EURAC research

LULCL II 2008

Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics

Bozen-Bolzano, 13th - 14th November 2008

REGIONE AUTONOMA TRENTINO-ALTO ADIGE
AUTONOME REGION TRENTINO-SÜDTIROL
REGION AUTONÓMA TRENTIN-SÜDTIROL

The initiative is co-financed by the Autonomous Region Trentino – South Tyrol,
Department III - Linguistic Minorities and European Integration,
Office for Linguistic Minorities

Bestellungen bei:

Europäische Akademie Bozen
Drususallee 1
39100 Bozen – Italien
Tel. +39 0471 055033
Fax +39 0471 055099
E-mail: press@eurac.edu

Ordinare Libro:

Accademia Europea Bolzano
Viale Druso, 1
39100 Bolzano – Italia
Tel. +39 0471 055033
Fax +39 0471 055099
E-mail: press@eurac.edu

Verantwortlicher Direktor: Stephan Ortner
Herausgeber: Verena Lyding
Redaktion / Koordination: Verena Lyding
Revision: Peter Farbridge
Druckvorstufe: Typoplus, BZ
Druck: ESPERIA srl

Direttore responsabile: Stephan Ortner
Curatore: Verena Lyding
Redazione / Coordinazione: Verena Lyding
Revisione: Peter Farbridge
Prestampa: Typoplus, BZ
Stampa: ESPERIA srl

An online version of the proceedings is available at:
http://www.eurac.edu/Org/LanguageLaw/Multilingualism/index

Last modified: August 2009

ISBN 978-88-88906-52-2

LULCL II 2008

Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (LULCL II)

"Combining efforts to foster computational support of minority languages"

Bozen-Bolzano, 13th-14th November 2008

Verena Lyding (Ed.)

Index

Preface

The publication at hand presents the contributions to the Colloquium on Lesser Used Languages and Computer Linguistics II (LULCL II) held at the European Academy of Bolzano-Bozen (EURAC) on the 13th and 14th of November 2008.

The Institute for Specialised Communication and Multilingualism organized a first conference on Lesser Used Languages and Computer Linguistics in 2005 that aimed to provide an overview of the state-of-the-art in research on lesser-used languages with special reference to their computational support. Apart from the methodological issues, the promotion of networking among researchers was central to the conference. The overwhelmingly positive feedback of the participants encouraged us to follow up on the initiative.

Three years later we were pleased to invite researchers for a second time to a scientific meeting on computational approaches for lesser-used languages. LULCL II extended the scope of the first conference to include—in addition to research on lesser-used languages—research initiatives on language varieties, sign languages, learner language and any type of spoken language. When it comes to the automatic processing of language and the building up of computational resources, sparseness of resources, a lower degree of standardization and smaller research communities are common issues facing researchers of all these fields. Under the motto "Combining efforts to foster computational support of minority languages", the second LULCL colloquium put special focus on bringing together these research communities in order to combine best practices, approaches and techniques, and to add value to individual initiatives.

Through individual talks and poster presentations, over the two day meeting in November 2008, general challenges and a wide variety of projects and initiatives were presented on different aspects of applying and making use of language technologies in the context of lesser-used languages, language varieties, learner language, spoken language and sign language. Thanks to the commitment of the authors, articles on the majority of the presented topics are included in the LULCL II proceedings.

The proceedings are structured according to the schedule of the conference, which devoted the first day to projects on minority languages and language varieties, and continued on the second day with projects on learner language, spoken language and sign language.

The first contribution, by Paul Videsott, is concerned with the research scenery on the local minority language Ladin, with a particular focus on the work carried out at the Free University of Bolzano/Bozen (FUB). In addition to discussing the current research policy within the Ladin community (spread over five valleys), Videsott presents the Ladin Section of the FUB and gives an overview on current and planned research projects, including the creation of bibliographies, dictionaries and a corpus of Ladin. Giovanni Mischì provides a complementary view of the work on Ladin by describing the compilation of a modern Ladin dictionary, with special focus on the creation of neologisms. Close collaboration with a recognized authority and adherence to a clear and consistent plan are key methodologies for creating dictionaries that provide for the practical needs and uses in a modern context. Computer applications like online dictionaries and spellcheckers add to the user-friendliness and attractiveness of the Welsh language. Delyth Prys' contribution on the Welsh language supports the above idea. The availability of electronic resources has had a great impact on the acceptance of the language in terms of perceived modernity, as well as on productivity and the confidence of the language's users. This is strikingly shown via usage statistics and surveys of a number of electronic resources created at the Language Technologies Unit at Canolfan Bedwyr, Bangor University, including spelling and grammar checkers, terminological databases, and text-to-speech synthesis.

Amàlia Mendes and the research group of the *Centre of Linguistics of the University of Lisbon* report on their experiences creating corpora for five African varieties of Portuguese, which serve as a fundamental basis for contrastive linguistic analyses. The authors discuss issues in designing, compiling and annotating a corpus for each variety and provide insights derived from initial contrastive studies focusing on differences in the varieties' lexicons. The Vis-À-Vis system presented by Stefanie Anstein aims at supporting this and similar types of contrastive studies on language varieties. By adopting tools for the processing of main varieties, the sketched system will provide support for the comparison of language varieties on different levels of linguistic description. Aleksey Andronov and Everita Andronova describe their work on a corpus for modern Latgalian, a Latvian variety used in the eastern part of Latvia (Latgale). They discuss the challenges of corpus creation with regard to size, representativeness and balance, which are ultimately related to the restricted usage of Latgalian (mainly to oral conversation and in a few spheres of social life). The inclusion of a Latgalian component in the Latvian National Corpus has the aim to strengthen the status of the language and facilitate its linguistic research.

Using the example of several African languages, Guy De Pauw and Gilles-Maurice de Schryver show how data-driven methods can support the development of NLP applications for lesser-resourced languages. Based on research efforts on automatic diacritic correction, the development of a robust POS-tagger and a tool for morphologic analysis, the authors show how minority languages can benefit from the typical advantages of the data-driven approach, such as robustness, empiricism, language independence and development speed.

In an article related to her keynote speech, Karin Aijmer gives an overview on issues in learner corpus research. With reference to main international projects, as well as a case study on the use of the expression 'of course', she discusses the challenges of corpus compilation and annotation as well as methodological issues in the analysis of learner language. In a second contribution devoted to learner corpora, Elisa Corino describes the experience of creating and analyzing VALICO, the online corpus of learning varieties of the Italian language at Torino. Besides a discussion on general characteristics of learner corpora and design considerations, she pays particular attention to approaches for data elicitation.

Julianne Nyhan discusses the role of Digital Humanities in research on different aspects of the Irish language, including its literature, history, and language learning. In this context, three resources of Irish are presented in more detail: the Corpus of Electronic Texts (CELT), the Dictionary of the Irish Language (DIL) and Dinneen's dictionary.

Thomas Schmidt's contribution on spoken language corpora starts by listing a number of methodological issues that distinguish the work with spoken from that of written language data. Schmidt infers that working with spoken languages requires the researcher to be more theory-aware, to keep a close link between recording, transcription and meta-data, and to employ a data model which can represent parallel temporal relationships. Schmidt then discusses how the EXMARaLDA system of the University of Hamburg is responding to these requirements. Caren Brinckmann focuses on a specific task in processing spoken language data: the so-called 'transcription bottleneck'. She presents several possible alternatives to the very time consuming manual production of orthographic and phonetic/phonological transcriptions, including a community-driven approach, an automated approach and a query-driven approach.

For languages that do not possess a writing system, transcription poses serious problems, and for non-written languages (such as sign languages) other media like video may be the only means of authentically capturing the language. Eleni Efthimiou's contribution addresses this and related issues in the context of educational applications for the Greek Sign Language (GSL). As a minority non-written language, resources for education, including materials and tools, remain scarce. Taking into consideration the inherent difficulties in the production and presentation of linguistic content of GSL, Efthimiou discusses the design requirements and implementation issues of a platform for educational applications for deaf users.

The wide variety of contributions to these proceedings provides an impressive insight into the different facets of the effort into computational applications for lesser-used and lesser-resourced languages. Despite the diversity of projects described and perspectives taken by the authors, two core aspects are recurrent in all articles: (1) the practical issues of the creation of computational resources with respect to quality and efficiency; and (2) the impact of the availability of computational resources on the community of speakers and researchers.

Item (1) includes facilitating as well as obstructive factors related to the availability of theoretical fundamentals and of computational resources for language processing; the context of usage of the language; the political status of the language; and the existence of standards concerning the language. Item (2) refers to the possibilities for carrying out research, the usefulness of the language for the different contexts of life, the image of the language, and the general acceptance by the speaker community.

LULCL II aimed to raise questions surrounding how the aims, tasks and practices that we have in common as research community can be practically combined when working with lesser-used and lesser-resourced languages. Standardisation (not only of languages, but also of data formats and resources), metadata, evaluation, ethics (including partnership, payback, respect for data), networking and cooperation were some of the concepts that Dafydd Gibbon chose in the closing session to summarize the central topics of the conference. A network was created at the second LULCL meeting, and we hope that the proceedings can contribute to further distribute the assembled experiences, practices and resources.

Special thanks are due to all authors who contributed to these proceedings with interesting and meaningful articles. Furthermore, we would like to thank the Scientific Committee once more for their commitment. There are many more people who we would like to thank for contributing to LULCL II by helping to organize the conference and preparing the proceedings for print. Last but not least, our gratitude goes to all who contributed their inspiration by taking part at LULCL II in November 2008.

Verena Lyding

Ladinistische Forschungsprojekte an der Freien Universität Bozen

Paul Videsott

Our contribution consists of three parts. The first part discusses the current policy for research on Ladin. Particular consideration has been given to possible scientific activities and collaborations within the Ladin community of the Dolomites, as well as within the neighboring Rhaeto-Romance regions in Grisons and Friuli. The second part presents the Ladin Section of the Free University of Bolzano-Bozen (established in 1998), which is in charge of research and education for Ladin at the Faculty of Education in Bressanone-Brixen. The third part is concerned with the description of current and planned research projects at the Ladin Section of the Free University of Bolzano-Bozen. These projects comprise efforts towards the creation of scientific bibliographies for Ladin, Ladin corpora, monolingual dictionaries and literary studies on Ladin of the Dolomites, and include the participation in the European Project LINEE (Languages in an European Network of Excellence). The present article aims at facilitating the cooperation between the Ladin Section of the Free University of Bolzano-Bozen and other research institutions that are working in a similar research area with minority languages comparable to Ladin.

1. Einleitung

Seit 1998 besteht an der Fakultät für Bildungswissenschaften der Freien Universität Bozen eine eigene Ladinische Abteilung, an der im Oktober 2006 die weltweit einzige Professur für Romanische Philologie/Ladinistik besetzt wurde. Der folgende Beitrag stellt die an dieser Abteilung sich bereits in Durchführung befindlichen bzw. die dort für die nächsten Jahre geplanten Forschungsprojekte vor. Unser Ziel ist es, über diese Vorstellung Kontakte und Kooperationen zu Forschungseinrichtungen herzustellen, die ähnliche Ziele verfolgen, um dadurch mögliche Synergien produktiv zu nutzen.

2. Zur aktuellen forschungspolitischen Situation des Ladinischen

Unter *Ladinisch* (*Ladin*) ist hier der zentrale Bereich des rätoromanischen Sprachtypus gemeint, zu dem im Westen (Kanton Graubünden, Schweiz) noch das *Bündnerromanische* (*Rumantsch Grischun*) und im Osten (historische Region Friaul, politische Region Friaul-Julisch-Venetien) das *Friaulische* (*Furlan*) als Gebiete mit einem historisch gewachsenen Eigensprachlichkeitsbewusstsein und einer in unterschiedlichen Graden offizialisierten Sprachverwendung gehören.[1]

Nachdem die linguistische und soziolinguistische Situation in den drei rätoromanischen Teilgebieten bedeutende Parallelen aufweist[2], liegt eine Zusammenarbeit zwischen den politischen, kulturellen und wissenschaftlichen Institutionen der Bündnerromanen, Dolomitenladiner und Friauler nahe.[3] Tatsächlich hat jedoch eine politische Zusammenarbeit in nennenswertem Umfang nie stattgefunden, was mit der unterschiedlichen staatlichen und regionalen Zugehörigkeit und z.T. auch mit den unterschiedlichen Aspirationen der drei Gebiete erklärt werden mag.[4] Auf kultureller Ebene haben sich Phasen intensiverer Zusammenarbeit mit eher „isolationistischen“ Bestrebungen abgewechselt: erwähnt seinen z. B. die „Ladinischen Kongresse“ in den 60er Jahren oder das Forum *LaFuRum* in den 90er Jahren des 20. Jh. Seitdem ist aber auch auf kulturellem Gebiet die Zusammenarbeit zurückgegangen, wofür u. M. nach – neben dem inzwischen eingetretenen Generationenwechsel in den jeweiligen Leitungsfunktionen – einerseits die Zersplitterung der Dolomitenladiner selbst verantwortlich gemacht werden muss, andererseits die damit zusammenhängende, diametral entge-

[1] Wir verzichten hier auf eine Rekapitulation der gesamten “Questione Ladina”, bei der es im Wesentlichen darum ging, ob die Zusammenfassung der drei Teilbereiche Bündnerromanisch, Dolomitenladinisch und Friaulisch zu einem Geotyp höherer Ordnung („Rätoromanisch“ im Sinne von Th. Gartner [1883], „Ladinisch“ im Sinne von G. I. Ascoli [1873]) im Gegensatz zum benachbarten Geotyp „Italienisch“ gerechtfertigt ist oder nicht. Der im 20. Jh. durchgeführte Sprachausbau (Korpus und Status) verbunden mit der Anerkennung von Minderheitenrechten hat die Diskussion nicht nur de jure, sondern auch de facto überholt und überflüssig gemacht.

[2] Zu den historischen und ethnogenetischen Parallelen zwischen den drei Teilgebieten vgl. Goebl (2001) und Gsell (1990).

[3] Die Literatur zur linguistischen und soziolinguistischen Situation in den drei rätoromanischen Teilgebieten ist relativ umfangreich, wobei aufgrund der raschen Veränderungen in den letzten Jahren nicht alle Daten immer den neuesten Stand widerspiegeln. Wir verweisen deswegen auf die regelmäßigen bibliographischen Aktualisierungen in der Zeitschrift *Sociolinguistica*.

[4] Ein makroskopisches Beispiel: Die Bündnerromanen gehören im Gegensatz zu den Dolomitenladinern und den Friaulern nicht zur EU! Jedoch sei zumindest ein ganz rezentes Beispiel politisch angeregter Zusammenarbeit zwischen Dolomitenladinern und Bündnerromanen erwähnt: die 5-sprachige Internetseite *filcultural* (www.filcultural.info/).

gengesetzte Sprachpolitik, welche die offiziellen Institutionen in den drei Teilgebieten im letzten Jahrzehnt verfolgt haben: Eine Politik der Domänenerweiterung für die Minderheitensprache mittels Anwendung der jeweiligen Dachsprachen „Rumantsch Grischun" und „Koiné furlana" in Graubünden und Friaul, hingegen eine eher konservative Abdeckung weniger Kernbereiche unter gezieltem Einsatz von Lokalvarianten (Talidiomen) vor allem in den ladinischen Tälern Südtirols, welche den numerisch, linguistisch und kulturell stärksten Teil der Dolomitenladiner darstellen.

Bekanntlich sind die Dolomitenladiner seit 1927 administrativ dreigeteilt: Die acht ladinischen Gemeinden des Grödner- und Gadertales gehören zur Autonomen Provinz Bozen; die sieben Gemeinden des Fassatals zur Autonomen Provinz Trient (beide Autonomen Provinzen bilden gemeinsam die Autonome Region Trentino-Südtirol); die drei Gemeinden Buchensteins und Ampezzos gehören schließlich zur Provinz Belluno, welche Teil des Veneto (italienische Region mit Normalstatut) ist. Diese Dreiteilung hat höchst unterschiedliche Schutzmechanismen für das Ladinische zur Folge, die sich direkt auf dessen sprachliche Überlebensfähigkeit auswirken.[5]

Südtirol	Einw. 1971	Einw. 1981	Einw. 1991	Einw. 2001	Einw. 2007	Diff. 71–07
Mareo	2.377	2.413	2.574	2.682	2.828	**+ 18,97 %**
S. Martin	1.374	1.427	1.495	1.690	1.731	**+ 25,98 %**
La Val	1.069	1.143	1.199	1.232	1.263	**+ 18,15 %**
Badia	2.271	2.575	2.722	3.015	3.241	**+ 42,71 %**
Corvara	951	1.183	1.236	1.266	1.270	**+ 33,54 %**
Sëlva	2.137	2.294	2.394	2.513	2.590	**+ 21,20 %**
S. Cristina	1.494	1.567	1.598	1.738	1.843	**+ 23,36 %**
Urtijëi	3.949	4.080	4.226	4.480	4.558	**+ 15,42 %**
Trentino						
Cianacei	1.447	1.608	1.730	1.818	1.844	**+ 27,44 %**
Ciampedel	588	653	708	732	743	**+ 26,36 %**
Mazin	355	379	422	440	451	**+ 27,04 %**
Poza	1.426	1.621	1.668	1.787	1.785	**+ 25,18 %**
Vich	815	883	936	1.073	1.054	**+ 29,33 %**
Soraga	440	519	590	673	677	**+ 53,86 %**
Moena	2.688	2.583	2.567	2.602	2.597	**– 3,39 %**
Belluno						
Cortina	8.499	8.109	7.109	6.085	6.190	**– 27,17 %**
Col	603	556	480	418	400	**– 33,67 %**
Fodom	1.718	1.576	1.440	1.417	1.436	**– 16,41 %**

Tabelle 1: Die Bevölkerungsentwicklung der ladinischen Gebiete

[5] Besonders deutlich wird das Gesagte auch anhand der demographischen Entwicklung (vgl. Tabellen 1 und 2).

	Val Badia	Gherdëina	Fascia	Fodom	Ampezzo	Insgesamt
Einwohner	10.311	10.126	9.348	1.844	6.175	37.804
Ladinisch-sprecher	9.720	8.676	7.740	ca. 1.660	ca. 2.470	ca. 30.250
Anteil	94,27 %	85,68 %	82,80 %	ca. 90 % (keine offiziellen Angaben)	ca. 40 % (keine offiziellen Angaben)	

Tabelle 2: Prozentuale Anteile von Ladinischsprechern in der Bevölkerung (Daten von 2007)

Drei besonders prägnante Beispiele mögen das verdeutlichen:

Präsenz des Ladinischen in der Schule: Im Gadertal und Gröden ist der Unterricht des Ladinischen seit 1948 gesetzlich verpflichtend vorgesehen. Bei einem ansonsten paritätisch deutsch-italienischen Schulsystem ist der Ladinischunterricht im Ausmaß von zwei Wochenstunden in der Pflichtschule und einer Wochenstunde in den Oberschulen vorgeschrieben. Im Fassatal war der Unterricht des Ladinischen zwar seit 1970 fakultativ möglich, ist aber erst 1994 in das Pflichtkurrikulum im Ausmaß von mindestens einer Wochenstunde aufgenommen worden bei einer ansonsten (für den Sachunterricht) fast ausschließlichen Verwendung des Italienischen als Unterrichtssprache.[6] In Buchenstein und Ampezzo war bis 1999 überhaupt kein Ladinischunterricht an der Schule vorgesehen; seit 1999 ist er zugelassen, sofern die Eltern der Schüler und das Lehrerkollegium einer Klasse damit einverstanden sind. Diese einhellige Zustimmung war in Ampezzo bisher nicht zu erreichen.[7]

Offizielle Anerkennung der Ladiner als sprachliche Minderheit: In der Provinz Bozen erfolgte diese Anerkennung bereits 1951, in der Provinz Trient erst 1976, in der Provinz Belluno schließlich erst 1999; hier jedoch völlig undifferenziert zwischen „Alt-" und „Neoladinern".[8]

Erhebung des Ladinischen zur offiziellen Verwaltungssprache: In der Provinz Bozen seit 1989, in der Provinz Trient seit 1994, in der Provinz Belluno in dieser Form nicht vorgesehen (vgl. Mischì 1994).

[6] Im Fassatal besteht zusätzlich die Möglichkeit, im Ausmaß von bis zu drei Wochenstunden auch ein Fach auf Ladinisch zu unterrichten. Die praktische Umsetzung dieser Möglichkeit stößt aber aufgrund fehlender Lehrmittel auf Ladinisch und aufgrund des Mangels an LehrerInnen mit entsprechender sprachlicher Vorbereitung auf Schwierigkeiten (vgl. Santuari 2006: 223–225).

[7] Ein Überblick über die wichtigsten Bestimmungen zum Schulunterricht in den ladinischen Tälern gibt Verra (2000). Die für Buchenstein und Ampezzo relevante Bestimmung findet sich im Art. 4 des Gesetzes 482 vom 15.12.1999: „[…] le istituzioni scolastiche elementari e secondarie di primo grado […] deliberano, anche sulla base delle richieste dei genitori degli alunni, le modalità di svolgimento delle attività di insegnamento della lingua e delle tradizioni culturali delle comunità locali […]".

[8] Zur Geschichte der Anerkennung des Ladinischen vgl. Richebuono (1982), zur Problematik der „Neoladiner" vgl. Goebl (1997).

Es ist offensichtlich, dass unter diesen Bedingungen die *zentrifugalen* Kräfte in der Dolomitenladinia besonders stark sind und deswegen durch gezielte *zentripetale* Aktionen zumindest abgemildert werden müssten. Dies müsste insbesondere im Bereich der Sprache passieren, da diese ja das wichtigste einigende Band zwischen den einzelnen Talschaften sowie das wichtigste distinktive Merkmal der Ladiner überhaupt darstellt. Doch der Versuch der Einführung einer gemeinsamen ladinischen Schriftsprache, dem nach dem Vorbild des „Rumantsch Grischun" von Heinrich Schmid ab 1988 konzipierten „Ladin Dolomitan" (vgl. Schmid 1998), wurde – nach einer anfänglichen passiven Toleranz – 2003 in der Provinz Bozen zugunsten einer „abwechselnden Verwendung der Talidiome Gadertalisch und Grödnerisch" unterbunden.[9] Gleichzeitig wurde durch eine Orthographie-Revision vor allem das Grödnerische stärker an die lokale Phonetik gebunden, was es für eine überlokale Verwendung in der Praxis nur erschwert brauchbar macht. Die Abkehr vom Ziel eines gemeinsamen Sprachausbaus hat in Südtirol zu einigen krassen Beispielen von Talisolationismus geführt, welche es den übrigen Ladinern (vor allem den rechtlich kaum geschützten Buchensteinern und Ampezzanern) noch schwerer als bisher machen, aktiver und gleichberechtigter Teil eines gemeinsamen Sprach- und Kulturraumes „Dolomitenladinia" zu sein.[10]

Auf wissenschaftlichem Gebiet (universitäres oder vergleichbares Niveau) müssen sich Kooperationen zwischen den drei rätoromanischen Gebieten ebenfalls erst etablieren, weil bis vor Kurzem die entsprechenden Institutionen alle außerhalb des rätoromanischen Sprachraumes lagen und eine evtl. Zusammenarbeit deswegen meistens zwischen nicht-rätoromanischen Wissenschaftlern, die (vor allem als Romanisten) zum

[9] Vgl. das entsprechende Gesetzesdekret unter www.noeles.net/modules.php?name=News&file=article&sid=173. Zu den Intentionen des "Ladin Dolomitan" vgl. Videsott (1998).

[10] Im Bereich der Orthographie wurde festgelegt (vgl. Forni 2001, 2002), intervokalisch bei Präsenz eines stimmhaften [s] im Italienischen <j> zu favorisieren (grd. *fantajia*, *poejia*, *ijula* vs. gad. *fantasia*, *poesia*, *isola*); den vokalischen Anlaut bei Internationalismen mit der Silbe *in-* nicht mehr zu schreiben (grd. *nteressant*, *ndicatif*, *ndustria* vs. gad. *interessant*, *indicatif*, *industria*); bei Internationalismen *u* statt *o* im Vorton zu schreiben (grd. *pruvinzia*, *cungiuntif*, *culaborazion* vs. gad. *provinzia*, *congiuntif*, *colaborazion*); die italienischen Kombinationen *-rs-* uns *-ns-* im Gegensatz zum Gadertal mit mit *-rsc-*/*-nsc-* wiederzugeben (grd. *verscion*, *perverscion*, *cumprenscion* vs. gad. *verjiun*, *perverjiun*, *comprenjiun*), die Präposition *en* nicht mehr vom unbestimmten Artikel zu unterscheiden (grd. *n* vs. gad. *n* [Art.]/*en* [Präp.]), den Nexus [ië] anders als im Gadertal <ie> zu schreiben (grd. *pazienza*, *sapienza*, *cuscienza* vs. gad. *paziënza*, *sapiënza, cosciënza*), der Einführung von neuen Akzentregeln (im grd. wird eine verbale Inversionsform nicht mehr akzentuiert, im Gadertal weiterhin schon) usw. Im Bereich des Lexikons wurden zahlreiche Neologismen im Gadertal und Gröden unterschiedlich gebildet, wie ein Vergleich der Wörterbücher von Forni (2002) und Mischì (2000) zeigt. Parallel dazu wird die Tendenz immer stärker, als Glottonym in Gröden nur mehr „gherdëina" zu verwenden und das übergeordnete „ladin" auf das Gadertal zu beschränken. Zur Gesamtproblematik vgl. Chiocchetti (2007).

Rätoromanischen arbeiteten, stattfand.[11] Im letzten Jahrzehnt haben jedoch alle drei Teilgebiete entsprechende Institutionen erhalten[12], die in der Regel auch mit Angehörigen der Sprachminderheit besetzt wurden, was in Zukunft die Zusammenarbeit zumindest erleichtern müsste.

3. Die Ladinische Abteilung der Freien Universität Bozen

Die Ladinische Abteilung der Freien Universität Bozen wurde 1998 im Zuge der Universitätsgründung eingerichtet. Sie ist an der Bildungswissenschaftlichen Fakultät in Brixen angesiedelt und hat sowohl didaktische wie auch wissenschaftliche Aufgaben.[13] Derzeit ist sie personell mit einer Professur, einer festen Mitarbeiterin für die Durchführung der Schulpraktika sowie mit Lehrbeauftragten und Projektmitarbeitern (letztere über Drittmittel finanziert) ausgestattet.

Die didaktischen Aufgaben liegen in der Organisation und Durchführung des ladinischen Teiles des Laureatsstudienganges für den Primarbereich / Ladinische Sektion. An der Fakultät für Bildungswissenschaften sind im Primarbereich (Kindergarten und Grundschule) zwei Laureatsstudiengänge eingerichtet: eine deutsche Sektion für die zukünftigen Lehrkräfte an den deutschen Schulen und eine entsprechende italienische Sektion für die italienischen Schulen. Nachdem aber in Südtirol daneben auch noch der paritätische ladinische Schultyp existiert, ist für die Ausbildung dieser Lehrkräfte die ladinische Sektion eingerichtet worden. Die Studierenden an dieser Sektion laufen formell unter dem deutschen Laureatsstudiengang, dürfen aber ihre Lehrveranstaltungen während des gesamten Studiums vollkommen frei aus dem Angebot der deutschen und der italienischen Sektion wählen, wobei aber auf ein ausgewogenes Verhältnis der beiden Sprachen geachtet wird. Zusätzlich absolvieren sie ca. 20 % ihrer Stundenanzahl an der ladinischen Abteilung. Deren Lehrveranstaltungen haben das Ladinische

[11] Um das Rätoromanische besonders verdient gemacht haben sich u.a. die Universitäten Genf (www.unige.ch/lettres/roman/index.html), Freiburg (www.unifr.ch/rheto/), Zürich (www.rose.uzh.ch/studium/faecher/raetorom.html), Innsbruck (www.uibk.ac.at/romanistik/), Salzburg (ald.sbg.ac.at/ald/default%20ald%20 1%202.htm), Eichstätt, Padua, Udine und Triest (www2.units.it/~clettere/pellegri.htm). Schöne rezente Beispiele von Kontakten auf universitärem Niveau waren die bisher vier „Colloquiums Retoromanistics“, die 1996 von Univ. Prof. Dieter Kattenbusch initiiert wurden (vgl. Kattenbusch [1999]; Ladinia 21 [1997]; Annalas da la Societad Retorumantscha 113 [2000]; Ladinia 26–27 [2002–03] sowie Vicario [2007]).

[12] Vgl. die Pädagogische Hochschule Chur (www.phgr.ch/), die Ladinische Abteilung der Freien Universität Bozen (www.unibz.it), das Interfakultäre Zentrum für Friaulisch an der Universität Udine (www.uniud.it/cirf).

[13] Zur Gründungsgeschichte der Abteilung vgl. Mischì (2008).

nicht nur zum Thema, sondern werden zum Großteil auch auf Ladinisch gehalten.[14] Diese systematische Verwendung des Ladinischen auf universitärem Niveau ist einmalig und bedarf natürlich der entsprechenden Fachterminologie, die in der Regel erst gebildet werden muss.

Die Ansiedlung der ladinischen Abteilung an der Fakultät für Bildungswissenschaften und ihre Bindung an den Laureatsstudiengang für den Primarbereich hat zur Folge, dass sie vor allem von zukünftigen Lehrkräften der Primarschulen des Gadertals und Grödens besucht wird, die ja als einzige Schulen in Ladinien paritätisch geführt werden. Die ladinische Abteilung ist deswegen nicht als ausschließlich linguistische Abteilung konzipiert, wie es bei vergleichbaren Institutionen, die in der Regel an philologisch-sprachwissenschaftliche Fakultäten angeschlossen sind, oft der Fall ist. Auch die im Studienmanifest der ladinischen Abteilung vorgesehene Pflicht zur Dreisprachigkeit und die spezifische Ausbildung für den paritätischen Unterricht bilden eine Einstiegshürde für Studierende aus den anderen ladinischen Tälern, wo der Unterricht des Deutschen einen vollkommen anderen Stellenwert hat. In diesem Kontext und im derzeitigen Klima der wirtschaftlichen Unsicherheit sei schließlich noch erwähnt, dass die Absolventinnen der Ladinischen Sektion nach ihrem Studium naturgemäß einen (sicheren) Arbeitsplatz an der Schule anstreben und weniger eine Anstellung als wissenschaftliche Mitarbeiterinnen eines befristeten Forschungsprojektes, für das aber wiederum Dreisprachigkeit und fundierte Kenntnisse des Ladinischen Voraussetzung sind.[15]

4. Ladinistische Projekte an der Ladinischen Abteilung der Freien Universität Bozen

Neben ihren Aufgaben im Bereich der Lehre hat die ladinische Abteilung auch den Auftrag, Forschung zum Dolomitenladinischen zu betreiben. Aufgrund ihrer Ausrichtung ergeben sich naturgemäß zwei Schwerpunkte: Forschung im *linguistischen* Bereich sowie im Bereich der *mehrsprachigen Didaktik*. Nachdem die Didaktik aber derzeit, im Gegensatz zur Linguistik, nur durch Lehraufträge vertreten ist, seien im Folgenden nur die Forschungsprojekte mit linguistischem Schwerpunkt vorgestellt.

[14] Zur Beschreibung der Lehre an der ladinischen Abteilung vgl. Videsott (2008).

[15] Als sich anbahnende Kooperation zwischen den Dolomiten und Graubünden im Bereich der Lehre sei auf die Möglichkeit verwiesen, Schulpraktika im jeweils anderen rätoromanischen Gebiet zu absolvieren. Bündnerromanische Studierende können somit das dreisprachige ladinische Schulsystem und umgekehrt ladinische Studierende die Realität der einsprachig rätoromanischen Schulen kennenlernen.

4. 1 Wissenschaftliche Bibliographien zum Ladinischen

4. 1. 1 Rätoromanische Linguistische Bibliographie

Bibliographien gehören zu den wichtigsten Hilfsmitteln des Wissenschaftsbetriebs überhaupt. Je besser eine Sprache bibliographisch erschlossen ist, desto eher ist sie für die (nationale und internationale) Forschung zugänglich.

An einer solchen Bibliographie für das Rätoromanische wird (unter Verwendung von Vorarbeiten von Maria Iliescu und Heidi Siller-Runggaldier, vgl. Iliescu & Siller-Runggaldier [1985]; Siller-Runggaldier & Videsott [1998]) an der Ladinischen Abteilung unter dem Werktitel *„Rätoromanische Linguistische Bibliographie (von den Anfängen bis 2010)"* gearbeitet. Diese Bibliographie soll alle sprachwissenschaftlichen Arbeiten *zum* Rätoromanischen, von Beginn der Rätoromanistik als wissenschaftlicher Disziplin Ende des 19. Jahrhunderts bis zum heutigen Tag, enthalten (ca. 4500 Einträge).

Die bereits in Buchform veröffentlichten Teilbereiche dieser Bibliographie wurden von der internationalen Wissenschaftsgemeinschaft überaus positiv aufgenommen, wie die zahlreichen dazu erschienenen Rezensionen beweisen.[16] Mit der Vervollständigung dieses Materials bis 2009 und der Gesamtveröffentlichung sowohl in Buchform als auch als Online-Bibliographie (für zukünftige Ergänzungen und Aktualisierungen) wird der internationalen (Räto-)Romanistik ein eminent wichtiges Forschungswerkzeug zur Verfügung stehen.

Einen besonderen dokumentarischen und praktischen Wert wird die Bibliographie durch das Vorhaben erhalten, alle Texte, die aufgrund ihres Alters nicht mehr urheberrechtlich geschützt sind, zu digitalisieren und als Anlage zum jeweiligen bibliographischen Eintrag über Internet zugänglich zu machen. Damit würden diese Texte – zum Großteil Unikate oder Rarissima – wieder der internationalen Forschungsgemeinschaft zugänglich gemacht werden.

4. 1. 2 Ladinische Nationalbibliographie

Ein zweites, ebenso wichtiges Hilfsmittel stellt eine Bibliographie zu den (gedruckten) Werken dar, die bisher *auf* Dolomitenladinisch erschienen sind *(Ladinische Nationalbibliographie)*. Eine solche Bibliographie fehlt derzeit (im Gegensatz zu Graubünden; vgl. *Bibliografia Retoromontscha* 1938, 1956) für den dolomitenladinischen

[16] Insgesamt 15 Rezensionen zum ersten und 11 zum zweiten Band, vgl. Siller-Runggaldier & Videsott (1998: 2).

Bereich vollkommen und wird deswegen als wichtiges Desideratum angesehen. Abgesehen von der dokumentarischen Nützlichkeit einer solchen Bibliographie an sich (insbesondere für die Zeit vor der Gründung der ladinischen Kulturinstitute 1975 bzw. 1976, da bis dahin sehr viele ladinische Publikationen im Eigenverlag erschienen und deswegen kaum auffindbar sind), ist eine solches Forschungsinstrument besonders hilfreich, um die Schriftproduktion auf Ladinisch quantifizieren zu können (etwa in Bezug auf die einzelnen Talschaftsidiome), ihr ein chronologisches Relief zu geben oder um einen Überblick über jene Bereiche zu bekommen, in denen das Ladinische bereits eingesetzt worden ist. Die „Ladinische Nationalbibliographie" stellt auch ein extrem wichtiges Hilfsmittel für alle korpusbasierten Studien zum Ladinischen dar.

Einen besonderen dokumentarischen und praktischen Wert wird die Bibliographie wiederum durch das Vorhaben erhalten, alle Texte, die aufgrund ihres Alters nicht mehr urheberrechtlich geschützt sind, zu digitalisieren und als Anlage zum jeweiligen bibliographischen Eintrag über Internet zugänglich zu machen.

4. 2 Aufbau von Korpora zum Ladinischen

Digitale Textkorpora sind heutzutage die unerlässliche Grundlage für alle sprachbeschreibenden und sprachnormierenden Arbeiten, die überindividuelle Gültigkeit beanspruchen (insbesondere normative Grammatiken und Wörterbücher). Ein solches Korpus für das Dolomitenladinische ist deswegen nicht nur ein wichtiges Desiderat der (räto-)romanischen Sprachwissenschaft, sondern auch ein eminent nützliches Hilfsmittel für die ladinischen sprachpflegenden und sprachfördernden Institutionen.

Seit der Erstellung der ersten Korpora in den 60er Jahren im Bereich der Angewandten Linguistik hat die Anzahl der Projekte zur Sprachforschung mittels Textkorpora stetig zugenommen. Aber erst in den letzten Jahren wurden davon auch Minderheitensprachen erfasst. Für Sprachen wie das Katalanische, das Baskische, aber auch das Irische oder das Walisische wurden und werden Korpora erstellt, um breit angelegte empirische Sprachstudien durchzuführen. Insbesondere mit diesen Minderheiten streben wir im Rahmen des Projektes den Ausbau einschlägiger Kontakte und Kooperationen an.[17]

Vom ladinischen Kulturinstitut „Majon di Fascegn" wurde das Korpus „TALL" (*Tratament Automatich dl Lingaz Ladin*)[18] initiiert. In dieses elektronische Korpus soll idealerweise die gesamte ladinische Schriftproduktion einfließen. Im Rahmen unseres

[17] Für die Bündnerromanen und Friauler kann hingegen das Projekt vorbildhafte Funktion haben.

[18] Vgl. http://corpuslad.tall.smallcodes.org/applications/textanalysis/sitecorpuslad/index.jsp

Teilprojektes *Corpus Ladin leterar*[19]ergänzen wir dieses bisher vor allem auf administrativen Übersetzungen beruhende Korpus mit solchen Texten, die aufgrund ihrer linguistischen Qualität als beispielhaft gelten können (d.h. vorrangig mit literarischen Texten). Damit soll eine digitale Materialbasis bereitstehen, die aufgrund ihrer Größe und statistischen Auswertbarkeit verlässliche Angaben zum Gebrauch und zum Kontext jener Formen im geschriebenen Ladinischen liefern kann, die gerade Gegenstand einer (orthographischen, morphologischen, lexikalischen oder semantischen) Untersuchung sind. Das Textkorpus soll so aufbereitet werden, dass die Gesamtheit der darin enthaltenen lexikalischen Elemente nach den Parametern Diatopie / Diachronie / Diastratie eindeutig ausgezeichnet werden kann.

Derzeit ist die Menge an digital zugänglicher ladinischer Sprache (Internetseiten, digitale Zeitungen, digitalisierte Bücher und Texte usw.) selbst im Vergleich zu anderen Minderheitensprachen minimal. Der Mangel an Daten stellt für die Sprachbeschreibung und -analyse des Ladinischen eine große Hürde dar. Dies führt unter anderem dazu, dass sich selbst Autoren von normativen Schulgrammatiken bei der Formulierung von Regeln auf ihr eigenes Sprachgefühl verlassen oder sich ihre Beispiele mühsam aus gedruckten Werken zusammentragen müssen, ohne sich sicher sein zu können, ob es sich dabei um ein *Hapax legomenon* (Einzelbeleg) eines einzelnen Autors, um einen Druckfehler oder um eine bisher nicht beschriebene, aber korrekte Ausdrucksweise handelt. Bei entsprechenden Nachfragen werden die Autoren meist unsicher und verweisen auf „andere“ Bezugspersonen, die man zusätzlich befragen sollte. Mit einem digitalen Textkorpus wird allein die Masse an verfügbaren Beispielen viele Fragen bezüglich der Akzeptabelität/Inakzeptabelität oder Korrektheit/Unkorrektheit von zahlreichen ladinischen Formen und Wendungen beantworten können.[20]

4. 3 Einsprachige Wörterbücher zum Ladinischen

Während in den letzten Jahren die zweisprachige ladinische Lexikographie große Fortschritte gemacht hat und mittlerweile für alle Idiome moderne Wörterbücher mit der Zielsprache Ladinisch vorliegen, war es den ladinischen Institutionen aus Zeit- und Personalmangel bisher nicht möglich, im Bereich der *einsprachigen* ladinischen Lexi-

[19] Vgl. http://vll.tall.smallcodes.org/applications/textanalysis/sitebolzano/index.jsp?_VP_V_ID=27392748

[20] Das ladinische Korpus kann auf folgende Vor- bzw. Parallelarbeiten zurückgreifen: das bereits erwähnte Korpus TALL; das CLE-Corpus (*Corpus Ladin dl'Eurac*, vgl. Streiter et al. [2004]); das *CATEx-Korpus* (*Computer Assisted Terminology Extraction,* vgl. www.eurac.edu/Press/Academia/14/Artikel2.asp), das *LexALP-Korpus* (vgl. http://217.199.4.152:8080/htdocs2/lexalp/corp_lexalp/search_corp.php) sowie das *Korpus Südtirol* (vgl. www.korpus-suedtirol.it/index_de).

kographie zu arbeiten. In Absprache mit den genannten ladinischen Kulturinstituten will deswegen die ladinische Abteilung dieses Arbeitsgebiet schwerpunktmäßig bearbeiten und ausbauen.

Ein erstes Projekt hat zum Ziel, ein *einsprachiges Wörterbuch des literarischen Ladinisch* zu erstellen. Unter „Ladinisch“ sind die Varietäten der brixnerisch-tirolerischen Ladinia intendiert, unter „literarisch“ jene Werke, die in hohem Maße ästhetisch-rhetorische Absichten verfolgen (wobei diese Grenze im Ladinischen naturgemäß tiefer anzusetzen ist als in den großen Literatursprachen). Abgedeckt werden soll ein repräsentativer Querschnitt der ladinischen Literatur, beginnend bei Angelo Trebo († 1888) bis herauf zu den modernen Autoren der Gegenwart. Lexikographisch soll nach dem Vorbild von gleich gearteten Wörterbüchern zum Deutschen und Italienischen vorgegangen werden, insbesondere aber nach dem Vorbild des Wörterbuchs zum literarischen Friaulischen von G. Faggin (vgl. Faggin [1985, 1989]).

Der Nutzen von einsprachigen Wörterbüchern ist unbestritten. Mit der Bearbeitung des literarischen Ladinischen wird zudem besonderes Augenmerk auf jene Sprachproduktion gelegt, die als besonders „vorbildhaft“ in Lexik, Semantik und Neologismenbildung gilt. Was Konzeption und Inhalt des Wörterbuchs betrifft, wird für das Ladinische absolutes Neuland betreten.

Das Wörterbuch soll als „Work in progress“ zuerst als Internet-taugliche Datenbank bereitgestellt werden, mit Abschluss der Arbeit soll auch eine Druckversion erscheinen. Die Größe des Wörterbuchs wird auf ca. 20.000 Lemmata veranschlagt.

4. 4 Dolomitenladinische Literaturgeschichte

Seit Januar 2009 läuft an der ladinischen Abteilung der Freien Universität Bozen das auf zwei Jahre angelegte Projekt „Ladinische Literaturgeschichte“. Es hat folgende Zielsetzungen: (a) die wichtigsten “literarischen” (wird im Zuge der Arbeit genauer definiert) ladinischen Texte aus den Dolomiten sollen erfasst werden; (b) es sollen originale Textbeispiele (im jeweiligen Idiom) mit Übersetzungen ins Deutsche und/oder Erläuterungen Platz finden; (c) es sollen weiters Autorenbiographien gegeben und Epochen-Einteilungen (Anfänge, Klassik, Moderne usw.) zur Diskussion gestellt werden; und schließlich (d) sollen Register der Autorennamen und der Titel der literarischen Texte erstellt werden.

Ansatzweise soll auch eine beschreibende Literatur-“Kritik” der Texte versucht werden. Die literarische Produktion in den Dolomiten soll nicht nur in ihren lokalen, sondern auch in den regionalen und europäischen Kontext gestellt werden.

4. 5 LINEE-Projekt

Die Ladinische Abteilung ist auch am EU-finanzierten Projekt LINEE (*Languages in an European Network of Excellence*[21]) beteiligt. Ziel des Workpackages W9a (*Language use and language values in minority school settings*), das in Zusammenarbeit mit der Universität Szeged (Ungarn) durchgeführt wird, ist es, die Einstellungen und Attitüden von ladinischen Schülern gegenüber den jeweiligen Unterrichtssprachen zu erheben und diese mit jenen der Schüler von ungarischen Minderheitenschulen in mehreren osteuropäischen Staaten zu vergleichen. Auch wird die Verwendung der unterschiedlichen Sprachen (Minderheitensprache, Mehrheitssprache, Fremdsprache; Muttersprache vs. Landessprache vs. internationaler Sprache etc.) im Fremdsprachenunterricht (insbesondere im Englischunterricht) in diesen Schulen umfassend analysiert. Es wird dabei die Sprachwahl und Sprachverwendung während des Englischunterrichts erfasst und es werden die amtlichen Vorgaben mit der schulischen Realität verglichen. Schließlich wird untersucht, durch welche externen Faktoren (sozialer/politischer Kontext, Ideologien, etc.) die Sprachenwahl beeinflusst wird und inwieweit Sprachattitüden ein Anzeichen für die Überlebensfähigkeit der jeweiligen Minderheitensprache darstellen können.

5. Ausblick

Die genannten Projekte sehen sich nicht als „isolierte“ und „elfenbeintürmene“ Vorhaben, sondern sind Teil einer strategischen Positionierung der ladinischen Abteilung als Forschungszentrum zur Ladinistik, das sowohl für die internationale Wissenschaftsgemeinschaft als auch für die lokale Bevölkerung grundlegende Instrumente zur Verfügung stellt, die einerseits einer noch besseren Erforschung der sprachlichen Gegebenheiten in den Dolomiten dienen, andererseits auch den Fortbestand der Sprache selbst erleichtern sollen.

[21] Vgl. www.linee.info/

Bibliographie

Annalas da la Societad Retorumantsch, 113, 2000. Chur: Societad Retorumantscha.

Ascoli, G. I. (1873). "Saggi Ladini", *Archivio Glottologico Italiano*, 1, 1–556.

Bibliografia Retoromantscha, Chur: Ligia Romontscha (1938), Vol. I, 1552–1930; Chur: Ligia Romontscha (1956), Vol. II, 1931–1952.

Chiocchetti, F. (2007). "È (ancora) possibile una politica linguistica nelle Valli Ladine?", *Mondo Ladino*, 31, 285–295.

Faggin, G. (1985). *Vocabolario della lingua friulana*. Udine: Del Bianco.

Faggin, G. (1989). "Aggiunte e correzioni al «Vocabolario della lingua friulana»", *Stud. Goriziani*, 70, 101–125.

Forni, M. (2001). *La ortografia dl Ladin de Gherdëina (cun i ponc dla ortografia che ie unic scemplifichei)*. San Martin de Tor: Istitut Cultural Ladin «Micurà de Rü».

Forni, M. (2002). *Wörterbuch Deutsch – Grödner Ladinisch. Vocabuler tudësch - ladin de Gherdëina*. San Martin de Tor: Istitut Cultural Ladin «Micurà de Rü».

Gartner, T. (1883). *Raetoromanische Grammatik*. Heilbronn: Henninger.

Goebl, H. (1997). „Der Neoladinitätsdiskurs in der Provinz Belluno", *Ladinia*, 21, 5–57.

Goebl, H. (2000–01). „Externe Sprachgeschichte des Rätoromanischen (Bündnerromanisch, Dolomitenladinisch, Friaulisch): ein Überblick"*, Ladinia*, 24–25, 199–249.

Gsell, O. (1990). „Die Kirchen und die romanischen Minderheiten von Graubünden bis Friaul" in Dahmen, W. / Holtus, G. / Kramer, J. / Metzeltin, M. (Hrsg.) (1990). *Die romanischen Sprachen und die Kirchen*. Romanistisches Kolloquium III. Tübinger Beiträge zur Linguistik, 343. Tübingen: Narr, 125–143.

Iliescu, M. / Siller-Runggaldier, H. (1985). *Rätoromanische Bibliographie*. Romanica Ænipontana, 13. Innsbruck: Institut für Romanistik.

Kattenbusch, D. (Hrsg.) (1999). *Studis romontschs.* Beiträge des Rätoromanischen Kolloquiums, Gießen / Rauischholzhausen, 21.-24.3.1996. Wilhelmsfeld: Egert.

Ladinia, 26-27, 2002-03. San Martin de Tor: Istitut Cultural Ladin «Micurà de Rü».

Mischì, G. (1994). „Der Weg des Ladinischen in den Stand der Amtssprache"*, Der Schlern*, 68, 337–341.

Mischì, G. (2000). *Wörterbuch Deutsch – Gadertalisch / Vocabolar Todësch - Ladin (Val Badia)*. San Martin de Tor: Istitut Cultural Ladin «Micurà de Rü».

Mischì, G. (2008). „Ein bisschen Ladinisch. Die universitäre Beheimatung einer Ur- und Randsprache" in Peterlini, H. K. (Hrsg.) (2008). *Universitas est. Essays zur Bildungsgeschichte in Tirol/Südtirol vom Mittelalter bis zur Freien Universität Bozen*. Bozen: Bolzano University Press, 574–585.

Richebuono, G. (1982). "La presa di coscienza dei Ladini"*, Ladinia*, 6, 95–154.

Santuari, A. (2006). "Esperienze sul territorio: il ladino in Val di Fassa", in *Il futuro si chiama CLIL. Una ricerca interregionale sull' insegnamento veicolare*. Trento: IPRASE Trentino, 219–226.

Schmid, H. (1998). *Wegleitung für den Aufbau einer gemeinsamen Schriftsprache der Dolomitenladiner*. San Martin de Tor: Istitut Cultural Ladin «Micurà de Rü»; Vich de Fascia: Istitut Cultural Ladin «Majon di Fascegn».

Siller-Runggaldier, H. / Videsott, P. (1998). *Rätoromanische Bibliographie 1985–1997*. Romanica Ænipontana, 17. Innsbruck: Institut für Romanistik.

Sociolinguistica. Internationales Jahrbuch für Europäische Soziolinguistik. Tübingen: Niemeyer.

Streiter, O. / Stuflesser, M. / Ties, I. (2004). "CLE, an aligned Tri-lingual Ladin-Italian-German Corpus. Corpus Design and Interface". Workshop on *First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*. Lisbon, May 24, 2004. Lisbon: LREC 2004.

Verra, R. (Hrsg.) (2000). *La minoranza ladina. Cultura, Lingua, Scuola*. Bolzano: Istituto Pedagogico Ladino; Intendenza per la Scuola delle Località Ladine.

