in the foreseeable future.

---

14 E.g. at the presentation of the DCRCS Vol. I. at the International Slavistic Congress in Ljubljana 2003, also Грковић-Мејџор (2007: 187).

## 6 Conclusion

The CCS corpus is an indispensable tool for the research of the CCS idiom as well as the prime-quality source for all the scholars who for various reasons need to consult the CCS texts. Its present two formats (paper card-file and JPEG) seek for a thorough digital conversion of the corpus, which is on hold at the moment due primarily to the financial reasons.

The DCRCS with its forthcoming 20<sup>th</sup> fascicle is in progress, although at a moderate pace, due to the limited number of available lexicographers. Hopefully, the compiling will be intensified in foreseeable future.

Despite the difficulties just mentioned, the work on the CCS corpus and the DCRCS compiling should be continued due to its major importance in the context of Croatian as well as (paleo)slavistic studies.

## 7 References

- ARj = *Rječnik hrvatskoga ili srpskoga jezika*, Vol. I-XXIII. Zagreb: JAZU. (1880-1976).
- Бончев 2002-2012 = архимандрит Анастасий Бончев. *Речник на църковнославянския език*. Т. 1. (А-О; 2002), Т. 2. (П-Я; 2012). София: Народна библиотека «Св. Св. Кирил и Методий».
- Borin, L. (ed.). (2002). *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999*. Amsterdam: Rodopi.
- Bratulić, J. (1981). Ediciona praksa hrvatskih istraživača i izdavača srednjovjekovnih tekstova u XIX i XX stoljeću (Historijski prikaz). In Д. Богдановић (ур.) *Меѓународни научни скуп: Текстологија средњовековних јужнословенских књижевности, 14-16. новембра 1977*. Београд: САНУ, pp.137-147.
- Brozović, D. (1970[1967]). *Standardni jezik*. Zagreb: Matica hrvatska.
- DCRCS 2000 = *Rječnik crkvenoslavenskoga jezika hrvatske redakcije*. Vol. I. (a-vrêdb.). 2000. Zagreb: Staroslavenski zavod Hrvatskoga filološkog instituta.
- Димитров, К. (2010). Речник - индекс на Словата на авва Доротеј (По ръкопис 1054 от сбирката на М. П. Погодин). Велико Търново: Университетско издателство “Св. св. Кирил и Методий”.
- Димитров, К. (2013). Авва Доротеј. Слова. Среднобългарски превод. Гръцко-български словоуказател. Велико Търново: Университетско издателство “Св. св. Кирил и Методий”.
- Etymologický = Etymologický slovník jazyka staroslověnského*. E. Havlová, A. Erhart & I. Janyšková (red.). Praha, Brno: Akademie věd České Republiky, Academia, Tribun EU. (1989-).
- Grabar, B., Mareš, F.V. & Mulc, I. (1991). Oblikovanje i sastav natuknice. In *Rječnik crkvenoslavenskoga jezika hrvatske redakcije. (sveščić 1., Uvod)*. Zagreb: Staroslavenski zavod Hrvatskoga filološkog instituta. pp. VIII-XVIII. (transl. in Eng. pp. XIX-XXX.)
- Grabar, B., Nazor, A. & Pantelić, M. (1973). *Missale Hervoiæ ducis Spalatensis croatico-glagoliticum = Hrvatskoglagoljskimisal Hrvoja Vukčića Hrvatinića = Croato-glagolitic Missal of Hrvoje, Duke of Split: transcriptio et commentarium*. V. Štefanić (red.). Zagreb, Ljubljana, Graz: Staroslavenski institut, Mladinska knjiga, Akademische Druck- u. Verlagsanstalt. + Faximil.
- Грковић-Мејуор, Ј. (2007). Рјечник црквенославенскога језика хрватске редакције, I. сvezak (a-vrêdb), Staroslavenski institut, Zagreb, 2000. *Прилози за књижевност, језик, историју и фолклор*, LXXII/1-4, pp.185-187.

- Jagić, V. (1879). *Quattuor evangeliorum codex glagoliticus olim Zographensis nunc Petropolitanus. Characteribus cyrillicis transcriptum notis criticis prologomenis appendicibus auctum adiuvante summi ministerii Borussici liberalitate edidit V. Jagi. Berolini: Apud Weidmannos.*
- Илиева, Т. (2013а). Старобългарският превод на Стария завет, том 3: Старобългарско-гръцки словоуказател към Книгата на пророк Иезекиил. София: Кирило-Методиевският научен център, Българска академия на науките.
- Илиева, Т. (2013б). *Терминологичната лексика в Йоан-Екзарховия превод на "De Fide orthodoxa"*. София: Самиздател.
- Katičić, R. (1992). 'Slověnski' i 'hrvatski' kao zamjenjivi nazivi jezika hrvatske književnosti. In *Novi jezikoslovni pogledi*. Zagreb: Školska knjiga, pp. 312-328.
- Mareš, F.V. (2007[1962]). *Návrh přípravných prací pro slovník jazyka církevněslovanského. Církevněslovanská lexikografie 2006*. In Václav Čermák (sestavil); E. Bláhová, E. Šlaufová & V. Čermák (eds.) Praha: Slovanský ústav AV ČR, Euroslavica, 64-84. (Russian translation: Мареш, Ф.В. (1966). Проект подготовки словаря церковнославянского языка. *Вопросы языкознания XV*. Москва: Академия наук СССР, Институт языкознания, pp. 86-99.)
- Mihaljević, M. (2010). *Položaj crkvenoslavenskoga jezika u hrvatskoj srednjovjekovnoj kulturi*. Свети Наум Охридски и словенската духовна, културна и писмена традиција (организиран по повод 1100-годишнината од смртта на св. Наум Охридски). Зборник на трудови од Меѓународниот научен собир. Охрид, 4-7 ноември. Скопје: Универзитет „Св. Кирил и Методиј“, pp. 229-238.
- Miklosich F. (1862-1865). *Lexicon palaeoslovenico-graeco-latinum*. Vienna: Guilelmus Braumueller.
- Nazor, A. (1991). *Uvod; Popis izvora; Navedena literatura*. In *Rječnik crkvenoslavenskoga jezika hrvatske redakcije. (sveščić 1., Uvod)*. Zagreb: Staroslavenski zavod Hrvatskoga filološkog instituta. pp. I-III (transl. in Eng.: pp. IV-VII.). XXXI-XXXVI. XXXVII-XXXIX.
- Nazor, Anica. (2008). *Rječnik crkvenoslavenskoga jezika iniciran na IV. međunarodnom slavističkom kongresu u Moskvi 1958. godine (u povodu 50. obljetnice inicijative)*. In M. Samardžija (ed.) *Vidjeti Ohrid. Referati hrvatskih sudionika i sudionika za XIV. međunarodni slavistički kongres (Ohrid, 10.-16. rujna 2008.)*. Zagreb: Hrvatsko filološko društvo, Hrvatska sveučilišna naklada, pp. 65-82.
- Pantelić, M., Nazor, A. (1977). *Uvod; Bibliografija*. In *Drugi novljanski brevijar: hrvatskoglagoljski rukopis iz 1495*. Phototypical edition. Zagreb: Staroslavenski institut "Svetozar Ritig", Turistkomerc, pp. 7-37.
- Řeckostaroslověnsky = Řeckostaroslověnsky index. *Index verborum graeco-palaeoslovenicus*. Praha: Slovanský ústav AV ČR, Euroslavica. (2008-).
- Речник = Речник на црковнословенскиот јазик од македонска редакција*. Том I. Вовед. А–Б. Скопје: Институт за македонски јазик. (2006).
- Rosenwein, B.H. (2006). *Emotional Communities in the Early Middle Ages*. Ithaca: Cornell University Press.
- Sinclair, J. (2003). *Corpora for lexicography*. In P. van Sterkenburg, (ed.) *A Practical Guide to Lexicography*. Amsterdam, Philadelphia: Benjamins Publishing Company, pp. 167-178.
- Словарь = Словарь древнерусского языка (XI–XIV вв.)*. В 10 т. Москва: Институт русского языка Российской академии наук. (1988-).
- Slovník = Slovník jazyka staroslověnskeho. Lexicon Linguae Palaeoslovenicae. I.-IV.* Praha: ČSAV Slovanský ústav. (1958-1997).
- Slovník V. = Slovník jazyka staroslověnského. Sv. V. (Addenda et Corrigenda)*. Praha: Euroslavica. (2010-).
- Српскословенски = Српскословенски речник јеванђеља*. Огледна свеска. Саставио: Виктор Савић. Уредник: Гордана Јовановић. Београд: Институт за српски језик САНУ. (2007).
- Старославянский = Старославянский словарь (по рукописям X–XI веков)*. Под редакцией Р.М.Цейтлин, Р.Вечерки и Э.Благовой. Москва: Славянский институт академии наук Чешской республики, Институт славяноведения и балканистики Российской академии наук, "Русский язык". (1994).
- Svensén, B. (2009). *A Handbook of Lexicography*. Cambridge: Cambridge University Press.

- Štefanić, V. (1962). Problem rječnika južnoslavenskih redakcija staroslavenskog jezika.** In *Slovo*, 11-12, pp. 181-187.
- Тасева, Ј. (2010). Триодните синаксари в средновековната славянска книжнина. Текстологично изследване. Издание на Закхеевия превод. Словоуказатели (Monumenta linguae slavicae dialecti veteris LIV). Freiburg im Briesgau: Weiher Verlag.
- Totomanova, A-M. (2012). Digital Presentation of Bulgarian Lexical Heritage. Towards an Electronic Historical Dictionary. In *Studia Ceranea*, 2, pp. 219-229.
- Vrtič, I. (2009). *Sintaksa Kašičeva prijevoda Svetoga pisma*. PhD. thesis. Filozofski fakultet Sveučilišta u Zagrebu, Zagreb, Croatia.
- Vukoja, V. (2012). O korpusu Rječnika crkvenoslavenskoga jezika hrvatske redakcije i njegovu odnosu prema korpusima hrvatskoga jezika. In *Filologija*, 59, pp. 207-229.
- Weinreich, U. (1954). Is a structural dialectology possible?. In *Word*, 10, pp. 388-400.
- Zgusta, L. (1971). *Manual of Lexicography*. The Hague, Paris – Prague: Mouton – Academia.





# Others



# Considerations about Gender Symmetry in the Dictionary of Bavarian Dialects in Austria

Isabella Flucher, Eveline Wandl-Vogt, Thierry Declerck  
Austrian Academy of Sciences, ICLTT, Austria;  
DFKI GmbH, Language Technology Lab, Germany  
bella\_flu@yahoo.de; eveline.wandl-vogt@oeaw.ac.at; declerck@dfki.de

## Abstract

This poster summarizes the first results of a study that has been pursued as part of an internship at the Austrian Academy of Sciences. The aim was to investigate cases of gender symmetry or asymmetry in a dictionary. As case study we focused on a traditional dialectal dictionary. An annotation schema has been developed and first natural language processing steps have been established. In this poster, forms of gender symmetry as well as asymmetry are presented, on the basis of analysis of the vocabulary employed in example sentences or excerpts used in the dictionary. The analysis of gender asymmetry was also based on the consideration of a selection of derogatory names. The work described in this poster provides a critical insight into lexicographical work, design and implementation from a feminist perspective and opens new perspectives for the development of gender-symmetric lexicographic works.