Vicario, F. (Hrsg.) (2007). *Ladine loqui. 4° Colloquium retoromanistich,* San Daniele, 26–27 agosto 2005. Udine: Società Filologica Friulana.

Videsott, P. (1998). „Ladin Dolomitan. Die dolomitenladinischen Idiome auf dem Weg zu einer gemeinsamen Schriftsprache", *Der Schlern*, 72, 169–187.

Videsott, P. (2008). „Die mehrsprachige Ausbildung der ladinischen Primarschullehrerinnen an der Freien Universität Bozen" in Frings, M. / Vetter, E. (Hrsg.) (2008). *Mehrsprachigkeit als Schlüsselkompetenz: Theorie und Praxis in Lehr- und Lernkontexten*. Akten zur gleichnamigen Sektion des XXX. Deutschen Romanistentages an der Universität Wien, 22.–27. September 2007. Romanische Sprachen und ihre Didaktik, 17. Stuttgart: ibidem, 307–322.

Das Ladinische auf dem Weg eines zeitgemäßen Ausbaus

Giovanni Mischì

The main focus on the expansion of Ladin currently is in neologisms (vocabulary expansion) and language planning (standardisation and normalisation). The considerable differences between the current and past language stages of Ladin is due to the rapid development of all social, political and economic aspects of life. As a result of these changes, dictionaries and grammars were compiled and the most important resources became available in electronic format. In 1989, the development of Ladin was given a boost when it was declared an official language—never before had a legal provision so promoted the lexical development of Ladin.

The coining of new words and expressions (especially special language terms) must follow a clear and consistent plan and be carried out under the scientific direction and supervision of a recognised authority. For neologisms to be seamlessly integrated into a language system it is essential that committed speakers and writers diffuse them through all possible channels.

Thanks to the development of dictionaries and specialised glossaries of Ladin as well as the hiring of official translators, such efforts have reached a remarkably high standard of quality. These experts' proposals have been widely disseminated and are now consolidated.

On the basis of a lexical database, in this paper I will present our work on the dictionary. The present dictionary was compiled with a view towards practical needs and uses; it attempts to collect and describe in a clear and simple way the vocabulary of modern Val Badia Ladin.

Als die sprachliche Forschung im Bereich des Ladinischen mit G. I. Ascoli und Th. Gartner vor etwa 160 Jahren einsetzte, richtete sich das Interesse in erster Linie auf die einzelnen ladinischen Idiome und deren phonologische und morphologische Eigenheiten sowie auf die Untersuchung älterer ladinischer Texte. Die Beschäftigung mit Sprachausbau, Wortschatzerweiterung bzw. ganz allgemein mit Wortbildung im heutigen Sinne schien hingegen auf wenig Interesse zu stoßen, im Gegenteil, man war teilweise sogar der Ansicht, Neuprägungen seien der Sprache alles andere als zuträglich (Gartner 1883).

Wer heute in Ladinien damit beschäftigt ist, Wörterbücher für die eine oder andere Talvariante oder auch für das Einheitsladinische zu erarbeiten, hat freilich eine andere Sicht der Dinge in Bezug auf die Wortbildung. Das gilt auch für Übersetzer von Amtstexten und für Journalisten aber auch für all jene, die mündlich oder telefonisch befragt werden, wie man dies oder jenes auf Ladinisch ausdrücken oder benennen könnte. Aus diesem Grund ist es heute dringender denn je, den entstandenen Benennungsbedarf wettzumachen.

Im Ladinischen hat sich das wesentliche Hauptaugenmerk auf die Wortschöpfung (Wortschatzerweiterung) und auf die Sprachplanung (Standardisierung und Normierung) verlagert. In den letzen Jahren entstanden Grammatiken und Wörterbücher, wobei der Zugriff auf letztere längst auch über Computer und Internet möglich ist.

Bezüglich der bisher unternommenen Bemühungen, das Ladinische in den Stand einer zeitgemäßen Sprache zu heben, konnten bezeichnende Ergebnisse erzielt werden (Fachglossare, Grundwortschatz wurde ausgebaut, wichtige Alltagsdomänen wurden sprachlich erschlossen).

Die zunehmenden Ausdrucksbedürfnisse machen heute nicht nur im Ladinischen, sondern auch in den Sprachen mit einer weitaus längeren Tradition wie dem Deutschen, dem Italienischen, dem Englischen, dem Spanischen oder dem Französischen, kurzum in allen europäischen und übereuropäischen Sprachen einen Ausbau des Wortschatzes unabdingbar.

In der heutigen Zeit ist jede Sprache gezwungen, neue sprachliche Zeichen zu schaffen und ihre Ausdrucksmöglichkeiten zu erweitern. Der gravierende Unterschied gegenüber älteren Sprachstadien besteht insbesondere darin, dass wir heute einer rasanten Entwicklung in allen Bereichen des kulturellen, sozialen, politischen und wirtschaftlichen Lebens ausgesetzt sind.

Die letzten vierzig bis fünfzig Jahre haben eine Flut technischer und anderer Neuerungen hervorgebracht, für die das Ladinische freilich keine zutreffenden Ausdrücke besitzt. Es wäre aber abwegig zu glauben, nur wir Ladiner hätten in diesen Belangen lexikalische Lücken aufzuweisen. Ausdrücke wie *promoziun dl esport* (Exportförderung), *ofize de spediziun* (Abfertigungsstelle), *elemënt-fostü* (Spurenelement), *zertificat de zitadinanza* (Staatsbürgerschaftsbescheinigung), *sozieté finanziara* (Holdinggesellschaft), *plann provinzial de svilup* (Landesentwicklungsplan), *fić de nuzaziun* (Nutzungszinsen), *cunt di resć* (Rückständerechnung), *injunta integrativa speziala* (Sonderergänzungszulage), *defraldaziun fiscala* (Steuerhinterziehung), *indenité por l'alzada di* *prisc* (Teuerungszuschlag) und hunderte und aberhunderte weitere wird man in älteren Wörterbüchern vergebens suchen, aber genauso vergebens wird man nach Begriffen wie *Aufbereitungsanlage*, *Hightechindustrie* oder *Darmspiegelung* in vor lediglich etwa dreißig bis vierzig Jahren erschienenen deutschen Wörterbüchern Ausschau halten.

Eines steht daher allemal fest: Allein mit dem überlieferten traditionellen Wortschatz lässt sich heute keine Sprache erhalten, und am wenigsten wohl eine Minderheitensprache.

Angeregt durch meine tägliche Erfahrung als Übersetzer am Ladinischen Kulturinstitut kann ich daher jeden Versuch zur Prägung einer neuen ladinischen Wortschöpfung und jeden konkreten Vorschlag nur wärmstens empfehlen und dankbar begrüßen.

Es mag zwar stimmen, dass das Ladinische aufgrund der geringen Ausdehnung seines Sprachraumes nur über einen entsprechend begrenzten Kommunikationsradius verfügt. Dies beeinträchtigt meines Erachtens seine Entfaltungsmöglichkeiten aber nur indirekt und lediglich bis zu einem gewissen Grad: Es hat sich gezeigt, dass im Ladinischen neben der quantitativen Erweiterung und dem qualitativen Ausbau des Wortschatzes auch eine verstärkte sprachliche Differenzierung in vielen Fach- und Sachbereichen möglich ist. Das Ladinische hat bisher in allen wichtigen Domänen des Alltagslebens (Verwaltung, Schule, Kirche usw. – und auch in den abstrakten Bereichen[1]) einen beachtlichen Ausbau erfahren und verfügt damit über die nötigen Voraussetzungen zur Produktion unterschiedlicher Textsorten.[2]

Die Veränderung sprachlicher Verhältnisse – dessen muss man sich bewusst sein – ist vielfach auch als Folge der Veränderung geschichtlicher Verhältnisse zu sehen. Die Metaphern vom kontinuierlichen ‚Wachsen' und ‚Aufblühen' einer Sprache sind daher nicht immer angebracht. Es gibt Zeiten, in denen sich geschichtlich und gesellschaftlich mehr ereignet als in anderen. Dieses Mehr oder Weniger wirkt sich auf die verschiedenen Bereiche der Sprache aus und manifestiert sich entweder in der Gestalt von Veränderungen bestehender Strukturen und Formen oder aber in deren Beibehaltung. Einen besonderen Niederschlag findet dieser Wechsel im Wortschatz. Im Unterschied zu den relativ geschlossenen und stabilen Systemen der Grammatik gilt der Wortschatz im Allgemeinen als „offen" und „unstabil". Nirgends ist die Sprache so

[1] Vgl. z.B. ladinische Fachtexte im Bereich der Erziehung (Castlunger Ellecosta 2004).

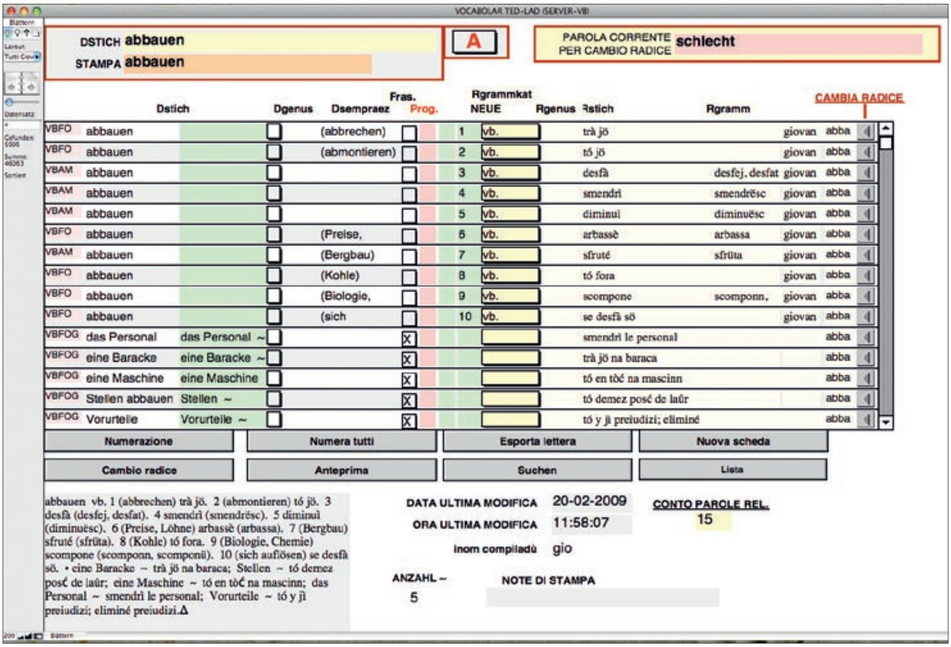
[2] Gemessen an seiner Schrifttradition ist das Ladinische eine sehr junge Sprache. Was seine alltägliche Verwendung betrifft, ist es eine anpassungsfähige und den beiden anderen Landessprachen (Deutsch und Italienisch) durch und durch ebenbürtige Sprache.

starken Veränderungen unterworfen wie in diesem Bereich. Dies gilt auch für das Ladinische. Am Ladinischen lässt sich dies sehr schön seit 1989 nachweisen, dem Jahr nämlich, in dem es zur offiziellen Amtssprache erhoben wurde. Nie zuvor war das Ladinische lexikalisch so stark herausgefordert wie nach Inkrafttreten dieser gesetzlichen Bestimmung[3].

Im Zusammenhang mit der schriftsprachlichen Aufwertung des Ladinischen in den letzten fünfzehn bis zwanzig Jahren zeichnet sich vor allem im Bereich des Wortschatzes mehr und mehr die Tendenz ab, verschiedene *Funktionalstile* aufzubauen und zu entwickeln (etwa den wissenschaftlichen und den umgangssprachlichen, den publizistischen und den künstlerisch-literarischen, den administrativen Stil usw.). Dies soll nicht als Ausdruck der Verwahrlosung und des Sprachverderbs oder als Beginn einer schleichenden Spracherosion betrachtet und gedeutet werden, wie gelegentlich kritisch bemerkt wird, sondern als Anbahnung eines Strukturwandels, einer Erscheinung, die in allen modernen Sprachen zu beobachten ist.

Wenn sich das Schriftladinische auf lexikalischer und phraseologischer Ebene von der gesprochenen Sprache etwas abhebt, so ist dies eine ganz „normale" Erscheinung. Man ziehe zum Vergleich z.B. einen deutschen oder italienischen Fach- bzw. Amtstext heran und wir werden auch hier Begriffen begegnen, die auf Anhieb nur schwerlich, wenn überhaupt zu verstehen sind – verfügt man nicht über ganz spezifische Fach- und Sachkenntnisse. Wer kann schon auf Anhieb eine erschöpfende Erklärung für Begriffe wie *Schuldwechselbuch/libro degli effetti passivi, Enteignungsanordnung/ordinanza di espropriazione, Ergebnisabführungsvertrag/contratto di cessione degli utili, Erlassantrag/domanda di remissione, Erlebensfallkapital/capitale pagabile in caso di vita, Konkursausfallgeldversicherung/assicurazione dei crediti di lavoro nell'insolvenza dell'impresa, Umschuldungskredit/credito per la conversione di debiti* oder *Steuerkosten/costo delle imposte* liefern?

[3] Der Rechtsanspruch der Ladiner auf die Verwendung der eigenen Muttersprache fand im Dekret des Präsidenten der Republik Nr. 574 vom 17.01.1989 *„Durchführungsbestimmungen zum Sonderstatut für die Region Trentino-Südtirol über den Gebrauch der deutschen und der ladinischen Sprache im Verkehr der Bürger mit der öffentlichen Verwaltung und in den Gerichtsverfahren"* ihren ersehnten Niederschlag. Der Art. 32, Absatz 2 und 3, enthält folgende Bestimmung: *„Die öffentlichen Verwaltungen in den ladinischen Ortschaften und jene Landesämter, die sich ausschließlich oder vorwiegend mit ladinischen Angelegenheiten zu befassen haben, sind verpflichtet, im mündlichen Verkehr auf Ladinisch zu antworten, im schriftlichen Verkehr hingegen auf Italienisch und Deutsch, mit anschließendem ladinischen Text."* Damit wurden erstmals auch die Grundfeste für den amtlichen Gebrauch des Ladinischen gestellt und verankert.



DSTICH abbauen
STAMPA abbauen

PAROLA CORRENTE PER CAMBIO RADICE schlecht

	Dstich		Dgenus	Dsempraez	Fras. Prog.	Rgrammkat NEUE		Rgenus	Rstich	Rgramm		
VBFO	abbauen		□	(abbrechen)	□	1	vb.		trà jö		giovan	abba
VBFO	abbauen		□	(abmontieren)	□	2	vb.		tó jö		giovan	abba
VBAM	abbauen		□		□	3	vb.		desfà	desfej, desfat	giovan	abba
VBAM	abbauen		□		□	4	vb.		smendrì	smendrësc	giovan	abba
VBAM	abbauen		□		□	5	vb.		diminuì	diminuësc	giovan	abba
VBFO	abbauen		□	(Preise,	□	6	vb.		arbassè	arbassa	giovan	abba
VBAM	abbauen		□	(Bergbau)	□	7	vb.		sfruté	sfrüta	giovan	abba
VBFO	abbauen		□	(Kohle)	□	8	vb.		tó fora		giovan	abba
VBFO	abbauen		□	(Biologie,	□	9	vb.		scompone	scomponn,	giovan	abba
VBFO	abbauen		□	(sich	□	10	vb.		se desfà sö		giovan	abba
VBFOG	das Personal	das Personal ~	□		☒				smendrì le personal			abba
VBFOG	eine Baracke	eine Baracke ~	□		☒				trà jö na baraca			abba
VBFOG	eine Maschine	eine Maschine	□		☒				tó en tòć na mascinn			abba
VBFOG	Stellen abbauen	Stellen ~	□		☒				tó demez posć de laûr			abba
VBFOG	Vorurteile	Vorurteile ~	□		☒				tó y jì preiudizi; eliminé			abba

abbauen vb. 1 (abbrechen) trà jö. 2 (abmontieren) tó jö. 3 desfà (desfej, desfat). 4 smendrì (smendrësc). 5 diminuì (diminuësc). 6 (Preise, Löhne) arbassè (arbassa). 7 (Bergbau) sfruté (sfrüta). 8 (Kohle) tó fora. 9 (Biologie, Chemie) scompone (scomponn, scomponü). 10 (sich auflösen) se desfà sö. • eine Baracke ~ trà jö na baraca; Stellen ~ tó demez posć de laûr; eine Maschine ~ tó en tòć na mascinn; das Personal ~ smendrì le personal; Vorurteile ~ tó y jì preiudizi; eliminé preiudizi.Δ

DATA ULTIMA MODIFICA 20-02-2009
ORA ULTIMA MODIFICA 11:58:07
Inom compiladù gio

CONTO PAROLE REL.
15

ANZAHL ~
5

NOTE DI STAMPA

Abbildung 1: Die elektronische Datenbank (Gadertalisch-Deutsch)

Der zunehmende Benennungsbedarf fordert nicht nur dem Ladinischen, sondern auch hoch entwickelten Nationalsprachen, ein hohes Maß an Flexibilität im Bereich des Wortschatzes ab. Ein wesentlicher Unterschied muss allerdings darin gesichtet werden, dass eine Kleinsprache wie das Ladinische nicht in der Lage ist, mit derselben Schnelligkeit wie Großsprachen auf äußere Veränderungen zu reagieren und mit den Veränderungen in der Gesellschaft Schritt zu halten. Das gilt für den Wortschatz in ganz besonderem Maße, weil dieser der sensibelste und äußeren Einflüssen am stärksten ausgesetzte Bereich der Sprache ist. Es wäre aber abwegig zu glauben, unsere Sprache besitze ein starres, unflexibles Lexikon und Wortbildungssystem. Einwände dieser Art gründen auf Vorurteilen, die sehr gerne gegenüber so genannten „niedrigeren“ Varietäten vorgebracht werden.

Die Ansicht, dass das Ladinische keine angemessenen Ausdrücke bereitzustellen vermag, darf uns nicht der Pflicht entheben, die wachsende Flut von Fremdwörtern einzudämmen, die unsere Sprache immer mehr durchsetzen und entstellen, sondern muss uns ermutigen die eigene Assimilation voranzutreiben. Wie alle übrigen Sprachgemeinschaften sind auch wir Ladiner dazu aufgefordert, ständig neue Wörter und neue Ausdrücke zu schaffen, damit unsere Sprache den Anforderungen von heute und morgen zu genügen vermag.

Gegenüber früher geht heute die Schaffung von Neubildungen wesentlich systematischer und koordinierter vor sich. Ladinische Neuprägungen sind keine Willkürprodukte der Phantasie, sondern sie entstehen auf der Basis bereits vorhandenen Sprachmaterials und bestehender/vorhandener Wortbildungsmuster – oder Wörter werden ganz einfach entlehnt und an das Ladinische angepasst. Hinter jeder Neubildung steckt daher etwas bereits Bekanntes und Vertrautes, weswegen man wohl vergeblich nach genuin „Neuartigem" Ausschau hält. Dass dabei bei bestimmten Neologismen, so etwa bei *avalianza de valüta* (Gleichwertigkeit), *inabilité al laûr* (Berufsunfähigkeit), *mosöra aministrativa* (Verwaltungsmaßnahme), *legislaziun fiscala* (Steuergesetzgebung), *arest preventif* (Untersuchungshaft), *costituzionalité* (Verfassungsmäßigkeit) der Eindruck einer gewissen Verwissenschaftlichung bzw. Terminologisierung entstehen kann, ist nicht ganz von der Hand zu weisen, liegt aber in der Natur der Sache; solche Produkte sollten deshalb nicht gleich als „künstlich" oder als „störend" empfunden und verworfen werden.

Mit der Schaffung von Wörterbüchern und Fachglossaren und durch die Anstellung offizieller Übersetzer haben in Ladinien die diesbezüglichen Bemühungen ein qualitativ bemerkenswertes Niveau erreicht. Zudem tun sich auch hier durch den Einsatz moderner Informationstechnologien allmählich neue Wege und Methoden auf: Wenn bis vor wenigen Jahren das Erlernen des Ladinischen praktisch nur über Bücher möglich war, bieten inzwischen Computeranwendungen und das Internet neue Möglichkeiten. So stellt z.B. auch das Istitut Cultural Ladin „Micurà de Rü" seit einigen Jahren ein Deutsch-Ladinisches Online-Wörterbuch[4] und eine erste (in vielen Punkten noch ergänzungs- und verbesserungswürdige) elektronische Rechtschreibhilfe[5] kostenlos zur Verfügung. Durch diese computerlinguistischen Möglichkeiten wird das Ladinische als Klein- und Randsprache um ein Vielfaches attraktiver und benutzerfreundlicher.

[4] http://www.micura.it/deu/695.html (Wörterbuch online). Die bezüglichen elektronischen Wörterbücher versuchen in verständlicher und allgemein zugänglicher Form den Wortschatz des heutigen Gadertaler-Ladinischen zu erfassen und darzustellen. Die deutsch-gadertalische Datenbank (vgl. Bild 1) umfasst 46.000 deutsche Eintragungen mit über 106.000 ladinischen Entsprechungen, die deutsch-grödnische 21.000 mit 67.000 Entsprechungen. Darin wird nicht nur der streng traditionelle Wortschatz berücksichtigt, sondern auch Wortgut, das über die reine Alltagssprache hinausgeht. Einen besonders breiten Raum nehmen Neuschöpfungen und Fachtermini aus den Bereichen *Verwaltung, Rechtswesen, Technik, Botanik, Zoologie, Medizin, Musik, Sport* u.a. ein. Diese beiden Instrumentarien sollen all jenen wertvolle Dienste erweisen, die sich in einem gehobenen und lexikalisch differenzierten Ladinisch ausdrücken wollen und müssen.

[5] http://www.micura.it/deu/695.html (Ladinische Korrekturhilfe), http://scl.ladinternet.smallcodes.org/applications/cort/site/index.jsp

Bibliographie

Castlunger Ellecosta, R. (2004). *L'educaziun é n ri mistier: racoiüda de consëis söl'educaziun por geniturs*. Uniun Ladins Val Badia, San Martin de Tor.

Forni, M. (2003). *Wörterbuch Deutsch–Grödner-Ladinisch/Vocabuler Tudësch – Ladin de Gherdëina*. St. Martin in Thurn: Istitut Cultural Ladin «Micurà de Rü»; CD-ROM: ibid. 2003.

Gartner, T. (1883). *Raetoromanische Grammatik.* Heilbronn: Vlg. Gebr. Henninger.

Mischì, G. (2000). *Wörterbuch Deutsch–Gadertalisch/Vocabolar Todësch – Ladin (Val Badia)*. San Martin de Tor: Istitut Cultural Ladin «Micurà de Rü»; CD-ROM: ibid. 2001.

Mischì, G. (1994). „Der Weg des Ladinischen in den Stand der Amtssprache", *Der Schlern*, 68, 337–341.

Pizzinini, A. (1966). *Parores ladines. Vokabulare badiot – tudësk*. Ergänzt und überarbeitet von Guntram Plangg. Innsbruck: Institut für Romanische Philologie der Leopold-Franzen-Universität.

Siller-Runggaldier, H. (2000). „Das Grödnerische: Sprache zwischen Idiom, Talvariante und dem Projekt Ladin Dolomitan" in Tagungsband *Das Werden einer Talschaft*. San Martin de Tor: Istitut Ladin «Micurà de Rü».

Siller-Runggaldier, H. (1994). „Probleme des Ladinischen heute am Beispiel der Wortschatzerweiterung" in F. Lanthaler (Hrsg.) (1994). *Dialekt und Mehrsprachigkeit/Dialetto e plurilinguismo*. Beiträge eines internationalen Symposiums/Atti di un simposio internazionle, Bozen/Bolzano, 1993. Meran: Alpha & Beta, 137–146.

SPELL: Dizionar dl ladin standard (2002). Urtijëi et al.

Videsott, P./Plangg, G. A. (1998). *Ennebergisches Wörterbuch – Vocabolar Mareo*. Innsbruck: Wagner. von Wartburg.

The Development and Acceptance of Electronic Resources for Welsh

Delyth Prys

It is often stated that development and use of electronic language resources are vital to the continued survival and revitalisation of minority languages. However, few studies have been conducted to measure the effect of such resources on individual languages in real-life settings. We postulate that the provision of electronic language resources has two effects on a minority language community. First is the practical effect of providing accessible language tools to write and use the language, thus increasing productivity and improving confidence in the use of the language. The second effect is harder to measure: changing the image of the language to one that is more contemporary and relevant to the twenty-first century, making it more attractive to the younger generation and acceptable as a vehicle for social interaction and use.

Many electronic resources have been developed for Welsh during recent years, including bilingual dictionaries, spelling and grammar checkers, speech and text corpora, educational language games, and text-to-speech synthesis. Taking case studies of the use of these new language tools, we will attempt to quantify their effects on the use of Welsh in recent years. We will ask and attempt to answer whether the provision of these electronic tools and resources has any impact on the perceived 'coolness' of the language and the desire of young people to use Welsh and transfer it to their children in due course.

1. Background

Welsh was a language without official status for much of the twentieth century. During this time, successive census figures revealed a decline in the number and percentage of Welsh speakers in Wales: from a peak of just under a million speakers in 1911 (43.5 % of the population of Wales), by 1991 the number of speakers had declined to 18.6 % of the population. However, by this time, there was also a widespread desire in the Wales population to ensure the survival of Welsh as well as to see the language revitalised as the national language of Wales. Ten years later, the returns for the 2001 census showed that the decline in the percentage of Welsh speakers had been halted, and even reversed, with the percentage of Welsh speakers over the age of 3 up to 20.5 % of the population. The trends for Welsh speakers between the ages of 3 and 15 were even more encouraging, promising further growth in the percentage of Welsh speakers in years to come. The total number of the population living in Wales in 2001 reporting that they were able to speak Welsh was 582,400 (*2001 Census.* 2003).

This reversal of fortunes in the history of the Welsh language has been attributed to a mixture of many different factors, including the growth in Welsh-medium education, Welsh-language media, improved legal status for the language, and the establishment in 1998 of the Welsh Assembly Government (Stevens 2009). In 2003, the Assembly published a far-reaching policy document, *Iaith Pawb: Everyone's Language* (2003), setting out a national action plan for a bilingual Wales. This document contained the explicit target of increasing the numbers of Welsh speakers by 5 percentage points by 2011, from those published from the 2001 census.

This ambitious target was to be reached through a number of means, but the one of interest to those involved in the creation of electronic resources for Welsh and other lesser-used languages was 'Language Tools' (section 4.44 of *Iaith Pawb*), which included terminology standardisation, a national database of terms, the development of lexicographical and machine translation aids for translators, and the creation of "an *ICT strategy* for increasing and facilitating the use of Welsh on the Internet and in IT packages to augment the existing list of computer resources available in Welsh."

This was in line with one of David Crystal's seven postulates for language revitalisation, which states that, "an endangered language will progress if its speakers can make use of electronic technologies" (Crystal 2000). This may sound like common sense, as there are practical considerations of language support when using computers and electronic media (for example, are there spelling and grammar checkers, etc. available for the language in question?), as well as issues of language status (if a language doesn't have electronic resources it must be a language which has no use and no status). However, measuring the effect of the availability of electronic resources on the use and status of a language is far more difficult, and very little work has been undertaken on this aspect of language revitalisation.

2. Language Technologies at Bangor University

The Language Technologies Unit at Canolfan Bedwyr, Bangor University, is the foremost developer of electronic resources for the Welsh language.[1] It has recently been trying to quantify the use made of the electronic resources it has developed for Welsh, in order to begin to measure their effectiveness in language revitalisation.

As the Unit is self-funded, it relies on a mixture of direct commissions, winning competitive tenders, research grants, and software sales to finance its activities. Due to this, the dissemination of electronic resources the Unit develops is often dictated by the requirements of the commissioning body or customer. Some may be for internal use only, and as such are not intended for wider distribution, but other resources are created with the intention of disseminating them as widely as possible, often in order to provide technical support for the electronic use of Welsh.

Important electronic resources created by the Unit in the period 2004–2008 include *Enwau Cymru*[2] (a database of Welsh place names), the *Welsh National Database of Terms*[3], *Y Termiadur: Standardized Terminology*[4] (a bilingual Welsh/English terminology dictionary) (cf. Prys et al. 2006), *WISPR*[5] (synthetic voices for text to speech applications), *Cysgliad 2004*[6] (a compendium of spelling and grammar checker), and a collection of bilingual Welsh/English electronic dictionaries. Logs of activity were available for some versions of these resources, and were therefore used to give some indication as to their popularity. In some instances, independent surveys had also reported on their use and were able to give additional feedback on the public's perception of their usefulness and user-friendliness.

2. 1 Enwau Cymru

Enwau Cymru was first published online as a searchable database of place names in Wales in 2005. Its aim is to provide guidance on the correct form of place names to be used in Welsh and in English. Targeted primarily at translators and public administrators who need to deal with both languages, the project was partly funded

[1] http://www.bangor.ac.uk/ar/cb/technolegau_iaith.php.en

[2] http://www.e-gymraeg.co.uk/enwaucymru/

[3] http://www.e-gymraeg.co.uk/bwrdd-yr-iaith/termau/Default.aspx

[4] http://geiriadur.bangor.ac.uk/termiadur

[5] http://www.e-gymraeg.org/wispr/index_en.htm

[6] http://www.e-gymraeg.org/cysgliad/

by a grant from the Welsh Language Board. The website may be accessed free of charge, and has been updated and expanded with links to Ordnance Survey, Streetmap.co.uk and Google Maps (and in some instances with sound files). It also has an email enquiry service, which provides answers on place names not included in the database. Emails to the service have shown great appreciation of the database, but comprehensive research into the number and nature of enquiries has not yet been conducted. Figures are however available for the number of searches of the database itself, and these reveal that there were 8500 searches in the six months between March and the end of November 2008. Allowing for the fact that an unknown proportion of these could have been from interested members of the public (rather than translation or administration professionals), this search activity shows a significant level of use for a specialist minority language database service, aimed at a small, specific sector.

2. 2 Welsh National Database of Terms

The *Welsh National Database of Terms* was a direct commission from the Welsh Language Board, launched in 2006. As well as allowing for online searches of a number of Welsh/English specialist dictionaries, one of its innovative features was that is also allows for downloads of these dictionaries in a format which is compatible for use by different translation memory systems such as Trados, Déjà Vu and Wordfast (Jones & Prys 2005). Again, this was targeted mainly at professional translators and bilingual administrators, rather than the general public, but the logs show a total of 35,611 searches in the six months between March and the end of November 2008. It was not possible to count the number of downloads separately; however, it can be assumed that those who download the dictionaries do not need to revisit the website for online searches. It should be remembered, therefore, that not all the website visits are for online searches, but also include one-off downloads. It was possible to count the number of visits to the website since the date it went live on March 23, 2006. The total number of visits to the end of November 2008 was 142,966, taking into account both online searches and downloads. Again this is a significant number of searches for a language community of little more than half a million.

2. 3 Termiadur: Standardized Terminology

This bilingual dictionary of standardised terminology contains specialist terms for all the main schools subjects covered by the educational system for 5 to 19-year-olds in Wales. It is intended to help support the development of teaching through the medium of Welsh in primary and secondary schools throughout Wales, which is one of the main pillars in the efforts to revitalise the Welsh language. The dictionary takes its origins from an original commission given by the UK Schools Curriculum Authority in 1993, which resulted in the publication of *Y Termiadur Ysgol: Standardized Terminology for the Schools of Wales* in 1998 (Prys & Jones 1998). This was published in book form, to be joined later by an electronic version, sold separately on CD. When the paper version was sold out, rather than reprint the original version, a new, enlarged and updated version was commissioned by ACCAC, the Qualifications, Curriculum and Assessment Authority for Wales (which has since become part of the Department for Education, Lifelong Learning and Skills of the Welsh Assembly Government). Realising its general value for translation professionals and bilingual administrators as well as the general public, the words *Ysgol* (School) and *for the Schools of Wales* were dropped from the title of the revised publication.

The new *Y Termiadur: Standardized Terminology* was published in 2006. This time the CD version was included at no extra charge with the book version. In addition, a searchable online version was also created, and a downloadable mobile phone version was also made available.[7] The combined book and CD version retail for a price of £15 sterling, but free copies have also been distributed to every publicly funded school in Wales (both Welsh and English medium schools), and there is no restriction on the free copying of the CD version. The online and mobile phone version are available free of charge in order to promote their widespread use. The logs show a total of 53,843 Web searches in the six months between March and the end of November 2008, and during the same period there were 200 downloads of the mobile phone version. These are considered to demonstrate substantial use of the dictionary, especially since the paper and CD versions are also in widespread circulation. This use may reflect the fact that this dictionary is targeted at schoolchildren, their parents, and the general public, as well as professional translators and bilingual administrators. It is likely also that familiarity with the product in one format actually encourages users to seek it out in other formats. Although the number of mobile phone versions during the six-month

[7] http://geiriadur.bangor.ac.uk/termiadur/

period analysed is only in the low hundreds, anecdotal evidence suggests that this format is popular with schoolchildren who are heavy users of mobile devices, and who prefer it to carrying a heavy paper dictionary around with them in their school bags.

2. 4 WISPR: Synthetic Welsh Voice

This is a free Welsh language synthetic voice for text-to-speech applications created during the Welsh and Irish Speech Processing Interreg-funded research project of 2004–2006. It is a basic quality diphone voice, working in a Windows environment. It is intended for use by visually-impaired people, to allow them access to emails, electronic documents, Web pages and similar texts written in Welsh. This application is different from those listed above, in that it is not provided in a form which may be used 'out-of-the-box' by ordinary users who are not technically competent. Rather, it is intended for use by software developers who can then integrate it into their software. Quoting from the Language Technologies Unit website: "The WISPR team members believe that the best way to disseminate speech technology in a minority language environment is to provide freely distributable tools and applications that are easy for the end user to use and liberally licensed to permit developers to integrate into their own software"[8]. A commercial version of the voice has also been licensed, which addresses customer requests for additional support and development. The commercial version may be heard on the websites of many public bodies in Wales, including those of the Countryside Commission for Wales and the North Wales Police Authority. The free version had been downloaded 720 times by the end of November 2008, again a significant number of downloads considering that it is only usable by technically competent developers.

2. 5 Cysgliad

The *Cysgliad* compendium for PCs contains two main programmes: a Welsh spelling and grammar checker called *Cysill*, and a collection of bilingual Welsh/English dictionaries called *Cysgeir*. Older versions of *Cysill* and *Cysgeir* were previously available separately, but in 2004 an updated and combined version was launched as a CD under the *Cysgliad* name. It is available as a stand-alone home application and also as a networked version for education and business. This is a commercial product that

[8] http://www.bangor.ac.uk/ar/cb/wispr.php.en

comes with software support and free updates. The standard price of a single licence is £55, with attractive discounts for multiple licences. The spellchecker part of the package is also available within Microsoft Word, and a free version of it may be obtained with OpenOffice. A free version for Mac users was sponsored by the Welsh Language Board and is downloadable from the Web.[9] Some of the dictionaries included in the *Cysgeir* compendium are also available in other formats, either as paper dictionaries or as free online ones. However, despite the cost of the *Cysgliad* compendium, and the fact that some of its components are also available in other formats (many of them free of charge), this is by far the most widely-used and appreciated of the electronic resources produced by the Language Technologies Unit at Bangor University. It is estimated that around 10,000 licences have been sold so far, and between March and the end of November 2008, 3300 Web updates were downloaded by home users.

In the case of *Cysgliad*, it is also possible to obtain independent corroborating evidence from other surveys on the use and popularity of the software. In 2007, the Welsh Language Board published a *Survey of Promoting Technology in Welsh* (Jones 2007), where it asked a wide range of institutions, schools, local companies and individuals in parts of Anglesey (a predominantly Welsh-speaking area in North West Wales) about their use of different office software. Of the respondents, 22 % reported that they used the *Cysgliad* compendium, a highly significant percentage, as none of the respondents used the Welsh language interface for Microsoft software on their computers. A follow-up project in 2008 revisited the 2007 report (Jones & Hughes 2008), and—in addition to reporting a 25 % increase in those who had installed the Welsh Interface / Office Software on their computers at work and a 24 % increase at home after seeing the presentation and receiving information the previous year—there was a specific reference to the usefulness of *Cysgliad*: "'Cysgliad' is great—It should be on every computer in schools in Wales." (Jones & Hughes 2008: 11).

Other comments recorded during this survey cast further light on the relationship between resources such as *Cysgliad* as practical language tools and the broader question of language status and, indirectly, of language preservation and revitalisation. One comment in particular made the connection between Welsh-medium Information Technology and the status of the language, saying: "More use should be made of IT through the medium of Welsh to give the language a higher status." Another comment referred to the way Welsh in a computing environment seemed more meaningful to a

[9] http://e-gymraeg.org/cysgliadmac/

Welsh-speaker, even though the Welsh-speaker would also have been fluent in English: "I enjoy seeing everything in Welsh—it doesn't feel as threatening as English. The Welsh has more meaning." (Jones & Hughes 2008: 11).

Another similar survey in Gwynedd and Conwy, two other predominantly Welsh-speaking areas in North West Wales, also gave some feed-back on Welsh language software and tools, after local businesses received help in installing it on their computers:

> The overall response by the majority of business was very positive with those that agreed to the installation extremely pleased with the software, amazed that will enable them to use Welsh on their computers, increase their confidence in using the language and improve the standard and quality of their work. It is also encouraging that so many of those contacted welcomed and appreciated the opportunity to have a Welsh language version on their computer and that there is so much general interest in improving access to and in using technology through the medium of Welsh. (Deudraeth 2008: 11)

3. Conclusions

This brief overview on the use of electronic resources produced by the Language Technologies Unit, Bangor University, to support Welsh in computing, online and multimodal environments confirms that these tools are welcomed and used by their intended targets. Quantifying the number of searches, downloads and sales of such products affords a quick and easy measurement of the use of electronic resources and technology in a minority language. Measuring their impact on attitudes towards the language is harder to achieve, and requires further surveys on the perceptions and attitudes of the target populations. However, if present trends towards the revitalisation of the Welsh language continue (with a further percentage gain recorded in the next census survey due to take place in 2011), developers of electronic technology for Welsh will be able to claim a small part of the credit, not only for the production of useful resources, but also for helping create the image of a contemporary, vibrant language, which provides its community with a positive and attractive self-image.

References

2001 Census - Main Statistics about Welsh (2003). Welsh Language Board. Retrieved May 27, 2009, from http://www.byig-wlb.org.uk

Crystal, D. (2000). *Language Death.* Cambridge University Press.

Deudraeth Cyf. (2008). *Report on the Project to Promote Welsh Technology and the Welsh Language Control Centre to Businesses and Organizations in Gwynedd and Conwy.* Welsh Language Board. Retrieved May 27, 2009, from http://www.byig-wlb.org.uk

Iaith Pawb: A National Bilingual Plan for a Bilingual Wales (2003). Welsh Assembly Government. Retrieved May 27, 2009, from http://wales.gov.uk/depc/publications/welshlanguage/iaithpawb/iaithpawbe.pdf?lang=en

Jones, D. B. / Prys, D. (2005). "The Welsh National Online Terminology Database" in *Proceedings of the Lesser Used Languages and Computer Linguistics Conference* (LULCL), October, 27–28, 2005. Bozen-Bolzano, Italy. 149–169.

Jones, E. W. (2007). *Survey of Promoting Technology in Welsh*: *Experimental Project—Anglesey.* Welsh Language Board. Retrieved May 27, 2009, from http://www.byig-wlb.org.uk

Jones, E. W. / Hughes, E. (2008). *Revisiting the 2007 Promoting Technology Report—Anglesey*. Welsh Language Board. Retrieved May 27, 2009, from http://www.byig-wlb.org.uk

Prys, D. / Davies, O. / Jones, J. M. J. / Prys, G. (2006). *Y Termiadur: standardized terminology*. Cardiff: ACCAC.

Prys, D. / Jones, J. P. M. (1998). *Y Termiadur Ysgol: Standardized Terminology for the Schools of Wales*. Cardiff: ACCAC

Stevens, C. (2009). *Telling the Story of Welsh.* Llandysul: Gomer Press.

African Varieties of Portuguese: Corpus Constitution and Lexical Analysis

Maria Fernanda Bacelar do Nascimento, Antónia Estrela,
Amália Mendes, Luísa Pereira and Rita Veloso

This paper presents the results of our recent experiences in establishing fundamental linguistic resources for contrastive linguistic analyses of the African varieties of Portuguese (AVP), namely Angola, Cape Verde, Guinea-Bissau, Mozambique, and Sao Tome and Principe. We discuss the difficulties involved in the compilation of corpora for each variety, as well as their annotation with PoS information and lemmatisation. Five contrastive lexicons have been extracted in order to establish a core and peripheral vocabulary for each variety, as well as to study AVP specific morphological processes. Some preliminary results on the syntactic properties of AVP are also discussed.

1. Introduction

Portuguese is the official language of five African countries—Cape Verde, Guinea-Bissau, S. Tome and Principe, Angola and Mozambique—although it is mainly spoken as a second language. In Cape Verde, Guinea-Bissau and S. Tome and Principe, Creole languages emerged and are widely used (which accounts for the fact that Portuguese is spoken by a minority), while in Angola and Mozambique, where there is a large diversity of Bantu languages and no Creole, Portuguese has come to establish itself as an agent of national unity. These African varieties of Portuguese show properties that differ from European Portuguese, although these are not systematic among the speakers and are still features of emerging grammars.

The shortage of studies on African varieties of Portuguese, especially when compared with the quantity of empirical studies on European and Brazilian Portuguese developed from corpora or lexicons, can be seen as a result of the lack of similar resources for these more recent varieties. The case of Mozambique is, however, an exception, since a spoken corpus has been compiled and several studies have been undertaken and published (Stroud & Gonçalves 1997a, 1997b; Gonçalves & Stroud 1998, 2000, 2002; Duarte et al. 1999; Gonçalves 1996, 1997).

To address this issue, five corpora for these African varieties of Portuguese (AVP) have recently been compiled (Bacelar do Nascimento et al. 2006, 2008a, 2008b), constituting the Africa Corpus, and have enabled initial contrastive studies on the lexicon and on the syntax, pointing to specific linguistic aspects where AVP differ from the European Portuguese (EP) norm.

2. Corpora of African Varieties of Portuguese

The Africa Corpus has around 640,000 words for each variety, divided by the same percentage for spoken and written texts (c. 25,000 spoken words [4 %] and c. 615,000 written words). The internal constitution of the corpus is further described in Table 1.

Internal Constitution				
Spoken corpus	Informal discourse	45 men	80 % high educated 20 % medium educated	4 %
		35 women		
Written corpus		Literary book		19 %
		Newspaper		52 %
		Miscellaneous		25 %

Table 1: Internal constitution of the Africa corpus

The reduced dimension of the spoken corpus in comparison to the written one is due to the high costs related to the collection and recording of the materials and their subsequent transcription. The proportion of spoken versus written presents, nevertheless, a high consistency in the following: on the one hand, the massive use of Portuguese in African countries occurred only after their independence, in the second half of the 20th century, and this recent generalisation results in a great instability of the spoken language; on the other hand, there are cases where the Portuguese variety is used mostly in formal and institutional situations, while the Creole languages are commonly used (in Cape Verde, Guinea-Bissau and S. Tome and Principe). This accounts for the importance of a written corpus to verify the strength and stability of the phenomena occurring in these varieties.

The five corpora are thus comparable in size, chronology, and broad types and genres. However, due to time constraints and the difficulty of finding and compiling adequate written materials (given, for instance, that the texts considered must be written by native people still living in those countries), it was not possible to ensure comparability at a more granular level. A follow-up of this work is under way and will assure broader coverage and a more fine-grained comparability of the five corpora.

For the task of corpus constitution, some samples of written and spoken materials of already existing corpora compiled at the Linguistics Center of Lisbon University (CLUL) during the last 30 years were reused, as well as new recordings specifically made for this project and new collections of texts. In terms of the spoken corpus, we could benefit from some materials from the previous project *Português Falado: Variedades Geográficas e Sociais* ('Spoken Portuguese: Geographic and Social Varieties')[1] and from materials of the Mozambican project *Panorama do Português Oral de Maputo* ('Overview of the Spoken Portuguese from Maputo', [Stroud & Gonçalves 1997a, 1997b])[2]. Nevertheless, new recordings were made, some by researchers and teachers resident in the five African countries, and some by our own team.

The newspapers selected are publications with wide national coverage. With regard to literature, poetry was avoided, as well as authors with strong stylistic marks, and only native authors or authors who had lived all their lives in the countries were selected. The category 'Miscellaneous' comprises texts from very different subtypes and genres that are difficult to place under specific categories, and that correspond, in fact, to a large collection of heterogeneous texts of different kinds, such as literary or social magazines, computer policies, official documents, religious discourse, political interventions, tourism information, university webpages, academic works, law, national constitution, army information and some short poetic texts. This broad category came to represent a large percentage of the written corpus. The authors of these types of texts were also born, and were residents of, the African country in study.

The spoken corpus includes recordings (dialogues and conversations) of spontaneous language on widely diversified topics, as well as recordings from TV and radio programs. According to our objectives, the transcription of these recordings follows

[1] The results of this project consist in 4 CD-ROMs with a corpus of 86 recordings (around 9h of speech), covering Portugal (30 texts), Brazil (20), Angola (5), Mozambique (5), Guinea-Bissau (5), Cape Verde (5), S. Tome and Principe (5), Macau (5), East Timor (3) and Goa (3). The CDs contain the transcriptions of the recordings, the sound files, and software that allows the users to see the text and sound alignment. It was published in 2001 by Instituto Camões and Centro de Linguística da Universidade de Lisboa. The different sources of these materials are identified in the CDs (Bacelar do Nascimento 2001; Bettencourt Gonçalves & Veloso 2000).

[2] The spoken corpus of the Mozambican variety also includes texts compiled by Perpétua Gonçalves (Gonçalves 1990).

orthographic conventions, including punctuation signs, which received the same value they have in writing; nevertheless, special importance was given to their prosodic function, to assure the inclusion of some information—even if only rudimentary—on the spoken language rhythm.

New words that were formed under regular patterns of derivation posed no problems for transcription (Rio-Torto 2007), and other cases were transcribed as closely as possible to their pronunciation. The transcription respects the Portuguese orthography (which follows etymological parameters), and was, in some cases, confirmed with native speakers. When foreign words were pronounced closely to the original pronunciation, they were transcribed in the original language orthography. When they were adapted to the Portuguese pronunciation, they were transcribed according to the entries of reference dictionaries or according to the orthography adopted in those dictionaries for similar cases. Mispronunciations were registered only if the speaker immediately corrected it, and in this case both spellings have been transcribed. But if the speaker misspelled a word and pursued his speech without any correction, only the standard spelling of the word was kept in the transcription. Paralinguistic forms and onomatopoeia not registered in the reference dictionaries were transcribed as closely as possible to the sound produced.

Concerning the dialogue representation, when the interviewer was not a speaker of the variety in study, his/her dialogical turns were excluded from the files intended for extractions for lexicon, morphosyntactic and syntactic analysis. (A complete file was kept, however, for further exploitation.) Overlapping between speakers was not encoded.

We believe that the disparity in language resources between European or Brazilian Portuguese varieties and African varieties has been partially reduced with the compilation of these comparable corpora, which are the first step towards the development of linguistic studies of the Portuguese varieties of the African countries where Portuguese is the official language and is taught as a second language (Bacelar do Nascimento et al. 2006, 2008a, 2008b). It will also allow the extension of empirical studies on lexical and grammatical properties to African varieties other than the Mozambican one (a variety which has undergone several studies based on corpora in the last 20 years [cf. Stroud & Gonçalves 1997a, 1997b; Gonçalves & Stroud 1998, 2000, 2002]).

In order to perform these studies, the Africa Corpus has been automatically annotated with Parts-of-Speech information using a tagger previously trained over a written and spoken Portuguese corpus of 250,000 words, morphosyntactically annotated and manually revised (Eric Brill's tagger [Brill 1993]). The corpus was further lemmatised.

The initial tagset for the morphosyntactic annotation of the corpus covered the main PoS categories (noun, verb, adjective, etc.) and secondary ones (mood and tense, conjunction type, proper and common noun, variable *vs.* invariable pronouns, auxiliary *vs.* main verbs, etc.). Person, gender and number categories were not included.

3. Contrastive Lexicons of AVP

The first studies undertaken based on the five comparable corpora are centred on the contrastive properties of each variety's lexicon: contrastive lexicons of the main PoS categories, nuclear versus peripheral vocabulary, and divergent derivational processes.

Five lexicons had been extracted from the corpora, one per each variety, comprising lexical items from the main categories of Common Name, Adjective and Verb, as well as a category for Foreign Words. For each lexical item, the following information is given: PoS, lemma and index of frequency of occurrence in the corpus. A total number of 25,523 lemmas have been described: 14,666 (57 %) nouns, 6,268 (25 %) adjectives, 4,292 (17 %) verbs and 297 (1 %) foreign words.

The lexicons of the different varieties have been compared and treated statistically, in the form of contrastive lists, with data of frequency and distribution, and are also available at CLUL's webpage for online query.

One of the most important aspects of the contrastive studies on corpora of varieties of a given language, especially languages such as Portuguese, English, Spanish or French (which are spoken in a great diversity of countries), is to establish the grammatical and vocabulary nucleus of all the varieties (Greenbaum 1996). This cohesion will assure the mutual understanding among the speakers of these varieties.

In what concerns English, Quirk et al. (1985) agree that: "A common core or nucleus is present in all varieties so that, however esoteric a variety may be, it has running through it a set of grammatical and other characteristics that are present in all the others. It is this fact that justifies the application of the name 'English' to all the varieties." (Quirk et al. 1985, *apud* Nelson 2006: 115).

Using the terminology in Nelson (2006), we have extracted the core vocabulary or nucleus of the five corpora (i.e. the common lexicon to all five varieties), as well as the peripheral vocabulary (i.e. that area of the lexicon where, in the corpus, overlapping between varieties does not occur). The common core data are completely reliable, but even in corpora with bigger dimensions (such as the International Corpus of English, where each variety is 1 million words) it is difficult to consider non-overlapping lexical items as definitively specific of one variety, since many situational and contextual factors may determine the occurrence—or lack thereof—of lexical items in one subcorpus. Nevertheless, the results of the peripheral vocabulary must be taken into consideration as being an important contribution to our lexical knowledge of AVP, even though they ought to be validated in corpora of bigger dimensions.