**Keywords:** Dialectal lexicography; Gender asymmetry; Gender symmetry; Austrian dialects

## 1 Introduction

This paper presents work achieved in the context of a practical training at the Austrian Academy of Sciences (ÖAW) in summer 2013<sup>1</sup>, and which has been pursued afterwards as part of a university seminar. A task designed for this internship was to analyze a traditional dialectal dictionary along the lines of gender-specific criteria.

While we know that the primary goal of a dialectal dictionary is to describe the authentic language use in a certain region, we consider this investigation on gender symmetry (or asymmetry) to be well motivated since the foundations of the dictionary we are considering were laid down well before any feministic concerns in the field of lexicography have been raised. In this paper describing the poster, we give first a brief description of the dictionary we have been selecting for the investigation on gender asymmetry, before presenting a selection of results.

---

1 In the context of this internship, Isabella Flucher, the main author of this poster, was collaborating with 3 other students, namely Nathan Balaz, Magdalena Schwarz and Andrea Steiner.

## 2 The Dictionary of Bavarian Dialects in Austria

For our work, we focused on language data contained in a traditional dialectal dictionary: The dictionary of Bavarian dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich, WBÖ). This traditional scientific territorial dictionary was developed since the early 20th century: The main part of the collection took place between 1915 and 1950; the dictionary itself is published since the early sixties. More recently, in the early nineties, a project has been established for developing and maintaining an integrated database (Database of Bavarian dialects in Austria; Datenbank der bairischen Mundarten in Österreich, DBÖ<sup>2</sup>). This database supports the storage, visualization und querying of a variety of dialectal language data and related information sources. Since 2004 the dictionary is build up as a digital platform and is now also available online via the Austrian Academy Press. Since 2013 we are developing a machine readable version (using SKOS<sup>3</sup> as the basic representation formalism), connecting also the data to the LOD.<sup>4</sup>

The whole project related to WBÖ gives thus an example for a transformation process of an encyclopedic dictionary type into the framework of cyberscience and digital humanities. .

### 2.1 The Project Framework at the Austrian Academy of Sciences

The investigation on gender (a)symmetry is embedded into the digital dictionary project, which we very briefly described above. In the framework of this project we are working on / with different methods and interdisciplinary knowledge to improve data access, enrichment and re-usability of data. In order to investigate if we could reduplicate results of the analysis on gender symmetry applied to the WBÖ, we decided to create a corpus containing annotation about the (semantic) genders and the type of vocabularies used. The development of this corpus is also aiming at supporting the development of natural language processing tools that could be trained on this set of annotations. Results of this work will be described in future publications, while we concentrate in this poster on the result of manual analysis applied to the annotated corpus.

### 2.2 The Gold Corpus

A basis for our work on gender symmetry is a WBÖ-XML-gold corpus, methodologically discussed and completely manually annotated: We were using 4 reference supplements of the WBÖ, namely 33-36. The first step of annotation included annotating headwords as well as reference entries that are

---

2 The DBÖ collection of mushrooms and related lexicographical materials is available online; see (Wandl-Vogt 2010).

3 SKOS stands for "Simple Knowledge Organization System" (<http://www.w3.org/2004/02/skos/>)

4 See (Wandl-Vogt 2008) and (Wandl-Vogt & Declerck 2013). LOD stands for Linked Open Data (<http://linked-data.org/>)

connected to the concept “person”. We annotated real persons as well as figures, such as legendary creatures or fairy-tale figures. The (informal) schema for the annotation was:

<h> </h> - Mensch (*human*)<sup>5</sup>

<f> </f> - feminine (*feminine*)

<m> </m> - maskulin (*masculine*)

<Bsp> </Bsp> - Beispiel (immer kursiv- dialektale Ausdrücke) (*marking an example in the dictionary, which is carrying a gender information*)

<Bed> </Bed> - Bedeutung (*same as above, but for a text span dealing with a definition*)

<hist> </hist> - historisch (*same as above, but marking an historical context*)

<gauspr> </gauspr> - gaunersprachlich (*language of the „crooks“*)

<geschlT> </geschlT> - Geschlechtsteil (*words on genital parts*)

A few examples of annotation are given below:

- <Bsp> es mit <m>einem</m>/<f>einer </f> tun| </Bsp> (*to do it [namely: have sex] with someone [namely: him or her]*)
- <Bed> es treibt <f> ihr</f>(vor Scham) d. Röte ins Gesicht</Bed> (*she is blushing*)
- <Bsp>¿ <m>e </m>verdrosch seine<f>Mutter</f> </Bsp> (*he was beating his mother*)
- <m><geschlT>Penis</geschlT></m>

The manual analysis of the annotated data was done on the base of theoretical aspects described in the next section.

### 3 Gender Symmetry and Gender Asymmetry

#### 3.1 Linguistic Perspective

The feminist linguistics aims to make visible an androcentric predominance, which is resulting in a critique of the language system on the one hand and of the language use on the other hand.<sup>6</sup>

The semantic field “human” is dominated by men. In that context, Pusch (1984) describes the woman as a “subclass of the men’s class” and she goes even further when she says: “Human is the man”.<sup>7</sup> Often, by “people” or “human” only man is meant, therefore it seems necessary to identify the woman with the female attribution. Thus, masculinity is enhanced in contrast to femininity, when the man is represented as a human being.<sup>8</sup> The attribution of the female will be used, because otherwise only the male group would be targeted, although it is a gender- neutral term.<sup>9</sup>

5 Annotation of this concept has been performed by Natahn Balasz.

6 See (Kollmann 2010: p. 14).

7 See (Pusch 1984: p. 17).

8 See (Poerber 2007: pp. 408-409).

9 See (Breiner 1996: p. 79).

In terms of the article entries of the WBÖ one notices that definitions which include the keyword “human” often contain a definition of woman. What looks at first glance like an imbalance in favor of the female sex, after closer analysis, is to be interpreted as the opposite phenomenon. The woman is called that often rather for the reason of a differentiation, whereas the man must not also be named because he embodies the prototype of human, the “universal human”.

### 3.2 Examples for Gender Symmetry and Gender Asymmetry

Aiming at gender symmetry means not to reverse a possible androcentric language in a gynocentric one, but to reach a state in which the same criteria are applied to both sexes.

A positive example of a WBO entry to illustrate the case of symmetry is the following: The verb “trackeln, -gg-” means being stupid. Its derivatives are symmetrical in relation to the two sexes, both in the length of the entry as well as their semantic shape: “Trackle, -gg-”: stupid woman [..] and “Trackler, -g (g)-”: stupid man.<sup>10</sup>

A counterexample is “Trab” and “traben”.<sup>11</sup> “Traben” has besides its first meaning of a horse’s walk two other meanings in their sample-sentences, where a clear gender asymmetry is found. Namely, when *he* goes on the “Trab”, he is sent abroad, but when *she* goes on the “Trab”, she works as a prostitute. This difference is confirmed by the term “Trab”, as the male form “Traber” is referred to herein and the trotting horse, whereas the female “Traberin” stands for a prostitute.<sup>12</sup>

Gender asymmetry can emerge as well from unequal treatment of the length of masculine and feminine entries, for example when the male thief “Dieb” has six columns and the female thief “Diebin” has only a slim column.<sup>13</sup> Also within an entry the example sentences in relation to each other manifest a different rating of the sexes. For example “Ferdienst” has such various ratings of the sexes: three example-sentences reflect the man as an appreciative, rewarding power, whereas the fourth rating, which refers to the woman, is the earning that comes from a “dirty business”, which could refer to prostitution again.<sup>14</sup>

Another form of asymmetry is to be shown on the basis of the lemmas “Trantsch”, “Träntsche” and “Träntscher”.<sup>15</sup> “Trantsch” and “Träntsche” provide both in its primary meaning an insult mainly for women. The term “Trantsch” has six female classifications and four neutral, which means related on both gender. Thus, referring to this, most often the woman is meant with this insult, the man is not explicitly marked, only indirectly by writing “human”. “Träntsche” has the same meaning, this time

---

10 WBÖ, S. 234.

11 WBÖ, S. 220-221.

12 WBÖ, S. 221.

13 WBÖ, S. 35-43.

14 WBÖ, S. 55-56.

15 WBÖ, S. 314-316.

with the mention of “woman” and “person”.<sup>16</sup> Even “Träntscher”, the male modification of “Trantsch”, refers only to “human” and one time to a “person”. These examples are unbalanced from a gender perspective, because the gender-classification of these three abusive terms, which belong together, is much more female-oriented.

### 3.3 Distribution of Gender (a)symmetric Cases across Topics

Current work is dedicated to the establishment of classes of topics in which a gender (a)symmetry in WBÖ can be established, looking at examples used for illustrating the meanings of entries. The topics we are studying are for now are “alcohol”, “talkativeness” and “violence”. The aim is to examine, if there are female- and male-specific categories of meaning in the WBÖ, and if so to analyse them in terms of gender criteria.

#### 3.3.1 Alcohol

The word field of drunkenness is clearly dominated by men. Alcohol use of women is only manifested in the individual cases as “Trinkerin”<sup>17</sup> (female form of drinker), “Alkoholikerin”<sup>18</sup> (a female alcohol-addict) and “Schnapsdorothea”<sup>19</sup> (“Schnaps” as a sort of strong alcohol combined with the woman’s name Dorothea, means a woman, that drinks a lot). But what is significant, is that there exist as good as no female example-sentences referring to alcohol consumption.

- Der Lump, .. der sein ganz’s Geld versauft<sup>20</sup> (*a man, who spends all his money for alcohol*)
- ols a nüachta [nüchtern] is er eh recht sölt n aonztreffn<sup>21</sup> (*he is rarely sober*)
- er hat a weng z’ tief ins Glasl g’schaut er hat einen Rausch<sup>22</sup> (*he is drunk*)

#### 3.3.2 Talkativeness

Talkativeness is the one meaning category, which is attributed widely to women. It is also remarkable that the context of meaning differs regarding the sexes, where “Tratschweib” (a talkative woman in a negative sense) faces the “Maulheld” (a talkative man, but who is called a hero, a boaster).

- *Postentragerin*: Frau, die andere Personen ausrichtet, abschwächend für Verleumderin<sup>23</sup> (*a woman who speaks about a person in a defaming way*)
- die ist eine rechte/alte Tratsche<sup>24</sup> (*she talks a lot and she is old*)

---

16 To which extent the term “person” suggests “woman” as well is debatable, but this is beyond the scope of this paper. However, “person” appears increasingly in association with “woman” and “man” in connection with “human”.