Lexical indexes gave us information on the lemmas that constitute the common nucleus of the five subcorpora and on those that had occurred in four, three, two or only one of the subcorpora. We present in Table 2 the quantitative results, in percentile terms, of these occurrences. As we can see, the percentage of lemmas common to the five corpora is lower than the lemmas that have occurred in only one of the subcorpora.

That common nucleus contains the lemmas with a larger frequency of occurrence in the corpus and it can be considered the Basic Vocabulary of the Africa Corpus.

Core lexicon	Common to 5 varieties	26 %
Lexicon From core to periphery	Common to 4 varieties	11 %
	Common to 3 varieties	11 %
	Common to 2 varieties	15 %
Peripheral lexicon	Specific to 1 variety	37 %

Table 2: Core and peripheral vocabulary in AVP

This vocabulary common to all five corpora (26 % of the lemmas) corresponds to 91.75 % of occurrences in the corpus. The lemmas that occurred in just one of the corpora present low frequencies or are *hapax legomena*, and are in fact more representative cases of lexical change of the lexicon of the Portuguese language. Table 3 below presents some examples of nouns, verbs and adjectives that do not occur in EP and occur in a single AVP, with the exception of *desconseguir,* ‘to unachieve’, which is present in the corpora of two varieties (Angola and Mozambique).

	Angola	Cape Verde	Guinea-Bissau	Mozambique	S. Tome and Principe
Nouns	desatracção desinteriorização	desaculturação descrucificação descravização	desfeitura	descamponês desemergência destriunfo	
Verbs	desconseguir desestrelar	desbaralhar		desconseguir descosturar destrabalhar	
Adjectives	descrispado	desapontador desmamentado		desapetitoso	desarrazoável

Table 3: Neologisms with prefix des ‘un’ in AVP

In many cases, word forms that seemed specific to the AVP turned out to be attested forms in EP dictionaries, although seldom occurring in contemporary Portuguese. With this in mind, lexical items were marked as neologisms, specific to the AVP, only after confirmation that they did not occur in two reference dictionaries of EP (cf. Gonçalves 1966; *Grande Dicionário da Língua Portuguesa* 2004) and that were not labelled as an *africanism*.

A comparison of the most frequent lexical items in AVP and EP was undertaken, by comparing, on the one hand, the spoken subpart of the Africa Corpus to a spoken corpus of EP (the frequency corpus of *Português Fundamental*) and, on the other hand, by comparing the total Africa Corpus to a spoken and written corpus of EP (the Corlex corpus of 16 million words). In both cases, the comparative analysis isolated a subset of 35 items common to AVP and to EP, out of the total 50 lexical items, accounting for 70 % (in spite of the different size of the corpora of AVP and EP). This high frequency lexical subset points to a nuclear and homogeneous vocabulary of both AVP and EP.

4. Morphological and Syntactic Analysis

A preliminary observation of the five corpora pointed to several aspects where AVP differ from the European norm. We will mention some of these properties.

4. 1 Gender and number marking in nominal phrases

In many cases, there are different gender marks from EP showing on the determiner and/or the adjective co-occurring with the noun, as in (1) and (2) below. In (1), the article occurring in the nominal phrase is used in its feminine form, though the noun *problemas* 'problems' is masculine in EP. The opposite situation occurs in (2), where *associação* 'association', a feminine noun in EP, occurs in the corpus of Mozambique with the masculine form of the article and of the adjective.

(1) "quais são *as* *principais* *problemas* da juventude
aqui do teu bairro?" AN (EP: os principais problemas)
What are the-fem main problems of the youth
here from your neighbourhood? (EP: the-masc main problems)
'What are the main problems of young people in your neighbourhood?'

(2) "andei estudar aqui *no* *associação africano*" MO
(EP: na associação africana)
I have been studying here at-the-masc association African-masc
(EP: the-fem African-fem association)
'I have been studying here at the African Association'

Contexts with an absence of number agreement between the head noun and the determinants and adjectives of the nominal phrase are extremely frequent in the Africa Corpus. Three examples are presented in (3)-(5). In (3), the plural head noun *situações* 'situations' does not trigger number agreement with the adjective that follows. In the two other examples the adjective is marked with plural features while the head noun is singular (and also the determinant in (5)). These contexts are less clear-cut cases when compared to the European norm: either the head noun should be marked with plural features, just as the adjective, or the adjective should be in the singular, agreeing with the noun. Both possibilities are indicated below for EP.

(3) "enfrentou *situações* *difícil*" CV (EP: enfrentou *situações difíceis*)
faced situations-pl difficult-sg
'[He/she] faced difficult situations'

(4) "como é possível sair-se desta situação, assim, sem
grandes problema?" AN (EP: sem grandes problemas / sem grande problema)
'how is it possible to get out of this situation, that way, without
big-pl problem-sg (EP: without big-sg problem-sg / without big-pl problems-pl)

(5) "sem haver *o* *transporte* *suficientes*" MO
(EP: transporte suficiente / os transportes suficientes)
without having the-sg transportation-sg sufficient-pl
'without the necessary transportation'

4. 2 Person and number agreement between subject and predicate

The verbal forms, in the contexts observed in the Africa Corpus, frequently lack person and number agreement with the subject, and are usually reduced to the 3rd person singular:

(6) "*nós* *vai* continuar o meu trabalho" CV; (EP: nós vamos continuar)
we going-3sg continue the my work (EP: we going-1pl continue)
'We are going to continue my work'

(7) "tudo *eles* *leva*" AN; (EP: tudo eles levam)
everything they take-3sg (EP: everything they take-3pl)
'They take everything'

4. 3 Verb complementation

Verb complementation is another aspect where AVP diverge from EP. Our preliminary analysis of the five corpora shows that each AVP presents internal variation regarding verb complementation, either converging or diverging from EP patterns. However, data from the five comparable corpora point to several general tendencies.

Prepositional or indirect objects in EP realised as direct objects in AVP:

(8) "peguei *a* *panela*" CV; (EP: peguei na panela)
[I] took the pot-dirOBJ (EP: took the pot-prepOBJ)
'I took the pot'

(9) "O doutor Mondlane assistiu *o nosso jogo*" MO; (EP: assistiu ao nosso jogo)
Doctor Mondlane watched our game-dirOBJ (EP: watched to our game-prepOBJ)

Direct objects in EP realised as prepositional or indirect objects in AVP:

(10) "o marido abandonou-*lhe*" MO; (EP: abandonou-a)
The husband left-her-indOBJ (EP: left-her-dirOBJ)
'Her husband left her'

(11) "para combater *com a deliquência*" GB;
(EP: combater a delinquência)
to fight with the delinquency-prepOBJ
(EP: fight the delinquency-dirOBJ)
'to fight the delinquency'

Prepositional objects in EP realised with different prepositions in AVP: AVP present a more limited range of prepositions, and some prepositions are extensively used in contexts which would show up as a large variation in EP, such as the case of the preposition *em* 'in' (see examples (12) and (13)) which covers different semantic values.

(12) "levamos *no hospital*" GB; (EP: levamos ao hospital)
[we] take in the hospital (EP: take to the hospital)
'We take (someone) to the hospital'

(13) "quando chego *em casa*" CV; (EP: chego a casa)
when [I] arrive in home (EP: arrive to home)
'When I arrive home'

4. 4 Pronominal constructions

The first observation of the data shows that the absence of clitic in constructions that would be pronominal in EP occurs with relative frequency in all five varieties[3], even in the Portuguese variety of Cape Verde, which in general presents much fewer diverging properties than the others.

This is illustrated in (14) with pronominal verbs such as *lembrar-se* 'to remember', *esquecer-se* 'to forget', *levantar-se* 'to get up':

(14) a. "até eu *lembro* uma vez… O doutor Mondlane… assistiu o nosso jogo…"
MO (EP: até eu *me lembro* uma vez)
'even I remember once… Doctor Mondlane…watched our game…'
(EP: even I CL remember)

[3] A more detailed contrastive analysis of pronominal constructions in AVP and EP is presented in Mendes and Estrela (forthcoming).

b. "os pais (…) ficam com ideia só em ganhar dinheiro para dar os filhos de comer e *esquecem* da responsabilidade moral." GB (EP: *esquecem-se* da responsabilidade)
'parents (…) think only about earning money to give their children something to eat and they forget their moral responsibility' (EP: they forget-CL)

c. "fomos ao mato, eu levantei e tal" ST (EP: eu *levantei-me*)
'we went to the bush, I got up and so on' (EP: I got up-CL)

It is important to note that the absence of clitic is not an established pattern in AVP. Although it occurs frequently, the same verb can be used as pronominal and as non-pronominal, sometimes even by the same speaker. This is the case of (15) where the verb *aproximar* 'to approach' is first used as non-pronominal, contrary to the EP norm *(aproximar-se)*, and immediately after as pronominal *(tenha aproximado-se)*.

(15) "tivemos casos de tubarões que tem[…] que os nossos homens não *aproximam* muito à praia. já tivemos um único caso de tubarão *tenha aproximado-se* da praia" ST (EP: os nossos homens não se aproximam muito da praia)
'we've had cases of sharks that […] that our men do not approach the beach much. we had a single case of shark had approached-CL the beach'

The absence of clitic occurs in other constructions, as in example (16), where the impersonal pronominal construction is used in EP, and the expression *como se costuma dizer* 'as is commonly said' is somewhat lexicalised.

(16) "quem tivesse idade para ir à soirée, como *costuma dizer*" GB
'the ones who were old enough to go to the soirée, as is commonly said'

It also occurs, although much less frequently, in cases of pronominal anticausative alternation[4]. The alternation causative / anticausative would be expressed in EP with a transitive / pronominal structure, as *Eles alteraram o programa* 'They changed the programme' / *O programa alterou-se* 'the programme changed-CL'. In the AVP, the anticausative can occur as a non-pronominal structure, as in the example given in (17), with the predicate *alterar* 'to change'.

[4] For an overview of the pronominal anticausative alternation and other pronominal constructions, see Kemmer (1993), Martins (2003), Ruwet (1972) and Zubizarreta (1985, 1992).

(17) "deixei de estudar porque, (…) depois de passar de classe, da décima classe, começaram a complicar, o programa alterou" AN (EP: o programa alterou-se)
'I quit studying because after finishing the tenth grade, [they] started complicating, the programme changed'

There are few reflexive constructions in the corpus where the clitic pronoun is in fact omitted. However, in the five cases registered, the structure is still understood as having a reflexive reading even without the coreferent direct object, as in (18), essentially due to the previous contexts leading the listener/reader to presume that the speaker is also the affected direct object.

(18) a. "Fui dar aulas. Uhm. Não, eh… *Inscrevi* num projecto de… de alfabetização." CV (EP: inscrevi-*me* num projecto)
'I started teaching. No, I applied [myself] to a literacy project'

There are also contexts where a clitic element is added to a verbal predicate. In these cases, the typical interpretation is that the verb is reanalysed as an intrinsically pronominal one, involving essentially revisions at a lexical level, as in example (19):

(19) "tudo isso contribuiu-se para" GB (EP: contribuiu para)
'all this contributed-CL to'

The total number of diverging contexts regarding pronominal constructions in AVP is presented in Table 4, below.

Absence of clitic	**AN**	**CV**	**GB**	**MO**	**ST**	**Total**
Intrinsically pronominal verbs	10	14	25	19	35	103
Reflexive / reciprocal	0	1	3	1	0	5
Impersonal (impersonal / passive)	1	4	3	4	6	18
Anticausative	1	0	2	0	0	3
Total—absence of clitic	**12**	**19**	**33**	**24**	**41**	**129**
Insertion of clitic	**4**	**3**	**9**	**10**	**3**	**29**

Table 4: Total occurrences of pronominal constructions differing from EP

The insertion of clitic does not show any consistency as an emerging pattern, but points rather to the instability in the use of pronominal constructions, in many cases as the result of hypercorrection (clitic elements in non-intrinsically pronominal verbs).

5. Conclusion

We presented a recently compiled corpus of the African varieties of Portuguese and some aspects of a preliminary corpus-driven contrastive analysis between these varieties and EP.

The Africa Corpus, automatically tagged and lemmatised, has been the source of a contrastive lexicon which establishes a common core vocabulary, as well as peripheral lexical sets for each variety. Some morphological, lexical and syntactic properties where AVP diverge widely from EP have been outlined and are the focus of ongoing contrastive studies. However, these require larger corpora in order to reach confident observations regarding the evolution of AVP and their relationship with EP. In fact, since most of the properties where AVP differ from the EP norm are still emerging and show strong variation inside each variety, it is essential to rely on balanced and comparable corpora.

This will enable us to establish more stable tendencies of linguistic change across the varieties and point to the process of identification and understanding of the unity and diversity factors that are at stake between European, Brazilian and African varieties of Portuguese.

References

Bacelar do Nascimento, M. F. (coord.) (2001). *Português Falado—Documentos Autênticos: Gravações Audio com transcrição alinhada*. 4 CD-ROM, Instituto Camões and Centro de Linguística da Universidade de Lisboa, Lisbon.

Bacelar do Nascimento, M. F./Bettencourt Gonçalves, J./Pereira, L. A. S./Estrela, A./Oliveira, S. M./Santos, R. (2008a). "Aspectos de unidade e diversidade do português: as variedades africanas face à variedade europeia", *Revista Veredas*, São Paulo.

Bacelar do Nascimento, M. F./Estrela, A./Mendes, A./Pereira, L. (2008b). "On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications" in *Proceedings of the VI Language Resources and Evaluation Conference—LREC2008, Workshop on Comparable Corpora*, Marrakesh, 39–46.

Bacelar do Nascimento, M. F./Pereira, L. A. S./Estrela, A./Bettencourt Gonçalves, J./Oliveira, S. M./Santos R. (2006). "The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon" in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, May 24–26, Genoa, Italy, 1791–1794.

Bettencourt Gonçalves, J./Veloso, R. (2000). "Spoken Portuguese: Geographic and Social Varieties" in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000)*, Vol. II. Athens, May 31–June 2, 905–908.

Duarte, I./Gonçalves, A./Miguel, M./Mota, M. A. M. (1999). "*Não cheguei de aprender nada* - Áreas de Variação e Tendências de Mudança no Português de Moçambique" in Lopes, A. C./Martins, C. (eds.) (1999). *Actas do XIV Encontro Nacional da Associação Portuguesa de Linguística, Volume I*. Braga. Associação Portuguesa de Linguística, 477–493.

Gonçalves, P. (1990). *A Construção de uma Gramática de Português em Moçambique: Aspectos da Estrutura Argumental dos Verbos*. PhD Thesis. Universidade de Lisboa, Faculdade de Letras.

Gonçalves, P. (1996). "Aspectos da sintaxe do Português de Moçambique" in Faria, I. et al. (eds.) (1996). *Introdução à Linguística Geral e Portuguesa*. Lisboa, Caminho, 313–322.

Gonçalves, P. (1997). "Tipologia de Erros" in Stroud, C./Gonçalves P. (eds.) (1997b). *Panorama do Português Oral de Maputo*. Vol. 2—*A Construção de um Banco de "Erros"*. Cadernos de Pesquisa nº 24. Maputo, INDE.

Gonçalves, P./Stroud, C. (eds.) (1998). *Panorama do Português Oral de Maputo*, Vol. 3—*Estruturas Gramaticais do Português: Problemas e Exercícios*. Cadernos de Pesquisa nº 27. Maputo, INDE.

Gonçalves, P./Stroud, C. (eds.) (2000). *Panorama do Português Oral de Maputo*, Vol. 4—*Vocabulário Básico do Português (espaço, tempo e quantidade) Contextos e Prática Pedagógica*. Cadernos de Pesquisa nº 36. Maputo, INDE.

Gonçalves, P., Stroud, C. (eds.) (2002). *Panorama do Português Oral de Maputo*, Vol. 5—*Vocabulário Básico do Português, Dicionário de Regências*. Cadernos de Pesquisa nº 41. Maputo, INDE.

Gonçalves, R. (1966). *Vocabulário da Língua Portuguesa*. Coimbra: Coimbra Editora.

Grande Dicionário da Língua Portuguesa (2004). Porto: Porto Editora.

Greenbaum, S. (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Kemmer, S. (1993). *The Middle Voice*. Amsterdam/Philadelphia: Benjamins.

Martins, A. M. (2003). "Construções com *se*: Mudança e variação no português europeu" in Castro, I. / Duarte, I. (eds.) (2003). *Razões e Emoção: Miscelânea de estudos em Homenagem a Maria Helena Mateus*, Vol. 2. Lisboa: Imprensa Nacional—Casa da Moeda, 163–178.

Mendes, A. / Estrela A. (forthcoming). *Pronominal constructions in African varieties of Portuguese*. *Phrasis*.

Nelson, G. (2006). "The Core and the Periphery of World Englishes: a Corpus-based Exploration", *World Englishes*, 25(1), 15–129.

Quirk, R. / Greenbaum S. / Leech G. / Svartvik J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Rio-Torto, G. (2007). "Caminhos de Renovação Lexical: Fronteiras do Possível" in Isquerdo, A. N. / Ieda, M. A. (eds.) (2007). *As Ciências do Léxico, Lexicologia, Lexicografia, Terminologia*, Vol. 3. Campo Grande, MS, Ed. UFMS, São Paulo, Humanitas.

Ruwet, N. (1972). "Les constructions pronominales neutres et moyennes", *Théorie Syntaxique et Syntaxe du Français*. Paris, Seuil, 87–125.

Stroud, C. / Gonçalves, P. (eds.) (1997a). *Panorama do Português Oral de Maputo*, Vol. 1—*Objectivos e Métodos.* Cadernos de Pesquisa nº 22. Maputo, INDE.

Stroud, C. / Gonçalves, P. (eds.) (1997b). *Panorama do Português Oral de Maputo*, Vol. 2—*A Construção de um Banco de «Erros»*. Cadernos de Pesquisa nº 24. Maputo, INDE.

Zubizarreta, M. L. (1985). "The Relation between Morphophonology and Morphosyntax: the Case of Romance Causatives", *Linguistic Inquiry*, 16(2), 247–289.

Zubizarreta, M. L. (1992). "The Lexical Encoding of Scope Relations among Arguments" in Stowell, T. / Wherli, E. (eds.) (1992). *Syntax and Semantics*, 26, *Syntax and the Lexicon*. San Diego: Academic Press, 211–258.

Vis-À-Vis: A System for the Comparison of Linguistic Varieties on the Basis of Corpora

Stefanie Anstein

Varieties of plurilinguistic languages are each investigated and described to different extents, where the 'main' varieties usually have a much better NLP coverage than their lesser-used counterparts. The toolkit described in this contribution aims at supporting linguists in their comparison of language varieties, which in most cases have many similar linguistic characteristics. For their differences, even subtle ones, text corpora are to be used as a basis to semi-automatically extract particularities on different levels of linguistic description. Existing and adapted as well as new tools will be used for the extraction, and the toolkit will provide 'candidate lists' to reduce the efforts of experts. The system will be first evaluated with German varieties, but it is intended to be transferable to other languages as well.

1. Introduction

Varieties of plurilinguistic languages such as English, French or German (cf. Ammon 1997) are each investigated and described to different extents, where the 'main' varieties usually have a much better NLP coverage than their lesser-used counterparts. Even though such varieties in most cases do have many similar linguistic characteristics, there are still differences that are to be extracted, for example, for variety documentation and variant lexicography (cf. Ammon et al. 2004); for standardising lesser-used varieties; or for aiding language teaching and learning.

Starting from the lexical level, differences usually consist in one-to-one equivalents such as *Kondominium* ('apartment building') in South Tyrolean German versus *Mehrfamilienhaus* in Germany or in many-to-one equivalents, such as *provisorische Ausfahrt* ('temporary exit') and *Behelfsausfahrt*, respectively. More complex phenomena such as differing collocations or subtle semantic differences or even pragmatic particularities of a variety are more difficult to find, for example, the additional meaning of *Mobilität* in South Tyrol, which means a kind of 'unemployment' as well as 'mobility'.

The toolkit Vis-À-Vis is being developed in a doctoral thesis in the framework of the project 'Korpus Südtirol'. A description of its first approaches can be found in Abel and Anstein (2008). It will be evaluated mainly with German varieties, but the resulting system aims to be language-independent.

2. Background and Motivation

Corpora for the different varieties of a language are a valuable basis for finding relevant particularities and differences. They are being compiled in projects such as the 'International Corpus of English' (ICE[1]) or, for instance, for use in work on Portuguese varieties by Bacelar do Nascimento et al. (2006). Also, an initiative of research centres in Basel[2], Berlin[3], Bolzano[4] and Vienna[5] called 'C4' is developing variety corpora for German that are comparable with respect to content and size.

To annotate corpora without specifically adapted or trained tools and complete lexica (e.g. for a language variety), tools for a similar language (mostly a main variety one) can be initially used. Their output is then manually corrected and inserted again as a basis for new rules or training, similar to what is sometimes done for unexpected input. In addition, lesser-used languages often have specific word lists that can be exploited, such as place or person names (available from maps or statistics offices). Again, these will usually be enhanced by findings of respective variety corpora.

Related work has been done in diachronic linguistics on the comparison of language over time (cf. Janda & Joseph 2004), of originals and translations (cf. Baroni & Bernardini 2006), of native and learner languages (cf. Netzel et al. 2003), and so forth. Many of the earlier studies were conducted manually, often for very specific phenomena. In addition, there are now more statistical approaches to data extraction from parallel or even from unrelated monolingual corpora (e.g. Nazar 2008).

[1] http://www.ucl.ac.uk/english-usage/ice; see also bibliography list for specific variety studies.

[2] http://www.schweizer-texkorpus.ch

[3] http://www.dwds.de

[4] http://www.korpus-suedtirol.it

[5] http://www.aac.ac.at

For a systematic and comprehensive comparison of corpora on different levels of linguistic description, semi-automatic tools are needed, since manual evaluation is time-consuming and costly. Resources such as corpora have to be compared directly according to regular patterns with statistical counts, and on different levels, where the comparability of the contents and of the corpora in general has to be taken into account (cf. Kilgarriff 2001; Gries 2007). Automatic filtering of statistically-produced lists containing suggested 'candidates' for differences or peculiarities reduces manual work and supports experts in their evaluation. In such a process, trivial characteristics of a variety or knowledge already investigated and confirmed (e.g. collections of proper names or regionalisms such as the above-mentioned *Kondominium*) can be automatically removed from candidate lists. Experts can then concentrate on the evaluation and interpretation of the remaining (mostly) new phenomena, which nevertheless will always have to be done manually.

3. System Sketch

Vis-À-Vis tools aim at providing support to linguists for the systematic comparison of varieties on the basis of corpora. This support consists of methods to filter huge amounts of data and to select for the expert only the material that is probably relevant. With this approach, less manual work is necessary and quantitative methods can be combined with qualitative ones. In addition, the data to be evaluated manually are presented in a user-friendly and intuitive way to facilitate the interpretation and further processing. In Figure 1, the overall Vis-À-Vis preliminary architecture and workflow is demonstrated.

As input, users give Vis-À-Vis the corpora to be compared as well as, if available, lists with previous knowledge as described above. The corpora are then annotated with standard tools, during which difficult cases for the tools (or errors produced by them) can identify the first set of candidates for special variety characteristics, given that the tools are usually created for the main varieties. In the next modules, the corpora are analysed and compared with a combination of existing as well as new or adapted tools, for example, concordancing or frequency statistics. The lexical level is the first and most promising linguistic area to explore; further studies will elaborate on collocations (e.g. using a system as described in Heid and Ritz [2005]) and phrases that include more subtle semantics (e.g. using Semantic Vectors [cf. Widdows & Ferraro 2008]) or pragmatic differences.

Wherever possible, knowledge about the variety is taken into account in all the modules. As a result, Vis-À-Vis produces filtered lists of probably relevant differences between the varieties for manual evaluation. It is also possible for the user to search directly in the relevant corpora for sentence contexts of ambiguous or other difficult cases. In a further step, the findings can again be used for the annotation of approved special vocabulary or more complex phenomena in other corpora of that variety to be compared.

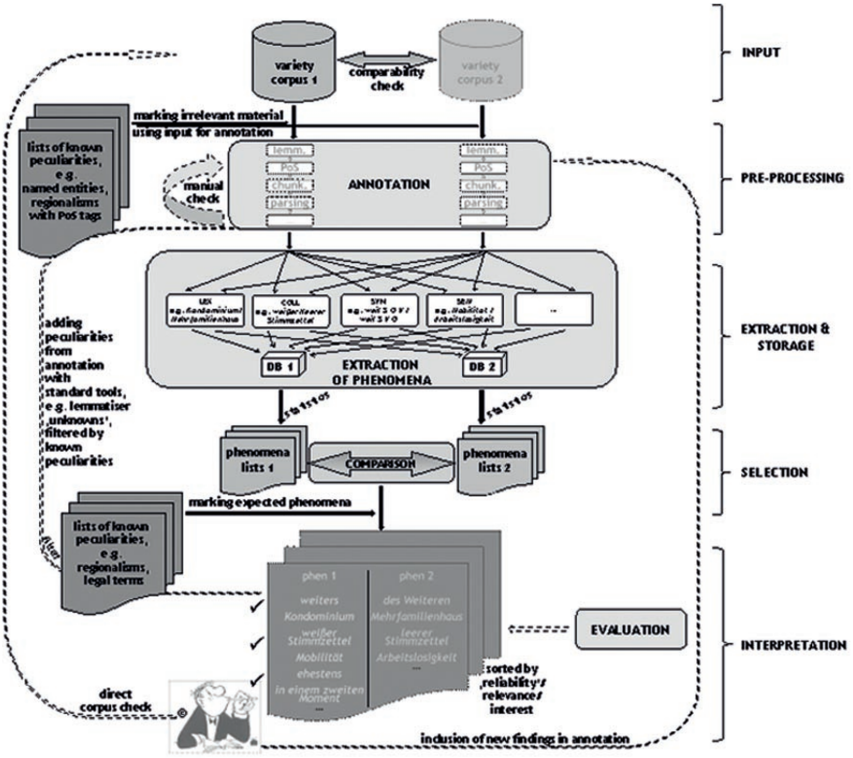

variety corpus 1
comparability check
variety corpus 2
INPUT
marking irrelevant material
using input for annotation
lists of known peculiarities, e.g. named entities, regionalisms with PoS tags
lemm.
PoS
chunk.
parsing
manual check
ANNOTATION
PRE-PROCESSING
LEX e.g. Kondominium/ Mehrfamilienhaus
COLL e.g. weißer/leerer Stimmzettel
SYN e.g. weil S O V / weil S V O
SEM e.g. Mobilität / Arbeitslosigkeit
...
DB 1
DB 2
EXTRACTION OF PHENOMENA
EXTRACTION & STORAGE
adding peculiarities from annotation with standard tools, e.g. lemmatiser ‚unknowns', filtered by known peculiarities
statistics
statistics
phenomena lists 1
COMPARISON
phenomena lists 2
SELECTION
marking expected phenomena
lists of known peculiarities, e.g. regionalisms, legal terms
filter
phen 1
weiters
Kondominium
weißer Stimmzettel
Mobilität
ehestens
in einem zweiten Moment
...
phen 2
des Weiteren
Mehrfamilienhaus
leerer Stimmzettel
Arbeitslosigkeit
...
✓
✓
✓
sorted by ‚reliability'/ relevance/ interest
EVALUATION
INTERPRETATION
direct corpus check
inclusion of new findings in annotation

Figure 1: Vis-À-Vis preliminary overall architecture

4. Conclusion

This paper briefly describes on-going work on a comprehensive system to compare language varieties on the basis of corpora, in which manual expert work will be supported semi-automatically. Possible alternative detail solutions are still to be discussed and decided upon. As first results, prototype modules and processing pipelines will be presented in future contributions. Parallelly, concrete results on differences between German varieties will be obtained and published.

References

Abel, A. / Anstein, S. (2008). "Approaches to Computational Lexicography for German Varieties" in *Proceedings of the XIIIth Euralex International Congress*, Barcelona, 251–260.

Ammon, U. (1997). "Nationale Varietäten des Deutschen" in *Studienbibliographien Sprachwissenschaft*, Vol. 19. Julius Groos: Heidelberg.

Ammon, U. / Bickel H. / Ebner, J. / Esterhammer, R. / Gasser, M. / Hofer, L. / Kellermeier-Rehbein, B. / Löffler, H. / Mangott, D. / Moser, H. / Schläpfer, R. / Schloßmacher, M. / Schmidlin, R. / Vallaster, G. (2004). *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtiro*l. Berlin/New York: Walter de Gruyter.

Bacelar do Nascimento, M. F. / Gonalves, J. B. / Pereira, L. / Estrela, A. / Pereira, A. / Santos, R. / Oliveira, S. M. (2006). "The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon" in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC),* Genoa, Italy, 1791–1794.

Baroni, M. / Bernardini, S. (2006). "A new approach to the study of translationese: Machine-learning the difference between original and translated text", *Literary and Linguistic Computing*, 21(3), 259–274.

Gries, S. T. (2007). "Exploring variability within and between corpora: Some methodological considerations," *Corpora*, 1(2), 109–151.

Heid, U. / Ritz, J. (2005). "Extracting collocations and their contexts from corpora" in *Papers in computational lexicography - COMPLEX*, Linguistics Institute, Hungarian Academy of Sciences. Budapest, 107–121.

Janda, R. D. / Joseph, B. D. (eds.) (2004). *The Handbook of Historical Linguistics.* Blackwell.

Kilgarriff, A. (2001). "Comparing corpora", *International Journal of Corpus Linguistics*, 6(1), 1–37.

Nazar, R. (2008). "Bilingual terminology acquisition from unrelated corpora" in *Actas del XIII Congreso Internacional Euralex*. European Association for Lexicography, Universidad Pompeu Fabra.

Netzel, R. / Perez-Iratxeta, C. / Bork, P. / Andrade, M. A. (2003). "The way we write". *EMBO Reports*, 4(5), 446–451.

Widdows, D. / Ferraro, K. (2008). "Semantic vectors: a scalable open source package and online technology management application" in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

The Latgalian Component in the Latvian National Corpus

Aleksey Andronov and Everita Andronova

The paper deals with the Latgalian written language used by a part of Latvians living mostly in the eastern Latvia. There is an initiative to start compiling a Latvian National Corpus, which is supposed to be a long-term activity to cover all printed Latvian texts and a wide scope of the speech. Since there are two standardised varieties of Latvian: the Latvian literary language and the Latgalian written language (with a tradition of more than 250 years), Latgalian should be represented in the Latvian National Corpus. The authors describe the first experience with compiling a corpus of Latgalian texts published in Soviet Russia (1917–1937) and detect some possible issues of compiling a corpus of Modern Latgalian, which can be delimited by the National Awakening and the reestablishment of the Republic of Latvia in 1991. Although Modern Latgalian is characterised by restricted usage and the lack of some text types, there are rich Latgalian data from other time periods and regions, and this would make it possible to develop some specialised corpora in future. Various linguistic resources and tools should be developed for Latgalian in order to raise its prestige.

1. The Latvian Corpus: the State-of-the-Art

Nowadays language corpora are a clear prerequisite for a comprehensive study of a language and its very existence in a global high-tech society ("linguistic corpora are intended to be the basis for the analysis and description of the structure and use of languages and for various applications" [Kennedy 1998: 60]). In these terms, the Latvian language could be considered as a lesser-used language, because there is still a lack of corpus resources and the community of linguists is still not very enthusiastic about using modern technologies in their everyday research. There is a certain gap observed between language resource developers and users. There is a rather long-term tradition of collecting Latvian texts in electronic form, dating back to the beginning of the 90s (Milčonoka et al. 2004; Grūzītis et al. 2004). Today the main language resource developers are the Institute of Mathematics and Computer Science at the University of Latvia (henceforth IMCS, UL), the National Library of Latvia, the IT company 'Tilde', as well as some academic institutions. The corpus activities were started earlier this century—a pilot morphological annotation has been performed (Levāne et al. 2000) and some studies of a parallel English-Latvian corpus have been carried out (Skadiņa 2005; Milčonoka 2001). Apart from this, in 2003 a diachronic Corpus of Early Written Latvian was launched, and its development is in progress (Andronova 2007). In 2005, a design of the Latvian corpus was developed by the IMCS, UL (Koncepcija 2005). In 2007, a one-million-token balanced corpus of Modern Latvian was compiled according to the guidelines set in this design. The compilation was supported by the State Language Agency and done at the IMCS, UL[1]. In 2009, the size of the corpus will be extended by another 2.5 million tokens from balanced and representative Latvian texts. There are some activities carried out towards an unbalanced large corpus from texts available on the Web (Džeriņš et al. 2007). There are a number of experimental language tools developed at the IMCS, although they are still not available for off-the-shelf use; there are plans to provide a graphical corpus interface for a semi-automatic morphological and syntactic analysis within the SEMTI-Kamols project[2].

The Latvian State Language Commission, established in 2002, aims to study "the situation of Latvian as the country's state language and to draft recommendations on how to strengthen its status and develop it further" (The State Language Commission). The year 2008 was marked by several initiatives supported by the State Language Commission: in April, a wider researcher community was introduced to the concept of a language corpus, and for the first time the idea of a Latvian National Corpus came up. Later the initiative was taken by the National Library of Latvia, which is the leader of the National Digital Library project[3] and inspired the Agreement of Intention between the main language resource developers and holders in Latvia, both academic and industry partners. In November, an international workshop was held in Riga, organised by the Latvian State Language Commission; the IMCS, UL; and the National Library to get acquainted with the practice of the Czech National Corpus and to set some further tasks. Unfortunately, the initiative is now slightly slowed down.

The Latvian National Corpus is supposed to be a long-term activity to cover all printed Latvian texts and a wide scope of the speech (cf. Vasiļjevs 2008). It will be a

[1] It is now available from www.korpuss.lv via the Manatee platform http://www.textforge.cz/products

[2] www.semti-kamols.lv

[3] www.lnb.lv/lv/digitala-biblioteka

complex system of separate sub-corpora, both synchronic and diachronic, monolingual and multilingual, developed by different partners such as the University of Latvia, the Institute of the Latvian Language, the Institute of Mathematics and Computer Science, the National Library, and so forth.

The National Corpus is expected to represent the Latvian language in full. According to the State Language Law, there are two standardised varieties of Latvian: the Latvian literary language and the Latgalian written language. The law states: "The official language in the Republic of Latvia is the Latvian language (§3.1). The State shall ensure the maintenance, protection and development of the Latgalian written language as a historical variant of the Latvian language (§3.4)" (VVL 1999). Thus, the Latvian National Corpus cannot be considered complete without the Latgalian component, which is the topic of the present report.

2. What is Latgalian?

Latvia is historically divided into several ethnographic regions; the eastmost of them is Latgale, former Polish Livonia. Latvian has been divided into three dialects: the Central dialect, the Tamian dialect and the High Latvian Dialect (for more detailed information see Balode et al. [2001]).

The Latgalian written language is a standardised variety of the language used by a part of Latvians, living mostly in eastern Latvia (Latgale). Due to the history of the region the native population of Latgale differs from other Latvians, not only in language, but also in ethnography, cultural life and religion (Latgalians are mostly Roman Catholics, while other Latvians are mostly Protestants). For almost three centuries (1629–1917), Latgale was separated from the rest of Latvia. In the 17th century its territory came under the rule of the Polish–Lithuanian Commonwealth, and it was known as Polish Inflantia or Polish Livonia. In 1772 it was incorporated into Vitebsk Province of the Russian Empire. Therefore, the Latgalian language has been exposed to influence from Polish and East Slavonic (Russian and Belarussian).

There is no common agreement on the linguistic status of the language spoken in Latgale: it is considered either one of the three main dialects of the Latvian language or a separate Baltic language on equal terms with Latvian and Lithuanian (Brejdak 2006: 195). Some linguists try to achieve for Latgalian the status of regional language in Latgale. Further in the text it will be referred to as just Latgalian, and its standardised variety as Standard Latgalian.

The linguistic distinction between Standard Latgalian and Standard Latvian is large enough to complicate mutual understanding. The differences are mainly found in the phonological system, as well as in the vocabulary, but certain important deviations exist also in morphology and syntax. See the text of prayer 'Pater noster' in Latvian and in Latgalian in Table 1.

Latvian	Latgalian
Mūsu Tēvs debesīs! Svētīts lai top Tavs vārds. Lai nāk Tava valstība. Tavs prāts lai notiek kā debesīs, tā arī virs zemes. Mūsu dienišķo maizi dod mums šodien. Un piedod mums mūsu parādus, Kā arī mēs piedodam saviem parādniekiem. Un neieved mūs kārdināšanā. Bet atpestī mūs no ļauna. [Jo Tev pieder valstība, spēks un gods mūžīgi mūžos.] Āmen.	Tāvs myusu, kas esi debesīs, svieteits lai tūp Tovs vuords, lai atīt Tova vaļsteiba, Tova vaļa lai nūteik kai debesīs, tai ari viers zemis. Myusu dīniškū maizi dūd mums šudiņ un atlaid mums myusu poruodus, kai ari mes atlaižam sovim poruodnīkim, un naīved myusu kārdynuošonā, bet atpestej myus nu ļauna. Amen.

Table 1: Example of 'Pater noster' in Latvian and Latgalian

Latgalian has a well-established written tradition dating back to 1753 (cf. Leikuma 2008); it has experienced the ban of publishing books in Latin script during the process of Russification as a part of Russia's anti-Polish policy (1865–1904). Probably more than 750 books have been published till now (cf. Seiļ 1936). There are several linguistic descriptions of Latgalian (practical grammars and dictionaries) reflecting deliberate work on developing a literary norm. A precise statistic evaluation is difficult, but according to the Research Institute of Latgale some 150–200,000 people speak Latgalian in their everyday life.[4] It is used not only at home, but also has a notable place in public life, cultural events, local authorities' work, and Catholic church services. The amount of the linguistic and social linguistic problems to be commented on in a comprehensive description of Latgalian corresponds to the frame of a language, not a dialect.

Among the 'alternative' languages of the Baltic States (compared to Võro in Estonia and Samogitian in Lithuania), Latgalian is the most prominent and fully-fledged. If we look at the *Unesco Digital Atlas of the World's Languages in Danger of Disappearing*, Latgalian is marked as unsafe (UNESCO 2008).

[4] Retrieved May 15, 2009, from http://dau.lv/ld/latgale(english).html

3. Why is the Modern Latgalian Corpus Necessary?

In spite of considerable usage of Latgalian in fiction and mass media (including radio broadcasting and the Internet)[5], the government pays no special attention to it, and it lacks linguistic research, thus making the language endangered in Latvia today. There are several courses on the Latgalian written language and its history at universities (in Rezekne Higher Education Institution, Daugavpils University and the University of Latvia), but the practical language is not taught at schools.

Several linguistic resources and tools should be developed for Latgalian in order to raise its prestige and to ensure its development. A standard dictionary and grammar, schoolbooks and readers, a spell-checker and a morphological analyser, together with a linguistic corpus are necessary.

A modern language corpus would serve as a basis for other resources. The modern language period began together with the National Awakening and the reestablishment of the Republic of Latvia in 1991, which gave a new impulse to the rebirth of Latgalian after its being almost neglected during the years of the Soviet rule. In June 1990, a public non-profitable organisation, The Latgalian Culture Centre[6] was established, and its publishing house is the main publisher of books of different genres in Latgalian today.

4. The first experience with compiling the Corpus of Latgalian texts published in Soviet Russia (1917–1937)

At the end of the 19th century, there was an organised movement to get free land in the Russian Empire, especially in areas of Siberia. Thousands of Latgalian people moved to Russia. A pioneer initiative has been started at St. Petersburg State University in close cooperation with the National Library of Russia to compile a corpus of the Latgalian texts published in Soviet Russia (Andronov et al. 2008). Concerning this corpus, all the texts (100 books and 11 periodicals) published during the period 1917–1937 are representative to generalise the language of that period as a whole, as we do

[5] There is a blogger in a daily newspaper of Latvia writing in Latgalian (http://www.diena.lv/lat/tautas_balss/blog/saprge), ‘Latgales Radio’ (which has existed since 2006) broadcasts mostly in Latgalian (http://www.lr.lv/).

[6] http://www.lkcizdevnieciba.lv/

not have any other sources to be included in the corpus. The issue of representativeness here might be associated with the argument concerning the case of diachronic corpora where "it can only be based on the body of preserved texts and the authenticity of those included in the corpus. However, the linking up of representativeness of diachronic corpora to the body of preserved texts means that the corpora reflect, in fact, the skewed stylistic, genre and other proportions in the body of texts rather than the characteristics of the real language of the time" (Kučera 2007: 1).

The corpus in process will be a static corpus, including full texts of newspapers, fiction (mostly translated from Russian and Latvian, but also original pieces), schoolbooks and social and political brochures. One of the data collection challenges in this case is the lack of some seven sources caused by the repressive policy of national minorities in Soviet Russia in 1937, when books in Latgalian were forbidden and destroyed. Therefore, one of the main tasks of this corpus is to provide researchers with unique data little explored till now. This, of course, requires a systematic and profound search of sources in the largest libraries and archives, and luckily there is still a chance to find lost books (Andronova et al. 2008). The task is to scan approximately 8000 pages and to ensure that facsimiles of the sources are made available on the Web via the server of the National Library of Russia[7]. At the moment 21 sources have been scanned and OCR has been carried out.

The first observations of the data reveal a great amount of linguistic variants in these sources. These are not only spelling versions (found both in the same text and in texts published by different authors, by different publishing houses), but also morphological versions which can be explained by the influence of the native spoken vernacular (since the settlers were from different parts of Latgale which have their own peculiarities) and syntactic versions influenced by the source language and Russian as a close contact language. For instance, different calqued constructions are observed: in Latgalian *jaunotne draudzejas ar komunarim un **jem nu jim lobu pimaru*** 'the youth make friends with Communards and **takes them as a good example**' there is a calqued construction from Russian *berët s nich primer*. This gives us an interesting picture of the language processes which were taking place in written Latgalian in Soviet Russia.

The corpus is supposed to provide several versions of the same text: an original form, a normalised orthography (removing imperfect spellings explained by the gradual adaptation of appropriate graphic means and lack of necessary letters in the typographies), a text with morphological annotation, and a lemma translation into Russian.

[7] http://www.nlr.ru/coll/onl/fonds_onl/latgalsk.htm

As there are no language processing tools for Latgalian, there are two ways to provide a morphological annotation of the text: either manually or by using a morphological analyser for Modern Latvian with some implemented transponation rules, which may be applied for a certain amount of the Latgalian lexicon. Although in this corpus there is obviously a rather high number of lemmas that are influenced by Russian. As for Latvian, there exists a lexicon-based morphological analyser developed at the Institute of Mathematics and Computer Science (IMCS) (Paikens 2007). If we want to make use of this analysis, we may add a specific Latgalian lexicon to the Latvian one.

Compilers of the Corpus of Latgalian texts published in Soviet Russia believe that in the future it will further foster the comparative studies of the varieties of Latgalian used in Latvia and Russia respectively.

5. Problems of the Corpus of Modern Standard Latgalian

There are common issues in corpus design and compilation that should be discussed before any activities are undertaken.

Today, the usage of Latgalian is restricted to a few spheres of social life. It is quite common in oral conversation, but its written form is less popular. The Corpus of Modern Standard Latgalian (CMSLg), a written synchronic corpus, will serve to strengthen the image and status of Standard Latgalian.

This restricted usage and lack of some text types and genres (Biber 1993: 244–245) affect the size, representativeness and balance of the CMSLg. To start with, some 2–5 million running words can be processed in the corpus, although estimating the size is problematic before one has compiled a complete list of sources and studied their availability (issues of authorship, etc.) and quality (see Table 2 below). Thus, composing a comprehensive bibliography of Modern Latgalian publications is a prerequisite, which can be a topic for a separate project. The main part of a corpus of modern language usually consists of texts from periodicals, but CMSLg is quite different in this respect because there are only few periodicals in Latgale publishing more or less sporadic articles in Latgalian ('Katōļu dzeive', 'Latgales Laiks', 'Vietējā Latgales Avīze', 'Rēzeknes Vēstis', 'Vaduguns'). There is an on-line newspaper, 'LaKuGa'[8], edited by the Latgalian Students Centre, which is also a good source for the corpus. Here we can find readers' commentaries, which are usually in a colloquial form; this will make

[8] Retrieved May 15, 2009, from http://www.lakuga.lv/lg/

the data of the corpus more varied. Seemingly, fiction (mainly original) will be the main source of data. An important publishing house is the Cultural Center of Latgale in Rēzekne, which prints fiction and poetry books in Latgalian as well as academic and popular studies in the cultural history of Latgale; a collection of scholarly articles in humanities, 'Acta Latgalica', is published annually by the Research Institute of Latgale in Daugavpils. In addition, we should not ignore the significant role of the Catholic Church in maintaining the Latgalian language both in printed religious texts and in public worship. Modern Latgalian lacks or has a very small amount of medical, juridical, business and technical texts.

Data acquisition and processing in the CMSLg can be solved on the same grounds as in the Latvian part of the National Corpus (Koncepcija 2005: 13, 75–88), although text selection and sampling procedures might differ. One should pay special attention to the input data quality. Many Latgalian texts are created just by mere phonetic transponations from Latvian according to sound correspondence rules, which gives an inadequate impression of the authentic lexicon, morphology and syntax. For instance, see Table 2 below:

Latvian	**so called 'Latgalian'**	**Latgalian**
Arī šorīt pamodos ļoti agri. Rūpēja ikdienas darbi. Kūtī brēca aitas, bubināja nedzirdītais kumeļš, māva neslauktā govs. Nelika mierā tās pašas domas, kuras mocīja jau vairākas nedēļas. Vai atradīs mana Anna cerēto laimi svešumā? Un kā tālāk dzīvot pašam?	Ari šūreit pamūdūs ļūti agri. Ryupēja ikdīnas dorbi. Kūtī brēce aitas, bubynōja nadzirdeitais kumeļš, mōve naslauktō gūvs. Nalyka mīrā tōs pošas dūmas, kuras mūceja jau vairōkas nedeļas. Voi atradeis muna Anna carātū laimi svešumā? Un kai tōļōk dzeivōt pošam?	I šūreit pasamūdu cīši agri. Pruotā stuovēja kasdīnys dorbi. Klāvā viekše vuškys, bubinēja nadzirdeitais kumeļs, bļuove naslauktuo gūvs. Nadeve mīra tuos pat dūmys, kuruos mūceja jau nazcik nedeļu. Voi atrass muna Ane īdūmuotū laimi svešumā? I kai tuoļuok dzeivuot pošam?

Table 2: Example of Latvian transponations into Latgalian[9]

An approximate translation would be as follows: *This morning I woke up early again. I was thinking about today's chores. In the byre sheep bleated, the horse, still unattended, neighed and the cow, still unmilked, mooed. The same thoughts that had bothered me already for some weeks, were again coming to my mind. Will my Anna find the happiness she hoped to get, there, in a foreign country? And how am I supposed to live further?*

Here we may see that instead of original Latgalian words (e.g. *i* 'again'; *cīši* 'very'; *pruotā stuovēja* 'the mind was occupied with', *kasdīnys* 'everyday') phonetically latgalianised forms of Latvian lexemes are used (cf. *ari* 'again'; *ļūti* 'very'; *ryupēja* 'concerned';

[9] Many thanks to Prof. Lidija Leikuma, University of Latvia, who composed the text samples.

ikdīnas 'everyday'). This transponation also concerns the morphology, for example, the original Latgalian reflexive verb form *pasamūdu* 'woke up' is replaced by the transponation of the Latvian form, where the reflexive marker in the prefixed verbs is placed at the end, not after the prefix as in Latgalian (*pamūdū*s vs. *pa*sa*mūdu*), and so forth.

Obviously, there is a question how to deal with such texts, that is, whether they can serve a source of the CMSLg or should be ignored.

Despite the publication of several practical grammars and a few dictionaries in the 20th century and the work of special commissions elaborating the literary norm, there is no generally accepted orthography, and a considerable variation is observed in the morphology and lexicon (not to mention the pronunciation, which is not yet even touched by the literary standard). The problem of mixing odd elements coming from the tradition and those promoted by the linguistic authorities should be solved to ensure the automatic processing of the corpus. An intelligent search engine is necessary to identify the spelling variants (cf. recent orthography rules—LPN 2008).

To sum up, there are two general problems complicating the development of the CMSLg: the objective peculiarities of a minor language and the lack of linguistic research of Latgalian. One should emphasise that developing a corpus will stimulate the research and language progress, contributing to the creation of a fully-fledged Latgalian literary language.

The IMCS together with partners at Rezekne Higher Education Institution are planning to start a compilation of corpus of Modern Latgalian.

6. Possible Types of Latgalian Corpora in the Future

Apart from the corpus of Modern Latgalian, which should be our first task to compile, we may consider the compilation of several possible specialised corpora in order to provide further resources and promote a deeper analysis of all aspects of Latgalian. The main emphasis here is placed on the monolingual corpora, as there are not many Latgalian original works translated into Latvian and vice versa. There are some activities that have been observed of the work of Latgalian authors being translated into Russian and vice versa. On the other hand, we cannot exclude the possibility of compiling a parallel Latgalian–Latvian or Latgalian–Russian corpus (or even other language pairs) in the future.

6. 1 Geographical varieties of Latgalian

On one hand, we should start with the modern Latgalian language spoken and written in Latvia. On the other hand, there is a pretty large Latgalian community that settled in Europe and the United States after the Second World War. While in Soviet Latvia Latgalian was used only as a colloquial language at home, a number of books and articles were printed in Germany by Vladislavs Locis' Press in 1945–1984. This might serve as a basis for the specialised corpus of the different varieties of Latgalian.

6. 2 Dialectal varieties of Latgalian

One should not exclude highly valuable data collected in Latvia during expeditions organised by the academic institutions after the Second World War: linguists, historians, folklorists and ethnologists have recorded Latgalian songs, narratives, and so forth, for more than 50 years. The data collection is still going on. These data are scattered all around Latvia, and the information about these collections and their distribution and characteristics is rather vague. Here, a question of a level of co-operation might rise. Hopefully, this might be partly solved within the CLARIN-Latvia framework, in which almost all Latvia's research institutions expressed their will to participate. Another question concerns the technical possibilities to digitalise these data from the old tape recorders. The usage of metadata is important to ensure the reusability of this valuable information.