17 WBÖ, S. 520, S. 523.

18 WBÖ, S. 523.

19 WBÖ, S. 189.

20 WBÖ, S. 238.

21 WBÖ, S. 373.

22 WBÖ, S. 45.

23 WBÖ, S. 290.

24 WBÖ, S. 332.

- Sie soll ná was drein rödn../ Da is má nôt bang, wir [werde] ihr´s Mudl schan tedten - / I kimm ihr schan grob gnua<sup>25</sup> (*she shouldn´t interfere my talk, if she does i will be aggressive against her*)

### 3.3.3 Violence

In the WBÖ-entries almost examples for words of violence are given by sentences that show violent-acting men. Women are more connected with a softened version of violence, as they are more often called böse (*evil*), launenhaften (*capricious*), zänkischen (*quarrelsome*), streitsüchtigen (*contentious*) women and wives.

- Töterling: grober Kerl, Raufbold<sup>26</sup> (*a rough man, sb.who likes to beat/fight*)
- töten ęa wiad des mäd (Mädchen) no ęwidra´n<sup>27</sup> (*to kill he will kill the girl*)
- daß sie von ihm noch ihre Treff (Schläge) kriegen werde<sup>28</sup> (*that she will get beaten up by him*)
- Drache: zänkische, herrschsüchtige Frau; streitlustige Ehefrau<sup>29</sup> (*dragon: a quarrelsome, dominant woman; a contentious wife*)

We are currently extending the list of concepts, either adding new ones, or further specifying existing ones. For this we are consulting work by Dornseiff (2004) and lexical resources like WordNet<sup>30</sup> or Wiktionary<sup>31</sup>, which are helping us in better classifying the words used in the dictionaries for describing the entries.

## 4 Conclusion

We presented actual work on gender (a)symmetry in the context of a dialectal dictionary. In the next future we will extend this work and also consider the underlying data of similar dictionaries, like collections of slips of papers, databases, or questionnaires to find out reasons of asymmetries. We also plan to focus more on natural language processing aspects and to be able to mark up relevant words in computational lexicons with this kind of gender interpretation, beyond the case of pure grammatical genders.

The work is to be continued, deepened, and extended as a research infrastructure for the comparison of lexicographical works as well as languages.

---

25 WBÖ, S. 210.

26 WBÖ, S. 212.

27 WBÖ, S. 248.

28 WBÖ, S. 369

29 WBÖ, S. 222-223.

30 <http://www.sfs.uni-tuebingen.de/GermaNet/>

31 <https://de.wiktionary.org/wiki/Wiktionary:Hauptseite>



## 5 References

- Barabas, Hareter-Kroiss, Hofstetter, Mayer, Piringer, Schwaiger. (2010). Digitalisierung handschriftlicher Mundartenbelege. Herausforderungen einer Datenbank. In: *Germanistische Linguistik 199-201*, Fokus Dialekt. Festschrift für Ingeborg Geyer zum 60. Geburtstag 2010, S. 47-64.
- Bayerisch-Österreichisches Wörterbuch, I. Österreich, Wörterbuch der bairischen Mundarten in Österreich, Institut für Dialekt- und Namenlexika DINAMLEX, Österreichische Akademie der Wissenschaften ÖAW (Hg.), Wien 1963.
- Breiner, I. (1996). Die Frau im deutschen Lexikon. Eine sprachpragmatische Untersuchung, Wien 1996.
- Dornseiff, F. (2004) *Der deutsche Wortschatz nach Sachgruppen*. De Gruyter, Berlin/Leipzig 1933-1940; 8. Auflage: De Gruyter, Berlin/New York.
- Hufeisen, B. (editor) (1993). „Das Weib soll schweigen...“ (I.Kor.14,34). Beiträge zur linguistischen Frauenforschung, Kasseler Arbeiten zur Sprache und Literatur, Band 19, Frankfurt am Main 1993.
- Eichhoff-Cyrus K.M. (2004), Adam, Eva und die Sprache. *Beiträge zur Geschlechterforschung, Thema Deutsch, Band 5*, Mannheim 2004.
- Flucher, I. (2013). Gendersymmetrie im WBÖ Wörterbuch der bairischen Mundarten in Österreich, Seminararbeit Uni Wien (Pober M.: „toller hengst“ : „läufige hündin“ - Zufall oder verborgenes Genderskript?
- Kollmann, S. (2010). Einstellungen zu geschlechtergerechtem Sprachgebrauch im Deutschen, Dip., Wien
- Pober, M. (2004). Überlegungen zur geschlechtersymmetrischen Struktur eines Genderwörterbuchs im Deutschen, Dissertation, Wien 2004.
- Pober M. (2007), Gendersymmetrie. Überlegungen zur geschlechtersymmetrischen Struktur eines Genderwörterbuchs im Deutschen, Würzburg 2007.
- Pusch, L.F. (1984). *Das Deutsche als Männersprache*. Frankfurt am Main.
- Wandl-Vogt, E. (2008). wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX). In: Ernst, P. (ed) 2008, *Bausteine zur Wissenschaftsgeschichte von Dialektologie / germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der internationalen Gesellschaft für Dialektologie des Deutschen*, Wien: 93-112.
- Wandl-Vogt, E. (2010). Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema). Accessed at: <http://wboe.oeaw.ac.at> [10/04/2014].
- Wandl-Vogt, E., Declerck, T. (2013). Mapping a Traditional Dialectal Dictionary with Linked Open Data. In: Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.
- Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1963-)*. Verlag der Österreichischen Akademie der Wissenschaften. Wien. Accessed at: <http://hw.oeaw.ac.at/cl?frames=yes> [10/04/2014].

### Acknowledgements

We would like to thank Nathan Balaz, Magdalena Schwarz and Andrea Steiner for their contribution to the internship, which constituted the basis for the work described in this poster.



# Advancing Search in the Algemeen Nederlands Woordenboek

Carole Tiberius, Jan Niestadt, Lut Colman, Boudewijn van den Berg  
Institute of Dutch Lexicology, belberg  
{carole.tiberius,jan.niestadt,lut.colman}@inl.nl, boudewijn@belberg.eu

## Abstract

In this paper, we will take a closer look at the advanced search option in online dictionaries from the perspective of the user. In the context of dictionaries, the advanced search allows users to retrieve sets of words which match a particular description, for example ‘archaic compounds consisting of three syllables’. However, the possible sets of words that could be retrieved are endless, and the challenge is how to present all these options to the user in a way that he can grasp and understand. Studies on dictionary use and log file analyses suggest that existing solutions offered by online dictionaries are not very successful as users do not really seem to use this search option. We will discuss different types of advanced search that can be distinguished and we will present a new, more user-friendly approach to advanced search in dictionaries using the *Algemeen Nederlands Woordenboek* as our test case.

**Keywords:** advanced search; online dictionaries; user-friendly.

## 1 Introduction

Comprehensive scholarly dictionaries contain a wealth of information. They do not only provide information on the meaning of a word, but they also contain information on morphology, pronunciation, etymology, pragmatics etc. With the online medium it is theoretically possible to let the user search on all the information available in the dictionary database. Most online dictionaries attempt to do this by offering an advanced search option allowing users to retrieve sets of words matching a particular description, for example ‘the set of Dutch nouns that can have more than one plural ending’ or ‘archaic compounds consisting of three syllables’. However, the sets of words that could possibly be retrieved are endless, and the challenge is how to present all these options to the user in a way that he can understand. Studies on dictionary use suggest that users do not really use the advanced search option. This is also the case for the *Algemeen Nederlands Woordenboek* (ANW) where log files show that the advanced search only accounts for 3% of all searches in the dictionary (Tiberius & Niestadt, To Appear). There may be different reasons for this: a) users do not use the advanced search as they are not familiar with such a search option in the context of dictionaries; b) users do not understand the interface that is used for the advanced search; or c) a mixture of these.

In this paper we will present an approach for a more user-friendly advanced search option for the ANW. First we define what advanced search is, then we will discuss different types of advanced search and finally we will present a new approach to advanced search in dictionaries.

## 2 Advanced search in electronic dictionaries: a definition

In order to define what is meant by advanced search and what user requirements it fulfils, we have looked at what different dictionaries have to say about their advanced search. The **Oxford English Dictionary** (OED) defines its advanced search as follows:

Advanced search is a full search of the entire dictionary text. It finds your term wherever it occurs in the dictionary. This could be in the form of an entry name, part of another word's definition, in a quotation, etc. An advanced search also allows you to search for words that occur near one another, such as bread before butter.<sup>1</sup>

The German **lexiko** dictionary offers an advanced search which allows users to search for words on the basis of specific criteria such as orthography, grammar and word family.

Die erweiterte Stichwortsuche in **lexiko** erlaubt dem Benutzer, Stichwörter mit bestimmten Kriterien aus den Bereichen Orthografie, Wortartzugehörigkeit, Grammatik, Sinnverwandtschaft oder Zugehörigkeit zu einer semantischen Klasse zu suchen.<sup>2</sup>

The **Trésor de la Langue Française Informatisé** (TLFi) offers what it calls an assisted search ('recherche assistée') and a complex search (recherche complexe'). The assisted search allows the user to search the dictionary articles on the basis of a number of criteria:

Permet de rechercher **à travers tout le TLF** les articles correspondant à **plusieurs critères**.

Quelques possibilités:

Quels sont les mots d'origine espagnole ?

Quels sont les exemples de Zola illustrant un sens ironique?

Quels sont les verbes utilisés dans la marine pour la manoeuvre des voiles?

Quelles sont les expressions contenant le mot singe?<sup>3</sup>

The complex search in the TLFi is similar to the assisted search, but is described as being even more powerful.

The **Algemeen Nederlands Woordenboek** does not employ the term advanced search, but offers four types of search of which the option to search from features to words is the most advanced.

---

1 <http://www.oed.com/public/advancedsearching/advanced-search/> [10/04/2014]

2 <http://www.owid.de/erweitert.jsp> [10/04/2014]

3 <http://atilf.atilf.fr/dendien/scripts/tlfiv5/showp.exe?13;s=2657178285;p=aide.htm> [10/04/2014]

This search option is the most advanced way of searching the ANW. Through different kinds of information that is stored in the dictionary articles you can search for words, idioms and proverbs. The possibilities to search for information are almost infinite. You can search for information which can occur anywhere in the article or you can search for information in a specific field.[...]

This is a search option which requires a certain amount of creativity from the user and works better the more one gets familiar with the system. We suggest that you take some time to try out this search option.<sup>4</sup> The last sentence shows clearly that we were rather idealistic when we first developed this search option for the ANW back in 2009.

Summarising, advanced search can be described as a powerful and complex search option that allows users to examine the dictionary using many different criteria.

### 3 Types of Advanced Search

Advanced search can be realised in different ways. We identify four types, i.e.