Apart from this, there are still Latgalians living in Russia, and serious fieldwork is in progress there. There are some research projects carried out by the University of Latvia and St. Petersburg State University to investigate linguistic, sociolinguistic, folklore and culture issues of Siberian Latgalians (cf. the Estonian-Latvian joint conference 'Compatriots in Siberia' held in Tartu 2008, where a number of papers on various topics were presented). One of the latest results is a Latgalian-Latvian-Russian phrase book (Andronovs et al. 2008). There are a number of recordings that have been collected during expeditions to Siberia organised in 2004–2009[10], and which in the future may serve as a solid ground for the spoken sources of the modern counterpart of corpus of the Latgalians in Russia. This, in turn, raises the question of the transcription principles to be used.

[10] This research has been financed by the Russian Foundation for Humanities (project no. 07–04-00208a) and University of Latvia (project no. 2007/ZP-38).

6. 3 Diachronic Corpus of Latgalian

Nonetheless, the written tradition stretching more than 250 years back provides a good foundation for developing a diachronic corpus in the future. The first printed Latgalian book which has survived till our days, 'Evangelia toto anno' (1753), is already included in the Corpus of Early Written Latvian[11] in order to give a complete picture of the texts from the 16–18th century. 'Evangelia toto anno' laid the foundation for a second written tradition of Latvian. While the orthography of the first Latvian printed sources was based on the German orthography, the Latgalian prints were using Polish orthography. The first period of written Latgalian ended in 1865, when the ban against printing Lithuanian and Latgalian books in Latin script came into force.

6. 4 Learner Corpus of Latgalian

One might consider the development of a Learner's Corpus. Considering that there are regular winter schools organised in Siberia and regular summer schools in Latgale ('Vosoruošona' and 'Atzolys'), there is a possibility to make a collection of essays written by learners with different backgrounds (national, educational, etc.). If Latgalian is taught as a facultative course in some Latgalian schools, their data is also very valuable source.

7. Conclusions

The restricted usage and lack of some linguistical text types and genres of Latgalian affect the size, representativeness and balance of the corpus of Modern Latgalian. One of the main issues to be dealt with is the input data quality, as there is a risk of 'noisy' texts, which are mere transponations from Latvian, but not Latgalian texts.

There are enough sources that can eventually turn into a number of sub-corpora (diachronic, regional, learner's, etc.). Different corpora will serve a basis for a profound linguistic analysis and will promote the further development of the language processing tools. This will counteract the present state, where Latgalian as a lesser-used language also has fewer resources.

Last, but not the least, the Latgalian corpora will be integrated into the Latvian National Corpus.

[11] www.korpuss.lv/senie

References

Andronov, A. / Andronova, E. (2008). "Latgalian in Soviet Russia: A Pilot Model of the Linguistic Corpus of Printed Texts", in *Proceedings of the 21st Conference on Baltic Studies 'Baltic Crossroads: Examining Cultural, Social, and Historical Diversity'*. Indiana University, Bloomington, Indiana. Retrieved May 15, 2009, from http://depts.washington.edu/aabs/documents/confabstr.pdf

Andronova, E. (2007). "The Corpus of Early Written Latvian: current state and future tasks" in *Proceedings of Corpus Linguistics 2007*. Birmingham, UK. Retrieved May 15, 2009, from http://ucrel.lancs.ac.uk/publications/CL2007/paper/245_Paper.pdf

Andronova, E. / Andronovs, A. / Leikuma, L. (2008). "«Mozī draugi Sibirī» (Tomska, 1918)—pirmā ābece Sibīrijas latgaliešiem un Krievijas latgaliešu valodas korpuss" in *J. Endzelīna 135. dzimšanas dienas atceres starptautiskās zinātniskās konferences 'No skaņas un burta līdz tekstam un korpusam tēžu krājums'*, Rīga: LU LVI, 3–6.

Andronovs, A. / Leikuma, L. (2008). *Latgalīšu-latvīšu-krīvu sarunu vuordineica*. Krasnojarskys nūvoda regionaluo sabīdriskuo organizaceja «Latgalīšu kulturys centrs» / Latvīšu volūdys apgivis vaļsts agentura. Ačynskys / Reiga.

Balode, L. / Holvoet, A. (2001). "The Latvian language and its dialects" in Dahl, Ö. / Koptjevskaja-Tamm, M. (eds.) (2001). *Circum-Baltic Languages.* Volume I: Past and Present. Amsterdam / Philadelphia: John Benjamins Publishing Company, 3–40.

Biber, D. (1993). "Representativeness in Corpus Design", *Literary and Linguistic Computing*, 8 (4), 243–257.

Brejdak, A. B. (2006). "Latgal'skij jazyk" in Toporov, V. N. / Zav'jalova, M. V. / Kibrik, A. A. et al. (eds.) (2006). *Jazyki mira: Baltijskie jazyki*. Moskva: Academia, 193–213.

Džeriņš, J. / Džonsons, K. (2007). "Harvesting National Language Text Corpora from the Web" in *Proceedings of the 3rd Baltic Conference on Human Language Technologies*. Kaunas, 87–94.

Grūzītis, N. / Auziņa, I. / Bērziņa-Reinsone, S. / Levāne-Petrova, K. / Milčonoka, E. / Nešpore, G. / Spektors, A. (2004). "Demonstration of resources and applications at the Artificial Intelligence Laboratory, IMCS, UL" in *Proceedings of the first Baltic conference 'Human Language Technologies—the Baltic Perspective'*. Riga, 38–42.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London / New York: Longman.

Koncepcija (2005). *Latviešu valodas korpusa koncepcija*. Unpublished. Rīga: Latvijas Universitātes Matemātikas un informātikas institūts. Retrieved May 15, 2009, from http://www.korpuss.lv/

Kučera, K. (2007). "Mapping the Time Continuum: A Major Raison D'être for Diachronic Corpora" in *Proceedings of Corpus Linguistics 2007*. Birmingham, UK. Retrieved May 15, 2009, from http://ucrel.lancs.ac.uk/publications/CL2007/paper/27_Paper.pdf

Leikuma, L. (2008). "The beginnings of written Latgalian" in Ross, K. / Vanags, P. (eds.) (2008). *Common Roots of the Latvian and Estonian Literary Languages*. Frankfurt am Main / Berlin / Bern / Bruxelles / New York / Oxford / Wien: Peter Lang, 211–233.

Levāne, K. / Spektors, A. (2000). "Morphemic Analysis and Morphological Tagging of Latvian Corpus" in *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece, May 31 - June 2, 2000. V. 2, 1095–1098.

LPN (2008). *Latgaliešu pareizrakstības noteikumi*. Tieslietu ministrijas Valsts valodas centrs. Rīga / Rēzekne.

Milčonoka, E. (2001). "Some observations about English-Latvian Translation Equivalents in a new Bible for Europe: A Study Based on the EU legislation and its translation" in *Proceedings of COMPLEX2001 6th Conference on Computational Lexicography and Corpus Research*. Birmingham, 175–187.

Milčonoka, E. / Grūzītis, N. / Spektors, A. (2004). "Natural language processing at the Institute of mathematics and computer science: 10 years later" in *Proceedings of the first Baltic conference 'Human Language Technologies—the Baltic Perspective'*. Riga, 6–11.

Paikens, P. (2007). "Lexicon-Based Morphological Analysis of Latvian Language" in *Proceedings of the 3rd Baltic Conference on Human Language Technologies*. Kaunas, 235–240.

Seiļ, V. (1936). *Grāmatas Latgales latviešiem*. Riga: Valtera un Rapas akc. sab. apgāds.

Skadiņa, I. (2005). "Studies of English-Latvian Legal texts for Machine Translation" in Barnbrook, G. / Danielsson, P. / Mahlberg M. (eds.) (2005). *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. Continuum, 188–195.

The State Language Commission. Retrieved May 15, 2009, from http://www.president.lv/pk/content/?cat_id=8

Unesco Digital Atlas of the World's Languages in Danger of Disappearing. Work in progress, beta 0.2 version October 8, 2008. Retrieved May 15, 2009, from http://www.unesco.org/culture/ich/atlas/.

Vasiļjevs, A. (2008). "Kā veidosim Latviešu valodas nacionālo korpusu?" Speech given at CLARIN project and the National Corpus workshop on November 3, 2008. Retrieved May 15, 2009, from http://www.clarin.lv/materiali/clarin-vasiljevs.ppt.

VVL (2000). "Valsts valodas likums", *Latvijas Vēstnesis* 428/433 (1888/1893), December 21, 1999. (The English translation is available at: http://isec.gov.lv/normdok/oflanglaw.htm—retrieved May 15, 2009.)

African Language Technology: The Data-Driven Perspective

Guy De Pauw and Gilles-Maurice de Schryver

In this paper we outline our recent research efforts, which introduce data-driven methods in the development of language technology components and applications for African languages. Rather than hard-coding the solution to a particular linguistic problem in a set of hand-crafted rules, data-driven methods try to extract the required linguistic classification properties from annotated corpora of the language in question. We describe our efforts to collect and annotate corpora for African languages and show how one can maximise the usability of the (often limited) data with which we are presented. The case studies presented in this paper illustrate the typical advantages of using data-driven methods in the context of natural language processing, namely language independence, development speed, robustness and empiricism.

1. Introduction

Most research efforts in the field of natural language processing (NLP) for African languages are still firmly rooted in the rule-based paradigm. Language technology components in this sense are usually straight implementations of insights derived from grammarians. While the rule-based approach definitely has its merits (particularly in terms of design transparency) it has the distinct disadvantage of being highly language-dependent and costly to develop, as it typically involves a lot of expert manual effort.

Furthermore, many of these systems are decidedly *competence*-based. The systems are often tweaked and tuned towards a small set of ideal sample words or sentences, ignoring the fact that real-world language technology applications have to be principally able to handle the *performance* aspect of language. Many researchers in the field of African language technology are quite rightly growing weary of publications that ignore quantitative evaluation on real-world data or that report unreasonably high accuracy scores, excused by the erroneously perceived regularity of African languages.

In a linguistically diverse and increasingly computerised continent such as Africa, the need for a more economical approach to language technology is high. In this paper we outline our recent research efforts, which introduce data-driven methods in the development of language technology components and applications for African languages. Rather than hard-coding the solution to a particular NLP problem in a set of hand-crafted rules, these data-driven methods try to automatically extract the required linguistic classification properties from large, annotated corpora of natural language.

We describe our efforts to collect and annotate these corpora and show how one can maximise the usability of the (often limited) data with which we are presented. We focus on the following aspects of using data-driven approaches to NLP for African languages, and illustrate them on the basis of a few cases studies:

- **Language independence:** we show how the same technique can be used to perform diacritic restoration for a wide variety of resource-scarce African languages (Ciluba, Gikuyu, Kikamba, Maa, Northern Sotho, Venda and Yoruba).
- **Development Speed:** we illustrate how a small, annotated corpus can be used to develop a robust and accurate part-of-speech tagger for Northern Sotho.
- **Robustness:** our case study of Swahili memory-based lemmatisation shows that a data-driven technique can rival a rule-based approach not only in terms of development speed, but also in terms of classification accuracy.
- **Empiricism:** all three case studies show how language technology components can be simultaneously developed and evaluated using real-world data, offering a more realistic estimation of their usability in a practical setting.

2. Corpus Collection and Normalisation: A Language-Independent Approach to Automatic Diacritic Correction

Early work in computational linguistics was burdened by the practical limitations of computational power and storage, preventing the use of large, annotated corpora. This all changed in the late 1980s when researchers started unearthing the full use of the language corpus, using statistical approaches and machine-learning techniques. In a matter of years, rule-based approaches had fallen out of favour in the research community and the new language-independent *performance* models had taken over most of the publications in the field.

2. 1 Corpus collection

While the corpus-based approaches were readily applicable to the world's most commercially interesting languages, resource-scarce languages were left behind. By definition, these languages are low on linguistic resources, with very few digital corpora available to them, let alone annotated data. For a long time, this forced researchers working on such languages to stick to the empirically less demanding rule-based paradigm, further alienating them from the main scientific current in NLP. This is even the case for a language like Swahili: despite being spoken by more than fifty million people, it is still a lesser-used language from a language technological point of view.

The proliferation of the Internet in the urban areas of Africa, however, meant that more and more vernacular language data became available in a digital format. This not only increases the visibility of African languages in the world, but now also enables the collection of large corpora, through web crawling the available content on the Internet (de Schryver 2002).

2. 2 Corpus normalisation

Unfortunately this type of user-generated corpus material comes at a cost, since its consistency and cleanliness cannot be guaranteed. This poses a particular problem for languages that have diacritically marked characters in their orthography. Despite an increasing awareness of encoding issues and the development of specialised fonts and computer keyboards (ANLoc 2009), many digital language resources do not use the proper orthography of the language in question, with accented characters represented by their unmarked equivalents. While language users can often perform real-time disambiguation of unmarked text while reading, a lot of phonological, morphological and lexical information is lost in this way–information that could be useful in the context of language technology.

Most automatic diacritic restoration methods tackle both the actual task of retrieving diacritics of unmarked text and the related tasks of part-of-speech tagging and word-sense disambiguation (e.g. Yarowski 1994). Although complete diacritic restoration ideally involves a large amount of syntactic and semantic disambiguation, this type of analysis can typically not be done for resource-scarce languages. Moreover, these methods rely heavily on lookup procedures in large lexicons, which are usually not available for such languages.

2. 3 Grapheme-based diacritic correction

One of the first applications of machine-learning techniques to an African language technology problem was presented in (Wagacha et al. 2006) for Gikuyu and expanded in (De Pauw et al. 2007) for a wider range of African languages. The basic method, adapted from (Mihalcea 2002), uses an alternative approach to diacritic restoration: it uses a machine-learning technique operating on the level of the grapheme. The general idea of the approach is that local orthographic context encodes enough information to solve the disambiguation problem. By backing off the problem from the word level to the grapheme level, it opens up the possibility of diacritic restoration for languages that have no electronic word lists available.

The training material for our approach is a word list for the language in question that contains all the proper diacritics. This word list can be the result of selecting properly encoded documents from a web crawling session. We then identify for each language the *confusables:* those characters that can occur with or without diacritics.

The diacritic correction task is identified as a classic machine-learning task, where we associate a number of features with a given class. This is illustrated in Table 1 for the Gikuyu word *mbũri*. We first strip the word of all its diacritics. Then, for each character in the word (F), we identify a window of five characters to the left (L) and five characters to the right (R). Finally, these features are associated with a class (C), which features the correct character. Instance 3 in Table 1, for example, describes the confusable *u*, which in Gikuyu orthography can be either *u* or *ũ*. In this case, the correct class is ũ. Similarly in Instance 5, the confusable *i*, should be represented as *i* instead of ũ.

	L1	L2	L3	L4	L5	F	R1	R2	R3	R4	R5	C
1	-	-	-	-	-	**m**	b	u	r	i	-	**m**
2	-	-	-	-	m	**b**	u	r	i	-	-	**b**
3	-	-	-	m	b	**u**	r	i	-	-	-	**ũ**
4	-	-	m	b	u	**r**	i	-	-	-	-	**r**
5	-	m	b	u	r	**i**	-	-	-	-	-	**i**

Table 1: Instances for Gikuyu diacritic restoration task

Instances are extracted for each character in each word in the word list and presented to the memory-based learner TiMBL (Daelemans et al. 2004) as training material. This data is stored in memory. Diacritics can now be restored for previously unseen words by deconstructing the word in the same vein. The second confusable in the word *umbũre* for example, is represented in Table 2. Its class is unknown, but it shares nine features with Instance 3 in Table 1 (namely L1, L2, L4, L5, F, R1, R3, R4 and R5). If Instance 3 turns out to be the most similar entry in memory, its class is extrapolated and suggested as the class for the instance in Table 2.

L1	L2	L3	L4	L5	F	R1	R2	R3	R4	R5	C
-	-	u	m	b	**u**	r	e	-	-	-	**???**

Table 2: Classification of new Gikuyu word

We compiled data for a wide range of African languages that have diacritically marked characters in their orthography: the Bantu languages Ciluba (Congo), Gikuyu, Kikamba (Kenya), Northern Sotho, Venda (South Africa), the Nilotic language Maa (Kenya) and the Defoid language Yoruba (Nigeria). We applied the exact same machine-learning technique to all of the languages to perform diacritic restoration.

The experimental results are displayed in Table 3. We evaluate the performance of our system (**MBL**) on a portion of the corpus that was not used in the training of the system. We compare our results to that of a lexicon lookup approach (**LLU**), which retrieves the diacritically marked variant of a word from the lexicon induced from the training set. Whereas the LLU approach by definition fails on previously unseen words, the memory-based approach working on the grapheme level is always equipped to make a calculated guess.

Language	Types	LLU	MBL
Ciluba	20.0k	77.0	85.3
Gikuyu	9.1k	77.3	92.4
Kikamba	9.7k	79.4	91.6
Maa	22.2k	66.7	75.5
Northern Sotho	157.8k	97.6	99.2
Tshivenda	9.6k	97.7	99.4
Yoruba	4.2k	67.8	76.8

Table 3: Diacritic restoration results

The results indeed show that the memory-based approach significantly outperforms a lexicon lookup method for all of the languages, sometimes with as few as 10,000 words in the training data. This is not surprising, given the morphological richness of these languages and consequently the high number of previously unseen words in the test data. While for some languages (e.g. Northern Sotho) diacritic restoration is close to a solved problem, the restoration of tonal diacritics appeared to be more problematic on the basis of graphemic data for others (e.g. Yoruba and Maa). Nevertheless, this research showed that the same, relatively simple set of preprocessing scripts and the same machine-learning technique, can be employed to a wide range of languages on the African continent, even when relatively little data is available.[1]

3. Corpus Annotation: Rapid Development of a Robust Part-of-Speech Tagger for Northern Sotho

Our work on diacritic restoration was one of the first published attempts at applying machine-learning techniques to African language technology. Previously we had described experiments with data-driven part-of-speech (POS) taggers for Swahili (De Pauw et al. 2006), trained and evaluated on the three million-word POS-tagged part of the Helsinki Corpus of Swahili (Hurskainen 2004a).

Many researchers assume that data-driven approaches to NLP require hundreds of thousands of annotated tokens. Inspired by the encouraging results that even smaller data sets had yielded, as seen in Table 3, we decided to build a small, manually POS-tagged corpus of Northern Sotho and develop a data-driven tagger on the basis of this data (de Schryver & De Pauw 2007).

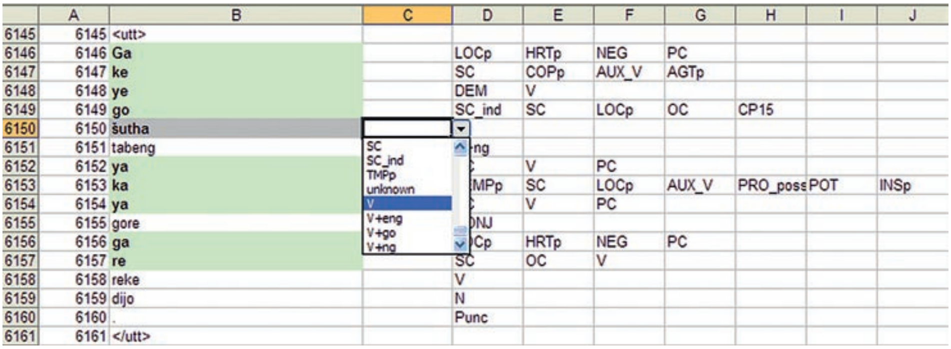
3. 1 Data annotation

The annotation was set up as an exercise in development speed. We pre-defined a list of around 50 PoS tags for Northern Sotho, but allowed annotators to refine the protocol during annotation. This on-the-fly approach enabled the organic construction of a consistent tag set, grounded in linguistic, corpus-based evidence.

Furthermore, despite the availability of dedicated annotation tools, we used Microsoft Excel as the annotation environment of choice. Installation is trivial and most computer-literate users are familiar with the Microsoft Office Suite, so that the learning curve for the annotators is favourable. While annotation is an unlikely use of a spreadsheet, Excel's cell-based approach can significantly speed up an annotation task such as PoS-tagging, also allowing on-line adjustments of the tagging protocol.

[1] A demonstration system for diacritic restoration of the languages in Table 3 can be found at http://aflat.org/?q=node/184

The annotation environment is illustrated in Figure 1: Column B contains the word to be tagged, while columns D, E, F, etc. provide the possible tags for the word, as retrieved from the TshwaneDJe HLT Northern Sotho lexical database. The correct tag for ambiguous words (highlighted in light-grey) is selected by the annotator. An additional drop-down box in Column C is available if the correct tag is not featured in columns D, E, F, etc. This is by definition the case for previously unseen words (highlighted in dark-grey). Should earlier annotations need to be adjusted for some reason, Column B can be sorted alphabetically, while the indices in Column A ensure the original order of the document can be restored.



	A	B	C	D	E	F	G	H	I	J
6145	6145	<utt>								
6146	6146	Ga		LOCp	HRTp	NEG	PC			
6147	6147	ke		SC	COPp	AUX_V	AGTp			
6148	6148	ye		DEM	V					
6149	6149	go		SC_ind	SC	LOCp	OC	CP15		
6150	6150	šutha								
6151	6151	tabeng		ng						
6152	6152	ya		C	V	PC				
6153	6153	ka		MPp	SC	LOCp	AUX_V	PRO_poss	POT	INSp
6154	6154	ya		C	V	PC				
6155	6155	gore		NJ						
6156	6156	ga		Cp	HRTp	NEG	PC			
6157	6157	re		SC	OC	V				
6158	6158	reke		V						
6159	6159	dijo		N						
6160	6160	.		Punc						
6161	6161	</utt>								

Figure 1: Excel sheet containing the initial POS-tagging material for the annotator

Restricted to a total annotation time of a mere 10 person-hours, the design of the annotation environment nevertheless maximised the amount of annotated data. After post-processing the data, we obtained a manually tagged corpus of more than 10,000 words, in a format ready to be used as training material for a data-driven tagger:

(1) Ke_SC a_PRES eletša_V ._Punc

While a 10,000-word-tagged corpus is indeed modest, compared to the million-word corpora available for English (Marcus et al. 1993), the experimental results (Section 3.3) show that even a small annotated data set can yield an accurate data-driven PoS-tagger.

3. 2 MaxTag

We used the annotated data to train and evaluate a POS-tagger based on the machine-learning technique of maximum entropy (Berger et al. 1996). Rather than the stock maximum entropy tagger, MXPOST (Ratnaparkhi 1996), we used a self-constructed POS-tagger, called MaxTag, which acts as a front-end to the general machine-learning package Maxent (Le 2004).

MaxTag takes as its input POS-tagged data (e.g. example (1)) and extracts for each word in the corpus a number of features that are possibly relevant to the disambiguation problem. Similar to the diacritic restoration approach, MaxTag uses a windowing approach to describe linguistic events. Instead of working on the character level however, MaxTag describes the problem on the word level, extracting for each word in the corpus an instance that contains both contextual and orthographic features. Each instance is then associated with the POS-tag for that word.

	Instance	**Tag**
1	[‚W-1=#‘,‚T-1=#‘, **‚FW=Ke‘**,‚FT=SC_COPp‘,‚W+1=a‘,‚T+1=SC_PRES_PC_DEM_OC_HRTp‘,‚P1=K‘,‚S1=e‘, ‚P2=Ke‘,‚S2=Ke‘,‚CAP‘]	**SC**
2	[‚W-1=Ke‘,‚T-1=SC‘, **‚FW=a‘**, FT=SC_PRES_PC_DEM_OC_HRTp‘,‚W+1=eletša‘,‚T+1=V‘,‚P1=a‘,‚S1=a‘]	**PRES**
3	[‚W-1=a‘,‚T-1=PRES‘, **‚FW=eletša‘**,‚FT=V‘,‚W+1=.‘,‚T+1=Punc‘,‚P1=e‘,‚S1=a‘,‚P2=el‘,‚S2=ša‘,‚P3=ele‘, ‚S3=tša‘]	**V**
4	[‚W-1=eletša‘,‚T-1=V‘, **‚FW=.‘**,‚FT=Punc‘,‚W+1=#‘,‚T+1=#‘,‚P1=.‘,‚S1=.‘]	**Punc**

Table 4: Four instances for Example (1)

Example instances are displayed in Table 4. The focus word (**FW=**) is associated with previous and subsequent words (**W±*n*=**) and tags (**T±*n*=**) and a list of possible tags for the word itself (**FT=**). MaxTag also allows for the inclusion of character clusters as morphological features towards disambiguation, which is valuable for processing morphologically rich languages.

On the basis of these instances, the maximum entropy machine learner constructs a statistical model that optimally relates features to classes. The advantage of using maximum entropy for this problem is that instances do not need to have a value for all features, making it more robust for sparse data sets.

3. 3 Experimental results

To evaluate the tagger, we performed 10-fold cross validation. The corpus is divided into 10 slices. In each experiment, one slice is used as the evaluation set, while the other nine are used as training data. We distinguish between known words (words in the evaluation set that are present in the training data) and unknown words (words in the evaluation set, not occurring in the training data). The latter category of words averages to about 8 % of the words in a typical evaluation set. The experimental results (Table 5) show that, despite the minimal amount of training data, the tagger is able to significantly outperform a baseline tagger (unigram probabilistic tagger). It achieves an overall tagging accuracy of 93.5 %. Nevertheless, the more modest score for unknown words indicates that more annotated data can still significantly improve the accuracy of the tagger.

	Known	**Unknown**	**Total**
Baseline	75.8	35.1	73.5
MaxTag	95.1	78.9	93.5

Table 5: Accuracy scores for the baseline method and MaxTag (all values in %)

This is corroborated by the learning curve experiments we conducted. In these experiments we started out with a POS-tagger trained on just 1/10 of the available training data (roughly 1000 words) and added 1/10 of the training data in each subsequent experiment. The result for this experiment can be found in the learning curve graph, displayed in Figure 2. The graph shows that the learning curve is still linear and that we can still gain quite a bit of tagging accuracy by collecting more annotated data. Future research will therefore concentrate on the semi-automatic development of new annotated data. The more-than-encouraging experimental results show that the Northern Sotho version of MaxTag can provide an invaluable tool in this endeavour.[2]

[2] A demonstration system of the Northern Sotho POS-tagger can be found at http://www.aflat.org/?q=node/177

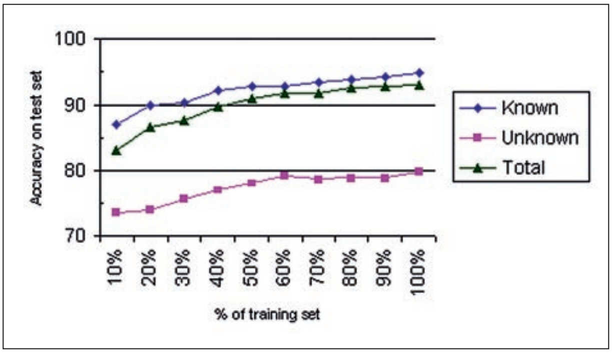


Figure 2: Graph for learning curve experiments

4. Corpus Extraction: Robust and Accurate Morphological Analysis of Swahili

The Helsinki Corpus of Swahili, HCS (Hurskainen 2004a) is the only large-scale annotated corpus of a Bantu language available to date. Annotation layers include POS-tagging and lemmatisation. It is important to note, however, that these have been generated by an automatic finite-state method, SALAMA (Hurskainen 2004b) and have not been manually cross-checked nor empirically evaluated. In this section, which draws on the work of De Pauw and de Schryver (2008), we present experiments that allow for a direct comparison between a meticulously designed rule-based approach to morphological analysis (SALAMA) and an alternative based on the machine-learning technique of memory-based learning (MBSMA).

4. 1 Data preparation

To construct a memory-based system for morphological analysis, we require morphologically annotated data. This is however not available for Swahili, but we can go a long way by extracting the necessary information from HCS, lemmatised using the SALAMA morphological analyser.

In HCS, every word is lemmatised, as illustrated in examples (2) and (3):

(2)	ulikanusha	kanusha
(3)	ulikonzia	anza

We can use this information to perform pattern-matching and match the lemma to the word form. Through this operation we can automatically induce a morphologically segmented surface and lexical representation of the word form, in which we distinguish a prefix-group ([P]), the root morpheme ([R]) and a suffix group ([S]). In some cases, this is straightforward, like for the entry in example (2) which can easily be transformed into example (4):

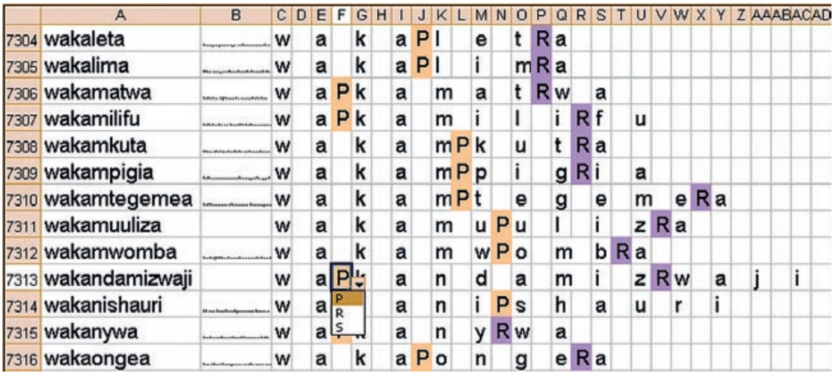
(4) ulikanusha kanusha → Surface: uli[P] + kanusha[R]
→ Lexical: uli[P] + kanusha[R]

For the entry in example (3), this leads to the creation of a bound root morpheme *anz-* in the surface representation, associated with the full lemma *anza* in the lexical representation:

(5) ulikonzia anza → Surface: uliko[P] + anz[R] + ia[S]
→ Lexical: uliko[P] + anza[R] + ia[S]

Using this method we automatically extracted a morphological database of 97,000 entries from HCS. Since HCS has been lemmatised using an automated method, quite a few erroneous and inconsistent lemmatisations can be observed in the data. We therefore randomly extracted 10 % of the data from the morphological database and had it manually annotated according to the prefix-root-suffix protocol illustrated in examples (4) and (5). The availability of this manually annotated, gold-standard evaluation set does not only allow us to cross-check the accuracy of our system on clean data, but also enables a post-hoc quantitative evaluation of the rule-based approach used to annotate HCS.

Similarly to the annotation approach described in Section 3.1, we again used Microsoft Excel as the annotation environment. The annotation sheet seen in Figure 3, lists each word on a separate row. The word form itself is listed in Column A. Column B contains a sentence extracted from HCS, illustrating that word form in context. The minimised sentence can be displayed in full by double-clicking on the cell. Columns C and onwards list the individual characters of the word form from Column A, separated by blank cells. Each blank cell has a drop-down box available with three options: P (end of prefix group), R (end of root group) and S (end of suffix group). The annotator can quickly move through the annotation process using only the keyboard or mouse clicks.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
7304	wakaleta		w		a		k		a	P	l		e		t	R	a													
7305	wakalima		w		a		k		a	P	l		i		m	R	a													
7306	wakamatwa		w		a	P	k		a		m		a		t	R	w		a											
7307	wakamilifu		w		a	P	k		a		m		i		l		i	R	f		u									
7308	wakamkuta		w		a		k		a		m	P	k		u		t	R	a											
7309	wakampigia		w		a		k		a		m	P	p		i		g	R	i		a									
7310	wakamtegemea		w		a		k		a		m	P	t		e		g		e		m		e	R	a					
7311	wakamuuliza		w		a		k		a		m		u	P	u		l		i		z	R	a							
7312	wakamwomba		w		a		k		a		m		w	P	o		m		b	R	a									
7313	wakandamizwaji		w		a	P	k		a		n		d		a		m		i		z	R	w		a		j		i	
7314	wakanishauri		w		a	P R S			a		n		i	P	s		h		a		u		r		i					
7315	wakanywa		w		a				a		n		y	R	w		a													
7316	wakaongea		w		a		k		a	P	o		n		g		e	R	a											

Figure 3: Excel sheet containing the morphological material for the annotator

In this way, the surface representation of the morpheme boundaries is annotated. In a second annotation step, the lexical representations of the roots, thus the actual lemmas, are double-checked and corrected where necessary.

4. 2 Memory-based morphological analysis

The memory-based Swahili morphological analyser reuses and adapts the basic methodology coined in van den Bosch and Daelemans (1999), which has been successfully applied to morphologically rich(er) languages such as Dutch (De Pauw et al. 2004) and Arabic (van den Bosch et al. 2007).

We use the dataset described in Section 4.1 as our primary information source. Analogous to the method described in Sections 2.3 and 3.2, we use a windowing approach to represent the data. Instead of using characters (Section 2.3) or words and tags (Section 3.2), we describe this problem at the level of the syllable. This is a more appropriate level of description when dealing with Bantu morphology than the character level, originally used in van den Bosch and Daelemans (1999).

We describe each syllable in the word, associated with a context on the left-hand side and a context on the right-hand side. This is illustrated for the word *ulikoanzia* in Table 6. Here each syllable is linked to a class. The syllable *ko* in Instance 3 marks the end of the prefix-group, while syllable *a* in Instance 7 marks the end of the suffix group. The syllable *zi* (Instance 6) marks the end of the root group and receives an extra instruction that the root needs to be repaired to the full lemma *anza* by deleting the *i* (part of the suffix group) and adding an *a* to the bound morpheme *anz-*. As was the case for the diacritic restoration task, new instances are classified by comparing them to the ones in memory and extrapolating the class of the most similar instance.

	L	L	L	L	L	F	R	R	R	R	R	CLASS
1	–	–	–	–	–	**u**	li	ko	a	nzi	a	0
2	–	–	–	–	u	**li**	ko	a	n	zi	a	0
3	–	–	–	u	li	**ko**	a	n	zi	a	–	P
4	–	–	u	li	ko	**a**	n	zi	a	–	–	0
5	–	u	li	ko	a	**n**	zi	a	–	–	–	0
6	u	li	ko	a	n	**zi**						R+a
7	li	ko	a	n	zi	**a**	–	–	–	–	–	S

Table 6: Syllable-based instances extracted from the morphological database

4. 3 Experimental results

We are most interested in the accuracy of the morphological analyser on previously unseen words: how well is the system able to morphologically segment and lemmatise unknown word forms? To investigate this, we perform blind testing, withholding a 10 % partition of the data to evaluate the system. As the evaluation set, we naturally use the manually annotated gold standard evaluation set, described in Section 4.1.

The gold-standard evaluation set also allows us to quantify the accuracy of SALAMA, the rule-based approach used to lemmatise the Helsinki Corpus of Swahili. We will follow the standard approach of using word-error rate (WER) as our primary evaluation metric. It expresses the accuracy on the word-level, that is, how many words have not been completely correctly segmented and lemmatised. In other words, the lower the word-error rate (WER), the better the system.

The experimental results in Table 7 show that the memory-based approach (**MBSMA**) can be observed to outperform SALAMA, establishing a small reduction in WER on surface-level segmentation and a more substantial reduction for full lemmatisation (i.e. full restoration of the underlying lemma).

	Segmentation of the surface representation	Further lemmatisation
	WER	**WER**
SALAMA	11.7%	12.0%
MBSMA	11.6%	11.7%

Table 7: Accuracy scores for SALAMA and MBSMA on the manually annotated evaluation set

This result may be surprising: how can a data-driven approach outperform the system that was used to create its information source? The answer to this question lies in the generalisation capabilities of the machine-learning technique. As previously mentioned, and as further illustrated by the SALAMA results in Table 7, quite a few erroneous analyses can be found in the annotation of HCS. Rather than completely mimicking the properties of the data the machine-learning approach uses to train its model, it implicitly generalises over the data and filters out the noise. This eventually generates a more accurate lemmatiser for the data in question.

The biggest advantage is the robustness of the memory-based approach: it does not rely on any kind of underlying lexicon of root forms or lemmas. When presented with an unknown word form or even a word form for a previously unseen lemma, the memory-based approach will *degrade gracefully* and guess the lemma with a surprisingly high degree of accuracy.

To the best of our knowledge, the research results presented in this section describe the first attempt at building a data-driven morphological analyser for a Bantu language. We have demonstrated how this system can be properly and quantitatively evaluated with relatively little manual effort, and experimental results show that the method compares favourably to a meticulously designed rule-based technique, even when it is trained on the basis of its output. Defining the problem of data-driven morphological analysis on the level of the syllable, rather than on the character level, furthermore showed how techniques typically designed with Indo-European language processing in mind, can be adjusted to work for Bantu languages as well.[3]

5. Conclusion: An Empirical Approach to African Language Technology

In this paper we presented an overview of on-going work on applying data-driven techniques to natural language processing of African languages. We demonstrated the **language-independent** aspects of data-driven NLP by applying the same technique to the problem of diacritic correction of a varied array of African languages. The goal of such a system goes well beyond simple diacritic restoration: the orthography of most African languages is (morpho-)phonological in nature with a mostly unambiguous mapping between phoneme and grapheme. A good diacritic restoration method, in other words, basically amounts to a robust grapheme-to-phoneme conversion method that can be used as a front-end for text-to-speech systems.

[3] A demonstration system of the Swahili lemmatiser can be found at http://www.aflat.org/?q=node/241

We then demonstrated how data-driven techniques can result in the **rapid development** of a part-of-speech tagger for Northern Sotho. Rather than investing time in designing extensive tagging protocols and painstakingly implementing expert knowledge, we showed how a small, annotated data set, constructed in about 10 hours, can already yield an accurate part-of-speech tagger. This tagger can then serve as the basis of future annotation efforts, further unlocking the language technology potential of this resource-scarce language.

We finally showed how a **robust** memory-based lemmatiser can be constructed on the basis of automatically annotated data. This research showed how previous rule-based efforts can go hand in hand with a data-driven approach and help construct a more accurate lemmatiser that is inherently capable of analysing previously unseen word forms, even when the underlying lemma is unknown. The lemmatiser is currently being used as a preprocessing module in the context of machine translation for the language pair English—Swahili (De Pauw et al. 2009).

Possibly the most important aspect of our research from a methodological point of view is its inherent **empiricism**. Working in the data-driven paradigm automatically enables quantification of research results, something that up to now has all too often been ignored in research efforts in the field of computational linguistics for African languages. Our experimental results do not only serve to showcase the strength of our approach, but more importantly help to indicate those areas that need to be further developed. This has served to create a more competitive research field, with many recent publications adapting and improving on the approaches described in this paper (Faaß et al. 2009; Groenewald 2009).

Our research efforts constitute the first thorough exploration of data-driven methods for African language technology, and experimental results show that this is indeed the way forward if we are to re-introduce African languages on the scientific agenda of the NLP research community. Furthermore, in the context of Africa's linguistic diversity, as well as the resource-scarceness of the languages in question, we believe the data-driven paradigm, with its language independence, fast development phase and its focus on creating robust *performance* models of language, is the most appropriate approach to African language technology.

Acknowledgements

The research presented in this paper was partly funded by the SAWA BOF UA-2007 project. The first author is funded as a Postdoctoral Fellow of the Research Foundation - Flanders (FWO). We wish to thank Peter Waiganjo Wagacha for his continued scientific input, as well as the annotators without whom data-driven NLP would be impossible.

References

ANLoc (2009). Retrieved February 24, 2009, from http://www.africanlocalisation.net

Berger, A. L. / Della Pietra, S. / Della Pietra, V. J. (1996). "A Maximum Entropy Approach to Natural Language Processing", *Computational Linguistics*, 22(1), 39–71.

Daelemans, W. / Zavrel, J. / van den Bosch, A. / van der Sloot, K. (2004). *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide*. ILK Technical Report 04–02. Tilburg University.

De Pauw, G. / de Schryver, G.-M. (2008). "Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes", *Lexikos* 18, 303–318.

De Pauw, G. / de Schryver, G.-M. / Wagacha P. W. (2006). "Data-driven part-of-speech tagging of Kiswahili" in *Proceedings of Text, Speech and Dialogue, 9th International Conference*. Berlin: Springer Verlag, 197–204.

De Pauw, G. / Laureys, T. / Daelemans, W. / Van Hamme, H. (2004). "A comparison of two different approaches to morphological analysis of Dutch" in *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology.* Barcelona: ACL, 62–69.

De Pauw, G. / Wagacha, P. W. / de Schryver, G.-M. (2007). "Automatic diacritic restoration for resource-scarce languages" in *Proceedings of Text, Speech and Dialogue, 10th International Conference.* Berlin: Springer Verlag, 170–179.

De Pauw, G. / Wagacha, P. W. / de Schryver, G.-M. (2009). "The SAWA Corpus: a Parallel Corpus English—Swahili" in *Proceedings of the First Workshop on Language Technologies for African Languages*. Athens: ACL, 9–16.

de Schryver, G.-M. (2002). "Web for/as Corpus: A Perspective for the African Languages", *Nordic Journal of African Studies*, 11(2), 266–282.

de Schryver, G.-M. / De Pauw, G. (2007). "Dictionary Writing System (DWS) + Corpus Query Package (CQP): The case of TshwaneLex", *Lexikos*, 17, 226–246.

Faaß, G. / Heid, U. / Taljard, E. / Prinsloo, D. J. (2009). "Part-of-Speech Tagging of Northern Sotho: Disambiguating Polysemous Function Words" in *Proceedings of the First Workshop on Language Technologies for African Languages*. Athens: ACL, 38–45.

Groenewald, H. J. (2009). "Using Technology Transfer to Advance Automatic Lemmatisation for Setswana" in *Proceedings of the First Workshop on Language Technologies for African Languages*. Athens: ACL, 32–37.

Hurskainen, A. (2004a). *HCS 2004 - Helsinki Corpus of Swahili*. Compilers: Institute for Asian and African Studies. University of Helsinki and CSC.

Hurskainen, A. (2004b). "Swahili Language Manager: A Storehouse for Developing Multiple Computational Applications", *Nordic Journal of African Studies*, 13(3), 363–397.

Le, Z. (2004). *Maximum Entropy Modeling Toolkit for Python and C++*. Technical Report. Retrieved February 24, 2009, from http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

Marcus, M. / Santorini, B. / Marcinkiewicz, B. (1993). "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*, 19(2), 313–330.

Microsoft. (2009). Retrieved February 24, 2009, from http://www.microsoft.com/

Mihalcea, R. F. (2002). “Diacritics restoration: Learning from letters versus learning from words” in *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics.* Berlin: Springer Verlag, 339–348.

Ratnaparkhi, A. (1996). “A Maximum Entropy Model for Part-of-Speech Tagging” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Somerset: EMNLP, 133–142.

TshwaneDJe HLT. (2009). Retrieved February 24, 2009, from http://tshwanedje.com/

van den Bosch, A. / Daelemans, W. (1999). “Memory-based morphological analysis” in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Maryland: ACL, 285–292.

van den Bosch, A. / Marsi, E. / Soudi, A. (2007). “Memory-based morphological analysis and part-of-speech tagging of Arabic” in *Arabic computational morphology: Knowledge-based and empirical methods*. Berlin: Springer Verlag, 203–219.

Yarowsky, D. (1994). “A comparison of corpus-based techniques for restoring accents in Spanish and French text” in *Proceedings of the Second Annual Workshop on Very Large Corpora*. Kyoto: COLING, 19–32.

Wagacha, P. / De Pauw, G. / Githinji, P. (2006) “A grapheme-based approach for accent restoration in Gĩkũyũ” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa: ELRA, 1937–1940.

The Learner Corpus: Description and Research

Karin Aijmer

> *Corpora of the language of children, geriatrics, non-native speakers, users of extreme dialects and very specialised areas of communication (like the heraldic blazon, the knitting pattern or the auctioneer's patter) should […] be designated special corpora because of the unrepresentative nature of the language involved. (Sinclair 1995: 24)*

1. Introduction

Geoffrey Leech has characterised the concept of the learner corpus as "an idea whose hour has come" (Leech 1998: 16). The corpus revolution has now also reached the pedagogical sphere and the start of the collection of the language written and spoken by learners. We can expect such 'lesser-used languages' to provide a challenge for the corpus linguist because they contain a number of unusual features. The object of this presentation is to discuss some issues raised by the advent of learner corpora on the corpus scene. So what is a learner corpus and what can one do with it?

A (computer) learner corpus is designed according to special criteria that make it suitable for research in second language acquisition and for teaching purposes, for example, to develop pedagogical tools and to improve curriculum design. Granger (2002: 7) suggest the following definition of learner corpora, which builds on Sinclair's (1996) definition of corpora:

> Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance.

Learner corpora provide a number of challenges both for the corpus linguist and for researchers in second language acquisition (SLA) who still have to learn to use them. The methods used in learner corpus research have 'come under fire' from SLA researchers, and the globalisation of English has led to a sometimes heated discussion about norms and the identification and the role of the native speaker (cf. Granger 2009: 19).

The discussion will focus on the following questions:

- What problems are associated with compiling and designing a corpus of learner English?
- How should we understand notions such as the native speaker, non-native speaker, advanced learner?
- Is the comparison with a norm in learner corpus analysis a strength or a weakness?
- Learner corpora can be both written and spoken. What are the special problems involved in building a learner corpus of spoken English?

The discussion is organised as follows. Section 2 looks at the building of learner corpora within different international projects. Section 3 discusses the identification of the native (and non-native) speaker in a globalised world. Section 4 discusses the annotation of learner corpora. Section 5 deals with the comparative methodology used in learner corpus research and the choice of a comparable native speaker corpus. Section 6 illustrates the methodology (contrastive interlanguage analysis) on the basis of a case study and Section 7 links learner corpus research to contrastive analysis with special reference to transfer. Section 8 deals with the compilation and design of spoken learner corpora. Section 9 (the conclusion) takes a look at the future and the possibility to build up new and bigger learner corpora.

2. Compiling and Designing a Learner Corpus

The best known learner corpus project is probably the international project International Corpus of Learner English, initiated by Professor Sylviane Granger at the University of Louvain-la-neuve in 1990 (Granger 1998, 2002). The learners in the project are advanced learners of English. The advanced learner is operationalised as a learner with three or four years' studies of English at the university level (for the

Swedish students the advanced learner has two years or four terms of higher studies). The data are argumentative essays. As a result of the research carried out by the project's members, we are now beginning to get a good picture of 'learner style' and the problems encountered by learners.

Moreover, much effort has been made to develop new corpus tools tailored for learner corpora, in particular, error tagging systems. As Granger writes (2009: 28), "It is time to start thinking seriously of a standardised mark-up and annotation system and a purpose-built architecture for storing, annotating and searching learner corpora, which should be both powerful and user-friendly."

Unlike many other learner corpora, the ICLE Corpus is generally available to the academic community. The first release of the corpus on CD-ROM (which also contains a handbook and manual) contains subcorpora (of roughly 200,000 words) representing learners whose native language is French, German, Bulgarian, Japanese, Dutch, Czech, Finnish, Italian, Polish, Russian, Spanish or Swedish. There are 3,640 essays and a total of 2,500,353 words. A new edition of the corpus is planned for 2009 with contributions from additional national subgroups.

The Swedish Component of the ICLE corpus (SWICLE) was compiled as part of the ICLE project at the universities of Lund and Göteborg by Bengt Altenberg and myself. There is also a Swedish spoken learner corpus. I will later discuss the compilation and design of spoken learner corpora, which provide different problems from written learner corpora (Section 8).

There are other learner corpora, mostly of English as the interlanguage, although access to them is restricted. Large learner corpora have, for example, been collected by publishing companies such as Longman and Cambridge University Press. The Cambridge Learners' Corpus consists of texts produced by learners with different national backgrounds and is coded for errors and inappropriate uses. It has, for instance, been used in the writing of *The Cambridge Grammar of English* (Carter & McCarthy 2006). The Longman Learners' Corpus is the basis for warning notes in advanced learners' dictionaries (such as the Common Error Study Notes in *Longman Exams Dictionary* 2006). In addition, learner or non-native speaker corpora have been compiled in many Asian countries such as Hong Kong, Taiwan, China and Japan. Useful overviews of available learner corpora are found in Pravec (2002) and Nesselhauf (2004).

3. Learner Corpora and the Native Speaker

Learner corpora having English as their target cover EFL and ESL. However, English is also spoken as a non-native language in Singapore, India, Kenya, and so forth (the New Englishes), and by people who have different native languages as a new international language. The English spoken by a person from Singapore differs from the Standard, but can be regarded as a variety of its own.

Other challenges come from the increased use of English by non-native speakers of English. According to Seidlhofer (2000), we need to pay more attention to English as a *lingua franca*, that is, English as a means of communication among people who have different mother tongue backgrounds. How does English as a *lingua franca* affect the role of the native speaker as a benchmark for the knowledge of English the learner should achieve? The discussion about a native speaker norm easily spills over into a political issue.

For Graddol (1999: 68) the concept of 'native speaker' is no longer relevant "[…] large numbers of people will learn English as a foreign language in the 21st century […]. But will they continue to look towards the native speaker for authoritative norms of usage."

On the other hand, as Davies (2004: 441) points out, "[t]he theoretical debate about native speakers may be unresolved, but in the daily life of language teaching and testing resolution is necessary and agreement on a model and a goal is required." However, what model?

Great care must be taken in the design of the learner corpus and in the choice of the pedagogical model or a norm with which learners can be compared. These issues will be discussed below under the following headings:

- annotating the learner corpus and the use of software (Section 4)
- learner corpus variables and comparability with a native speaker corpus (Section 5)

4. Learner Corpora, Annotation and Software Tools

An advantage of corpora is that there is text retrieval software such as WordSmith-Tools (Scott 1999) that can be used for searches as well as for sorting the search terms according to their frequency and distribution in the text.

However, using these tools to analyse a learner corpus is not without its problems. As Granger (2002: 15) writes, "[…] researchers should be aware that when using non-native language data, some degree of caution should be exercised with these tools, whether they are used to analyse lexis or grammar."

An example is part-of-speech (POS) tagging. The results of the automatic tagging must be treated with caution since the outcome of the tagging is affected by the number of errors made by the learners. Automatic searches using a POS tagger are even more unreliable if the corpus consists spoken English, due to hesitations, pauses, overlapping and unclear speech.

Syntactic tagging or parsing of the corpus may encounter even greater problems (Granger 1984: 21–25). At present it is therefore not possible to automatically search for syntactic constructions such as passives or clefts. Many of the searches carried out can be characterised as semi-automatic (a combination of automatic searches and manual weeding out). For example, in Annelie Ädel's study of metadiscourse (Ädel 2006), she made an automatic search for the personal pronouns *I, we, you* that explicitly indicate the writer's and reader's presence in the text, and manually removed examples that were not metatextual. Another possibility is to use pre-established lists for example of conjuncts such as *however* or *in any case* (Altenberg & Tapper 1984) or of intensifiers (Lorenz 1999). Such lists (based on frequencies) can for instance be found in reference grammars such as Biber et al. (1999) or Quirk et al. (1985).

The disadvantage of the latter approach is that only the most frequent examples are included (or only the examples referred to in the grammars and handbooks) and that only phenomena which can be treated as lexical are searchable in this way. An advantage is that the frequencies and distribution of a lexical element can easily be compared across different corpora.

What makes learner corpora special is that they contain many more errors than corpora representing native speakers. In addition to other types of annotation, learner corpora therefore need to be error-tagged. Error-tagging provides the first step for error analysis (analysing, classifying and evaluating errors).

One part of the Swedish ICLE Corpus (50,000 words) has been error-tagged automatically using the Louvain Error Editor: UCLEE (1.2) (Dagneaux et al. 1998;

Eriksson & Mondor 2005). The process involves two steps. First, native speakers manually corrected the texts and proposed an alternative to the misused word or phrase. The Swedish project members inserted a tag for each error from a fixed error tag set, together with the proposed correction. The appropriate error tag was inserted in the text by clicking at the appropriate tag on the error-tagging menu. The error tag is inserted before the error and the correct target item is added afterwards within dollar signs. The following is an example from the SWICLE Corpus analysed by the Swedish team:

> At some level, Sweden has probably always had immigrants. Therefore, it is important that we work out careful plans of how to deal with the situation so it does not get out of hand, but according to some it already has. The media sometimes (GVN[1]) gives $give$ you the impression that most Swedes would rather have no immigrants at all interfering with our business. Whether this is true or not is difficult to tell. Sweden has decided to accept a certain (XNUC[2]) amount $number$ of immigrants, and the way they are introduced into our society (GVAUX[3]) is going to $will$ decide whether they will be seen as a 'problem' or no.

An advantage of the error-tagging system is that one can search for a single word (e.g. *its*) and retrieve both the correct and erroneous examples:

- …litation, especially if (FS) its $it is$ used very early in a (GNC) persons…
- On the other hand if (FS) its $it is$ put (FS) in to $into$ action too…
- deals with America and its Jewish population, it contains a lot of…

A problem with the tagger has been that many errors could be analysed in different ways: as syntactic, stylistic, or a word error. Moreover the project is time-consuming; it is dependent on having native speaker judges and has therefore so far only been carried out on a small part of the corpus.

[1] GVN: code for number errors

[2] XNUC: code for (lexico-grammatical) errors relating to the count/non-count status of nouns

[3] GVAUX: code for (Grammar Verb) auxiliary errors

5. Corpus Comparability

Much research on advanced learners' language has been carried out in areas such as metadiscourse, pragmatics and text organisation. In these areas, the differences between native and non-native speakers are easy to overlook because they are not always to be regarded as errors. According to Pawley and Syder (1983: 91), "[s]peakers and writers have their own ways of speaking and writing about things, people and events which may involve words or constructions which are not only grammatical but also native-like." 'Non-native-like selections' can be uncovered by comparing native and non-native speakers. The results of the L1/L2 comparison ('contrastive interlanguage analysis' Granger 1998: 12ff) are discussed in terms of overuse (L2 speakers use a word or construction more frequently than L1 speakers), underuse (L2 speakers use a word or construction less frequently than L1 speakers) and misuse. The ICLE corpus has also been designed to enable comparisons of learners with different mother tongue backgrounds (L1, L2, L3, etc.). However, so far there are few studies of this kind.

Within the ICLE project the learner data are compared with data from LOCNESS (Louvain Corpus of Native English Essays), a corpus of native English essays functioning as a reference norm for the different subcorpora within the ICLE project. However, it is difficult to get a perfect match for the ICLE corpora. The LOCNESS corpus is, for instance, made up of British (LOCNESS British) and (LOCNESS American) data. The essays are both literary and argumentative and are therefore associated with different genre conventions.

The essay writers both in the ICLE corpora and LOCNESS are students mostly in their twenties. Is it perhaps more appropriate to use the writings of professional experts as a descriptive and prescriptive norm than the data provided by the novice writers in LOCNESS? Alternatively, the learners' argumentative texts in the ICLE corpus could be compared with newspaper editorials produced by professionals. On the other hand, it can be argued that it would be unfair as well as unrealistic to use the writings by professional writers as a reference norm (Lorenz 1999; Swales 1990; cf. also Ädel 2006: 205).

One of the strengths of learner corpus research is that different standards of comparison can be chosen depending on one's pedagogical goal. No corpus has it all, but all comparisons bring out some characteristic features of learner English. We can, for instance, also make quantitative comparisons with the major standard corpora of written English. Moreover, in *the Longman Grammar of Written and Spoken English* (Biber et al. 1999) we find corpus-based quantitative information about a number of lexical or grammatical phenomena in different genres, such as academic text, journalism and conversation in both British and American English.

5. 1 Comparing apples with apples and pears with pears

In Granger's words, "the usefulness of a learner corpus is proportional to the care that has been exerted in controlling and encoding the variables" (Granger 2004: 9). Such variables are, for instance, the learning setting and the linguistic proficiency of the learner.

The information about the learners and the task situation is stored in special files linked to the text and can be called up if one wants to study the influence of a particular variable. In the ICLE project, 20 different variables related to the learner and to the task conditions are included in the so-called 'learner profile'. The variables have to do with the characteristics of language learners (mother tongue, other foreign languages, level of proficiency) and the task-setting situation (time limit, use of reference tools, exam).

Unless we consider all the relevant factors in the writing situation that can affect the results, we may draw the wrong or too sweeping conclusions about the learners' writing style. Many features are shared by native speakers and non-native speakers, for instance age, learning context, level, medium and genre. There are also differences having to do with the learners' knowledge of other foreign languages and practical experience of the language.

Moreover, learners represent many different educational practices, as Osborne (2008: 145) suggests:

> But given the range of educational practices in different countries, the value placed on language learning, individual motivation, opportunities for using the language outside the classroom, and so on, the fact that two learners have apparently similar profiles does not guarantee that they have reached similar levels of competence. It is helpful therefore to have as much information as possible about their linguistic proficiency.

The importance of considering the impact of the task-setting situation can be illustrated with a Swedish example. Swedish 'learner style of writing' has been described as 'over-emphasis' or 'over-involvement', reflected in the overuse of personal pronouns and questions. Ädel (2008) has, however, shown that the learners' writing style and strategies used should be explained by the task-setting conditions under which the writing takes place. Ädel sums up (2008: 153):

> We must conclude that the extreme role of involvement found in the SWICLE Corpus is due to the fact of the writers not having enough time to make the text formal or 'written-like' and not having access to model texts.

6. *Of course*—a case study of overuse and underuse

Of course is one of the most frequent adverbs expressing modality ('self-evidence') and it is highly multifunctional. It is overused by Swedish learners in their argumentative writing (Aijmer 2005) and is therefore a good example of non-nativeness. Thus the non-native speakers (represented in SWICLE) used *of course* 188 times to be compared with native speakers (in LOCNESS) who used it only 58 times. However, the quantitative data need to be complemented by a qualitative analysis (a more detailed description of the differences and their causes).

When we analyse the functions of *of course* we get a different picture of its overuse and underuse. Although *of course* was generally overused by learners, they used the following pattern much less frequently than native speakers.

> Of course, a lot more could be said about how to create best-sellers but time and space does not allow that. (SWICLE)

In the LOCNESS Corpus 57 % of the examples were of this kind (i.e. used with a concessive function), compared with only 27 % in the learner corpus.

On the other hand, the learners overused structures such as *must of course, may of course, should of course, would of course* (the emphatic *of course*), a construction which is typical of conversation only.

Moreover, *of course* can sound patronising or over-confidential in argumentative writing because of its implications that the writer 'knows best' (and that the reader should know it as well). In spoken language on the other hand, *of course* ('as you know', 'as we both know') has the interpersonal function to smooth the way for a socially harmonious relationship characterised by shared values and solidarity.

The example illustrates that we need to move from a quantitative to a qualitative analysis of the learner data. The qualitative analysis shows among other things that learners are not sensitive to differences between the written and the spoken mode.

7. The Learner Corpus and Transfer

Learner corpora have features characteristic both of the learner language in general and of the specific mother tongue spoken by learners. We can therefore expect learner corpus research corpora to throw light on what has been referred to as 'the transfer mystery' (Gilquin 2008: 4).

The transfer explanation is closely associated with contrastive similarities and differences between languages. In the 1950s and 1960s, transfer was used as a factor predicting and explaining learner errors. However, the link between contrastive analysis and 'errors' in second language acquisition came to be regarded as spurious when it appeared that we cannot predict learners' errors on the basis of the contrastive analysis (Ringbom 1994). However, transfer analysis based on learner corpora stands on firmer ground and can be combined with a contrastive corpus-based analysis that makes use of parallel corpora.

A parallel corpus is a bilingual (or multilingual) corpus consisting of translations in two directions. The English-Swedish Parallel Corpus (Altenberg & Aijmer 2000), for instance, contains translations from English to Swedish and from Swedish to English in the same proportions. The contrastive analysis shows the correspondences between lexical elements and constructions in the two languages.

Transfer is a possible explanation not only for 'errors' but also for overuse and underuse. Altenberg (2002) has for instance suggested that transfer may be involved in Swedish learners' overuse of the analytic causative construction with *make* in English (make something happen) and the corresponding underuse of a synthetic causative verb (e.g. break NP). The transfer explanation is suggested by the frequency of the analytic construction in Swedish and is especially likely if other alternatives are more marked (synthetic causative verbs or other causative constructions) and therefore less familiar.

8. Compiling a Corpus of Spoken Learner Language

Corpora of spoken learner language are newcomers in learner language research. They take a longer time to collect; the actual recording is a cumbersome process and it is time-consuming and sometimes difficult to transcribe the texts. An interview will for instance take about five hours to transcribe even when a simple, mainly orthographic transcription system is used. Spoken learner corpora are therefore in general smaller than the written ones.

The LINDSEI project (Louvain International Database of Spoken English Interlanguage) is an umbrella project incorporating several spoken learner corpora. The project was launched at the University of Louvain-la-Neuve in 1995, and is the result of the cooperation between several universities (Belgium, Spain, Germany, Japan, Italy, China, Sweden) (de Cock 2004; de Cock et al. 2006).

Like its written sister corpus, the LINDSEI Corpus is targeted at advanced learners. Each national corpus consists of 50 interviews (a little more than 100,000 words) conducted by a native speaker on a topic such as a country the interviewee has visited, or a film he/she has seen which has made a great impression. In addition, the interviewee comments on a series of pictures which could be interpreted as a story. The information about the learner is given in a general learner profile and can be used to restrict the searches to a particular category of speakers (see Appendix 1). The transcription includes symbols for speaker turns, overlapping speech, (filled and unfilled) pauses, unclear passages, truncated words, coughing and laughter. (see Appendix 2). Moreover, pairs such as *going to/gonna* and *don't know/dunno*, where one of the forms is non-standard, have been distinguished in the transcription.

We are fortunate to have a parallel corpus consisting of interviews with students at the University of Lancaster (the LOCNEC Corpus[4], compiled by Sylvie de Cock according to the same principles as the LINDSEI Corpus) (de Cock 2004). It is therefore possible to contrast native and non-native speakers under comparable conditions. Comparisons can also be made with native speakers in the London-Lund Corpus of Spoken English (LLC) or the Bergen Corpus of London Teenager language (the COLT Corpus). The LOCNEC consists of 170,000 words, including the interviewer. The Swedish component of the LINDSEI Corpus consists of 104,000 words. Of these, 70,000 are provided by the interviewee. The access to spoken learner corpora is restricted at the moment to the teams working on the corpora. However, both the native and the non-native corpus will be available on CD in 2009.

[4] 'Louvain Corpus of Native English Conversation'

The LINDSEI Corpus makes it possible to study phenomena which are characteristic of spoken English, such as hesitation or other dysfluencies and pragmatic markers in learners' spoken production. Hesitation signals and pragmatic markers contribute to fluency and near-nativeness and are therefore important in an SLA or FLT perspective (cf. also Gilquin 2008).

The differences between native and non-native speakers involve both frequencies of pragmatic markers or pauses and how they are used. Simone Müller (2005) has, for example, found that German learners use pragmatic markers less frequently and in different ways from English native speakers.

The learner corpus has above all been used as a resource for studying the collective attainment of learners in comparison with a native control group. However, the collective data from the learners' performance may hide large individual differences, and therefore not give a correct indication of how well a construction has been acquired by the learners. The following example is taken from a study where I looked both at learners as a group and at the individual variation (Aijmer, forthcoming). The purpose was to study the use of the fixed phrase (discourse marker) *I don't know* (or *dunno*) in spoken interlanguage compared with native speaker's use of the phrase. The use of *dunno* is illustrated in the following example:

> <B> so I was helping out there and then they gave me free food <laughs> and I just I dunno I just stayed and lived with them <\B>
> <A> what's a . Thai: breakfast look like <\A>
> <B> they: eat: sticky rice <\B>
> <A> for breakfast <\A>
> <B> yeah . with I dunno I can't remember what it's called some m= m= mashy thing with l= small fish <\B>
> <A> mhm <\A>
> <B> and . dunno cold<?> . with the sticky rice <\B>
> (LINDSEI)

As seen from the example, *dunno* (*I dunno)* is used to signal the speaker's hesitation. However, the use of *dunno* (or its more literal form 'I don't know') was unevenly distributed among the non-native speakers. The most interesting results can be summarised as follows. Some learners (9 learners out of 50) do not use *I don't know* (or *I dunno*) at all. Five learners stood out because they only used *I dunno* and 17 learners only used *I don't know.* Two speakers were in fact responsible for most of the examples of *dunno.* They were also the speakers who used the pragmatic marker most frequently (12 and 13 times respectively).

We can now also use learner corpora to compare native and non-native speakers in speech (LINDSEI/LOCNEC) and in (argumentative) writing (SWICLE/LOCNESS). Börjesson (forthcoming) found, for example, that intensifiers were used differently by native and non-native speakers both in writing and in speech. In writing (SWICLE), Swedish learners overused *totally, a bit, not so, more or less,* and *very much.* Many more intensifiers were underused (or not used at all). In the Swedish component of LINDSEI, on the other hand, the overused intensifiers were *so, kind of, extremely* and *a little bit*. *Really* was underused in relation to *very*.

9. Conclusion

To sum up, learner corpus analysis is both quantitative and qualitative. Advanced learners make different lexical and grammatical selections than native speakers, which can be described in quantitative terms as overuse and underuse. A closer qualitative analysis suggests for example that (Swedish) advanced learners are not aware of the differences between speech and writing or the linguistic requirements associated with genre; as a result, they make choices which make them sound overemphatic or patronising.

Learner corpora provide a number of challenges. For example, in the present-day globalised world it is not easy to define the non-native speaker. The non-native speaker may learn English in natural settings or in the classroom (or in both ways). The English spoken or written by non-native speakers also has many features in common with English as a *lingua franca* or with new Englishes. Also the native speaker's role as a standard or norm in learner corpus analysis has come under fire.

In principle, the samples of learner language can be analysed with the same tools as we use for standard corpora (e.g. POS taggers), but the results of the tagging may be less reliable because of the frequency of errors. Error-tagging annotation is specific to spoken and written learner corpora and has resulted in a revival for error analysis. However, spoken learner language is difficult to annotate because of the speakers' hesitations and uncertainties. On the other hand, such phenomena are interesting for second language acquisition research because they reflect the speaker's needs to search for words and plan what to say on-line.

Learner corpus research is heading off in several directions. In the Swedish project, we were interested from the beginning in the possibilities provided by compiling both a parallel (or translation) corpus and a learner corpus (or learner corpora). Linking the two types of research can contribute to the ongoing debate over the extent of transfer in learner language and how it should be identified and described.

To come back to the starting point of this article, the time has come to make use of the resources provided by learning corpora. Learner corpus analysis is still a young addition to other types of specialised corpus analyses. However there is no doubt that the use of learner corpora is now making inroads into second language acquisition research. This has benefited both SLA research and learner corpus analysis. However the comparative methodology will also probably be under fire from SLA researchers in the future, warning us that comparisons are at a risk unless the influencing factors are carefully controlled.

The focus in this presentation has been on advanced learners and near-nativeness in the Swedish ICLE corpus. This is a purely synchronous corpus covering only one variety of English. It is a fairly small corpus that is targeted at advanced learners, uncovering the features of near-nativeness. On my wish list for the future are interactive or multilingual learner corpora in addition to the corpora that are available. Moreover, we can hope that a closer collaboration between corpus linguistics and SLA theorists will spark off the compilation of new corpora which are longitudinal rather than synchronous, and which focus on young learners in addition to advanced ones.

We can also expect a broadening of learner corpus studies to other non-native speaker varieties. The English spoken by advanced learners should be studied in its own right and be compared with other varieties, such as English as *lingua franca* or the new varieties of English.

Appendix 1

<Learner profile>
Date of learner profile: 99–05-11
Text code:
Learner
Surname: X nr 1
First names: X
Age: 25
Sex: female
Nationality: Swedish
Native language: Swedish
Father's native language: Swedish
Mother's native language: Swedish
Language(s) spoken at home: Swedish
Percentage of each if more than one:
Education
Current studies: D-level English
Current year of study:
Institution: English Department Göteborg University
Medium of instruction
English only (yes/no): no
Swedish only (yes/no): no
Both (yes/no): yes
Other language(s) (specify):
Years of English at school: 11
Years of English at university: 2
Stay in an English-speaking country
Which: London
Where:
When: 1994–1995
How long: a year
Other languages (in decreasing order of proficiency): Spanish (OK) German (OK)
Danish (poor)

Appendix 2

<p nt=SW nr=SW021>

<A> give you a few minutes to look through them <\A>

<B> yeah <\B>

<A> and then you can er pick out<?> one and talk for about three or f=<stammer> four minutes and then we'll just carry on the conversation <\A>

<B> yeah <\B>

<A> and then I'm gonna show you just one or two pictures which you can just talk about <\A>

<B> all right <\B>

<A> so it's you know it's not a test at [all

<B> [no no <\B>

<A> it's just to get you to talk <\A>

<B> okay [yeah <\B>

<A> [okay good well have a look at er these three topics and choose one that you think you know you might be able to talk a little bit about and then we'll take it from there okay just spend a few minutes looking at that first <\A>

<B> can I take erm a book instead of a film [or a play <\B>

<A> [yeah that's fine [that's fine <\A>

<B> [good <breath> because em . erm on the C level er I think you was my teacher <\B>

<A> yes <\A>

<B> and we read some a book called .. <sigh> I don't remember what <X> the title but it [was about <\B>

<A> [wasn't Dirty Weekend is it <\A>

<B> no no [it wa= was it was about er twins

<A> [no <laugh> <\A>

<B> in i= . India India <\B>

<A> oh [that's right <\A>

<B> [two boy= boys t= <?> er a boy and a girl <\B>

<A> er it's the[i:] er God of Small Things <\A>

<B> yeah yeah that's cool<?> <\B>

<A> mhm <\A>

<B> and I'm a twin myself so I really liked it and <breath> i= it it em yeah yeah I <\B>

<A> perhaps you could talk about that you know why you thought it was particularly good or . erm <\A>

<B> mhm because I I like the way she had described how twins . talked and understood each other because it was exactly like me and my sister <\B>

<A> [uhum <\A>

<B> [so I just er . yeah it moved me in a way . yeah and it was er yeah it was nice written and everything so <\B>

<A> what did you recognise from your own childhood you know [er you being

<B> [erm <\B>

<A> a twin <\A>

<B> er it was mostly the thoughts . that they: they had s= same<?> the same thoughts and sh= she er the sister could understand er the boy and . yeah it was just erm .

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English.* Amsterdam/Philadelphia: Benjamins.

Ädel, A. (2008). "Involvement features in writing: do time and interaction trump register awareness?" in Gilquin, G. / Papp, S. / Belén Díez-Bedmar, M. (eds.) (2008). *Linking up contrastive and learner corpus research.* Amsterdam: Rodopi, 35–53.

Aijmer, K. (Forthcoming). "«So er I just sort I dunno I think it's just because…» - A corpus study of *I don't know* and *dunno* in learners' spoken English."

Altenberg, B. (2002). "Causative constructions in English and Swedish: A corpus-based contrastive study" in Altenberg, B. / Granger, S. (eds.) (2002). *Lexis in Contrast: Corpus-based approaches.* Amsterdam: Benjamins, 97–116.

Altenberg, B. / Aijmer, K. (2001). "The English-Swedish Parallel Corpus: A resource for contrastive research and translation studies" in Mair, C. / Hundt, M. (eds.) (2001). *Corpus linguistics and linguistic theory. Papers from the 20th International Conference on English Language Research on Computerized Corpora* (*ICAME 20*) *Freiburg im Breisgau 1999.* Amsterdam and Philadelphia: Rodopi.

Altenberg, B. / Tapper M. (1998). "The use of adverbial connectors in advanced Swedish learners' written English" in Granger, S. (ed.) (1998). *Learner English on Computer.* London and New York: Longman, 80–93.

Biber, D. / Johansson, S. / Leech, G. / Conrad, S. / Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* London: Harlow.

Börjesson, V. (Forthcoming). "Reinforcing and attenuating modifiers of adjectives in Swedish advanced learners' English: A comparison with native speakers".

Carter, R. / McCarthy, M. (2006). *Cambridge Grammar of English. A Comprehensive Guide. Spoken and Written English Grammar and Usage.* Cambridge: Cambridge University Press.

Dagneaux, E. / Denness, S. / Granger, S. (1998). "Computer-aided error analysis", *System. An International Journal of Educational Technology and Applied Linguistics,* 26(2): 163–174.

Davies, A. (2004). "The native speaker in applied linguistics" in Davies, A. / Elder, C. (eds.) (2004). *Handbook of Applied Linguistics.* Oxford: Blackwell, 431–450.

de Cock, S. (2004). "Preferred sequences of words in NS and NNS speech", *Belgian Journal of English Language and Literatures (BELL)*, New Series: 2, 225–246.

de Cock, S. / Granger, S. / Petch-Tyson, S. (2006). *The Louvain International Database of Spoken English Interlanguage- LINDSEI.* Retrieved June 15, 2009, from http://cecl.fltr.ucl.ac.be/Cecl-Projects/Lindsei/lindsei.htm

Eriksson, A. / Mondor, M. (2005). "Computer-aided error tagging: A description of a new type of error analysis" in Larsson Ringqvist, E. / Valfridsson, I. (eds.) (2005). *Forskning om undervisning i främmande språk. Rapport från workshop i Växjö. 10–11 juni 2004* (Research on teaching foreign languages. Proceedings from a workshop in Växjö, June 10–11, 2004). Växjö: Växjö University Press, 88–95.

Gilquin, G. (2008). "Hesitation markers among EFL learners: Pragmatic deficiency or difference?" in Romero-Trillo, J. (ed.) (2008). *Pragmatics and Corpus Linguistics: A Mutualistic Entente.* Berlin: Mouton, 119–149.

Gilquin, G. / Papp, S. / Belén Díez-Bedmar, M. (2008). *Linking Up Contrastive and Learner Corpus Research.* Amsterdam: Rodopi.

Graddol, D. (1999). "The decline of the native speaker" in Graddol, D. / Meinhof, U. H. (eds.) (1999). *English in a Changing World. The AILA Review* 13, 57–68.

Granger, S. (1996). "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora" in Aijmer, K. / Altenberg, B. / Johansson, M. (eds.) (1996). *Languages in contrast: Papers from a symposium on text-based cross-linguistic studies,* Lund 4–5 March 1994. Lund University Press: Lund, 37–51.

Granger, S. (1998). "The computer learner corpus: a versatile new source of data for SLA research" in Granger, S. (ed.) (1998). *Learner English on Computer*. London and New York: Longman, 3–18.

Granger, S. (2002). "A Bird's eye view of learner corpus reserach" in Granger, S. / Hung, J. / Petch-Tyson, S. (eds). (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam and Philadelphia: Benjamins.

Granger, S. (2009). "The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation" in Aijmer, K. (ed.) (2009). *Corpora and Language Teaching.* Amsterdam and Philadelphia: Benjamins, 13–32.

Leech, G. (1998). "Preface" in Granger, S. (ed.) 1998. *Learner English on Computer.* London and New York: Longman, XIV–XX11.

Longman Exams Dictionary. (2006). London: Longman.

Lorenz, G. (1999). *Adjective Intensification: Learners versus Native Speakers. A Corpus Study of Argumentative Writing*. Amsterdam and Atlanta: Rodopi.

Müller, S. (2005). *Discourse Markers in Native and Non-native English Discourse.* Amsterdam: Benjamins.

Nesselhauf, N. (2004). "Learner corpora and their potential for language teaching" in Sinclair, J. (ed.) (2004). *How to use Corpora in Language Teaching*. Amsterdam: Benjamins, 125–152.

Osborne, J. (2008). "Adverb placement in post-intermediate learner English: a contrastive study of learner corpora" in Gilquin, G. / Papp, S. / Belén Díez-Bedmar, M. (eds.) (2008). *Linking Up Contrastive and Learner Corpus Research.* Amsterdam: Rodopi, 127–146.

Pawley, A. / Frances S. (1983). "Two puzzles for linguistic theory: native-like selection and native-like fluency" in Richards, J. C. / Schmidt, R. (eds.) (1983). *Language and Communication.* London: Longman.

Pravec, N. A. (2002). "Survey of learner corpora", *ICAME Journal* 26, 81–114.

Quirk, R. / Greenbaum, S. / Leech, G. / Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Ringbom, H. (1994). "Contrastive Analysis" in Asher, R. Y. / Simpson, J. M. Y. (eds.) (1994). *Encyclopedia of Linguistics.* Oxford: Pergamon Press, 737–742.

Scott, M. (1999). *WordSmith Tools version 3*. Oxford: Oxford University Press.

Seidlhofer, B. (2000). "Mind the gap: English as a mother tongue versus English as a lingua franca", *Views* 9/1. University of Vienna Department of English, 51–68.

Sinclair, J. (1995). "Corpus typology- A framework for classification" in Melchers, G. / Warren, B. (eds.) (1995). *Studies in Linguistics.* Stockholm: Almquist & Wiksell International, 17–33.

Sinclair, J. (1996). "EAGLES". *Preliminary recommendation on corpus typology.* http://www.ilc.pi.it/EAGLES96/corpustyp/corpustyp.html

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Websites

International Corpus of Learner English (ICLE). http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm

Swedish Component of the International Component of English (SWICLE). http://www.englund.lu.se/corpus/corpus/swicle.html

The Louvain International Database of Spoken English Interlanguage (LINDSEI).http://cecl.fltr.ucl.ac.be/Cecl-Projects/Lindsei/lindsei.htm

VALICO: An Online Corpus of Learning Varieties of the Italian Language

Elisa Corino

Learner corpora are the product of methods and technologies used by traditional corpus linguistics that are applied to the study of a particular variety of language, that is, the interlanguage spoken by learners. A modern learner corpus should allow for queries of the learner's characteristics—mother tongue, age, other foreign languages spoken—and should facilitate the investigation of a single student's language, an entire class of students, or defined group of learners. Users of a learner corpus range from linguists to teachers; therefore, among its essential features are usability and accessibility. All the possible needs of each group of potential users should be taken into account, and—aside from the PoS research—users should have access to the entire set of texts that refer to the learner's profile, as well as to whatever materials that yield comparisons with native corpus. In our paper, we will discuss our experiences with the methodological and technical challenges in learner corpus research by outlining the critical history of VALICO (choice and elicitation of texts, tagging and architecture, metadata, and so forth) to give insight into the most current research, results and materials derived from its analysis and exploitation.

1. Corpus Linguistics and Learner Corpora

In recent years, corpus linguistics has become a well-established discipline within the general framework of linguistic studies. Its findings have contributed to a number of areas of study, from general language studies to lexicography, and the vast amount of empirical data which has been collected has given significant insight into the different aspects of language varieties. Corpus linguistics has also been enriched by a particular type of data collection dealing with the unique character of a particular type of speaker: *the learner*.

Recently, linguists have witnessed the flourishing of an empirical approach to the study of language. Learner corpora are the result of the methods and technologies used by traditional corpus linguistics and applied to the study of the interlanguage spoken by learners. The availability of these kinds of language resources has contributed significantly to linguistic theory because they have provided more precise and innovative descriptions of the language by highlighting what is difficult for the learner. This allows researchers to infer the characteristics of the language in a more systematic and reliable way, rather than through time-consuming manual search or (oftentimes) unreliable intuition. The creation of comparable corpora in several foreign languages may potentially shed light on what proves challenging for foreign language learners, irrespective of the specific foreign language they are studying.

Learner corpora whereas originally created through an English as a Foreign Language (EFL) environment, riding the wave of the ICLE (International Corpus of Learner English) project, which spread all over the world, thus becoming a potent model which offers standard patterns that can be followed by any researcher interested in the field of language learning and acquisition. Such corpora have affected syllabi and have contributed to making pedagogical materials more sensitive to learners' needs by taking their learning challenges into account. The essential vocabulary described within the Common European Framework of Reference (CEFR) is based on the evidence of learner corpora, and even dictionaries have profited from the findings of learner corpora-related research: notes based on learner corpora and typical mistakes made by learners have even been introduced in the latest edition of the Longman Dictionary of Contemporary English and the Cambridge Advanced Learner's Dictionary.

1. 1 Native corpora and learner corpora

Corpora generally differ according to some very broad variables, that is, who makes them, how they are made, and what they contain. Depending on the 'author', a corpus serves different purposes and is therefore designed according to a number of criteria; publishing companies and research groups interested in various aspects of a language may have very different requirements and therefore may need different corpora. Furthermore, a corpus can either be uploaded as a free online resource, or put on the market. It can be part of a set of corpora, or it can be a single corpus issued for a unique purpose. It can be 'user-friendly' or just 'linguist-friendly'. Content is an extremely relevant factor when building a corpus, as it can heavily influence results and findings. For instance, when collecting newspaper data, one should be certain that the data covers different themes and geographical areas and that it comes from different time periods, otherwise the results will yield altered extractions, displaying only a defined set of lexis directly linked to topical news (for example, elections or a current war). Some corpora are meant to be as representative as possible, and thus include both spoken and written materials, as well as scientific and daily language; others aim at studying only a particular set of linguistic features within a framework of Language for Specific Purposes (LSP).

Overall, corpus linguists agree on the essential characteristics[1] that a corpus should have, as well as how these features should be presented in the actual resources according to the research being carried out. Corpora give linguists the opportunity to observe a representative number of instances of a linguistic phenomenon through the PoS search and the use of concordances (the presentation of a word in its context). The first relevant feature one should consider is *size*. Size depends on the different research questions that are being asked, that is, what language genre or skill will be investigated, and what more or less frequently-occurring language phenomena will be in focus. The smaller the corpus, the more specialised and carefully controlled it should be. The *choice of topics* may also be an important variable to consider, as it affects the 'aboutness' of the text, the use of lexis, and morpho-syntactical structures. Finally, a corpus should be as *representative as possible* of the language being investigated.

As learner corpora focus on a particular variety of language—the learners' interlanguage(s)—they present particular challenges that need to be approached with proper methods. Many of the requirements that apply to native corpora also apply to learner corpora, even though some of these operations can be affected by the nature of interlanguage and by the presence of occurrences that deviate from the corresponding target language norm, for example, PoS tagging procedures.

'Representativeness' is the result of authentic texts produced during classroom activities and includes cross-sectional collections of materials that take into account various mother tongues and learner levels. Learners' characteristics, such as their levels of proficiency, contexts of learning, contexts of production, and mother tongue backgrounds, along with sociolinguistic information like age and sex, should be encoded as metadata so that they can be queried together with the linguistic search. Eventually a control or reference corpus against which learner production and errors can be compared may prove useful in the closer examination of some features of the language,

[1] All of these essential characteristics are thoroughly discussed in Barbera et al. (2007: 25–88). Definitions of 'corpus' are systematically examined and a complete analysis of how a good resource should be is illustrated as well.

providing a number of comparative dimensions, including the target language alongside the learners' mother tongue as well as several types of learner interlanguages.

Many learner corpora currently exist which take into account all these features, but the original model is the ICLE, which consists of over 3 million words of essay writing by advanced learners of English from 21 different mother tongue backgrounds. The ICLE respects almost all the prescriptions for a good learner corpus mentioned above: different mother tongues, encoded sociolinguistic variables, and a balanced set of texts. That said, it lacks a variety of levels: it considers only proficient learners and thus does not allow for a diachronic collection ranging from beginners to intermediate learners. In addition, it is based on a single textual genre, the essay (argumentative essays and literature examination papers), and focuses mainly on collocational patterns as far as query means are concerned.

Learners' interlanguages have been investigated previously by many research projects that have focussed on oral conversations, especially in the area of acquisition. The Italian *Progetto di Pavia* contains 120 hours of interviews that became the base for the development of acquisitional scales as well as for making observations about natural learning processes. A similar approach was followed by Child Language Data Exchange System (CHILDES), the part of the TalkBank database of conversations used to study language development in children. More recently Anke Lüdeling's group at the Humboldt University in Berlin has implemented FALKO (Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache[2]), collecting written essays and summaries by advanced learners of German as a foreign language together with a control corpus of German native speakers. Compared with ICLE, FALKO contains a considerable number of mother tongues (although they are not sufficiently balanced, and the corpus has only two textual genres), which make a kind of interlinguistic comparison possible.

1. 2 Planning a learner corpus

Considering all the variables featured in 1.1, when planning a learner corpus it is fundamental to outline some basic characteristics that make it consistent and reliable. Balancing in terms of represented languages, levels, and textual types is of critical importance to the influence it has on the results of any possible comparison. In addition, the dimension of the corpus should not be forgotten, and objectives for the corpus must be made clear. Researchers need to bear in mind their goals and what they want

[2] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

to elicit, so that they may carefully select the materials and design proper tasks and instructions that will provide the intended quantitative and qualitative information on the learners' interlanguage.

From a computational perspective, the architecture of the corpus is essential, as it determines what kind of research one can do with the corpus itself. As with a native corpus, a learner corpus employs all the commonly-used procedures of electronic acquisition, encoding and markup. However, they are considerably more complex due to the non-standard nature of the learners' interlanguage. Moreover, the potential for error-tagging makes the implementation of a good learner corpus even more complex, as it requires procedures necessary to preserve the textual characteristics as accurately as possible and enable them to be queried in a fruitful and effective way.

A fundamental problem when designing language resources is how to encode linguistic content that cannot be classified in a categorical way. Both 'canonical' and 'non-canonical' sentences and words need to be annotated but, as the non-canonical elements are ungrammatical, the computer program is unable to analyse them. As Lüdeling (2007) points out, most non-canonical structures in a learner corpus can be interpreted as errors and many non-canonical structures in a corpus of spoken language or computer-mediated communication may be considered interesting features of those varieties.

Therefore, the choice of architecture matters a great deal in the construction of the most complete and consistent resource possible. According to the chosen language, computational structures can differ for morphological and syntactic tagging and can vary according to the presence of error-tagging and search possibilities. Words and sentences deviate with respect to the corresponding target language in a number of ways: there may be no agreement within the SP or the VP, prepositions could be lacking, relative clauses may not be overtly coded, spelling mistakes are typical, and the word formation rules are not always observed.

Some learner corpora have now become a kind of 'standard' in the field of language acquisition and a model for many other electronic resources that follow the same patterns and use the same tools. ICLE is now one of the most well-known and widespread learner corpora, counting on a number of partners working in EFL all over the world. Aside from a structure based on a number of sociolinguistic data, it provides detailed error-tagging both for single syntactical and morphological errors. On the other hand, FALKO (Lüdeling et al. 2005; Siemen et al. 2006; Lüdeling et al. 2007) takes into account the syntactic level alongside the PoS tagging, applying a topological structure to a multi-layer model where 'target hypothesis' about the non-canonical sentences are put forward and an alternative annotation is suggested.

Metadata plays a fundamental role within the framework of corpora implementation, and even more so when dealing with learner corpora, since a well-defined learners' profile is essential for a better understanding of the features of the interlanguage. The more detailed the profile, the more careful the queries will be, and therefore the conclusions to be drawn from the results obtained when the corpus is analysed will be more detailed as well. Among the various aspects one should consider there is, of course, the mother tongue of the learners. This will make it possible to examine subsections, for example, Spanish mother tongue learners, learners who speak some English at home, learners for whom German is the second language, and so forth. A survey of other foreign languages spoken and of the contact with the target language makes it possible to draw conclusions about transfers and odd constructions. Age, gender, socio-economical status will also make it possible to examine more sociolinguistic aspects. Eventually the query interface should enable any user to find whatever they want to investigate among the multifarious opportunities offered by the materials. The design of a learner corpus should thus take into account all the possible different users of the resource apart from specialised linguists. For instance, teachers might seek information from the corpus in order to become acquainted with the linguistic production of a particular group of learners. Linguists might investigate a particular use and occurrence of some aspect of the interlanguage. Publishers often use learner corpora to detect the most frequently occurring mistakes in order to create exercises that can help prevent errors from occurring. Moreover, one has to plan in advance if the corpus is meant to be online or offline, as the latter choice implies specific needs in terms of hardware and software equipment.

2. The Example of VALICO

VALICO (*Varietà di Apprendimento della Lingua Italiana Corpus Online*, i.e. the Online Corpus of Learner Varieties of Italian) was conceived in Turin in 2003 with a purpose to serve as a free online Italian learner corpus of foreign and second-language learners for both linguists and other users. Being PoS tagged and connected with other native corpora, it includes all the sociolinguistic information required to provide researchers with new insights into variation and acquisition. In particular, it offers a survey of textual genres that teachers of Italian around the world have their students write and demonstrates how students of different ages and mother tongues write in Italian. Therefore it can also work as a roundtable for methods to prevent the most frequent errors and avoidance strategies.

VALICO is constantly increasing in size and is on the way to becoming a well-balanced corpus both as far as the L1s and the stimuli are concerned. As one can observe from Table 1 below, it consists of almost one million words (at present the online version is 700000 tokens, but, taking into account offline material, will soon reach the million-word mark), and it includes different textual genres, elicited initially from different tasks and instructions. With regard to the other corpora mentioned so far (and besides essays, questionnaires about literature and exams), VALICO elicits its data with the help of a series of pictures and comic strips, thus offering material that is interlinguistically comparable as far as vocabulary, syntax and textual structure are concerned.

	Falko	ICLE	VALICO
Tokens (02/2009)	ca. 37.000	ca. 2 Mio.	ca. 0.7 Mio.
Language	DE	EN	IT
L1	25	17	44
Essays	(literature)	Y	Y
Questionnaires	(literature)	–	Y
Picture descriptions	–	–	Y
Exams	Y	Y	Y

Table 1: Comparing different corpora

2. 1 Planning the corpus

Planning is an essential stage of corpora implementation. On the one hand it addresses the architecture and computational structure (together with the hardware and software requirements); on the other hand it takes into account all the possible needs of the users and all the possible linguistic variables that one might query when searching the corpus. As far as architecture is concerned, VALICO is implemented on a CQP/CWB platform developed at the IMS Stuttgart, the same one as all the corpora of the suite found at “corpora.unito”[3]. Being built on similar systems, all the corpora allow the same query syntax, which makes it possible to expand the research and compare their results.

[3] www.corpora.unito.it

The structure of the corpus has been modified more than once, due to the growing needs and requirements of its users. So from its beginnings as a first pure xml version, it has transformed into a more complex structure made of macro-, meso- and micro-levels where the text was incorporated in the metadata, finally evolving into a mixed system consisting of a database of sociolinguistic variables and a CQP-based query system for the PoS tagged text[4]. In this way it is easier to manage the querying process, as the PoS tagging is still based on the same system as the other corpora.

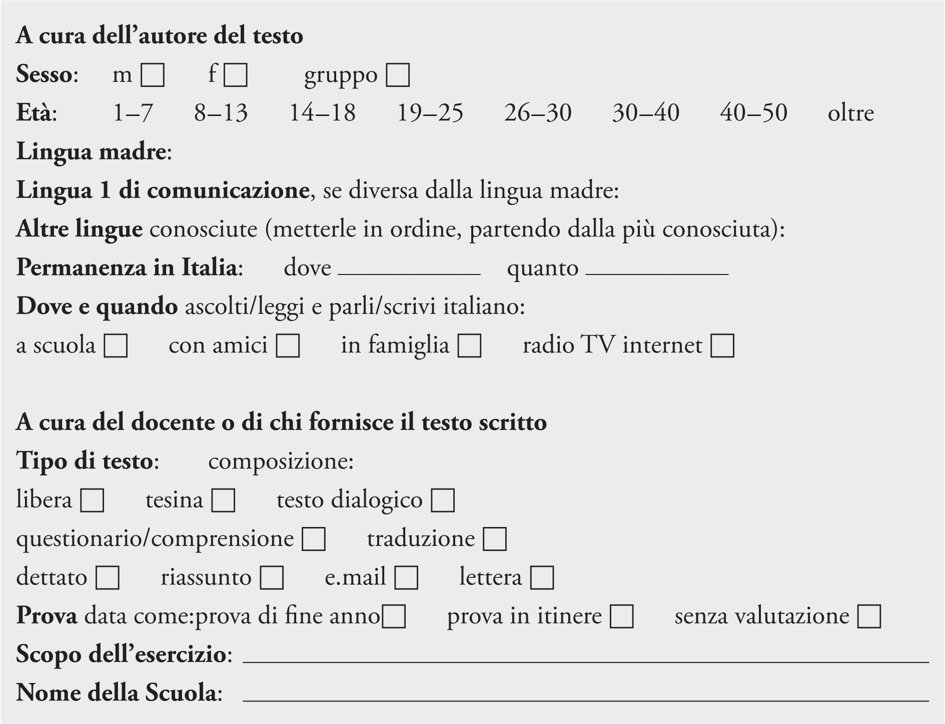
2. 1. 1 Sociolinguistic metadata

Sociolinguistic metadata are extremely important to help define the characteristics of the author, and they are far more significant when dealing with learners. Some information is essential for analysing and understanding the process of learning, the avoidance strategies, and the interference phenomena that are in progress when we consider learners' language. Furthermore, a good metadata collection makes it easier to compare different learner's profiles with respect to language level, mother tongue, and contact with the language.

VALICO relies on a large collection of sociolinguistic metadata, ranging from the author's characteristics to the textual genre. In particular, all the learners have to fill out a questionnaire (Figure 1) that requests information about their knowledge of the Italian language (including the contact they have or had with it, the places in Italy where they have lived for more than a month), and other languages they can speak. In addition, teachers are asked to give information about the assignment and the context in which the writing took place; they should inform if the task was given to be graded, if students could look up words in dictionaries, if a particular structure or set of vocabulary was suggested, and how much time they had to complete the task. This information provides a clear picture of the situation to be 'translated' into the corpus architecture in order that all users of the corpus will have it available to them.

It is also helpful to be able to query the corpus according to the name of the school or university, to find all the pieces of writing of a particular class, or even to find all the texts of a particular student. All this information is included into the header of each document (see example header in Figure 2) and a database is populated with the information.

[4] A more detailed description of the architecture can be found in Corino & Marello (forthcoming).



A cura dell'autore del testo

Sesso: m □ f □ gruppo □

Età: 1–7 8–13 14–18 19–25 26–30 30–40 40–50 oltre

Lingua madre:

Lingua 1 di comunicazione, se diversa dalla lingua madre:

Altre lingue conosciute (metterle in ordine, partendo dalla più conosciuta):

Permanenza in Italia: dove ____________ quanto ____________

Dove e quando ascolti/leggi e parli/scrivi italiano:

a scuola □ con amici □ in famiglia □ radio TV internet □

A cura del docente o di chi fornisce il testo scritto

Tipo di testo: composizione:

libera □ tesina □ testo dialogico □

questionario/comprensione □ traduzione □

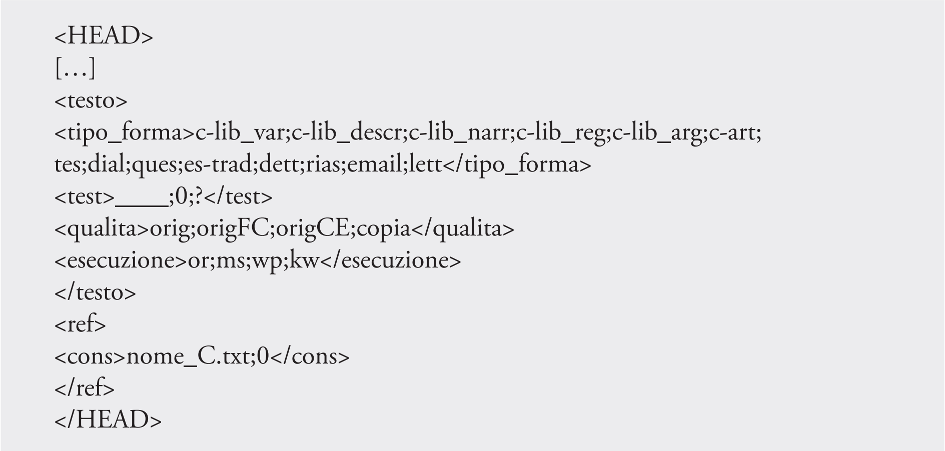
dettato □ riassunto □ e.mail □ lettera □

Prova data come:prova di fine anno□ prova in itinere □ senza valutazione □

Scopo dell'esercizio: ____

Nome della Scuola: ____

Figure 1: A segment of the questionnaire—author's characteristics



```
<HEAD>
[…]
<testo>
<tipo_forma>c-lib_var;c-lib_descr;c-lib_narr;c-lib_reg;c-lib_arg;c-art;
tes;dial;ques;es-trad;dett;rias;email;lett</tipo_forma>
<test>_____;0;?</test>
<qualita>orig;origFC;origCE;copia</qualita>
<esecuzione>or;ms;wp;kw</esecuzione>
</testo>
<ref>
<cons>nome_C.txt;0</cons>
</ref>
</HEAD>
```

Figure 2: Part of the header concerning the text's characteristics

2. 1. 2 Textual markup

VALICO relies on two kinds of transcriptions[5] (see Figure 3 for a comparison): the first, the *diplomatica* [TD], is a kind of record file where the text is reproduced as it has been written, that is, with the same new lines, blanks, indentations, and so forth. Self-corrections, insertions, illegible characters are marked out in curly brackets. This transcription is kept as a term of comparison in case doubts arise or computer problems make it difficult to work on the second version. The second transcription is the Tokenized and Marked-Up Transcription [TTM], which bears textual tags that mark VALICO as a model among the other learner corpora. For instance self-corrections <CORR> are reported and linguists can study the real process behind a linguistic performance. Data about the progression of interlanguage is thus more reliable, as one can carefully observe what the learners' doubts are and if they truly are familiar with a particular word or sentence structure. In this way one can observe the learners' uncertainties, at times understanding that what is considered as a 'horrible' mistake is in fact not as bad as it seems, because it is the result of a process of correction. It is not uncommon to discover that what is hidden under the correction is indeed correct.

Other features of the textual markup are insertions—<INS> (the parts of the text inserted between the words or over the line), anthroponyms—<anth>, toponyms—<topn> and all those characteristics dealing with the text structure, such as titles, direct discourse, and indentations.

Marta non era una ragazza qualunque… da che aveva cinque anni
abitava con suo padre e con la {sua} nuova compagna di questo {0} {alla quale} aveva odiato tante volte. [TD]

<titolo><docente><anth>Marta</anth> non era una ragazza qualunque …</docente></titolo> da che aveva cinque anni
abitava con suo padre e con la <CORR>sua</CORR> nuova compagna di questo <INS>di questo</INS> <CORR>alla quale</CORR> aveva
odiato tante volte . [TTM]

Figure 3: Textual markup [TD] vs [TTM]

[5] For the detailed transcription guidelines and textual markup, see http://www.bmanuel.org/projects/br-g00-IDX.html

All these features are then visualised according to a different key character in the result window. A significant problem when dealing with hand-written texts that have to be transcribed into electronic format is the material and markup errors. If the transcribers do not know the markup language well, they will tend to forget to properly close tags, or forget to detach punctuation; therefore, there may be a lot of *noise* in the corpus and many PoS will be declared *unknown*. To avoid such inconveniences, VALICO includes a transcription interface that helps to prevent misprint and markup errors through JavaScript and Perl control scripts. By choosing from a defined range of possibilities displayed in a drop-down menu format, and using buttons to insert the textual markup, the opportunity for making mistakes is significantly reduced.

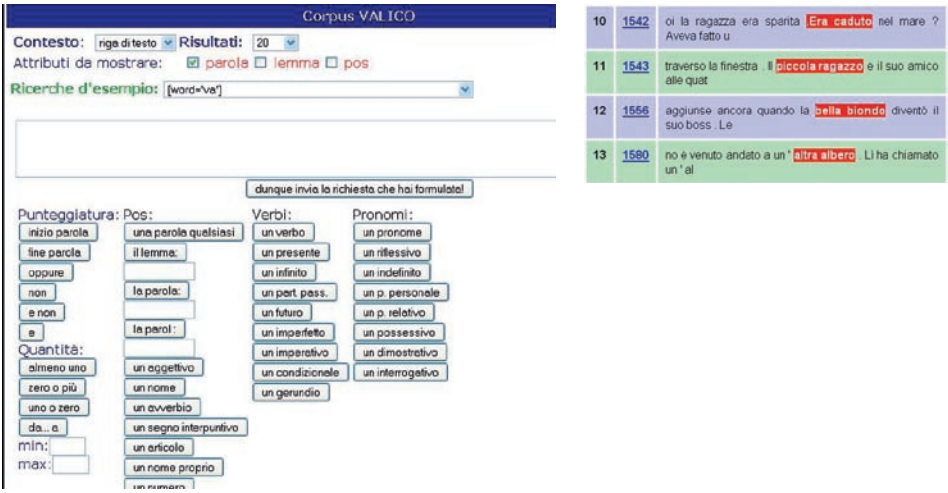
2. 1. 3 Querying the corpus

The metadata database together with the PoS tagging enable the user to cross-reference different variables within the same query, asking for a particular set of texts or a certain linguistic phenomenon. For instance, one could investigate morphological agreement, and ask for all the adjectives ending with 'a', followed by substantives ending with 'o', in texts of English female learners under fifteen years of age, who can speak French.