- **Classical/traditional advanced search:** where boxes and dropdown lists together form the search query;
- **Faceted search:** a step by step search where the user gradually refines the query by adding criteria (e.g. shopping websites);
- **Wizard search:** a step by step search where the user is given a sequence of questions and no intermediate results are shown (e.g. Foreign Labor Certification<sup>5</sup>);
- **Query language:** single search box which offers many possibilities, but which is hard to learn and remember.

A fifth type could be identified, i.e. natural language queries. However, we do not consider this option here, as we do not believe that the current state of technology is advanced enough for this to be a viable candidate in the context of online scholarly dictionaries.

From these four types, the classical advanced search is the most popular among scholarly e-dictionaries at the moment (cf. the ANW, the OED, *lexiko* and the complex search in the TLFi). The TLFi also offers a wizard-like search ('recherche assistée') by showing a sequence of questions which are to guide the user to an answer.<sup>6</sup>

The DWDS (das *Digitale Wörterbuch der deutschen Sprache*) is an example of a dictionary that does not offer a separate advanced search option but offers a query language to give the user more flexibility. For instance, the query "Stein with \$p=NN" searches for occurrences of the lemma *Stein* 'stone' as a noun.

For the ANW too, a custom query language, called FunQy ('functional query language'), was developed to power its traditional advanced search option (Niestadt et al. 2009). As an undocumented feature, users can play with this query language themselves. At one time we planned for this query language to

---

4 <http://anw.inl.nl/show?page=help#zoek3> [10/04/2014]

5 <http://www.flcdatcenter.com/OESWizardStart.aspx> [10/04/2014]

6 <http://atilf.atilf.fr/dendien/scripts/tlfiv5/showp.exe?30;s=1771472760;p=assiste.htm> [10/04/2014]

be used internally at the Institute of Dutch Lexicology (INL), but this never materialised. The FunQy query language may be too complex even for language professionals to use.

## 4 A new way of advanced search

We now turn to our approach to realise a more user-friendly advanced search option for scholarly dictionaries. We believe that it is a mistake to think that one single advanced search option can appeal to all users equally. A full-featured search may be convenient for frequent users, but new users will most likely be put off by its complexity. However, if you do your best to capture new users with a friendly, step-by-step approach, experts will get annoyed by how much clicking is required to perform common searches. This means there is no single best approach; you have to compromise, or develop separate interfaces for different users. We decided to clearly identify the target users of the ANW to be linguists and academics more generally, who want powerful search features, but are intimidated by our current interface (see Figure 1).

The features that can be searched for are presented in a tree structure on the left of the screen. This tree structure is the same as the one used to structure the dictionary articles (as seen on the article screen). It starts with syntactic category and then spelling and pronunciation, etc. The user starts with an empty query screen and is asked to select criteria from this tree structure on the left. As soon as the user selects a criterion, a query appears on the right-hand side of the screen. By default, the user searches for words, but it is also possible to search for proverbs or idioms. This will result in a tree structure with different criteria as only a subset of the criteria that can occur in a query for words apply to idioms and proverbs.

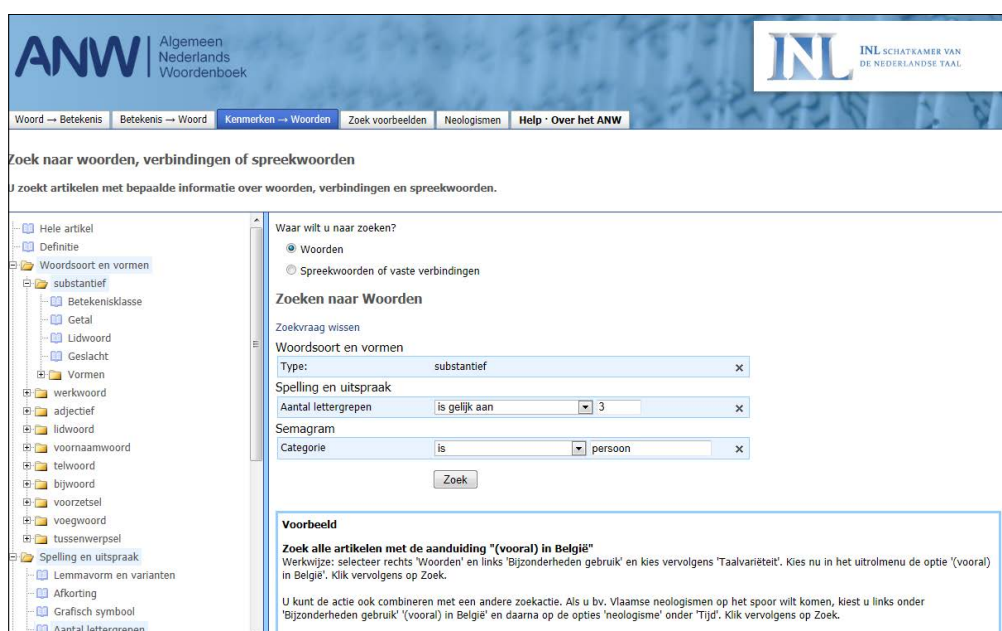
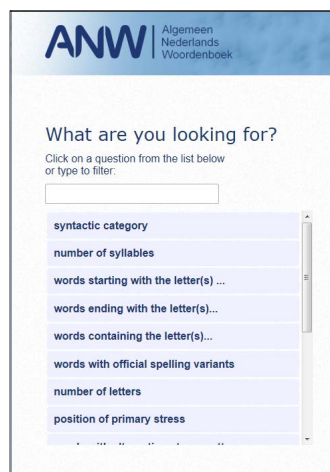


Figure 1: Screenshot Features → Words in the ANW.