So, after having set the sociolinguistic variables, this will be the syntax of the query:

[pos='ADJ' & word= '.*a'] [pos='NOM' & word= '.*o']

Of course the query syntax could appear quite difficult for non-linguists, therefore a more user-friendly interface has been developed for the PoS research. As Figure 4 shows, buttons point to the Parts-of-Speech and examples are given in order to make the query easier. The research can be done within a context ranging from 5 words up to 2000. Results are shown on the right part of the screen and one can trace back the learner's characteristics by clicking on the number near the text.



Contesto: riga di testo Risultati: 20

Attributi da mostrare: ☑ parola □ lemma □ pos

Ricerche d'esempio: [word='va']

10	1542	oi la ragazza era sparita .Era caduto nel mare ? Aveva fatto u
11	1543	traverso la finestra . Il piccola ragazzo e il suo amico alle quat
12	1556	aggiunse ancora quando la bella biondo diventò il suo boss . Le
13	1580	no è venuto andato a un ' altra albero . Li ha chiamato un ' al

Figure 4: PoS query interface

2. 2 Eliciting and collecting

When VALICO was created in 2003, we decided to collect any kind of text written by the learners of Italian as a foreign language. This allowed us to rely on a substantial source for our research, but it curtailed, in a way, the possibility to compare different levels and different languages on a common base. Therefore we created a special set of exercises, specially devised to elicit texts for the corpus. We opted for the iconic stimulus, that is, a picture-based task, and asked students to narrate the story they observed in the pictures.

Our model was derived from the research carried out by Berman and Slobin (1994) into language development in children and adults. They used the story by Mercer Mayer (1969) "Frog, Where Are You?", a series of 24 pictures describing the adventures of a child and his animals. We also took into account the work done by these same Danish scholars on the *Mr. Bean* corpus, which recorded the narration of some stories played by the British actor Rowan Atkinson. The drawer sketched four comic strips (see examples in Figure 5) based on the principle of misunderstanding: the opening circumstances are always reversed at the end.[6]

[6] For a more detailed description and analysis of the usefulness of pictures to elicit written texts by foreign learners see Corino and Marello (2009).

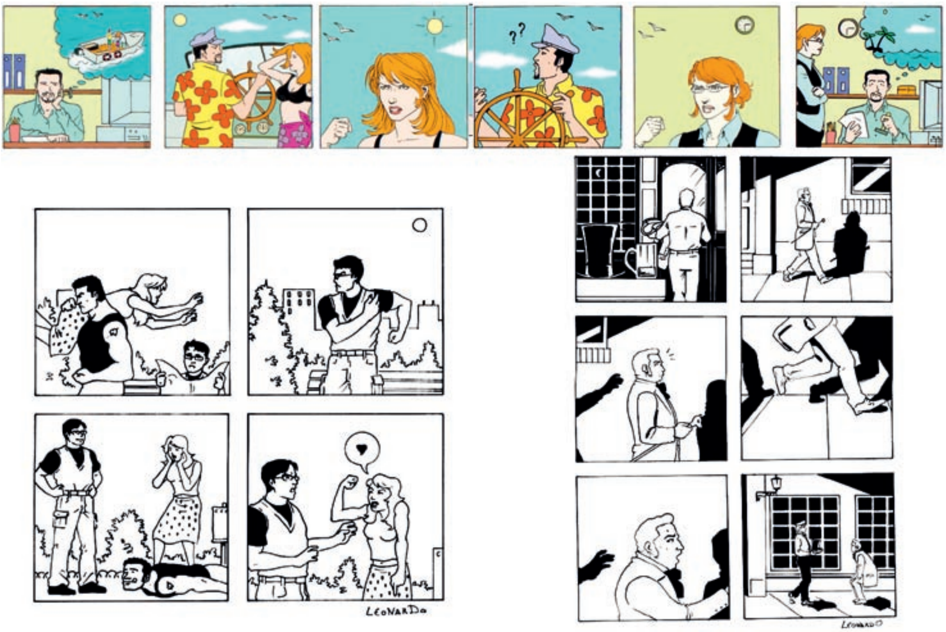
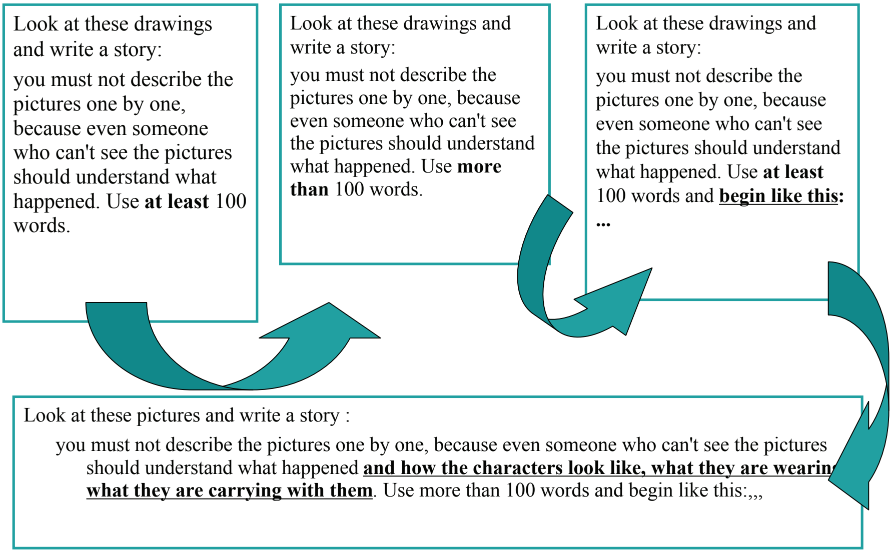


Figure 5: Some of the strips used to elicit the texts

In this way we meant to elicit adversative structures; furthermore we added an *incipit* to set the narration in the past, thus trying to bring forth past tenses.

Mindful of the *Progetto di Pavia* experience, where learners described the pictures one by one, numbered them, and produced texts that lacked cohesion markers, we asked students to narrate a story and at the same time avoid mentioning each single image. The expected length was at least 100 words. After a few texts we realised that many learners could not understand the word *almeno* ('at least') and, as it contains the word *meno* ('less'), they usually wrote less than 100 words. So we changed the instructions and we asked for *more than* 100 words. Eventually we added the request for more detailed descriptions, as learners tended to narrate the story without taking note of what was really happening in the pictures, filling their descriptions with background events and 'translating' actions that were not described by the drawings. Figures 6 and 7 illustrate the genesis of the instructions.

The process of writing the instructions is an integral part of planning a learner corpus and the results largely depend on the success of a well-formulated task. If the instructions are not clear, the students will not understand them, and the texts will be of no use.



Look at these drawings and write a story:

you must not describe the pictures one by one, because even someone who can't see the pictures should understand what happened. Use **at least** 100 words.

Look at these drawings and write a story:

you must not describe the pictures one by one, because even someone who can't see the pictures should understand what happened. Use **more than** 100 words.

Look at these drawings and write a story:

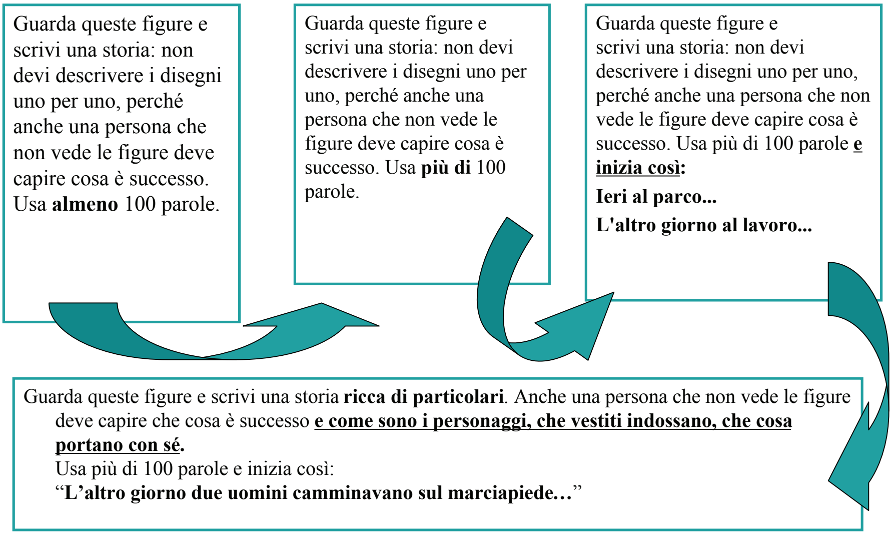
you must not describe the pictures one by one, because even someone who can't see the pictures should understand what happened. Use **at least** 100 words and **begin like this:**

...

Look at these pictures and write a story :

you must not describe the pictures one by one, because even someone who can't see the pictures should understand what happened **and how the characters look like, what they are wearin what they are carrying with them**. Use more than 100 words and begin like this:,,,

Figure 6: Genesis of the instructions (English translation)



Guarda queste figure e scrivi una storia: non devi descrivere i disegni uno per uno, perché anche una persona che non vede le figure deve capire cosa è successo. Usa **almeno** 100 parole.

Guarda queste figure e scrivi una storia: non devi descrivere i disegni uno per uno, perché anche una persona che non vede le figure deve capire cosa è successo. Usa **più di** 100 parole.

Guarda queste figure e scrivi una storia: non devi descrivere i disegni uno per uno, perché anche una persona che non vede le figure deve capire cosa è successo. Usa più di 100 parole **e inizia così:**

Ieri al parco...

L'altro giorno al lavoro...

Guarda queste figure e scrivi una storia **ricca di particolari**. Anche una persona che non vede le figure deve capire che cosa è successo **e come sono i personaggi, che vestiti indossano, che cosa portano con sé.**

Usa più di 100 parole e inizia così:

"**L'altro giorno due uomini camminavano sul marciapiede…**"

Figure 7: Genesis of the instructions (original version)

2. 3 Balancing and comparing

Balance is one of the essential features of a good corpus: it is even truer when dealing with a learner corpus. The possibility of relying on a set of texts of comparable size, textual genre, vocabulary and morphosyntactical structures is crucial for valid and reliable research. VALICO offers a great variety of texts ranging from questionnaires to argumentative essays. This makes VALICO a clear sample of what instructors of Italian as a foreign language teach their students all over the world (although the texts are not always comparable). As they are not all the same size, there is not always the same text for each language.

The comic strips answer the need for a common comparable base: potentially the same vocabulary, the same incipit, and the same textual structure. Therefore we decided to create a sub-corpus specifically devised to contain the balanced version of VALICO (it can be freely consulted at www.valico.org) based on the picture descriptions. Here we paid special attention to the size of each 'monolingual' section, building a corpus containing the same number of texts for each L1 group of learners and the same number of descriptions for each iconic stimulus.

As with ICLE and FALKO, VALICO also has a control corpus of native speakers. VINCA (Varietà di Italiano di Nativi Corpus Appaiato) is based on texts written by Italian mother tongue speakers and is elicited beginning with the same tasks. It has also been implemented with the same architecture as VALICO. The texts have been primarily collected in three Italian regions, which we believe could be representative of the northern, central, and southern varieties of Italian. The research on VINCA can be both 'intra-corpus based' or 'inter-corpus based', which means that VINCA has its own independent efficacy as a corpus, but can also be used as a measure of comparison for VALICO. In fact, some studies are currently in progress that compare the size of vocabulary used by Italian native speakers and advanced learners of Italian when describing the pictures.

2. 4 Current studies

If a corpus is used as a basis for developing specifically adapted teaching tools, then the potential advantages of resources such as VALICO are clear. Computer corpus linguistics has affected syllabi by providing more adequate and innovative descriptions of the language through highlighting the frequency of phenomena and the phraseological nature of language. Corpus-based descriptions are particularly useful for the building of syllabi for restricted languages and professional and academic genres.

As far as learner corpora are concerned, they have effectively revealed the occurrence of a number of errors, stressing the importance of being aware of the errors, but also of their overuse, underuse, as well as avoidance strategies of each specific language choices with reference to a selected norm. While most research on learner corpora has predominantly descriptive aims, some research serves as the basis for pedagogical applications both in terms of material design and classroom methodology. Native corpora have begun to be used, both to bring pedagogical materials in line with real usage, as well as to give the students direct access to language awareness activities. Learner corpora can make pedagogical materials more sensitive to learners' needs by taking their difficulties into account.

VALICO has been the source for the creation of language-specific multiple-choice exercises[7]. Starting from the observation and analysis of the corpus, we tried to work out the main morphosyntactical difficulties of some groups of learners at different levels; thus far we have been concentrating on English and Spanish mother tongue learners, but a similar study of French, German and Portuguese learners has been scheduled. Once the morphosyntactical difficulties were determined, we created multiple-choice exercises containing 'real' distractors that were extracted from the corpus itself, instead of being invented (and following an artificial pattern). This proved to be a practical and successful way of working with the corpus, which was particularly appreciated by students, who found the exercises more challenging, as well as teachers, who finally found tools to help them with the language specific problems they have to face when teaching to a particular group of students.

[7] A beta version of the exercises can be seen at www.valico.org

References

Adorno, C. / Rastelli, S. (2009). *Corpora di italiano L2: tecnologie, metodi, spunti tecnici*. Perugia: Guerra.

Bertol, A. / Díaz de Ilarraza, A. / Gojenola, K. / Maritxalar, M. / Oronoz, M. (2003). "A database system for storing second language learner corpora" in Archer D. / Rayson P. / Wilson A. / McEnery T. (eds.) (2003). *Proceedings of the 2003 Corpus Linguistics Conference, Lancaster,* 33(41). Technical Papers, Lancaster University.

Belz, J. A. (2004). "Learner Corpus Analysis and the Development of Foreign Language Proficiency" in *System: An International Journal of Educational Technology and Applied Linguistic*, 32(4), 577–591.

Barbera, M. / Corino, E. / Onesti, C. (eds.) (2007). *Corpora e linguistica in rete*. Perugia: Guerra.

Barbera, M. / Marello, C. (2004). "VALICO (Varietà di Apprendimento corpus Online)", *ITALS* anno II numero 4, 7–18.

Beißwenger, M. / Storrer, A. (2009). "Corpora of computer-mediated communication" in Lüdeling, A. / Kytö, M. (eds.) (2009). *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Corino, E. / Marello, C. (2009). "Elicitare scritti a partire da storie disegnate: il corpus di apprendenti VALICO" in Adorno, C. / Rastelli, S. (2009). *Corpora di italiano L2: tecnologie, metodi, spunti tecnici.* Perugia: Guerra, 113–137.

Dagneaux, E. / Denness, S. / Granger, S. (1998). "Computer-aided Error Analysis" in *System: An International Journal of Educational Technology and Applied Linguistics,* 26(2), 163–174.

Granger, S. (2003), "The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research", *TESOL Quarterly,* 37(3) (special issue on Corpus Linguistics), 538–546.

Granger, S. (2003). "A Multi-contrastive Approach to the use of Linkwords by Advanced Learners of English: Evidence from the International Corpus of Learner English". Paper presented at the '*Pragmatic Markers in Contrast*' *workshop* organized by the Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, Brussels, May 22–23, 2003.

Granger, S. (2004). "Practical Applications of Learner Corpora" in Lewandowska-Tomaszczyk B. (ed.) (2004). *Practical Applications in Language and Computers (PALC 2003)*. Frankfurt: Peter Lang, 291–301.

Lüdeling, A. / Walter, M. / Kroymann, E. / Adolphs, P. (2005). "Multi-level error annotation in learner corpora" in *Proceedings of Corpus Linguistics 2005*. Birmingham. Retrieved June 28, 2007 from http://www.corpus.bham.ac.uk/PCLC/

Lüdeling, A. (2008). "Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora" in Grommes, P. / Walter, M. (eds.) (2008). *Fortgeschrittene Lernervarietäten*. Tübingen: Niemeyer, 119–140.

Prat, M. T. (2004). *Computer Learner Corpora*. Alessandria: Edizioni dell'Orso.

Siemen, P. / Lüdeling, A. / Müller, F. H. (2006). "FALKO – ein fehlerannotiertes Lernerkorpus des Deutschen" in *Proceedings of Konvens 2006*, Konstanz. Retrieved from http://ling.unikonstanz.de/pages/conferences/konvens06/konvens_files/abstracts/siemenetal.pdf

Weinberger, U. (2002). *Error Analysis with Computer Learner Corpora. A corpus based study of errors in the written German of British University Students*. MS Thesis, Lancaster University.

Some Digital Humanities Methodologies and their Importance to Irish Studies

Julianne Nyhan

Here, I present an overview of a selection of scholarly resources for Old, Middle, Early Modern and Modern Irish and of the Digital Humanities methods that have informed aspects of their development. The importance of such methods to the future development of Irish studies is argued for in terms of both the new knowledge and new problems they can help to create or identify.

Beginning with a discussion of what Digital Humanities is and what some of its central methods are, I then present an overview of the Corpus of Electronic Texts (CELT), followed by a prototype of an electronic Lexicon of Old, Middle and Early Modern Irish, and finally an enhanced, retrodigitised edition of Patrick S. Dinneen's Foclóir Gaedhilge agus Béarla (Irish-English dictionary). In relation to Dinneen's Dictionary, I discuss research into and about the dictionary that Digital Humanities methods are supporting or enabling. I conclude with a few words about possible next steps.

1. Introductory Remarks

In this paper, I define the phrase 'lesser-used languages' broadly, so that it may include both the modern and historical phases of the Irish language. In line with this broad definition, I aim to present an overview of some scholarly, electronic resources of and for Old, Middle, Early Modern and Modern Irish that are either freely available on the Internet or are in preparation for publication on the Internet. My aim is not to present a systematic overview of every digital resource currently available for Irish, whether born digital, digitised or retrodigitised[1]. Rather, I limit my inquiry to Digital Humanities resources for learning and researching the Irish

[1] For a more comprehensive overview of digital resources available for early Irish studies see Moran (2009).

language, its literature and history. And I have furthermore filtered this set of resources to include only those that I have either created or made a contribution to in collaboration with my colleagues, because it is those I am most qualified to discuss in terms of their methodology.

2. Digital Humanities: Histories

As a discipline, or more accurately, an interdiscipline, Digital Humanities or Humanities Computing[2] is very much an emerging one. Widely agreed upon is that its modern origins can be traced to 1949, when Fr Roberto Busa, an Italian Jesuit priest, began work on an *index variorum* of some 11 million words of medieval Latin in the works of St. Thomas Aquinas and related authors (Hockey 2004).

Contributions and notes towards a history of the field have begun to appear in recent years (e.g. Fraser 1996; McCarty 2003b; Hockey 2004). The most recent, that of Hockey, is a chronological account that emphasises "landmarks where significant intellectual progress has been made or where work done within humanities computing has been adopted, developed or drawn on substantially within other disciplines" (Hockey 2004). As welcome and important as such contributions are, they neither are, nor aim to present, comprehensive histories of the field. Meanwhile, at an institutional, strategic and research policy level, support for Digital Humanities continues to grow.[3] Thus, the need for such a history, or histories, becomes all the more urgent not only in terms of research administration, management, networking, collaboration and foresight, but also in terms of the present understanding, future development and effective communication of the discipline. Indeed, as McCarty has asked (taking Collingwood's *Idea of History* as his guide) "without such an understanding how can Digital Humanities be *of* the Humanities as well as *in* the Humanities?" (McCarty 2008: 255).

[2] For the purposes of this essay, I will use these terms interchangeably, though a differentiation is often made between them by members of the wider community.

[3] Consider, for example, the office of Digital Humanities recently set up in the National Endowment for the Humanities (http://www.neh.gov/odh/); the Digital Humanities observatory in Ireland.(http://www.dho.ie/); and the DFG/NEH and JISC/NEH bilateral / transatlantic Digital Humanities funding programs, to mention but a few.

2. 1 Digital Humanities: definitions

Orlandi (2002) has argued that "part of the humanities was 'computed' well before computers were used." This differentiation between the process of computing and the use of a computer is an immensely important one ("the word 'computing' is a participle—a verbal adjective that turns things into algorithmic performances", writes McCarty [2008: 254–55]). Both Orlandi and Unsworth have also made clear that the use of computers in Humanities research does not necessarily Digital Humanities research make.

> One of the many things you can do with computers is something that I would call humanities computing, in which the computer is used as tool for modelling humanities data and our understanding of it, and that activity is entirely distinct from using the computer when it models the typewriter, or the telephone, or the phonograph, or any of the many other things it can be. (Unsworth 2002)

De Smedt's analogy is also helpful:

> The telescope was invented in 1608 and was initially thought useful in war. Galileo obtained one, improved it a little, and used it to challenge existing ideas about the Solar System. Although a magnificent new technology in itself, the telescope was hardly a scientific tool until Galileo used it to create new knowledge. (de Smedt 2002: 99)

Answering the question 'What is Digital Humanities?' continues to be a rich source of intellectual debate for scholars. Most recently, this question was explored as part of the 'Day in the Life of the Digital Humanities' community publication project that brought together Digital Humanists from around the world on March 18, 2009 to document their activities on that day.[4] Nevertheless, not only does a comprehensive definition appear to be impossible to formulate, when the breadth of work that is covered by a number of recent and forthcoming companions is considered (Sutherland 1997; Schreibman et al. 2004, 2008; Siemens et al. 2006; Deegan 2009; Crawford 2009; Greengrass 2009; Archer 2009), it might ultimately prove unproductive, by

[4] http://tapor.ualberta.ca/taporwiki/index.php/Day_in_the_Life_of_the_Digital_Humanities

fossilising an emerging field, and constraining new, boundary-pushing work. Terras (2006: 242) has asked whether a definition of the field is essential and reflected that such an absence may offer its practitioners additional freedom when deciding their research and career paths. This seems to hold true in Sinclair et al., and their discussion of the establishment of the Humanities Computing MA programme at the University of Alberta. In the context of designing the MA, they choose not to ask what Humanities Computing is, but rather "what do we want Humanities Computing to be…?" (2002: 168). McCarty meanwhile has argued for the fundamental importance of the self-reflection that appears so prominently in the literature of the field and comments: "What is Humanities Computing? This, for the humanities, is a question not to be answered but continually to be explored and refined" (2003b: 1233).

In the context of this essay, however, I will focus on one of the central methods, and thus defining aspects of Digital Humanities: 'modelling' or 'knowledge representation'. By far the most work on modelling, its intellectual and philosophical framework and potential, has been done by McCarty (1999, 2002a, 2002b, 2003b, 2003c, 2006, 2008; McCarty et al. 2002) and Unsworth (1997, 2002, 2002, 2006).

> First some basics. For a computer to do anything useful at all, it must have a formalised plan of action (which we call a ‚program') that represents in simplified form whatever object or process one is interested in. (McCarty 2008: 254).

McCarty describes this simplified form as a ‚model', and he has most recently classified modelling as *analytic*, *synthetic* and *improvisational*. Space limitations allow for but a brief and simplified description of analytic modelling only.[5]

Some aspects of the process of modelling are not new to Humanities researchers, and this is especially clear in McCarty's discussion of analytic modelling, which, he argues, involves analysing how something works "by taking it apart, for example, when a literary critic dissects a poem to understand how it works and what it does" (2008: 256). In Digital Humanities, this activity is extended and transformed when a computer is used to represent the scholars' (rarely fixed) understanding of how the poem works, or how it can productively be taken apart or explored. When using a computer to model a poem, for example, the model must be created within the constraints of computing technology and so complete explicitness and consistency are required. This

[5] A full exploration is to be found in McCarty (2005: 20–72).

can be a tall, if not impossible order for many aspects of works of the imagination and learning, as well as for the scholars who seek to model them.

Paradoxically, then, the greatest successes of such modelling are to be found in its failures, and McCarty argues that it illuminates through what he calls the *via negatia*. In discussing the *via negativa*, some interesting points of convergence and divergence are to be found in the observations of Santiago Ramón y Cajal (who won the Nobel Prize for his work on neurohistology) regarding the role of drawing in scientific research "as a tool in disciplining the eye" (de Rijcke 2008: 295). "[Drawing] forces us to examine the entire phenomenon, thus preventing the details that commonly go unnoticed in ordinary observation from escaping our attention" (de Rijcke 2008: 295, quoting Cajal). Analytic modelling does not necessarily force one to examine the entire phenomenon in equal measure, as some aspects of the work being modelled are necessarily brought into focus more sharply than others. However, by analytically "isolating what would not compute" (2008: 256), it does prevent the details that commonly go unnoticed from escaping our attention—the ways that we know in the Humanities. "It's great and revolutionary success for the humanities is to force the epistemological question—how is it that we know how we somehow know—and to give us an instrument for exploring that" (ibid.). In the words of Schreibman et al., talking about modelling,

> This method, or perhaps we should call it heuristic, discovers a new horizon for humanities scholarship, a paradigm as powerful as any that has arisen in any humanities discipline in the past, and indeed, maybe more powerful, because the rigour that it requires will bring to our attention to undocumented features of our own ideation. Coupled with enormous storage capacity and computational power, this heuristic presents us with patterns and connections in the human record that we would never otherwise have found or examined. (Schreibman et al. 2004)

The following is an overview of aspects of some resources for learning and researching Irish studies that have been modelled analytically. In the case of the Lexicon and Digital Dinneen, analytical modelling was carried out in order to, *inter alia*, enable particular structures and interrelationships to be represented, and thus interrogated, in computerised form. I aim to discuss both the new and changed ways of researching Irish Studies that such approaches are beginning to create and point to the many pathways that have yet to be traversed, at the time of writing.

3. CELT

Founded in 1992, the Corpus of Electronic Texts (CELT), in University College Cork is an online, multilingual textbase of Irish literature and history. It has been criticised as representing "a deeply unsystematic approach to digitising Ireland outside of the areas of Irish-language scholarship and nationalist historiography…" (Stewart 2004). Nevertheless, its declared mission is "to bring the wealth of Irish literary and historical culture (in Irish, Latin, Anglo-Norman French, and English) to the Internet in a rigorously scholarly and user-friendly project for the widest possible range of readers and researchers."[6] The texts of CELT are written in Irish of all periods, Latin, English, Hiberno-Norman French and Spanish and many English translations are available too. The Irish texts included cover all periods of the language. At present, the textbase comprises some 12 million words, perhaps not a huge collection in the context of corpora for mainstream languages, but it remains the largest Irish studies textbase of its kind. The working method of the textbase is to choose texts from the best available editions (with due regard to copyright), scan and proofread them. Texts are then modelled in terms of their structural and analytic features, according to the recommendations of the Text Encoding Initiative (TEI), the internationally-recognised standard for the encoding and interchange of electronic texts for scholarly research. The corpus is currently encoded in SGML, and its planned transformation to XML will be a welcome one that will enable some of the ‚proofs of concept' about electronic research infrastructures that have been carried out in recent years to be implemented (Nyhan 2008a).

While conversions of the TEI-SGML encoded master files to HTML are made for online reading, the master files are freely available to enable those with sufficient knowledge of SGML and SGML software to design or populate their own applications, or to convert the files to other formats, to support contextual searching, concordancing or other analyses. While a basic mechanism for searching personal names, place names, dates, numbers, occupations, and peoples and tribes is available on the site, the project's focus has been on the building of a textbase rather than on the analysis of that textbase. Thus, notwithstanding what some describe as the uneven coverage of the collection, the considerable opportunities that exist, for example, for data mining, authorship attribution and stylistic analysis, textual analysis, gender linguistics, visualisation, ‚sand-box' applications, and so forth, remain wholly unexploited and present a considerable opportunity for interested and suitably skilled researchers.

[6] Retrieved March 26, 2009, from http://www.ucc.ie/celt/about.html

4. A Prototype Lexicon of Medieval Irish

Of the Indo-European languages, Old Irish is one of the most complex and displays an especially difficult morphology and orthography. The only historical dictionary available for the language is the Royal Irish Academy's *Dictionary of the Irish Language* (hereafter *DIL*). An invaluable tool, it is widely accepted as the most authoritative reference work of its kind available for Old, Middle and Early Modern Irish. Nonetheless, it exhibits serious limitations, such as inadequate cross-referencing, use of subjective information labels and inadequate rendering of headwords.[7] An additional aspect of *DIL*—which tends to be problematic for students and sometimes scholars—is the need to know in advance the headword form of a word in order to find it. This is very much the norm for hardcopy and many electronic dictionaries, regardless of language; however, given the complexities of Old Irish morphology and orthography (e.g. some 696 forms of the verb *téit* 'proceeds, goes, goes forward, turns out' are included in the Lexicon), finding a word in DIL can be a considerable problem.

The main research question of my PhD was how a subset of the dictionary could be modelled and encoded in XML in order to enable the essential data encountered by the scholar or student in the first instance to be retrieved upon searching for any of the inflected forms of a headword contained in DIL. Other questions focused on the effective modelling of other categories of information; new insights into the dictionary that were generated when information could not be appropriately modelled in XML (Nyhan 2008b); and the history of *DIL* and information ordering in Irish glossaries and early print dictionaries.

Having determined the essential data to be included in the Lexicon (headword, grammatical information, etymology, definitions, and word forms), an XML structure was devised to describe this information and that could also be expanded in a number of ways to describe more detailed information or indeed to express caveats. I proceeded to work through the hardcopy version of *DIL* in order to select essential data (which I then typed up). This subset was then enhanced with additional information such as part-of-speech (absent, for the most part, from *DIL*) and the categorisation of sense types (to allow for thematic searches of definitions). Inflected forms and their variant spellings were grouped together and encoded as simple forms; the same process was followed in the case of compound forms, which were grouped together and encoded as compound nouns, verbs or adjectives; likewise, the same process was

[7] For a discussion of such limitations see Nyhan (2006b: 79–90).

followed for forms with emphasising suffixes or infixed pronouns, for example. The further modelling and XML encoding of this data was very much an iterative process that involved many layers of refinement and revision, and the resulting prototype is deeply encoded in XML.

That the prototype demonstrates a successful approach to answering my chief research question is demonstrated by reference to the digitised version of the Dictionary of the Irish language (eDIL). eDIL was digitised in line with the TEI guidelines by the team in the University of Ulster, Belfast, led by Professor Gregory Toner and with funding from the Arts and Humanities Research Council (AHRC) over a period of three years (Fomin et al. 2006). It was published in May of 2007. While eDIL is an electronic version of the complete Dictionary of the Irish Language (some 35,000 headwords), as already discussed, the Lexicon contains but a subset of that information. For example, the Lexicon does not reproduce the complete citations of *DIL*, but rather the word form in question is excerpted from its citation (a decision that was made on practical grounds as the Lexicon was the work of one person without access to a captured version of the hardcopy for this project). The word form is then categorised, described and modelled in XML. Nevertheless, the Lexicon contains (except where there has been an erroneous ommission) all of the word forms of the entries that are contained in eDIL/*DIL*. Despite this difference of content, the approach used in the Lexicon would work just as well by encoding the words within their citations accordingly.

In order to demonstrate the advance to the modelling and subsequent retrieval of historical Irish lexicography that the prototype Lexicon makes, I carried out a number of searches using both eDIL and the Lexicon. Using the 'quick search' mechanism in eDIL[8], and filtering the results using the 'most relevant first' option, I searched for the headword *téit* and retrieved 198 individual results (i.e. entries that the headword occurs in), the entry for *téit* being listed on page 18 of those results. I then changed the filter to 'headword first' and searched again, but this again returned 198 results, among which I could not find *téit* (the alphabetical ordering of the results being inconsistent). Next, I tried an inflected form of *téit*, *luid*. Using the 'most relevant first' filter this returned 313 results, the relevant entry being result number 287 on page 29 of 32 pages of results. Thus, as with *DIL*, at the present time specialist knowledge is required in order to use eDIL effectively. I then searched for the forms *téit* and *luid* in the prototype Lexicon and retrieved for entry for *téit* only on both occasions.[9] Then, consid-

[8] Retrieved March 23, 2009, from http://www.dil.ie/search-all.asp

[9] Retrieved March 23, 2009, from http://epu.ucc.ie/lexicon/entry

ering that the material available in the online prototype consists of the verb *téit*, all the word forms listed under B in *DIL*, apart from verbs, the definite article, the first person singular possessive and personal pronouns, I thought it necessary to search on nouns also. I randomly choose *buide*. Using the 'exact' rather than the 'contains' filter, as well as the 'lemmas' filters, the lexicon returns *buide* 1, 2, 3, only; eDIL returns 150 results (16 of them are words beginning with B), with the words in question listed on page four, beginning with result number 38. Experimenting with the 'Advanced search mechanism' did not enable me to refine these results.

Despite such observations, eDIL is without question a monumental contribution to Irish Studies present and future, and has made the most important reference work of Old, Middle and Early Modern Irish available in its entirety and to a very high standard. While it does require some specialist knowledge in order to use the online version effectively, it offers much for the student and specialist of Irish, especially in terms of the 'advanced' search mechanism. Indeed, non-headword searching was not an aim of the Editors of eDIL.[10] The team is presently at work on an update of *DIL* itself. Also, for fear of implying that Digital Humanities scholarship consists merely of sequences of different models that cannot be evaluated or assessed, the work of Unsworth is important here too. Drawing on Popper's criterion of falsification in particular, he has set out a number of evaluative criteria for Hypertext Projects including: "Does it declare in terms of its own success of failure?", "Does it formulate a methodology for solving the problem it addresses?", "Does it address (or generate) unsolved problems?", and "Can its solutions be generalized?" (1997).[11]

While seven letters of the Lexicon have been prepared, it has been possible thus far to make only a portion of this available, in what is very much a prototypical format, and with an experimental search interface that must be further refined and adapted, and it cannot be taken forward without adequate funding and resources. On the one hand, the possibilities for developing the Lexicon in conjunction with eDIL, of seeking to fuse aspects of the different models of both those works and preparing them for use with the CELT corpus, are clear, and some ‚proofs of concept' have been worked out (Nyhan 2008a). However, there is also much potential for using the Lexicon at an experimental level, as a tool for investigating and imagining both assumptions about

[10] Cf. http://www.dil.ie/about.asp, retrieved May 15, 2009.

[11] See also the materials made available for the MLA 'Evaluating Digital Work for Tenure and Promotion: A Workshop for Evaluators and Candidates' workshop, retrieved February 28, 2009, from http://www.mla.org/resources/documents/rep_it/dig_eval

the historical lexicography contained in *DIL*, and for testing hypotheses about that data within the context of the way it has been modelled.

The framework of the Lexicon successfully demonstrates the value of the application of XML to Irish historical lexicography in order to, *inter alia*, make a contribution to our understanding of how such complex and compressed information can effectively be modelled and interrogated in terms of, *inter alia*, its inflected forms. And in doing so it created a further set of research problems and questions. While the method developed is effective, it is intensely laborious. Discussing "Humanities Computing shaped by the need for human communication", Unsworth reflected: "A representation is the language in which we communicate; hence, we must be able to speak it without heroic effort" (Unsworth 2002). Thus, the effective modelling of the data in order to achieve a particular end resulted in a new set of research questions that are currently being explored in the electronic, enhanced edition of Dinneen's dictionary (discussed below).

Space will permit but a reference to the method that has been developed during work on Dinneen's dictionary. As we have seen in the examples above, when searching XML encoded dictionaries, it is extremely difficult to write a query capable of differentiating between words that occur in a citation in which they are being illustrated or that happen to occur in a citation that is illustrating another headword, unless such words have been differentiated with XML (or another suitable language). This has a direct bearing on the relevance of the results that are returned, as well as the level of knowledge that the end-user must have in order to use the work effectively. In contrast with the Lexicon, every word form has not been encoded in Dinneen. Rather, the text has been modelled so that entries can be weighted and returned based on the answers to a number of questions (that will be implemented via the search mechanism) about the model of the dictionary that has been developed. I hope to discuss the results of this approach in a future paper.

5. The Digital Dinneen

> Fr Pádraig Ua Duinnín was a titan in the pantheon of Irish scholarship. The range of his achievements is vast. His editions of the Munster poets and of Keating alone would have ensured an honoured place for him in the history of Irish scholarship but it is as a lexicographer that his name will be forever be anchored in the Irish psyche. (Uí Bheirn 2005: 102)

An introduction to the Digital Dinneen project and its proposed integration with CELT and eDIL has been published elsewhere (Nyhan 2008a), so I will not address these topics at length here.

With a research award from the Irish Research Council for the Humanities and Social Sciences (IRCHSS) to cover a period of three years, work on creating an enhanced, electronic edition of Patrick S. Dinneen's *Foclóir Gaedhilge agus Béarla* began in the CELT project in July 2006. Quoting from Uí Bheirn again,

> [Dinneen's dictionary] is a treasure house where the valuables are stored hiddledy-piggledy. […] [I]t will present a considerable challenge to digitise it in a coherent searchable way, because it is such a complex unstructured entity. (Uí Bheirn 2005: 106).

The most basic aim of the project was to allow Dinneen to be effectively searched or transformed in order to support a range of Basic Research. Following on from observations made by scholars about the potential of the electronic edition to contribute, for example, to "identifying unattributed sources" (Uí Bheirn 2005: 106) and "reconstructing obscure abbreviations" (Ua Súilleabháin 2005: 74), the work was, *inter alia*, encoded in XML (and is currently being transformed so that it is conformant with TEI P5), and additional information including homograph numbers, word roots and (in the case of two letters), the post-spelling reform forms of headwords were also added by the doctoral students who also worked on the project.[12] A brief overview of ongoing research and some research that the edition will support is presented below.

[12] Emer Purcell, Benjamin Hazard (both of the History Department, University College Cork) and Emma McCarth (University College Cork). For more information see here: http://www.ucc.ie/celt/people.html

5. 1 Modelling the dictionary so that it can be used in new ways and contribute to the creation of new knowledge about the Irish language.

In contrast to *DIL*, for example, Dinneen's consistency in terms of the mechanics of citing his sources is remarkable. Nevertheless, inconsistencies occur and can be observed especially in his use of abbreviations, place names and bibliographical citations. In the electronic edition, all bibliographical citations are encoded in XML and will be supplied (on an automated basis) with a canonical form in order to ensure consistent searching. A simple but important and powerful benefit is that word lists, which are important to, for example, Irish dialectical studies, but had been lost after their integration into the dictionary, can be reconstructed (or at least the portion of them cited in the Dictionary can be)—their recovery from a hard copy version of the dictionary being extremely difficult. For example, a list of words from an area of county Waterford, Ireland, was collected by the poet Riobaird Bheldon and Fr Mícheál MacCraith and is cited in Dinneen's dictionary as '*Cm.*'. Dinneen integrated it into the Dictionary, and the full list was also supposed to have been published, but was lost. It remained so until a version of it was published in 2007, after being easily extracted from the encoded version of Dinneen's dictionary and reassembled. The full range of dialectical information in the dictionary will also be available to researchers to develop further and amalgamate with other resources and information, for example, as content for interactive dialectical maps of the Irish language or for inclusion in a historical GIS application.

5. 2 Modelling the dictionary so as engage in new kinds of research about Dinneen and the dictionary itself.

A recent article by Coleman et al. (2009) sets out a range of evidence-based methodologies—statistical, textual, contextual and qualitative—that can be used in forensic dictionary analysis.

In a discussion of Dinneen's dictionary, O'Connell (1984) described examples of usage and idiom given by Dinneen "that have seduced the reader over the years into many fascinating byways." At present, the author is carrying out a forensic analysis of the dictionary, with regard to such byways, in order to establish whether new insights into Dinneen's working practices and techniques can be concluded.

Based on samples of dictionary entries occurring at the beginning, middle and end of a sample of letters, the present study is building on the XML already in place, describing senses and citations in order to model such digressions. The aim is to question how frequent they are, if there is a pattern in terms of where they occur in the work, and what their topics are. Of course, the process of modelling such digressions brings us back to a fundamental question of what a digression is (should it include possibly helpful, additional information related, loosely or not to the main sense of a word, as well as Dinneen's flights of fancy?). Notwithstanding the problems of deciding what a diversion is and is not, using XML to categorise and encode a sample of those diversions, and then producing a visualisation of, for example, where they tend to occur in the work, we may get a new perspective on Dinneen, his interests, informants and working methods.

6. Conclusion

This paper has treated of some work that is being done at the interface of Digital Humanities and Irish Studies, especially in terms of methodology. Due to space restrictions, it has not been possible to discuss theoretical aspects of Digital Humanities or, indeed, some of the theoretical problems and advantages of using XML in the modelling of Humanities data.

While I have not presented a comprehensive overview of all Digital Humanities resources for Irish Studies, I hope to have shown that Digital Humanities methods offer much that is important to Irish Studies, and by extension to other minority languages. McGann (2005: 181) has written, "In the coming decades—the process has already begun—the entirety of our cultural inheritance will be transformed and re-edited in digital forms. Do we understand what that means, what problems it brings, how they might be addressed?" It is clear that we are still in the earliest stages of such a journey; there is much to be done both in terms of making, extending, using, theorising and integrating such resources. Of vital importance to this will be working more and better together, regardless of disciplinary boundaries, in order to share and develop new methods and theories for the computation support of minority languages, and to formulate the new questions that such resources and the collaborative research involved in making those resources should help us to ask.

References

Archer, D. (2009). *What's in a word-list? Investigating word frequency and keyword extraction.* Ashgate: UK.

Beynon, M. / Russ, S. / McCarty, W. (2006). "Human Computing—Modelling with Meaning", *Literary and Linguistic Computing*, 21(2), 141–157.

Crawford, T. (2009). *Modern Methods for Musicology: Prospects, Proposals and Realities.* Ashgate: UK.

Deegan, M. / Sutherland, K. (2009). *Text Editing, Print, and the Digital World.* Ashgate: UK.

de Rijcke, S. (2008). "Drawing into abstraction. Practices of observation and visualization in the works of Santiago Ramón y Cajal", *Interdisciplinary Science Reviews*, 33(4), 287–311.

de Smedt, K. (2002). "Some Reflections on Studies in Humanities Computing", *Literary and Linguistic Computing,* 17(1), 89–101.

Fomin, M. / Toner, G. (2006). "Digitizing a Dictionary of Medieval Irish: the eDIL project", *Literary and Linguistic Computing*, 21(1), 83–90.

Fraser, M. (1996). *A Hypertextual History of Humanities Computing.* http://users.ox.ac.uk/~ctitext2/history/

Greengrass, M. / Hughes, L. (2008). *Virtual Representations of the Past.* Ashgate: UK.

Hockey, S. (2004). "The History of Humanities Computing" in Schreibman, S. et al. (eds.) (2004). *A Companion to Digital Humanities.* Oxford: Blackwell. http://www.digitalhumanities.org/companion/

Moran, P. (2009). "Irish glossaries and other digital resources for early Irish studies" in Rehbein, M. / Ryder, S. (eds.) (2009). *Jahrbuch für Computerphilologie*, 10. http://computerphilologie.de/jg08/moran.pdf

McCarty, W. (1998). "What is humanities computing? Toward a definition of the field." Liverpool, 20 February 1998. Reed College (Portland, Oregon, U.S.) and Stanford University (Palo Alto, California, U.S.), March 1998. Würzburg, Germany, July 1998. http://www.cch.kcl.ac.uk/legacy/staff/wlm/essays/what/

McCarty, W. (1999). "Humanities computing as interdiscipline" from the seminar series *'Is Humanities Computing an Academic discipline?'* (IATH), Virginia: University of Virginia. http://www.iath.virginia.edu/hcs/mccarty.html

McCarty, W. (2002a). "Humanities Computing: Essential Problems, Experimental Practice", *Literary and Linguistic Computing*, 17(1), 103–125.

McCarty, W. (2002b). "New Splashings in the Old Pond: The Cohesibility of Humanities Computing" in Braungart, G. / Eibl, K. / Jannidis, F. (eds.) (2002). *Jahrbuch für Computerphilologie 4.* Paderborn: mentis Verlag, 9–18.

McCarty, W. / Short, H. (2002). "A Roadmap for Humanities Computing", http://www.allc.org/reports/map/

McCarty, W. / Kirschenbaum, M. (2003a). "Institutional Models for Humanities Computing", *Literary and Linguistic Computing,* 18(4), 465–489.

McCarty, W. (2003b). "Humanities Computing" in *Encyclopedia of Library and Information Science*. New York: Marcel Dekker, 1224–1235.

McCarty, W. (2003c). "«Knowing true things by what their mockeries be»: Modelling in the Humanities", *Computing in the Humanities Working Papers*, A.24, jointly published with *TEXT Technology*, 12/1.

McCarty, W. (2004). "As it almost was: Historiography of recent things", *Literary and Linguistic Computing*, 19(2), 161–180.

McCarty, W. (2005). *Humanities Computing*. Basingstoke: Palgrave MacMillan.

McCarty, W. (2006). "Tree, Turf, Centre, Archipelago—or Wild Acre? Metaphors and Stories for Humanities Computing", *Literary and Linguistic Computing*, 21(1), 1–13.

McCarty, W. (2008). "What's going on?", *Literary and Linguistic Computing*, 23(3), 253–261.

McGann, J. (2005). "Culture and technology: the way we live now, what is to be done?", *Interdisciplinary Science Reviews,* 30(2), 179–188.

Nyhan, J. (2006a). "Findfhocla an Chomaraigh", *An Linn Bhuí: Iris Ghaeltacht na nDéise,* 10, 97–111.

Nyhan, J. (2006b). *The Application of XML to the Historical Lexicography of Old, Middle and Early Modern Irish: a Lexicon-based Analysis*. PhD thesis. University College Cork.

Nyhan, J. (2008a). "Developing Integrated Editions of Minority Language Dictionaries: the Irish Example", *Literary and Linguistic Computing,* 23(1), 3–12.

Nyhan, J. (2008b). "The problem of date and context in electronic editions of Irish historical dictionaries" in Mooijaart, M. / van der Wal, M. (eds.) (2008). *Yesterday's Words: Contemporary, Current and Future Lexicography*. Cambridge Scholars Press, 319–332.

O'Connell, N. (1984). *Father Dinneen—his Dictionary and the Gaelic Revival*. Irish Texts Society, Dublin. http://www.irishtextssociety.org/dinneen.htm

Orlandi, T. (2002). "Is Humanities Computing a discipline?" in Braungart, G. / Eibl, K. / Jannidis, F. (eds.) (2002). *Jahrbuch für Computerphilologie 4.* Paderborn: Mentis Verlag, 51–58.

Schreibman, S. / Siemens, R. / Unsworth, J. (eds.) (2004a). *A Companion to Digital Humanities*. Oxford: Blackwell. http://www.digitalhumanities.org/companion/

Schreibman, S. / Siemens, R. / Unsworth, J. (2004b). "The Digital Humanities and Humanities Computing: An Introduction" in Schreibman, S. et al. (eds.) (2004). *A Companion to Digital Humanities*. Oxford: Blackwell. http://www.digitalhumanities.org/companion/

Schreibman, S. / Siemens, R. (2008). *A Companion to Digital Literary Studies*. Oxford: Blackwell Publishing.

Siemens, R. / Moorman, D. (2006). *Mind Technologies. Humanities Computing and the Canadian Academic Community*. Calgary: University of Calgary Press.

Sinclair, S. / Gouglas, W. (2002). "A Theory into Practice: A Case Study of the Humanities Computing Master of Arts Programme at the University of Alberta", *Arts and Humanities in Higher Education*, 1(2), 167–183.

Stewart, B. (2004). "Medium and Message: Reflections on Irish Studies in the Informatics Age", *Journal of the Association for History and Computing,* 8(3). http://journals2.iranscience.net:800/mcel.pacificu.edu/mcel.pacificu.edu/JAHC/JAHCVII3/ARTICLES/stewart.html

Sutherland, K. (ed.) (1997). *Electronic Text: Investigations in Method and Theory.* Oxford: Clarendon Press, 107–126.

Terras, M. (2006). "Disciplined: Using Educational Studies to Analyse «Humanities Computing»", *Literary and Linguistic Computing,* 21(2), 229–246.

Unsworth, J. (1997). "Documenting the Reinvention of Text. The Importance of Failure", *The Journal of Electronic Publishing*, 3.2.

Unsworth, J. (2000). "Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?". Paper read at the *Symposium on 'Humanities Computing; formal methods, experimental practice'*, London: King's College, May 13, 2000.

Unsworth, J. (2002). "What is Humanities Computing, and What is it Not?" in Braungart, G. / Eibl, K. / Jannidis, F. (eds.) (2002). *Jahrbuch für Computerphilologie 4.* Paderborn: Mentis Verlag, 71–84. http://computerphilologie.uni-uenchen.de/jg02/unsworth.html

Ua Súilleabháin, S. (2005). "Dinneen's Dictionaries of 1904 and 1927: Background, use of Historical Dictionaries and of Contemporary Informants" in Riggs, P. (ed.) (2005). *Dinneen and the Dictionary 1904–2004*. Dublin: Irish Texts Society, 62–77.

Uí Bheirn, U. (2005). "The present state of Irish lexicography" in Riggs, P. (ed.) (2005). *Dinneen and the Dictionary 1904–2004*. Dublin: Irish Texts Society, 62–77.

Creating and Working with Spoken Language Corpora in EXMARaLDA

Thomas Schmidt

Spoken language corpora—as used in conversation analytic research, language acquisition studies and dialectology—pose a number of challenges that are rarely addressed by corpus linguistic methodology and technology. This paper starts by giving an overview of the most important methodological issues distinguishing spoken language corpus work from the work with written data. It then shows what technological challenges these methodological issues entail and demonstrates how they are dealt with in the architecture and tools of the EXMARaLDA system.

1. Introduction

At first glance, it may seem inappropriate to include a contribution on spoken language corpora in a colloquium on "Lesser Used Languages and Computer Linguistics" —after all, spoken language as such is certainly not a 'little used' kind of language. However, spoken language shares a fate with 'smaller' languages insofar as corpus and computational linguistics as fields of study have a definite bias, not only towards major, standardised languages, but also towards written language in general. Thus, large reference corpora usually consist entirely or to a great part of written texts, the greater part of corpus linguistic literature deals with phenomena of written language, and the technology for constructing and analysing language corpora is also much further developed and established for the written medium. The reasons for this may be of a theoretical nature—certainly, some research questions are best approached by looking at written data. Yet the prevailing cause for the dominance of written language in corpus linguistic seems to me to be a pragmatic one: whereas large amount of written text are easily available for integration into a corpus, spoken data has to be tediously recorded and transcribed; both processes involve difficult methodological challenges, and it is thus much harder to arrive at a reasonably large amount of spoken language material to address a certain research question than it is for written data.

However, it is also unquestionable that certain linguistic phenomena cannot be studied by looking at written language data alone. Child language acquisition, dialectal variation and, of course, the structure of face-to-face interaction (as studied for example by conversation analysis) are all cases in point.

This paper starts by discussing some of the methodological challenges involved in constructing spoken language corpora, comparing them to the corpus construction workflow for written data. It then proceeds to demonstrate how these challenges are addressed in the EXMARaLDA system.

2. Some Methodological Challenges for Spoken Language Corpora

2. 1 Primary and secondary data, modelling

It is common in corpus linguistics to draw a distinction between primary and secondary data. When we deal with written language, 'primary data' usually denotes the original text, as published and intended to be read by its audience—for example, a printed book or an electronic (e.g. PDF) document—while 'secondary data' means a derived representation of this text as included in the corpus (e.g. a simple text file or a document with TEI markup).[1] Getting from the primary to the secondary data almost always involves some kind of simplification, abstraction, interpretation and, possibly, purposeful modification. Thus, for most written language corpora, text layout and formatting of the original document are not represented (i.e. abstracted over) in its derived version, non-textual elements (pictures, diagrams, etc.) are also left out, and some normalisation (e.g. undoing hyphenation at the end of a line) is carried out. All these modifications are (or at least should be) justified with respect to the research question the corpus is meant to address. We can therefore regard this step from primary to secondary data as a kind of scientific modelling, because it has the three general properties that, for instance, (Stachowiak 1973) uses to characterise a scientific model:[2]

[1] In other contexts, though, 'primary data' means what is called 'secondary data' here, and 'secondary data' means analytic information ('annotation') added to this data.

[2] See Schmidt (2005a, b) for a comprehensive discussion of the modelling aspect in the work with linguistic data.

- the representation property (*Stellvertretermerkmal*) states that, in the process of scientific study, the model takes the place of the actual thing to be studied;
- the abstraction property (*Verkürzungsmerkmal*) states that a model is always a simplified representation of the thing modelled; and,
- the pragmatic property (*Pragmatisches Merkmal*) states that this simplification is motivated by a certain purpose.

When it comes to spoken language data, the first thing to notice is that we are dealing here not with one but two distinct steps from the original linguistic fact to its representation in a corpus. In analogy to the written language case, we should call the original interaction the primary data. However, this data is ephemeral—it is not available for systematic study unless it is made more permanent through the process of recording. One or more audio or video recordings of the interaction would thus have to be called the secondary data. In transcription, this secondary data is then transferred into a written representation, which, consistently, should then be denoted 'tertiary data'.

Without a doubt, both these steps—from interaction to recording and from recording to transcription—involve a fair amount of modelling. For instance:

- The choice of recording type (audio or video) has to be motivated by the intended research to be carried out. Since audio recording blinds out all visual aspects of interaction, the resulting corpus can only be used to study the audible aspects. The same can be said for certain parameters of the recording (e.g. how many cameras or microphones to use, where to put them).
- Transcription itself has always been characterised as a "selective process reflecting theoretical goals and definitions" (Ochs 1979: 44). It is hardly controversial that the process of transferring spoken language to the written medium can only be done on the basis of a theoretically-motivated decision about which aspects of the recording to include and which to leave out. The great number of existing transcription systems (e.g. HIAT, Rehbein et al. 2004; GAT, Selting et al. 1998; CHAT, MacWhinney 2001) and the different principles for transcript layout (e.g. musical score vs. line notation) testify to this.
- In contrast to written language as "the language of distance" (Koch & Oesterreicher 1995), spoken language is embedded in a specific situational context. Understanding, interpreting and analysing spoken language therefore also depends on the availability of information about the speech situation, such as time and place of

> the interaction, things that happened before the interaction started, the spatial constellation of speakers, and so forth. As with transcription itself, there is no independent external criterion for deciding which of this information to include and which to leave out. Again, the modelling of such metadata thus relies on theoretical considerations. The same holds for sociographic metadata about speakers, such as age, social status, language competence and so forth.

While the construction of both written and spoken language corpora thus involves a modelling step, this step can be said to be much more pronounced, that is, requiring more abstraction and theory-guided interpretation, for the latter type. For written language, at least the character table for the alphabet of a given language as well as the list of words defined by an established dictionary can be regarded as a common ground for all corpus modelling—meaning that, for a modern and standardised language, the mapping of these entities from original to corpus representation is usually *not* interpretative. For spoken language, there is no such common ground. A transcription of a spoken language recording is therefore a much less stable basis of analysis than, say, an ASCII text representation of a newspaper article. Hence, when working with a spoken language corpus, researchers will value the possibility to verify, and possibly revise, the transcriber's decisions by listening to the original recording.

2. 2 Data structures

When representing language data in the digital medium, a choice has to be made about general properties of data structures, that is, salient structural relations between entities that must be encoded in a file or database. Certainly, hierarchic inclusion (e.g. a paragraph being made up of sentences, which, in turn, are made up of words) and sequential ordering (e.g. the words in a sentence following one another), are two of the most important such relations in linguistic description. In fact, it has been argued in the famous OHCO thesis (De Rose et al. 1990: 6) that these two relations are sufficient to characterise the structure of a written text, or, in the author's words, that "text is an ordered hierarchy of content objects". Although this thesis has variously been refuted as being too strong (also by the authors themselves), OHCO remains the dominant modelling paradigm of many approaches to encoding corpus data, most notably the TEI guidelines. In these approaches, then, the primary data structure is a tree grouping smaller linguistic entities into larger ones, and all non-tree-like structures or overlapping hierarchies (e.g. the paragraph vs. page division of a text) are treated, if at all, as exceptions to the rule. As the success of TEI encoded corpora shows, this has proven a practicable way of handling written language data.

Spoken language, however, as it unfolds over time, exhibits many non-sequential, non-tree-like relations on the lowest structural level: speakers' utterances may overlap, verbal behaviour is accompanied by simultaneous gestures or facial expressions, and the verbalisations of one speaker are themselves made up of different aspects (e.g. lexical words and suprasegmental characteristics like modulation, voice quality), which may need to be described in 'parallel' structures. In spoken language, the exceptions to the OHCO assumption thus become the rule. While it is still possible to use OHCO-based paradigms to encode such data (see, for instance, Schmidt 2005c), any system claiming to be adequate for spoken language representation has to pay due attention to a consistent and practicable method for also encoding non-hierarchic structures. Bird/Liberman's (2001) annotation graph formalism—arguably one of the most influential proposals in the field in the last ten years—therefore radically emphasizes the temporal aspect of spoken language, suggesting using as the primary data structure an acyclic graph whose nodes can be anchored to a timeline.

2. 3 Size and balance, speed and efficiency

The size of a corpus determines to a great deal the empirical findings that may be derived from it. Any quantitative or statistical analysis of empirical data requires a critical mass so that regularities in the sample can be generalised to the population the sample represents. And even for purely qualitative analyses, only a sufficiently large corpus allows the researcher to judge the value of an individual example, because it is only in comparison to a reasonably big number of other examples that its uniqueness or 'prototypicalness' can be plausibly evaluated. Likewise, the concept of balance, that is, the property of a corpus to represent certain parameters (like genre, geographic origin etc.) in adequate, non-skewed proportions when compared to the entirety of linguistic facts, is a very important criterion for empirical investigations.

For written language corpora, both the problems of size and balance have been addressed in a satisfactory fashion. For example, the German reference corpus of the IDS[3] consists of no less than 3.6 billion words, and the fact that the WWW makes enormous amounts of text readily available for electronic search shows that there is, in principle, no upper limit to the size of a written language corpus. Convincing

[3] http://www.ids-mannheim.de/kl/projekte/korpora/

concepts for balanced corpus stratification have been applied, for instance, in the English BNC corpus[4] or the German DWDS corpus (Geyken 2009), both resources counting over 100 million words.

The situation is much different for spoken language corpora. One of the largest such resources, the Spoken Dutch Corpus (CGN, Oostdijk & Broeder 2003), although it counts an impressive 8 million transcribed words, is still one or two orders of magnitude smaller than the above-mentioned resources. Most other published spoken language corpora do not exceed the million-word boundary.

Yet, from a theoretical perspective it might be argued that spoken language corpora, in order to enable the same kind of generalisations, should actually be *larger* than their written counterparts. Since spoken language is less standardised and occurs in a much wider variety of circumstances, a really 'balanced' corpus would have to take into account a very large number of speaker and interaction types. For example, a reference corpus of spoken German would have to cover in comparable proportions dialectal, register and topic variation across such diverse interactions types as telephone calls, TV debates, service encounters, informal talk, classroom discourse and political speeches. Aside from the fact that no widely accepted model for such stratification exists (whereas, for example, Biber's [1993] ideas can be considered a kind of standard approach to written language stratification), practical reasons make it virtually impossible to construct spoken language corpora in the 100-million-word dimension.

This is so because spoken language corpus construction requires manual work for many steps that can be automated (or at least semi-automated) for written language data. In (primary) data acquisition, written texts can be harvested from the Web or otherwise be provided in electronic format, whereas spoken interaction has to be recorded in the field. In secondary (or tertiary) data creation, semi-automatic methods like HTML cleanup or Optical Character Recognition are only applicable to written language documents, whereas spoken language transcriptions have to undergo the extremely time-consuming process of manual transcription. Thus, in fields like conversation analysis with its fine-grained and detailed transcription procedure, it is not uncommon to estimate 100 hours of transcriber's time for one hour of recorded interaction. Given these drastic differences in the time and effort required to construct corpora, the speed and efficiency of corpus tools becomes a paramount concern when the size of a spoken language corpus is considered relevant in any way.

[4] http://www.natcorp.ox.ac.uk/

2.4 Summary

Summarising the preceding sections, a spoken language corpus, when compared to a written language corpus, poses three methodological challenges:

- Its base is more *instable* insofar as the modelling step between original data and corpus data is much more pronounced—far-reaching theoretical decisions have to be taken very early in the corpus construction process. The corpus data may therefore require corrections during analysis; in any case, a close link between recording (secondary data), transcription (tertiary data) and contextual information (metadata) is methodologically desirable
- Its base is *complex* insofar as it involves parallel relations on the lowest structural level. Standard OHCO processing is therefore not sufficient for spoken language corpora, a more complex data model is required
- Since time-consuming manual methods prevail in the construction of spoken language corpora, tools and workflows must be optimised for speed and efficiency in order to attain adequate corpus sizes.

Translating this into requirements for corpus technology for spoken language corpora, it can be said that such technology must:

- be theory aware;
- keep a close link between recording, transcription and meta-data;
- use a data model which can naturally represent parallel temporal relationships; and,
- consider questions of speed and efficiency.

The following section will demonstrate how these requirements are met in the EXMARaLDA system.

3. EXMARaLDA

EXMARaLDA (Extensible Markup Language for Discourse Analysis) is a system of data models, data formats and software tools for the construction and analysis of spoken language corpora. It has been under development since 2000 in a project at the Special Research Centre on Multilingualism at the University of Hamburg.

EXMARaLDA is a data-centric system, that is, it is designed and implemented around a central data model (rather than, say, a specific workflow or a specific piece of software). In accordance with the above-mentioned considerations, the data model is optimised for the representation of structural relations occurring in spoken language. It is based on Bird and Liberman's (2001) idea of annotation graphs, allowing an intuitive encoding of temporally or otherwise parallel structures. Moreover, EXMARaLDA draws a distinction between temporal and linguistic entities of description. Since the former are much less dependent on a specific theoretic approach, the system can provide a set of operations applicable across theories, imposing theory-dependent structure only where it is necessary.[5]

To ensure maximal reusability of data, EXMARaLDA uses open standards like XML and Unicode in its data formats, and it provides interfaces to the most important other systems (like Praat, ELAN, CHAT, TEI) as well as to standard desktop software (Microsoft Word, Internet Browsers) for optimal interoperability. Software tools are programmed in JAVA so that they can be used on all major operating systems (Windows, Linux, Mac).

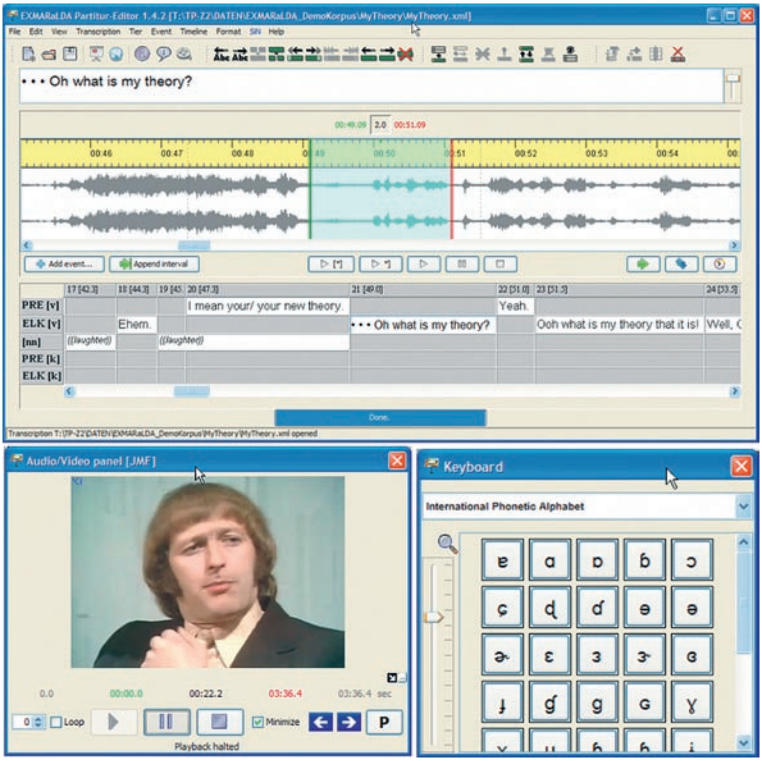
EXMARaLDA's main application areas are discourse and conversation analysis, first and second language acquisition studies, and dialectology. In addition to that, the system has also been used for multimodal analyses, phonological or phonetic studies and for the annotation of written data.

3. 1 Transcription: Partitur-Editor

EXMARaLDA's transcription tool is the Partitur-Editor (Figure 1), a tool for entering and editing transcriptions in musical score notation. During transcription or in a separate step, the transcribed text can be linked to the underlying audio or video file by setting appropriate timestamps in the transcription's timeline. The interface is

[5] In systems like CHAT (MacWhinney 2001) on the other hand, theory-dependent concepts like utterances are so central to the system‘s functionality that it is impossible to encode theory-independent structure separately.

based on the temporal structure relations alone so that the editor can be used independently of a specific theoretical approach.[6] After transcription has been completed, a separate processing step—'segmentation'—is used to calculate the linguistic structure, based on the regularities of established transcription systems (currently, HIAT, GAT, CHAT and DIDA are supported).



EXMARaLDA Partitur-Editor 1.4.2 [T:\TP-Z2\DATEN\EXMARaLDA_DemoKorpus\MyTheory\MyTheory.xml]

• • • Oh what is my theory?

	17 [42.3]	18 [44.3]	19 [45	20 [47.3]	21 [49.0]	22 [51.0]	23 [51.5]	24 [53.5]
PRE [v]				I mean your/ your new theory.		Yeah.		
ELK [v]		Ehem.			• • • Oh what is my theory?		Ooh what is my theory that it is!	Well, (
[nn]	((laughted))		((laughted))					
PRE [k]								
ELK [k]								

Transcription T:\TP-Z2\DATEN\EXMARaLDA_DemoKorpus\MyTheory\MyTheory.xml opened

Figure 1: User interface of the EXMARaLDA Partitur-Editor

[6] While the Partitur-Editor thus fulfils several of the above-stated requirements—it allows to keep a close link between transcription and recording, and it is 'theory-aware'—it should be noted that some of its flexibility is paid for in terms of speed and efficiency of transcribing. After testing the editor in a number of corpus construction scenarios, we found that its main drawback in this respect is its non-optimal use of screen real estate: like other multi-layer-tools (e.g. Praat and ELAN), the horizontal, musical-score-like layout of the interface means that transcribers can only look at small stretches of transcription text at a time and thus get a much less text-like experience than they normally have in the vertically organised layout of a standard word processor. The FOLKER transcription tool (http://agd.ids-mannheim.de/html/folker.shtml), developed on the basis of EXMARaLDA for the FOLK corpus of the IDS Mannheim, provides the user with the possibility to switch between a horizontally (musical score) and two (segment and contribution list) vertically-organised views. First experiments show that transcription can indeed be sped up considerably in this way.

3. 2 Metadata: Corpus Manager

In order to be able to deal with the various metadata requirements for spoken language corpora mentioned above, EXMARaLDA provides a separate tool, the "Corpus Manager" (CoMa), for bundling larger sets transcriptions into a corpus and describing its components through appropriate metadata sets. Coma models a corpus as a set of communications, each of which can consist of one or several recordings and corresponding transcriptions (Figure 2). Speakers are kept in a separate list and are assigned to communications in an n:m relation. In that way, it becomes possible to represent one speaker's participation in several communications as well as the fact that one communication usually involves more than one speaker, with the effect that unnecessary duplication of metadata is avoided.

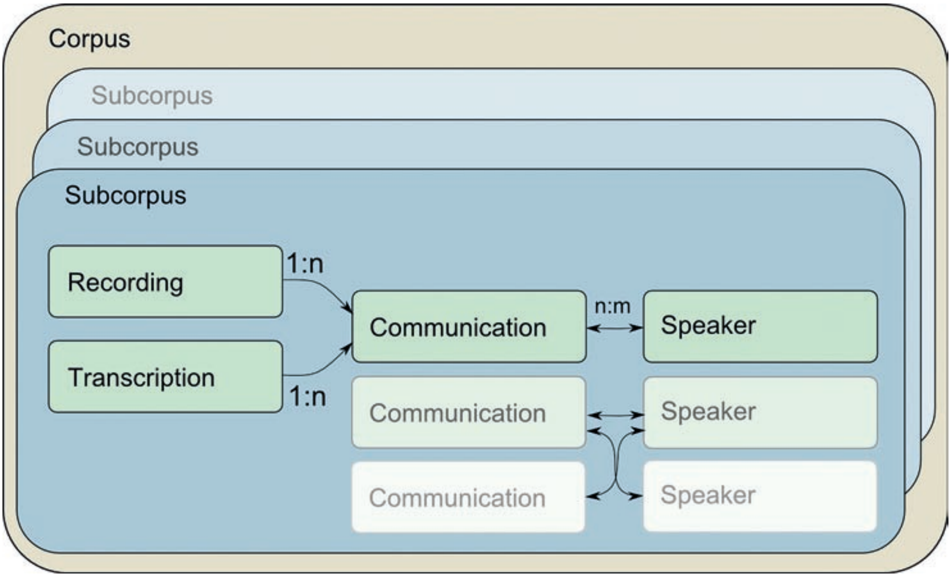
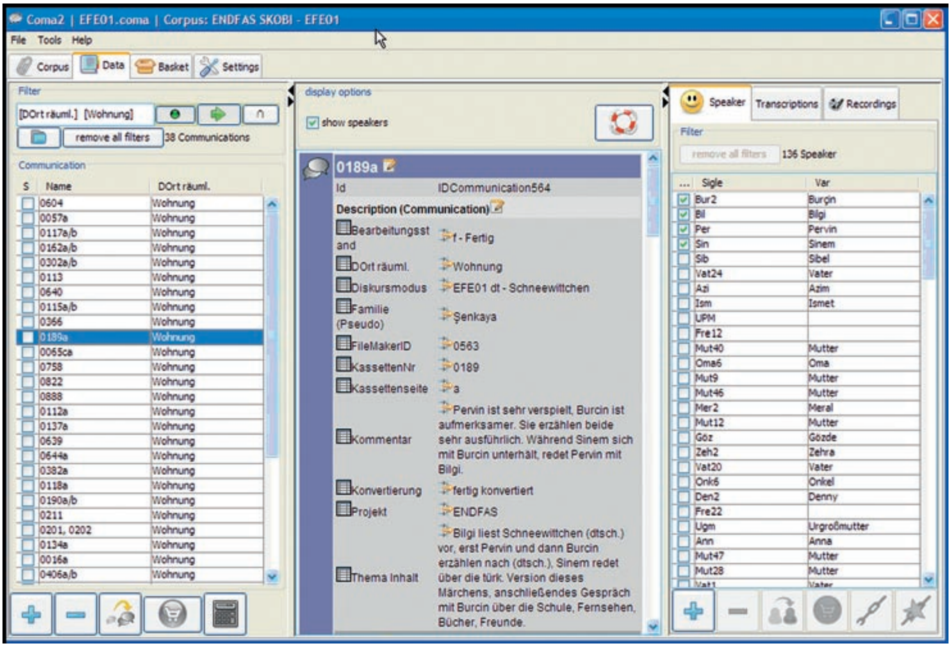


Figure 2: CoMa data model



[DOrt räuml.] [Wohnung]

38 Communications

S	Name	DOrt räuml.
□	0604	Wohnung
□	0057a	Wohnung
□	0117a/b	Wohnung
□	0162a/b	Wohnung
□	0302a/b	Wohnung
□	0113	Wohnung
□	0640	Wohnung
□	0115a/b	Wohnung
□	0366	Wohnung
□	0189a	Wohnung
□	0065ca	Wohnung
□	0758	Wohnung
□	0822	Wohnung
□	0888	Wohnung
□	0112a	Wohnung
□	0137a	Wohnung
□	0639	Wohnung
□	0644a	Wohnung
□	0382a	Wohnung
□	0118a	Wohnung
□	0190a/b	Wohnung
□	0211	Wohnung
□	0201, 0202	Wohnung
□	0134a	Wohnung
□	0016a	Wohnung
□	0406a/b	Wohnung

☑ show speakers

0189a

Id	IDCommunication564
Description (Communication)	
Bearbeitungsstand	f - Fertig
DOrt räuml.	Wohnung
Diskursmodus	EFE01 dt - Schneewittchen
Familie (Pseudo)	Şenkaya
FileMakerID	0563
KassettenNr	0189
Kassettenseite	a
Kommentar	Pervin ist sehr verspielt, Burcin ist aufmerksamer. Sie erzählen beide sehr ausführlich. Während Sinem sich mit Burcin unterhält, redet Pervin mit Bilgi.
Konvertierung	fertig konvertiert
Projekt	ENDFAS
Thema Inhalt	Bilgi liest Schneewittchen (dtsch.) vor, erst Pervin und dann Burcin erzählen nach (dtsch.), Sinem redet über die türk. Version dieses Märchens, anschließendes Gespräch mit Burcin über die Schule, Fernsehen, Bücher, Freunde.

136 Speaker

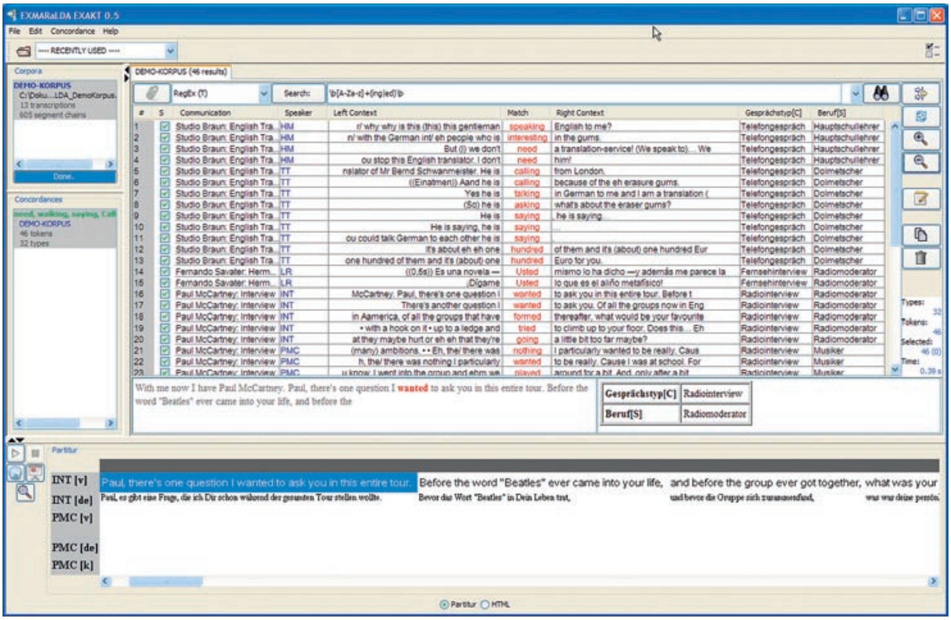
...	Sigle	Var
☑	Bur2	Burçin
☑	Bil	Bilgi
☑	Per	Pervin
☑	Sin	Sinem
□	Sib	Sibel
□	Vat24	Vater
□	Azi	Azim
□	Ism	Ismet
□	UPM	
□	Fre12	
□	Mut40	Mutter
□	Oma6	Oma
□	Mut9	Mutter
□	Mut46	Mutter
□	Mer2	Meral
□	Mut12	Mutter
□	Göz	Gözde
□	Zeh2	Zehra
□	Vat20	Vater
□	Onk6	Onkel
□	Den2	Denny
□	Fre22	
□	Ugm	Urgroßmutter
□	Ann	Anna
□	Mut47	Mutter
□	Mut28	Mutter
□	Vat1	Vater

Figure 3: User interface of the Corpus Manager

The Corpus Manager presents communications, speakers, recordings and transcriptions in a graphical user interface (Figure 3), allowing the user to enter metadata for each of them, either in the form of freely definable attribute-value pairs, or as one of several pre-defined data types (e.g. location, language). A filtering mechanism can be applied to select specific communications or speakers on the basis of the meta-data (e.g. all communications taking place in Turkey, only speakers younger than 20 years), and to extract a subcorpus for that selection.

3. 3 Query: EXAKT

For querying and analysing corpora, EXMARaLDA provides the "EXMARaLDA Analysis and Concordance Tool" (EXAKT). The basic functionality of EXAKT is modelled after the classical corpus analysis instrument: a KWIC (keyword in context) concordancer. After having loaded a corpus compiled in the Corpus Manager, users can enter a search expression. Several types of search expressions are offered, the most common of which is a regular expression, that is, a pattern specifying a string or a set of strings (e.g. "\b[A-Za-z]+(ing|ed)\b" for all words ending in 'ing' or 'ed').



DEMO-KORPUS
C:\Doku....LDA_DemoKorpus.
13 transcriptions
605 segment chains

need, walking, saying, Call
DEMO-KORPUS
46 tokens
32 types

DEMO-KORPUS (46 results)

RegEx (T) Search: \b[A-Za-z]+(ing|ed)\b

#	S	Communication	Speaker	Left Context	Match	Right Context	Gesprächstyp[C]	Beruf[S]
1	☑	Studio Braun: English Tra...	HM	r/ why why is this (this) this gentleman	speaking	English to me?	Telefongespräch	Hauptschullehrer
2	☑	Studio Braun: English Tra...	HM	n/ with the German int/ eh people who is	interesting	in the gums.	Telefongespräch	Hauptschullehrer
3	☑	Studio Braun: English Tra...	HM	But (I) we don't	need	a translation-service! (We speak to)... We	Telefongespräch	Hauptschullehrer
4	☑	Studio Braun: English Tra...	HM	ou stop this English translator. I don't	need	him!	Telefongespräch	Hauptschullehrer
5	☑	Studio Braun: English Tra...	TT	nslator of Mr Bernd Schwanmeister. He is	calling	from London.	Telefongespräch	Dolmetscher
6	☑	Studio Braun: English Tra...	TT	((Einatmen)) Aand he is	calling	because of the eh erasure gums.	Telefongespräch	Dolmetscher
7	☑	Studio Braun: English Tra...	TT	Yes he is	talking	in German to me and I am a translation (	Telefongespräch	Dolmetscher
8	☑	Studio Braun: English Tra...	TT	(So) he is	asking	what's about the eraser gums?	Telefongespräch	Dolmetscher
9	☑	Studio Braun: English Tra...	TT	He is	saying	, he is saying...	Telefongespräch	Dolmetscher
10	☑	Studio Braun: English Tra...	TT	He is saying, he is	saying	...	Telefongespräch	Dolmetscher
11	☑	Studio Braun: English Tra...	TT	ou could talk German to each other he is	saying	.	Telefongespräch	Dolmetscher
12	☑	Studio Braun: English Tra...	TT	It's about eh eh one	hundred	of them and it's (about) one hundred Eur	Telefongespräch	Dolmetscher
13	☑	Studio Braun: English Tra...	TT	one hundred of them and it's (about) one	hundred	Euro for you.	Telefongespräch	Dolmetscher
14	☑	Fernando Savater: Herm...	LR	((0,5s)) Es una novela —	Usted	mismo lo ha dicho —y además me parece la	Fernsehinterview	Radiomoderator
15	☑	Fernando Savater: Herm...	LR	¡Dígame	Usted	lo que es el aliño metafísico!	Fernsehinterview	Radiomoderator
16	☑	Paul McCartney: Interview	INT	McCartney. Paul, there's one question I	wanted	to ask you in this entire tour. Before t	Radiointerview	Radiomoderator
17	☑	Paul McCartney: Interview	INT	There's another question I	wanted	to ask you. Of all the groups now in Eng	Radiointerview	Radiomoderator
18	☑	Paul McCartney: Interview	INT	in Aamerica, of all the groups that have	formed	thereafter, what would be your favourite	Radiointerview	Radiomoderator
19	☑	Paul McCartney: Interview	INT	• with a hook on it • up to a ledge and	tried	to climb up to your floor. Does this... Eh	Radiointerview	Radiomoderator
20	☑	Paul McCartney: Interview	INT	at they maybe hurt or eh eh that they're	going	a little bit too far maybe?	Radiointerview	Radiomoderator
21	☑	Paul McCartney: Interview	PMC	(many) ambitions. • • Eh, the/ there was	nothing	I particularly wanted to be really. Caus	Radiointerview	Musiker
22	☑	Paul McCartney: Interview	PMC	h, the/ there was nothing I particularly	wanted	to be really. Cause I was at school. For	Radiointerview	Musiker
23	☑	Paul McCartney: Interview	PMC	u know. I went into the group and ehm we	played	around for a bit. And, only after a bit	Radiointerview	Musiker

Types: 32
Tokens: 46
Selected: 46 (0)
Time: 0.39 s

With me now I have Paul McCartney. Paul, there's one question I **wanted** to ask you in this entire tour. Before the word "Beatles" ever came into your life, and before the

Gesprächstyp[C]	Radiointerview
Beruf[S]	Radiomoderator

INT [v] Paul, there's one question I wanted to ask you in this entire tour. Before the word "Beatles" ever came into your life, and before the group ever got together, what was your
INT [de] Paul, es gibt eine Frage, die ich Dir schon während der gesamten Tour stellen wollte. Bevor das Wort "Beatles" in Dein Leben trat, und bevor die Gruppe sich zusammenfand, was war deine persön
PMC [v]
PMC [de]
PMC [k]

Figure 4: User interface of EXAKT

As Figure 4 demonstrates, the result of such a query is first presented as a keyword in context concordance, consisting of the matched expression itself with its immediately preceding and following context, typically the words uttered by the same speaker right before and after the word(s) matched by the search expression. As in other concordancing tools, this result can then be sorted by the left or right context column in order to facilitate the discovery of context regularities.

As discussed above, however, the analysis of spoken language data often requires additional context data, such as information about the place and time of the interaction or about a speaker's biography. For additional interactional context, EXAKT offers the possibility to display the corresponding part of a full musical score transcription (or the full transcription in some other layout) by double-clicking on any search result. Similarly, the corresponding part of the audio or video recording can be played back. In order to access meta-data about communications and speakers (as entered in CoMa), users can select arbitrary attributes to be displayed in additional columns of the KWIC table.

Spoken language research is often of an explorative nature, that is, researchers do not approach the data with an *a priori* hypothesis in mind, but rather derive their hypotheses through a step-by-step interaction with the data. EXAKT supports such a 'corpus-driven' (rather than just 'corpus-based') approach by allowing a stepwise filtering, manual annotation and selection, and combination of search results.

4. Conclusion

The first part of this paper has discussed a few requirements corpus technology must fulfil in order to be used for the creation and the work with spoken language corpora. In summary, these requirements are:

- theory-awareness, that is, the recognition of the fact that spoken language corpora, more than written language corpora, are theory-dependent models of linguistic facts;
- a data model which adequately deals with the special structural relations occurring in spoken language; and,
- a dedicated approach to supporting quick and efficient data creation.

The EXMARaLDA system demonstrated in the second part of this paper attempts to meet these requirements by:

- using a time-based, rather than a hierarchy-based data model;
- separating theory-dependent from theory-independent constructs in the transcription interface and data model;
- supporting several widely used transcription systems as the embodiment of different theoretical approaches to spoken language;
- keeping a close link between recordings, transcriptions and metadata; and,
- paying attention to speed and efficiency in the transcription process.

Hopefully, EXMARaLDA can thus make a contribution to prevent spoken language from remaining a 'lesser' studied type of language in corpus linguistics.

References

Biber, D. (1993). "Representativeness in Corpus Design", *Linguistic and Literary Computing,* 8(4), 243–257.

De Rose, S. / Durand, D. / Mylonas, E. / Renear, A. (1990). "What is Text, Really?", *Journal of Computing in Higher Education,* 1(2), 3–26.

Geyken, A. (2009). "The DWDS corpus: A reference corpus for the German language of the 20th century" in Fellbaum, C. (ed.) (2009*). Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies*. Continuum Press.

Koch, P. / Oesterreicher W. (1985). "Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgebrauch", *Romanistisches Jahrbuch,* 36 S, 15–43.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* Mahwah, NJ: Lawrence Erlbaum.

Ochs, E. (1979). "Transcription as Theory" in Ochs E. / Schieffelin B. B. (eds.) (1979*). Developmental Pragmatics*. New York, San Francisco, London: Academic Press, 43–72.

Oostdijk, N. / Broeder D. (2003). "The Spoken Dutch Corpus and Its Exploitation Environment" in *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora* (LINC-03), April, 14, 2003. Budapest, Hungary.

Rehbein, J. / Schmidt, T. / Meyer, B. / Watzke, F. / Herkenrath, A. (2004). "Handbuch für das computergestützte Transkribieren nach HIAT" in *Arbeiten zur Mehrsprachigkeit*, Folge B, 561ff.

Schmidt, T. (2005a). *Computergestützte Transkription – Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln*. Frankfurt a. M.: Peter Lang.

Schmidt, T. (2005b). "Modellbildung und Modellierungsparadigmen in der computergestützten Korpusanalyse" in Fisseni, B. / Schmitz, H. / Schröder, B. / Wagner, P. (eds.) (2005*) Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*. Beiträge zur GLDV-Tagung 2005 in Bonn. Sprache, Sprechen und Computer – Computer Studies in Language and Speech 8. Frankfurt a. M.: Peter Lang.

Schmidt, T. (2005c). "Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech" in *Arbeiten zur Mehrsprachigkeit*, Folge B, 62.

Selting, M. / Auer, P. / Barden, B. / Bergmann, J. / Couper-Kuhlen, E. / Günthner, S. / Meier, C. / Quasthoff, U. / Schlobinski, P. / Uhmann, S. (1998). "Gesprächsanalytisches Transkriptionssystem (GAT)", *Linguistische Berichte,* 173, 91–122.

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Wien u. a.: Springer.

Transcription Bottleneck of Speech Corpus Exploitation

Caren Brinckmann

While written corpora can be exploited without any linguistic annotations, speech corpora need at least a basic transcription to be of any use for linguistic research. The basic annotation of speech data usually consists of time-aligned orthographic transcriptions. To answer phonetic or phonological research questions, phonetic transcriptions are needed as well. However, manual annotation is very time-consuming and requires considerable skill and near-native competence. Therefore it can take years of speech corpus compilation and annotation before any analyses can be carried out. In this paper, approaches that address the transcription bottleneck of speech corpus exploitation are presented and discussed, including crowdsourcing the orthographic transcription, automatic phonetic alignment, and query-driven annotation. Currently, query-driven annotation and automatic phonetic alignment are being combined and applied in two speech research projects at the Institut für Deutsche Sprache (IDS), whereas crowdsourcing the orthographic transcription still awaits implementation.

1. Introduction

Generally, written corpora are more easily accessible and exploitable than speech corpora. There are many written corpora readily available for research, even very large corpora such as DeReKo[1] for German (3.4 billion words). Using the 'web as corpus', specialised corpora can be constructed with readily available tools[2]. Once compiled, written corpora can be exploited without the addition of any annotation. This annotation-free approach was applied for the extraction of higher-order collocations in the CCDB[3]. Compared to written corpora, far less speech corpora are available, and those that can be used for research are often rather small.

[1] http://www.ids-mannheim.de/kl/projekte/korpora

[2] E.g. Web as Corpus ToolKit: http://www.drni.de/wac-tk/

[3] http://corpora.ids-mannheim.de/ccdb/

To exploit speech corpora for linguistic research, at least a basic transcription is needed. This basic transcription is usually an orthographic transcription, often without any punctuation marks. For languages without a standardised orthography, however, it is not clear which form the basic transcription should take.

For accessibility and further analyses, it is essential that the transcription is aligned with the speech signal. Most of the widely used speech annotation tools such as ELAN[4] or Praat (Boersma & Weenink 2009) do not allow a transcription without text-to-audio alignment. In several speech corpus projects (e.g. Spoken Dutch Corpus CGN[5]), 2 to 3 second inter-pause stretches of running speech are segmented and transcribed.

Most experienced speech corpus builders agree that only near-native speakers can produce a reliable orthographic transcription. This poses a problem for minority languages and dialectal speech, if near-native speakers are not available at the corpus building institution. Crowdsourcing the orthographic transcription might be a solution to this transcriber-scarcity problem and is described in detail in Section 2.

Further types of annotations can be added, such as phonetic segmentations and transcriptions for phonetic and phonological research, or prosodic labels such as ToBI[6]. When the orthographic transcription is available, all types of annotations that can be added to written corpora can be added to speech corpora as well, for example, part-of-speech, information structure, or co-references.

Manual phonetic segmentation and transcription is a very time-consuming task (ca. 1:200, that is, at least 200 hours are needed for one hour of phonetically transcribed speech). Furthermore, it requires considerable skill and training. Automatic broad phonetic alignment (discussed in Section 3) and query-driven annotation (presented in Section 4) can facilitate this task.

The overall aim of this paper is to provide an overview of these three different approaches that can be combined to overcome the transcription bottleneck of speech corpus exploitation.

[4] http://www.lat-mpi.eu/tools/elan/

[5] http://lands.let.kun.nl/cgn/ehome.htm

[6] http://www.ling.ohio-state.edu/~tobi/

2. Crowdsourcing

The term 'crowdsourcing' was popularised by Howe (2006) and is a blend of 'crowd' and 'outsourcing'. Outsourcing is defined as subcontracting a task or process traditionally performed by an employee (such as product design or manufacturing) to a third-party company. Crowdsourcing means outsourcing a task to an undefined, generally large group of people. Even though crowdsourcing is often associated with Web 2.0, classical crowdsourcing has existed before the Internet. For example, at self-service restaurants, supermarkets, IKEA, automatic teller machines and ticket vending machines, customers perform tasks (sometimes with the help of machines) that formerly were carried out by an employee. In the Web 2.0 era, the Internet is used to publicise and manage crowdsourcing projects. The most popular crowdsourcing project is Wikipedia[7], where volunteers collaborate to write and constantly improve online lexica. Another related concept is the 'wisdom of crowds' (Surowiecki 2004), meaning that the aggregation of information in groups result in decisions that are often better than could have been made by any single member of the group.

2. 1 Examples of crowdsourcing

2. 1. 1 CastingWords @ Amazon Mechanical Turk

In 2005 Amazon launched a web-based service called the 'Amazon Mechanical Turk'[8]. The name is derived from Wolfgang von Kempelen's 18th-century chess-playing machine, which turned out to be hiding a human chess master inside. Anyone with a US-billing address can become a 'requester' and create so-called 'human intelligence tasks' (HITs). These tasks can usually be carried out easily by humans but are still hard to automatise. HITs are often small tasks worth less than a dollar, sometimes even only a few cents. Anyone with an Amazon account can become a 'turker', complete a task and claim the advertised reward.

The US company CastingWords offers transcription services and uses the Mechanical Turk to crowdsource all necessary tasks: "After a transcription assignment is accepted by a worker, and completed, it goes back out on Mturk.com for quality assurance, where another worker is paid a few cents to verify that it's a faithful transcript of the

[7] http://www.wikipedia.org/

[8] https://www.mturk.com/

audio. Then, the transcript goes back on Mturk.com a third time for editing, and even a fourth time for a quality assurance check" (Mieszkowski 2006: 2).

2. 1. 2 Distributed Proofreaders

Distributed Proofreaders[9] is a non-commercial platform that provides a Web-based method to support the conversion of public domain books from the Project Gutenberg[10] into electronic texts. Volunteers proofread pages that have been scanned and converted with optical character recognition (OCR) software. As of March 2009, more than 15,000 books have been converted into electronic texts.

2. 1. 3 Ph@ttSessionz

In order to document regional variation in speech, usually speakers from different dialectal regions who live near the recording site are recorded (cf. König 1989: 20), or the researchers visit the respective areas to carry out the recordings (e.g. 'German Today', see Section 4.3). This is a very time-consuming process. A different solution is provided by the software SpeechRecorder, which is now part of the WikiSpeech[11] system (Draxler & Jänsch 2008: 1647). SpeechRecorder was used for Ph@ttSessionz, a project carried out at the Institute of Phonetics and Speech Processing at LMU Munich. The aim of the project was to collect teenage speech from all over Germany. Participating schools were sent a recording device to be connected to a PC with broadband Internet connection. The recordings were prompted via the Web-based SpeechRecorder, which immediately uploaded the recorded speech to a central server. Thus, the speakers themselves carried out all recordings without the presence of a researcher or technician at the recording sites.

2. 1. 4 The ESP Game and Google Image Labeler

The aim of the ESP game (von Ahn & Dabbish 2004) is to collect labels for images, a task that is still hard for computers, but easy for humans. This rather dull task is turned into a game, where two persons are shown the same image and have to guess what the other person is typing. The players are given points for each image for which

[9] http://pgdp.net/

[10] http://www.gutenberg.org/

[11] http://wikispeech.org/

they agree on a word. Players can pass five skill levels based on the number of accumulated points, which adds an additional goal-based motivation to the game. The ESP game was licensed to Google for the Google image labeler[12] in 2006. As of July 2008, 200,000 players had contributed more than 50 million labels (von Ahn & Dabbish 2008: 60).

2. 1. 5 reCAPTCHA

Von Ahn and colleagues also invented reCAPTCHA[13] (von Ahn et al. 2008). A CAPTCHA is a test to determine whether the user is a human or a computer. Usually the user is asked to decipher several severely distorted characters—a task that is still hard for computers—, and many websites use CAPTCHAs as security measures against spam. The reCAPTCHA system presents one known distorted character string together with a word from scanned texts that could not be recognised by an OCR program. This way it helps to digitise old printed material. Recently, the reCAPTCHA system has been extended to include sound files, thus helping to transcribe historic recordings.[14]

2. 2 Guidelines for successful crowdsourcing

Of course it is not sufficient to set up a webpage describing the problem at hand, hoping that some volunteers will send an e-mail offering to help. There are some guidelines to follow for successful crowdsourcing (cf. Hempel 2006):

1. *Focus*: Vaguely defined problems get vague answers. Every task should be described as clearly as possible together with a set of rules. Often it is advisable to split up a large task into several smaller ones and to provide a suitable infrastructure (Web-based platform, software, etc.).
2. *Filter*: Use the crowd and experts to extract the best answers. Even though in many social networks only 1 % of the users generate original content, another 10 % comment on it or change it, and 89 % are just passive observers, crowdsourcing often produces a wealth of material that has to be filtered. While the final filtering can

[12] http://images.google.com/imagelabeler/

[13] http://recaptcha.net/

[14] http://blog.recaptcha.net/2008/12/new-audio-recaptcha.html

be done by the task requester, the crowd can and should be used for at least the first filtering step.

3. *Reward*: Since many of the crowd-sourced tasks are rather dull, it is indispensible to offer incentives. Depending on the task and the requester this can be money (mturk), recognition (dp, Ph@ttSessionz), or fun (ESP game).

2. 3 Possible application: Crowdsourcing the orthographic transcription of the speech corpus 'German Today'

The IDS speech corpus project 'German Today' aims at determining the amount of regional variation in (near-)standard German spoken by young and older educated adults and to identify and locate regional features (Brinckmann et al. 2008). To this end, secondary school students and 50-to-60-year-old locals were recorded in 195 cities throughout the German speaking area of Europe (Germany, Austria, Switzerland, Liechtenstein, Luxemburg, South Tyrol, and East Belgium). More than 800 speakers read a number of short texts and a word list, named pictures, translated words and sentences from English, answered questions in a sociobiographic interview, and took part in a Map Task experiment (Anderson et al. 1991). The resulting corpus comprises over 1200 hours of read and spontaneous speech.

Currently, only the read speech and the interview data are transcribed. The Map Task data contains some rather dialectal speech and can only be transcribed reliably by near-native speakers of the dialect who are not available at the IDS. Therefore, the Map Task data has currently not been processed at all, which is rather unfortunate.

The proposed system for crowdsourcing the orthographic transcription consists of a central database and a software application controlling the transcription process (see Figure 1). The database stores the speech signals, metadata, transcripts, and information about the transcribers and the transcription process.

- *Database of speech files and metadata*: As a first step, the corpus provider fills the central database with the speech signals that are to be transcribed and the corresponding metadata. It might be necessary to pre-process the speech signals before storing them in the database. For example, 'German Today' Map Task speech signals consist of two channels that have to be transcribed separately, and therefore split up into two separate mono files. The metadata should contain basic information about the speech file, for example, recorded task, number of speakers, sex and age of speakers, and place of recording. The latter is crucial for dialectal speech when the transcribers can opt for certain regions they are able to transcribe.

- *Task definition*: The database also contains clear and detailed task descriptions for each corpus, including conventions regarding transcription, grading and correction tasks. These conventions contain several examples and are presented to each registered transcriber of the corpus.
- *Process control*: The corpus providers have to define some parameters of the transcription process, for example, the maximum duration of presented speech signals, the number of human transcribers who have to transcribe each speech signal in parallel, grades for rating the quality of a transcription, and the awarded points for each task.
- *Database of human transcribers*: The human transcribers have to register before they can start transcribing, and for dialectal speech they are able to select certain regions by listening to examples. The database also stores which speech files each transcriber has already transcribed, graded or corrected, as well as the grades she/he received for transcriptions.
- *Transcription process*: When a transcriber is logged on, the process controlling software presents all available tasks (filtered by regions, where applicable):
  - o *Initial transcription*: The speech files are presented in inter-pause stretches that do not exceed the specified maximum duration (e.g. 3 seconds). The Web-based transcription tool can be implemented in a very simple fashion using a publicly available media player[15] and an input form for the transcription. Solutions that allow more control include WebTranscribe (Draxler 2005).
  - o *Grading*: Each transcription of an inter-pause stretch is graded by another human transcriber according to a predefined scale (e.g. 0=severe errors, 1=some error(s), 2=error-free)
  - o *Correction*: If parallel transcriptions differ or the received grade is not 'error-free', the transcription is corrected and graded again. If all (corrected) transcriptions are identical and graded as error-free, the transcription is stored as final version.
- *Rewards*: Transcribers earn points for each task modified by the received grade. These points are posted as high-score lists on the website, for example, as All-Time Top Transcribers or Today's Top Transcribers. Virtual titles can be awarded as well as real-world incentives (such as a visit to the corpus-providing research institute).

[15] For example, http://www.longtailvideo.com/players/jw-flv-player/

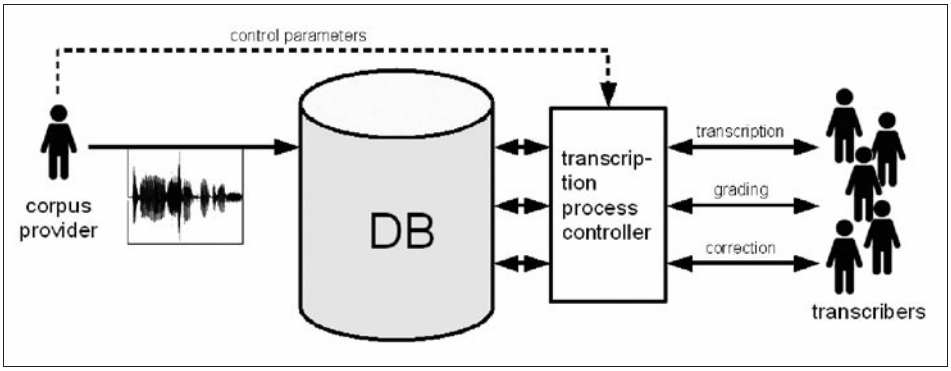


Figure 1: Simplified architecture of the proposed system for crowdsourcing transcriptions

Potential transcribers for 'German Today' could be recruited by contacting the schools where the original recordings were made. The students whose speech was recorded are usually interested in the project and have the time needed to participate as transcribers of regional speech. However, crowdsourcing the orthographic transcription is only feasible in countries with many broadband connections, which might not be the case for some lesser-used languages.

3. Automatic phonetic alignment

To facilitate (or even replace) manual phonetic segmentation and labeling, automatic phonetic alignment software has been developed. These systems usually produce time-aligned broad phonetic transcriptions (without fine phonetic detail) and need the following:

- speech signals
- orthographic transcriptions of the speech signals
- canonical/phonemic transcription of all words in the corpus: this can be provided by simple look-up in a pronunciation lexicon (e.g. CELEX: Baayen et al. 1995) or by a grapheme-to-phoneme converter. The former approach is only feasible for small corpora with a limited lexicon (e.g. read speech). Often, a grapheme-to-phoneme converter is combined with a lexicon of pronunciation exceptions.
- language-specific phoneme models (often trained HMMs), which are usually part of the alignment software.