Our aim was to combine the best parts of the different types of advanced search discussed in Section 3 in our solution. A very preliminary version of our approach has been presented at the eLex 2013 conference.<sup>7</sup> We are now in the process of developing a full-working prototype which will be accessible through the ANW website.

The opening page of the ANW remains as it is with a simple search box in the center of the screen for searching a word (or multi word expression). However, below the search box a link will be added for users who are looking for something else. After clicking on this link, users are asked what they would like to search for. They can choose a question or criterion from a scrollable list, or they can type in a word to filter the list. The current list covers spelling, pronunciation, combinations and pragmatics, and has been defined on the basis of the criteria offered in the current search interface of the ANW web application. In future, this list will be further expanded and refined.



**Figure 2: Prototype of list of questions users may like to ask.**

For convenience, the most frequently used questions will appear at the top of the list so that users see these first. To make the search as effective and user-friendly as possible (Lew 2012) functionalities such as autosuggest, fuzzy matching and smart filtering will be integrated. Thus, the filter box will also respond to terms that do not literally occur in the visible descriptions because hidden tags specifying the category to which questions belong have been added (e.g. typing ‘morphology’ also shows questions such as ‘words derived from ...’). Clicking on a question opens up the search form, with a single input box for that question. As soon as the user types something in this search box, the results are shown on the right-hand side of the screen together with the relevant information from the dictionary entry: for instance, if the user was searching for words consisting of four syllables, the syllable structure of the resulting words is shown. This direct feedback makes this ‘advanced’ search faster and less intimidating as the user has direct access to the information he is interested in.

---

7 <http://eki.ee/elex2013/> [10/04/2014]

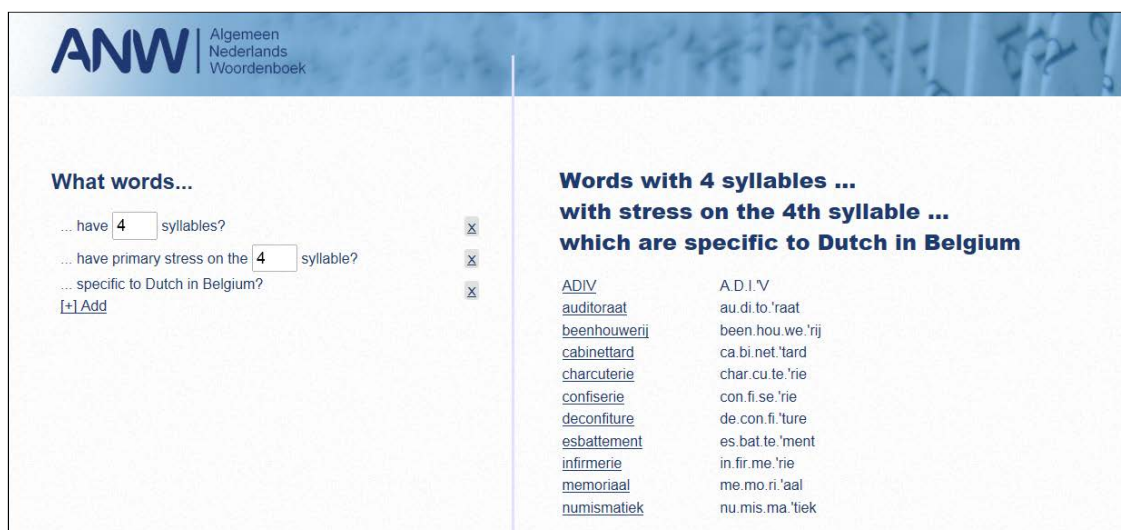


Figure 3: Prototype of what a search query plus results may look like.

If the user wants to refine his query, he simply clicks 'Add' and goes again to the list of search questions where he can select another question to add to his query. As a bonus, an expert who regularly uses the same criteria with different values can bookmark the page in its current state, so he can avoid the ten clicks required to get here.

## 5 Conclusion

We have presented an approach for a more user-friendly advanced search option in the ANW. It combines the best aspects of the advanced search types we have discussed in Section 3. It is like a wizard, because the user is guided through constructing a query. It also includes the advantages of a faceted search because there is direct feedback when you add or change something. The possibility to bookmark a search query makes it again very similar to the classical advanced search.

So far user research has not been carried out, but this is planned for the near future when the prototype is up and running.

## 6 References

- Algemeen Nederlands Woordenboek*. Accessed at: <http://anw.inl.nl> [10/04/2014].
- das Digitale Wörterbuch der deutschen Sprache*. Accessed at: <http://www.dwds.de> [10/04/2014].
- elexiko*. Accessed at: <http://www.owid.de/wb/elexiko/start.html> [10/04/2014].
- Lew, R. (2012) How can we make electronic dictionaries more effective? In Sylviane Granger and Magali Paquot (eds.) *Electronic lexicography*. Oxford. 343-361.



Niestadt, Jan, Carole Tiberius & Fons Moerdijk (2009). Searching the ANW dictionary. Poster presented at eLexicography in the 21st century. Louvain-la-Neuve.

*Oxford English Dictionary*. Accessed at: <http://oed.com> [10/04/2014].

Tiberius, C and J. Niestadt (To Appear) Dictionary Use: A Case Study of the ANW Dictionary. In: Tiberius, C. and C. Müller-Spitzer (eds.) *Wörterbuchbenutzungsforschung* 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“. Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik).

*le Trésor de la langue française informatisé*. Accessed at: <http://atilf.atilf.fr/> [10/04/2014].