Often, the automatic aligner can operate in two different modes: forced alignment and alignment with post-lexical phonological processes. In forced alignment, the given transcription of each word is not changed and 'forced' onto the speech signal. If a sound present in the transcription has been deleted by the speaker, the system will nonetheless label a part of the speech signal with the sound's symbol. The forced alignment mode is useful if one's analysis is based on the canonical transcription or if a detailed manual phonetic transcription is available that has to be time-aligned. The other mode tries to model post-lexical phonological processes (deletions, replacements, and insertions) and changes the given transcription accordingly. For all analyses that depend on the realised form, this is the preferred mode.

Van Bael et al. (2006, 2007) compared 10 aligners for Dutch with a manually-obtained reference transcription as 'gold-standard'. They found that a system that used canonical transcriptions and models post-lexical phonological processes with a decision tree performed best. Furthermore, the number of remaining disagreements with the reference transcription (14.6 % for spontaneous speech, 8.1 % for read speech) was only slightly higher than human inter-labeler disagreement scores reported in the literature.

For the project 'German Today' (see Section 4.3) and a research project on word-internal boundary effects, we carried out an informal task-based evaluation of the Munich Automatic Segmentation System MAUS[16] for German (Schiel 2004). During evaluation we found that MAUS produces some obvious errors such as extreme duration outliers, but these can easily be detected automatically and marked for manual correction. Apart from that, it depends very much on the task at hand whether an automatic segmentation with MAUS is useful or not:

- *Corpus access*: For the corpus 'German Today' MAUS proved to be especially useful for accessing specific portions of the speech signal for further manual annotation.
- *Analyses of segmental durations* can be based on the automatically aligned segment boundaries as well. However, we found that only large significant effects can be detected by relying on the automatic segmentation. Therefore, it is useful for a first gauge of the effect.
- For *analyses in the frequency domain* (e.g. formant slope), which depend on accurate segmental boundaries, the automatically set boundaries should always be corrected by hand.

[16] MAUS can be downloaded from http://www.phonetik.uni-muenchen.de/forschung/Verbmobil/VM14.7eng.html

While phonetic aligners are available for all major languages, this is not the case for less-resourced languages. One solution is to use an existing alignment system and train your own language-specific phoneme-models (e.g. with the Hidden Markov Model Toolkit HTK[17]). The catch is that at least one hour of phonetically segmented and labeled speech data is usually needed as training material. Another solution might be to find an alignment system for a language that is phonetically similar to the target language. Its pre-built phoneme-models could be used, adding a mapping between the phonemes of the target language and the language modeled by the existing aligner.

4. Query-driven Annotation

The traditional corpus annotation process consists of three tasks. First, an annotation schema is developed, then the actual annotation is performed and finally, when the whole corpus is annotated, the corpus can be queried and analysed (see Figure 2). One major problem of this sequential approach is that it is too time-consuming. Large corpora require many years of annotation work before the corpus can be exploited and any results can be published. Furthermore, due to coder drift during the long annotation process (see Gut & Bayerl 2004: 567), the reliability of the annotations can be rather limited. Finally, the corpus queries are restricted to those phenomena which have been annotated beforehand. Some queries might be impossible due to the structure of annotations, and it might be necessary to re-annotate the whole corpus.

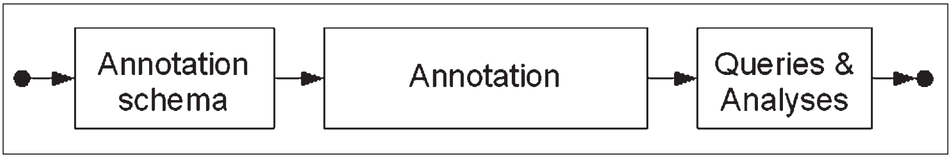


Figure 2: Traditional sequential process of corpus annotation

To overcome these problems, Voormann and Gut (2008) suggest an approach they call ‘agile corpus creation’, where the traditional corpus annotation process is replaced by a cyclic and iterative corpus annotation process. As shown in Figure 3, each annotation cycle starts with the formulation of a query. For example, the duration of schwa

[17] http://htk.eng.cam.ac.uk/

in word-final /C@n/[18] sequences shall be compared between German words ending in a suffix and those not ending in a suffix. Then the annotation schema necessary for this query is specified and the annotation is carried out. After a successful analysis, the next query is formulated and the cycle starts afresh. Sometimes it becomes clear during the annotation or the analysis that the annotation scheme has to be modified, and jumping back within the cycle becomes necessary.

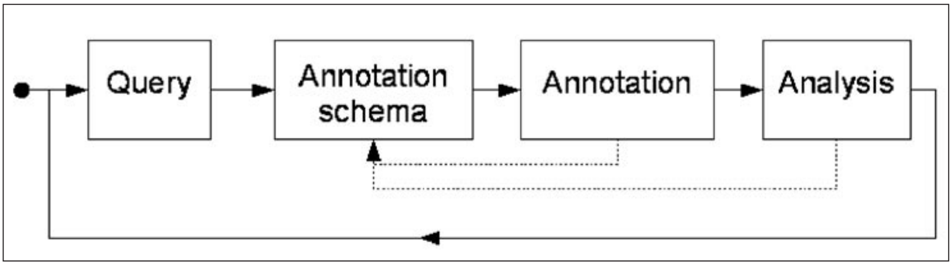


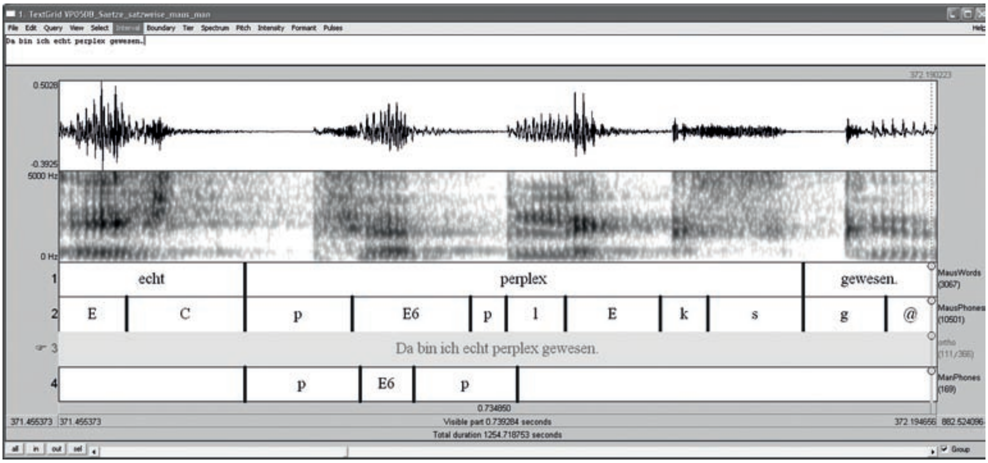
Figure 3: Query-driven corpus annotation process

In this query-driven approach, successive cycles improve the annotation schema and limit it to the elements necessary for the queries. Suitability and consistency of each addition to the annotation schema are immediately put to the test, so that only small amounts of data have to be re-annotated or discarded. Since the annotators focus on one particular annotation task within each cycle, coder drift is seldom a problem. Most importantly, results can be published shortly after the completion of the first annotation cycles. While the query-driven approach can be applied to a corpus of any size, for large corpora compiled for speech research there is almost no alternative.

To further speed up and facilitate the annotation process, the query-driven approach can be combined with an automatic phonetic alignment. First, the whole corpus is automatically segmented and labeled. These automatically set labels are then used to access those parts of the corpus that are to be annotated manually. Figure 4 shows a small excerpt from a corpus compiled for the study of word-internal boundary effects (Raffelsiefen & Brinckmann 2007: 1442). The third tier (shaded in grey) of the Praat TextGrid contains the orthographic transcription of the complete sentences that were read by the subjects. This orthographic transcription served as input for the phonetic aligner MAUS. Its output was converted to the Praat TextGrid format and is shown

[18] All phonetic symbols in this paper are given in SAMPA: http://www.phon.ucl.ac.uk/home/sampa/

in the first tier (word boundaries) and second tier (phoneme labels and boundaries). A customised Praat script allows the human annotator to comfortably access those parts of the corpus which match a certain query. In the displayed example, the annotator wants to correct the automatic phonetic annotation of all word-initial consonant-[E6]-consonant sequences. The Praat script repeatedly jumps to the respective parts of the corpus and copies the matching phoneme sequences to the fourth tier where the annotator can easily correct the location of the segment boundaries.



Da bin ich echt perplex gewesen.

Figure 4: Combining query-driven annotation and automatic phonetic alignment

5. Conclusion

Compared to written corpora, there are far fewer speech corpora available for linguistic research. Speech corpora are not only more difficult to compile, their exploitation is further hindered by the fact that at least a basic orthographic transcription has to be added to the speech signal. Manual transcription and further linguistic annotation tasks are very time-consuming and require considerable skill. Three approaches addressing this 'transcription bottleneck' have been presented in this paper: crowdsourcing the orthographic transcription, automatic phonetic alignment, and query-driven annotation.

Producing a reliable orthographic transcription is almost impossible for non-native speakers. For dialectal speech or minority languages, transcribers with near-native competence might not be available at the corpus producing institution. Crowdsourcing the orthographic transcription might prove a viable solution as it has been successfully applied for commercial applications. To my knowledge, however, for speech research corpora this is a concept still awaiting implementation.

Once the orthographic transcription is available, the corpus can be treated as a written corpus, adding annotations on different linguistic levels. Phonetic and phonological research requires annotations based on the speech signal, such as phonemic/phonetic segmentation and labeling, which is even more time-consuming than orthographic transcription. Here, an automatic phonetic alignment software is useful. For some tasks, the automatically set segmental boundaries can be used.; for others they have to be corrected manually, which is nonetheless faster than segmenting and labeling the signal from scratch.

For large corpora, the traditional sequential corpus annotation process requires many years of annotation work before any analyses can be carried out. Using the query-driven cyclic approach, results can be published shortly after the completion of the first annotation cycles. At the Institut für Deutsche Sprache, we combine the query-driven annotation process with automatic phonetic alignment in two speech-corpus-based research projects.

References

Anderson, A. / Bader, M. / Bard, E. / Boyle, E. / Doherty, G. / Garrod, S. / Isard, S. / Kowtko, J. / McAllister, J. / Miller, J. / Sotillo, C. / Thompson, H. S. / Weinert, R. (1991). "The HCRC Map Task Corpus", *Language and Speech,* 34 (4), 351–366.

Baayen, R. H. / Piepenbrock, R. / Gulikers, L. (1995). *The CELEX Lexical Database* (Release 2). CD-ROM: Linguistic Data Consortium, University of Pennsylvania, Philadelphia, USA [Distributor].

Boersma, P. / Weenink, D. (2009). *Praat: doing phonetics by computer* (Version 5.1.02) [Computer program]. Retrieved March 9, 2009 from http://www.praat.org/

Brinckmann, C. / Kleiner, S. / Knöbl, R. / Berend, N. (2008). "German Today: an areally extensive corpus of spoken Standard German" in *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Morocco. Retrieved March 9, 2009 from http://www.lrec-conf.org/proceedings/lrec2008/pdf/806_paper.pdf

Draxler, C. (2005). "WebTranscribe—an extensible web-based speech annotation framework" in *Proceedings of the 8th International Conference on Text, Speech and Dialogue* (TSD 2005), Karlovy Vary, Czech Republic, 61–68.

Draxler, C. / Jänsch, K. (2008). "WikiSpeech—A Content Management System for Speech Databases" in *Proceedings of Interspeech 2008*, Brisbane, Australia, 1646–1649.

Gut, U. / Bayerl, P. S. (2004). "Measuring the Reliability of Manual Annotations of Speech Corpora" in *Proceedings of Speech Prosody 2004*, Nara, Japan, 565–568. Retrieved March 9, 2009 from http://www.isca-speech.org/archive/sp2004/sp04_565.pdf

Hempel, J. (2006). "Crowdsourcing: Milk the masses for inspiration", *BusinessWeek*, September 25, 2006. Retrieved March 9, 2009 from http://www.businessweek.com/magazine/content/06_39/b4002422.htm

Howe, J. (2006). "The Rise of Crowdsourcing", *Wired,* 14.06. Retrieved March 9, 2009 from http://www.wired.com/wired/archive/14.06/crowds.html

Keibel, H. / Belica, C. (2007). "CCDB: a corpus-linguistic research and development workbench" in *Proceedings of Corpus Linguistics 2007*, Birmingham, United Kingdom. Retrieved March 9, 2009 from http://www.corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf

König, W. (1989). *Atlas zur Aussprache des Schriftdeutschen in der Bundesrepublik Deutschland. Band 1: Text.* Ismaning: Hueber.

Mieszkowski , K. (2006). "I make $1.45 a week and I love it". Retrieved March 9, 2009 from http://www.salon.com/tech/feature/2006/07/24/turks/

Raffelsiefen, R. / Brinckmann, C. (2007). "Evaluating phonological status: significance of paradigm uniformity vs. prosodic grouping effects" in *Proceedings of the 16th International Congress of Phonetic Sciences* (ICPhS XVI), Saarbrücken, Germany, 1441–1444. Retrieved March 9, 2009 from http://www.icphs2007.de/conference/Papers/1684/1684.pdf

Schiel, F. (2004). "MAUS Goes Iterative" in *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, Portugal, 1015–1018.

Surowiecki, J. (2004). *The Wisdom of Crowds*. Boston: Little, Brown.

Van Bael, C. / Boves, L. / van den Heuvel, H. / Strik, H. (2006). "Automatic phonetic transcription of large speech corpora" in *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006), Genoa, Italy, 4–11.

Van Bael, C. / Boves, L. / van den Heuvel, H. / Strik, H. (2007). "Automatic phonetic transcription of large speech corpora", *Computer Speech and Language*, 21 (4), 652–668.

von Ahn, L. / Dabbish, L. (2004). "Labeling Images with a Computer Game" in *Proceedings of the SIGCHI conference on Human factors in computing systems* (CHI 2004), Vienna, Austria, 319–326.

von Ahn, L. / Dabbish, L. (2008). "General Techniques for Designing Games with a Purpose", *Communications of the ACM*, 51 (8), 58–67.

von Ahn, L. / Maurer, B. / McMillen, C. / Abraham, D. / Blum, M. (2008). "reCAPTCHA: Human-Based Character Recognition via Web Security Measures", *Science* 321, 1465–1469.

Voormann, H. / Gut, U. (2008). "Agile corpus creation", *Corpus Linguistics and Linguistic Theory* 4 (2), 235–251.

Making Educational Content Accessible for the Deaf: The Development of a Multi-level Platform

Eleni Efthimiou

In this paper, we focus on design requirements and implementation issues underlying a platform environment that allows for the development of various educational applications that are fully accessible by deaf users. Subject to 'Design for All' principles, the environment to be discussed as a showcase is built on methodological principles that adopt sign language as the basic means for communication of linguistically-uttered educational content. It also makes extensive use of visual objects to support comprehension and navigation at all levels of human-computer interaction. Currently available instantiations of the environment have incorporated both video-for-content presentation and an avatar-based dynamic sign synthesis mechanism. The educational applications to be referred to when discussing the design principles and user requirements taken into account include Web-based and off-line Greek Sign Language (GSL) teaching tools for the same school level (early primary school) as well as a vocational training tool for adult users.

1. Greek Sign Language (GSL): an Official Minority Language

Sign languages (SLs) are natural languages articulated in 3D space, where the linguistic message is organised according to geometrical parameters for the expression of semantic-syntactic relations, visually represented on and with the signer's body or in the space in front of him/her (Stokoe 1978: 18; Sutton-Spence & Woll 1999; Neidle et al. 2000: 25).

Greek Sign Language (GSL) has been developed as a minority non-written language system in a socio-linguistic environment similar to those of most other known sign languages. It is a natural language system used as the mother language of the Greek deaf community, where estimates of GSL natural signers number at 40,600 (1986 survey, Gallaudet University). In addition to the above, there is also a large number of hearing non-native signers of GSL, mainly students of GSL and the families of deaf people. Records of the Hellenic Federation of the Deaf (HFD) show that in the last five years the demand for classes in GSL as a second language has radically increased (Karpouzis et al. 2007: 55). This is demonstrated by the recently increased demand for GSL knowledge in the local language market, following the recognition of GSL as an official language of the Hellenic State in the year 2000 (Law 2817/2000), with the direct consequence of the subsequent use of the language in education and official communication services.

Following the national policy for integration of people with disabilities into the society, a recent increase of deaf students in mainstream education has been observed. Nonetheless, a considerable proportion of the deaf student population still remains scattered in other institutions, including small local centres for the deaf and private class environments. Beyond legislative measures, integration of the deaf in society is heavily contingent upon the quality of education they receive, and although GSL is the official language for education of the deaf population in Greece, educational materials and tools still remain very poor. This is partially due to the widely-held misconception that, since deaf people are able to see, they are also able to access written material. However, individuals who are born without hearing find it extremely difficult to make associations between physical communication concepts and their written forms. This happens because the written form of an utterance is a convention for the representation of sounds, which is incomprehensible in the case where no perception of sound is possible. According to statistics of the Hellenic Pedagogical Institute, the average reading capability of deaf adults corresponds to the mid-primary school level. This fact is confirmed by measurements and estimates from other sign language studies as well (Huenerfauth 2004: 25). On the other hand, the nature of the language per se acts as an obstacle against the systematic production of representations of its linguistic content in huge quantities, if the requirement is to preserve quality of the delivered message to the native utterance level. This happens because video—though a rather static, not easily reusable source of linguistic content—is currently the only representational means that fully preserves the naturalness of the signing utterance. In the search for a solution to the limitations of video, and to satisfy the growing requirement for Universal Access (Sapountzaki et al. 2006: 161) in today's Information Society, the demand for efficient solutions to the problems of 'deaf human'-computer interaction that includes accessibility to e-content has led to research for the development of dynamic systems for the representation of sign language utterances by means of avatar technologies (Fotinea et al. 2005: V8; Glauert 2002: 21; Marshall & Safar 2002: 384).

This paper focuses on the architectural design and implementation of an educational environment, fully accessible by deaf users, that allows for the development of various educational tools, both Internet-based and off-line.

2. The Language Barrier in Education of the Deaf

Unrestricted integration to an open society is heavily contingent upon the quality of education received by different groups of a given population. In other words, an initial barrier to integration is related to discrimination with respect to one's level of literacy.

Most native speakers of minority languages experience the need to attend a school where the primary language of instruction is different than their 'mother' language. The case of deaf populations is even more complicated, since born-deaf individuals develop a specific type of bilingualism (Mayberry 1993: 62) derived from the demand to grow into a language environment that uses not only non-verbal (i.e. physical) means of articulation, but also uses structures for concepts and relationships of the linguistic message that operate on a completely different level than their particular sign language of origin.

From the legislative point of view, inclusion is supported by the incorporation of deaf students either directly into mainstream education or into special education units inside the mainstream schools, or even by modernising special schools for the deaf. In all cases, educational content should be delivered in sign language. Currently, sign language used in the classroom is supported by mature multimedia and digital image technologies, parallel to the development of user-friendly interfaces for end-users who may not be previously familiar with computer use (Efthimiou & Fotinea 2004: 184). Moreover, appropriate methodologies for content presentation allow for systematic development of tools for the deaf, on the basis of 'Design for All' principles, where the key characteristic of such systems is the use of sign language in order to convey meaning of linguistic content at any level of interaction with the user.

In this context, the teaching of a sign language (GSL in our case) to complete mastery becomes crucial, as it provides the necessary linguistic background for the full academic development of the deaf individual.

In line with the above, we discuss below the main characteristics of an environment that supports sign language teaching as well as teaching in GSL, incorporating educational material presentation and testing mechanisms that make it fully accessible by deaf end-users, both pupils and tutors (Efthimiou & Fotinea 2007: 124).

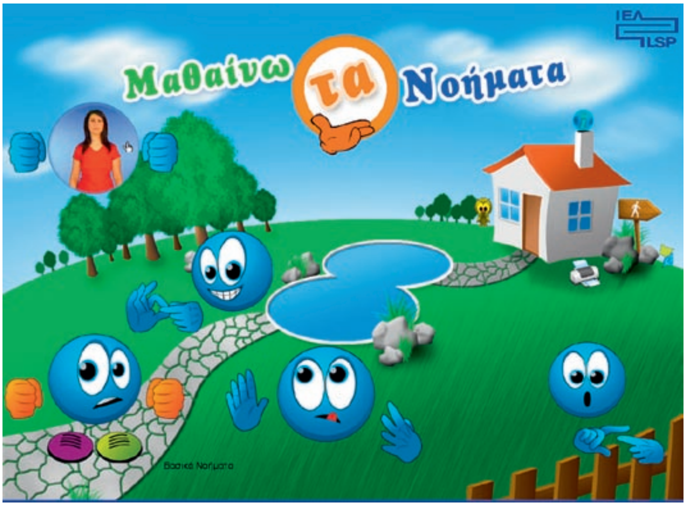


Figure 1: MTN 1 main menu screen

The environment instantiations used to exhibit multi-purpose implementation of the same basic architecture involve a range of educational software products that address the needs of language teaching in primary schools up to vocational training. For the purposes of discussing architectural principles, environment ergonomics and content creation methodology, we make extensive reference to "MTN 1" (from the initials of the full title "I learn the signs" in Greek) (Figures: 1, 2, 3, 4, 5). These are the first in a series of educational software products intended to teach GSL. The educational goal of this specific showcase is to introduce vocabulary primes at early primary school level.

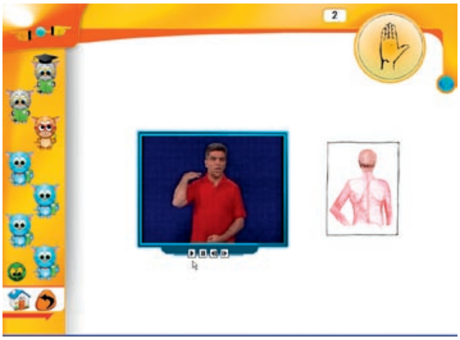


Figure 2: Handshape teaching

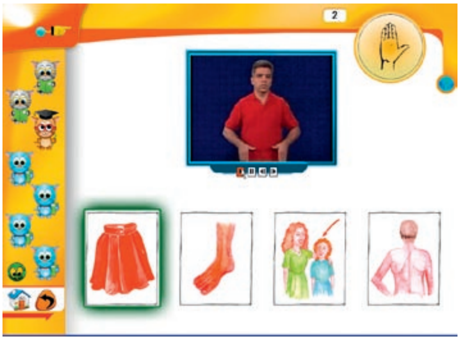


Figure 3: Handshape based interactive vocabulary presentation

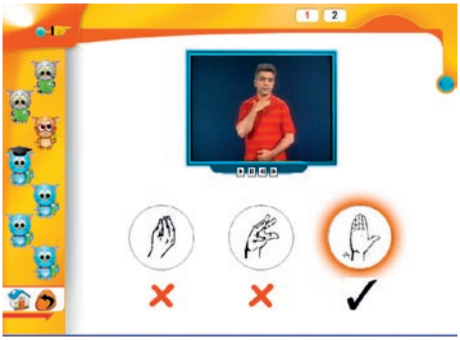


Figure 4: Handshape checking exercise

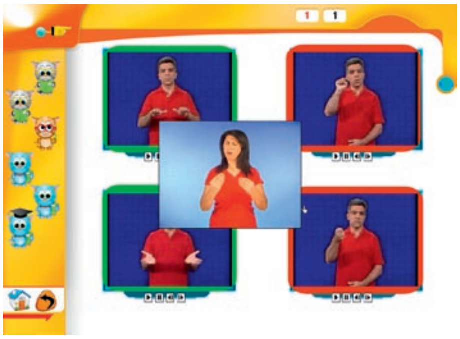


Figure 5: Commentary on exercise execution

The specific application has been also implemented as a bilingual Web-based educational tool. In its Web-based version, MTN 1, as depicted in Figures 6, 7, 8 and 9, has kept all major characteristics related to Human Computer Interaction (HCI) principles, content completeness and focus on sign language as the only means of conveying the teaching objective. This application, however, mainly addresses needs of learners who approach GSL as a second language, rather than native deaf signers.

The same educational goal is served by an experimental Web-based prototype system that allows for the structuring and presentation of GSL educational material and linguistic resources, addressing the needs of GSL grammar teaching to early primary school pupils. In this case, dynamic sign synthesis through the use of avatar technologies for 3D sign representation of linguistic content is tested as an alternative to the use of video (Figures 10, 11 and 12) (Karpouzis et al. 2007: 65; Efthimiou et al. 2006b: 4; Sapountzaki et al. 2006: 166).

3. Environment Design Principles and Implementation Instantiations

Basic environment architecture is built on general principles applied to the design of educational software, along with the presupposition that the linguistic content needs for the users should be represented three-dimensionally. Image or video (capture) are the only means for concept clarification, where decisions about the style of graphics and use of cartoons or photographs depends upon the age group of the users and the educational objective. Design modifications, for example, required for the development of MTN 1, were organised on the basis of the main educational goal and user group needs of very young pupils and their tutors; these goals and needs were discovered through extensive evaluation procedures of various prototype versions of the environment.

The use of graphics, bold colours and game-like screen layouts, as well as the incorporation of student awards for correct responses to lesson exercises and unit tests were derived from design principles applicable to small children. Similarly, user-need studies suggested that video should be the means of representation for GSL in this specific case, whereas other basic functionalities should be adopted, as the need for user-driven repetition of the educational linguistic message, as well as the need for navigation details and button functionalities, presented in GSL and combined with graphics. The use of graphics in this case follows an organisation with clear semantic structure.

Standard educational means for the teaching of GSL vocabulary incorporated into this software include drawings for visual representation of concepts and video for the accurate 3D representation of sign articulation. Video also provides clarifications and instructions in terms of the educational goals, execution of the exercises, and help at all levels of the software. Educational content is presented solely by a native-sign tutor. Presentation of all other linguistic information is provided by the tutor’s assistant, also in GSL.

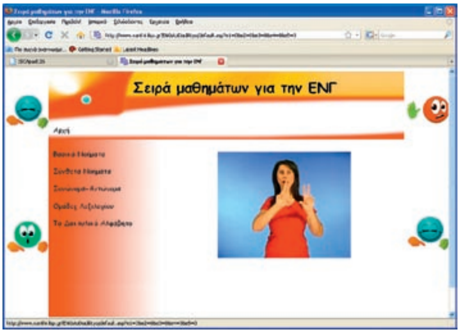


Figure 6: Chapter list presentation

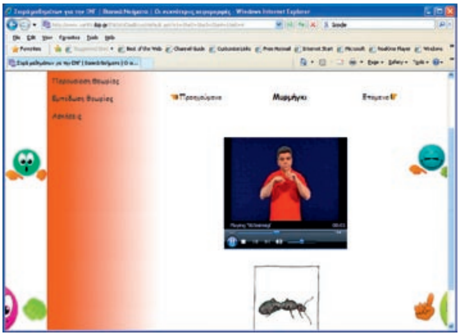


Figure 7: Lemma teaching

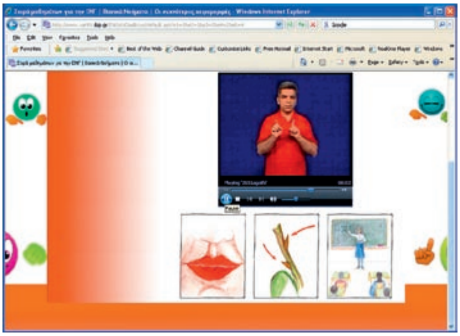


Figure 8: Handshape based interactive vocabulary presentation

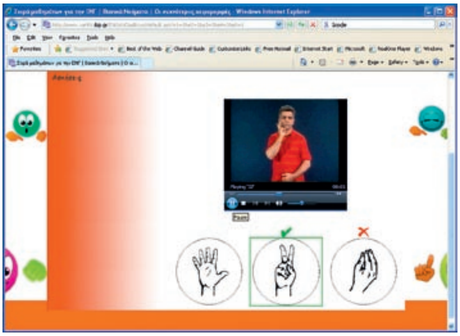


Figure 9: Exercise execution

The use of drawings for concept representation and graphical navigation are product-specific design characteristics. The use of drawings and the respective complete absence of photographs, in this case, are intended to relate a concept with a wider possible range of objects than one may find in the real world. The rationale for the presentation of all navigation functions by graphics is that graphical navigation has been proven worldwide to be a popular way for children to interact with computer environments. A general design characteristic is the total absence of written language for presentation of educational content (except in cases of multilingual terminology as

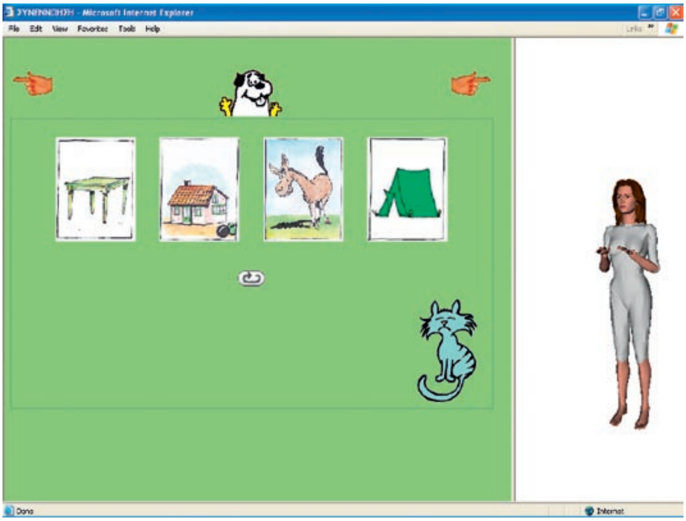

ΣΥΝΕΝΝΟΗΣΗ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links
Done
Internet

Figure 10: Full graphics-based navigation with avatar performed vocabulary presentation

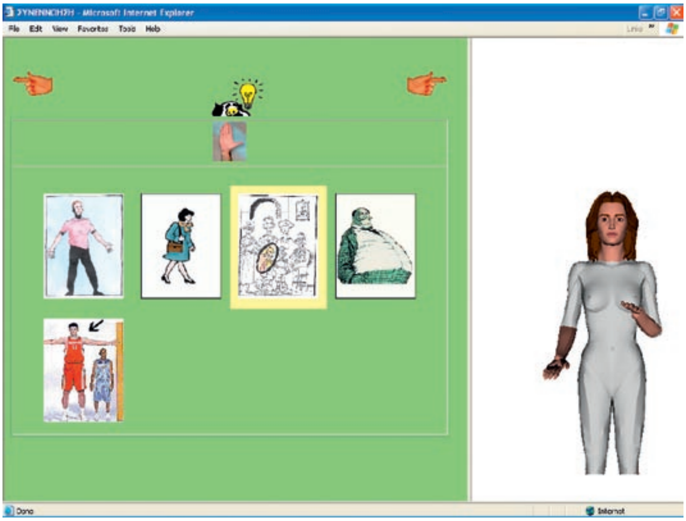

ΣΥΝΕΝΝΟΗΣΗ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Links
Done
Internet

Figure 11: Handshape-based vocabulary presentation with avatar in frontal view

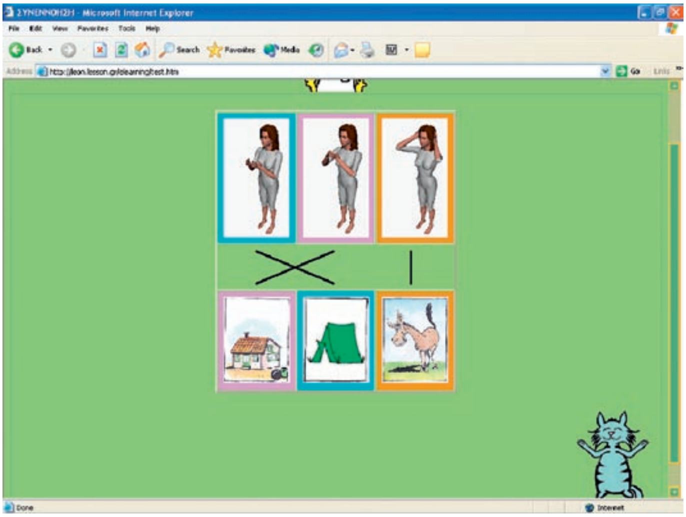

ΣΥΝΕΝΝΟΗΣΗ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://leon.lesson.gr/elearningtest.htm
Go
Links
Done
Internet

Figure 12: Color code-matching exercise

in Figure 13). Written language may appear only after selection of the relevant functionality, for use by hearing tutors when viewing lists such as tables of contents.

The educational software MTN 1 reflects the structure of GSL with respect to principles, features and rules applicable to the vocabulary of the language (Efthimiou & Fotinea 2004b: 821), while at the same time complying with methodological principles, thematic structure and educational activities proposed in the Analytical Curriculum for the teaching of GSL, set by the Hellenic Pedagogical Institute.

The MTN platform utilises GSL as the sole means of communication, excluding the use of written Greek in all levels of communication with the students. MTN supports personalised teaching of the language, adaptable to individual learner needs, to guarantee efficient comprehension of the teaching subject by the student, especially in classes where students present different levels of command of the language.

Implementation of the MTN educational platform was based on the available architecture (Efthimiou & Fotinea 2004: 183; Sapountzaki et al. 2006: 162), on the selection of appropriate educational content with respect to the age of the students, on user needs as reflected in the Hellenic Pedagogical Institute's Analytical Curriculum for GSL (http://www.pi-schools.gr/special_education/kofosi-a/), as well as on the needs/requirements set by user-needs studies of the specific target group (deaf students / GSL teachers) as recorded via informal evaluation procedures undertaken for educational prototypes (Efthimiou et al. 2006b: 8). In addition, for the development of the MTN platform, studies regarding teaching conditions of a sign language and worldwide creation of educational curricula have also been taken into account. The Internet-based prototype that exploits avatar potentials for sign representation adopts the same architecture with the goal to support distance learning.

4. GSL Educational Content

The development of educational software instantiations that makes use of the basic environment design, derives linguistic content from an open pool of reusable resources as regards to both vocabulary items and grammar rules of GSL.

Vocabulary resources are organised in a lexical database of signs (the lexical units of sign languages), where lexical entries are marked for phonological composition, grammar behaviour and semantic characteristics (Efthimiou & Fotinea 2004b: 822).

Grammar rules of GSL are coded in the form of a computational grammar, based on results of basic research on a representative sign language corpus (Efthimiou et. al 2006: 51).

In the case of the development of vocational training software, a further parameter of content creation is the demand for the introduction of GSL into the domain of terminology. Creation of new terms in any natural language follows standard procedures (ISO 1087–1:2000), (ISO/R 704:1968), (ISO 704). Moreover, since terminology is usually introduced to the receiver in the language in which it was originally created, terminological lists are multilingual in the default case. To fulfil this task, the environment is supported by a methodology for term creation in GSL (Efthimiou & Fotinea, 2004b: 822), which allows trilingual (GSL-Greek-English) representation of terminology items in terminology intensive educational applications. In this case, definitions of terminology concepts are visually available (window at the bottom right side of the screen in Figures: 13, 14, 15).

The currently available GSL resources have been used in a number of applications, which include—with the exception of the applications discussed in this paper—a bilingual dictionary of basic GSL vocabulary, with 3000 entries, a children's dictionary of 500 entries and an NLP-based conversion tool from written Greek to GSL.

As regards MTN 1, educational content is organised into five chapters with seventy lessons and 600 lemmata that teach the methodology of sign formation: basic signs, complex signs, synonyms-antonyms and word families.

The above-presented vocational training software provides educational content for basic computer skills (7 ECDL topics) with 350 trilingual terminology items and 650 demonstrators of term usage, covering all possible appearances of a term in the related thematic units.

The educational platform environment that underlies the applications discussed in this paper is subject to continuous evaluation cycles and optimisation, while the linguistic resources databases for GSL content are constantly updated.

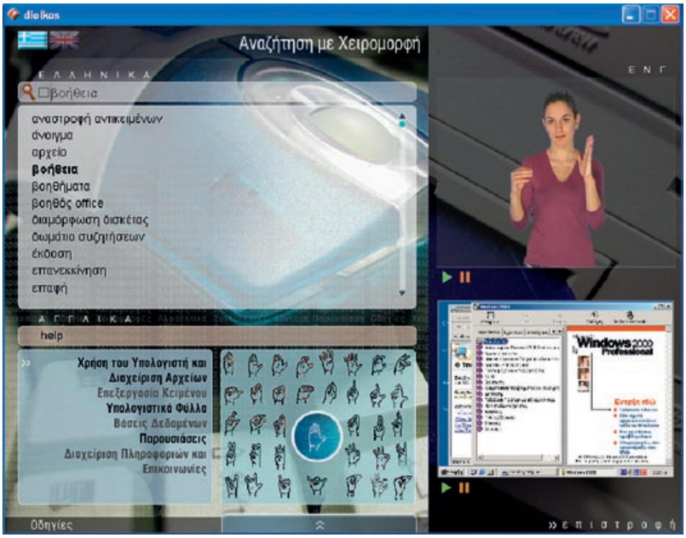


Figure 13: Trilingual terminology presentation for computer skills with term search by handshape and video capture example of term use

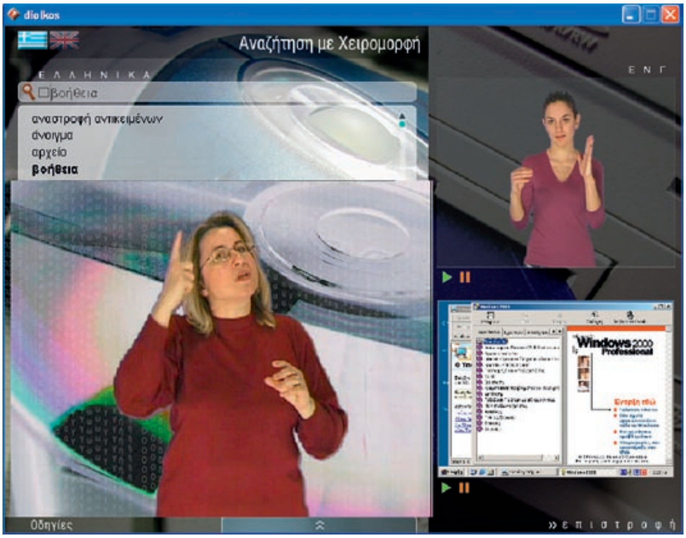


Figure 14: Instructions in GSL, supported by term presentation and example of use

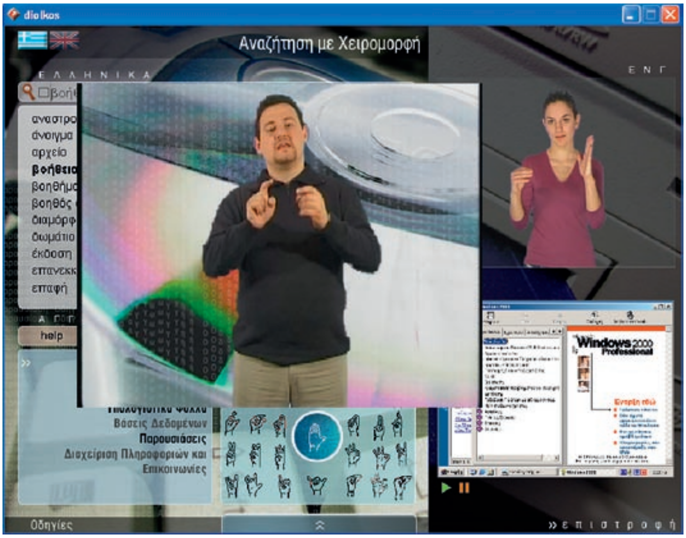


Figure 15: On-line help in GSL

References

Efthimiou, E. / Fotinea, S.-E. (2004). "An adaptation-based technique on current educational tools to support universal access: the case of a GSL e-Learning platform" in Courtiat J.-P. / Davarakis C. / Villemur T. (eds.) (2004). *Proceedings of the TeL'04 Workshop on Technology Enhanced Learning*, 22 August, Toulouse, France, Workshop to WCC-2004, 177–186.

Efthimiou, E. / Fotinea, S.-E. (2004b). "Multipurpose Design of Greek Sign Language Resources: A factor towards Universal Access" in *Proceedings of the 1st International Conference 'From Scientific Computing to Computational Engineering'* (IC-SCCE), Invited Session on Human Computer Interaction, Athens, 8–10 September, 2004, 820–827.

Efthimiou, E. / Fotinea, S.-E. (2007). "An Environment for Deaf Accessibility to Educational Content" in *Proceedings of the First International Conference on Information and Communication Technology and Accessibility*, April 12–14, Hammamet, Tunisia, 125–130.

Efthimiou, E. / Fotinea, S.-E. / Sapountzaki, G. (2006). "Processing linguistic data for GSL structure representation" in *Proceedings of the Workshop on the Representation and Processing of Sign Languages: Lexicographic matters and didactic scenarios*, Satellite Workshop to LREC-06, 49–54.

Efthimiou, E. / Fotinea, S.-E. / Sapountzaki, G. (2006b). "E-accessibility to educational content for the deaf", *European Journal of Open and Distance Learning (EURODL),* 2006/II, electronically available since December 15, 2006; Retrieved May 08, 2009, from http://www.eurodl.org/materials/contrib/2006/Eleni_Efthimiou.htm

Fotinea, S.-E. / Efthimiou, E. / Karpouzis, K. / Caridakis, G. (2005). "Dynamic GSL synthesis to support access to e-content" in *Proceedings of the 3rd International Conference on Universal Access in Human-Computer Interaction (UAHCI 05)*. Las Vegas, Nevada, USA.

Glauert, J. R. W. (2002). "ViSiCAST: Sign Language using Virtual Humans" in *International Conference on Assistive Technology (ICAT 2002)*, Derby, 2002, 21–33.

Huenerfauth, M. (2004). "Spatial representation of classifier predicates for machine translation into American Sign Language" in *Proceedings of the Workshop on the Representation and Processing of Signed Languages, 4th International Conference on Language Resources and Evaluation: LREC-04*, 24–31.

International Standard ISO 1087–1:2000 – "Terminology work—Vocabulary—Part 1: Theory and application".

International Standard ISO/R 704:1968 – "Naming principles".

International Standard ISO 704 – "Principles and methods of terminology".

Karpouzis, K. / Caridakis, G. / Fotinea, S.-E. / Efthimiou, E. (2007). "Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture", *Computers and Education* 49, 54–74.

Marshall, I. / Safar, E. (2002). "Sign Language Synthesis Using HPSG" in *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, Keihanna, Japan.

Mayberry, R. (1993). "First-Language acquisition after childhood differs from second language acquisition: The case of American Sign Language", *Journal of Speech Language and Hearing Research*, 36, 51–68.

Neidle, C. / Kegl, J. / MacLaughlin, D. / Bahan, B. / Lee, R. G. (2000). *The Syntax of American Sign Language*. Massachussets: MIT Press.

Sapountzaki G. / Efthimiou, E. / Fotinea, S.-E. (2006). “Digital technology in Greek Sign Language teaching in primary school curriculum” in *Proceedings of Hellenic Conference on Digital Educational Material: Issues of Creation, Exploitation and Evaluation, April 6–7*, 161–170.

Stokoe, W.C. (1978). *Sign Language Structure*. (2nd ed.). Silver Spring MD: Linstok Press.

Sutton-Spence, R. / Woll, B. (1999). *The Linguistics of British Sign Language. An Introduction*. Cambridge: University Press.

Authors

Karin Aijmer is a professor emerita at the University of Gothenburg. Among her publications are 'Conversational Routines in English: Convention and Creativity' (1996), 'English Discourse Particles: Evidence from a Corpus' (2002), and 'The Semantic Field of Modal Certainty: A Corpus-Based Study of English Adverbs' (with Anne-Marie Simon-Vandenbergen) (2007). She is co-editor of 'English Corpus Linguistics: Studies in Honour of Jan Svartvik' (1991), 'Discourse Patterns in Spoken and Written Corpora' (2004) and 'Pragmatic Markers in Contrast' (2006). Her research interests include pragmatic and discourse, spoken English, corpus linguistics, cross-cultural linguistics, and learner English.

Aleksey Andronov is an assistant professor at St. Petersburg State University in Russia. His main fields of research include Baltic philology, Latgalians in Russia, and lexicography.

Everita Andronova is a researcher at the Institute of Mathematics and Computer Science at the University of Latvia. Her main fields of research include corpus linguistics, language history and lexicography.

Stefanie Anstein, Dipl.-Ling., is a researcher at the Institute for Specialised Communication and Multilingualism of the European Academy Bozen/Bolzano. Her main fields of research are corpus linguistics and language varieties, and her major activity is in the initiative 'Korpus Südtirol'.

Maria Fernanda Bacelar do Nascimento is the main researcher at the Centre of Linguistics of the University of Lisbon. She has coordinated several projects regarding the development of Language Resources for Portuguese, such as the *Reference Corpus of Contemporary Portuguese*, a 350 million-word monitor corpus, and a corpus of the five African varieties of Portuguese. Her PhD focused on the syntactic properties of spoken discourse and her main areas of interest are the contrastive studies of Portuguese varieties, language resources compilation and exploitation, corpus linguistics and the lexicon.

Caren Brinckmann is a computational linguist and phonetician at the Institut für Deutsche Sprache (Institute for German Language, IDS) in Mannheim, Germany. Since 2006 she has been involved in two different research projects focussing on socio-phonetic variation and word phonology. Her current research interests include word-internal boundary effects, modeling regional variation with data mining techniques, and the efficient annotation of large speech corpora.

Elisa Corino, PhD, is a research assistant at the Department of Literary and Philological Science at the University of Turin. She is interested in translation studies and text linguistics. She has been dealing with applied and corpus linguistics since 2003 and has been coordinating the corpus VALICO together with Carla Marello since its inception. As a linguist with a training combining German and Italian studies, she has been working on both Italian and German morphology, syntax, history of grammar, and on the learner varieties of Italian as a foreign language. She is also interested in Netspeak, CMC and CMT.

Guy De Pauw is a senior researcher at the CNTS–Language Technology Group (University of Antwerp, Belgium) and a visiting lecturer at the School of Computing and Informatics (University of Nairobi). His research focuses on the application of machine-learning techniques to natural language processing (NLP) for African languages. A recurring theme in his research is how we can port current state-of-the-art NLP techniques (that are developed with Indo-European languages in mind) to Bantu languages. Recent publications concentrate on morphological and syntactic processing of Swahili and the development of resources for and components of a machine translation system for the language pair English–Swahili. He is also one of the founders of the AfLaT organization (http://aflat.org), which aims to bring together researchers on language technology for African languages.

Prof. Dr. Ing. **Gilles-Maurice de Schryver** is an assistant professor at The department of African Languages and Cultures at Ghent University, Belgium, and an extraordinary professor in the Xhosa Department at the University of the Western Cape, South Africa. He is also a founding member of African Language Technology, an elected board member of Afrilex, Euralex, Asialex as well as Australex, initiator of PangaeaLex, and co-founder of TshwaneDJe HLT. Gilles-Maurice is mainly interested in corpus and dictionary topics for the African languages, and increasingly also in language-independent lexicography software. His main research results are invariably published in the International Journal of Lexicography and Lexikos, while his own corpus-driven dictionaries for the Bantu languages are published by Oxford University Press.

Dr. **Eleni Efthimiou** is Research Director at the Institute for Language and Speech Processing (ILSP / R.C. "Athena"). She currently heads the Sign Language Technologies Lab and co-ordinates the Assistive Technology Group. In 1999 she founded the ILSP team for basic research and resources collection for Greek Sign Language (GSL). Ever since, she has dedicated major effort to the analysis of GSL and its introduction into various computational environments. Her research activities involve analysis of natural language, optimisation techniques of human-computer interaction and NLP-based prototype development. At the SLT Lab, her work focuses on creation of electronic SL resources (corpora, vocabularies and grammars), development of systems for SL synthesis via virtual agent (signing avatar), SL recognition, development of sign language educational and communication tools, as well as the implementation of principles of Universal Access for system architecture.

Antónia Estrela is a teacher of Portuguese Language and a PhD student in the area of syntax (use and acquisition of passive constructions). She collaborates at the Centre of Linguistics of the University of Lisbon (where she participated in the compilation of the corpora of African varieties of Portuguese), and at the Center of Linguistics of the New University of Lisbon. Her main areas of interest are the acquisition of language, writing and the African varieties of Portuguese.

Amália Mendes is researcher at the Center of Linguistics of the University of Lisbon (CLUL) and teaches linguistics courses at the Faculty of Letters of the University of Lisbon (FLUL). In 2004 she published her PhD dissertation, a corpus-based study of the lexical semantics of Portuguese psychological verbs. She has been involved since 1989 in several projects on the development of language resources for Portuguese at CLUL. Her main interests are the syntax and semantics interface, regular polysemy, collocations and idioms, and contrastive studies of Portuguese varieties.

Giovanni Mischì is a native Ladin from Lungiarü (Val Badia). He studied German philology and history at the University of Innsbruck. His background also includes paleography, diplomacy and archival science. Since 1990 he has been working as a research associate at the Ladin Institute "Micurà de Rü" in Martin de Tor / St. Martin in Thurn. His work focuses on lexicography, language development, language planning and translation. He is currently teaching at the Free University of Bozen–Bolzano.

Julianne Nyhan is a research associate in the *Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften* at the University of Trier in Germany. She is also the book reviews editor of *Interdisciplinary Science Reviews* and the Co-chair of the TEI Special Interest Group on Education. Ms. Nyhan's background is in Digital Humanities. She cut her teeth in the Corpus of Electronic Texts at the University College Cork in Ireland, where she completed her PhD on the application of XML to the historical lexicography of Old, Middle and Early Modern Irish. She also worked on numerous TEI encoded texts of Irish literature, history and politics. At present she is at work on a TEI conformant version of Patrick S. *Dinneen's Foclóir Gaedhilge agus Béarla* (an Irish-English Dictionary). Her other research interests include the history of Irish lexicography, forensic dictionary analysis and data visualisation.

Delyth Prys is an experienced terminologist who has edited over twenty bilingual Welsh/English specialist terminology dictionaries. She leads a mixed team of linguists, software developers and speech technologists in an innovative, multidisciplinary research unit at Canolfan Bedwyr, Bangor University. Her unit researches and develops speech and text applications to support Welsh and other languages in a multilingual electronic multimedia and multimodal environment. They have developed a philosophy of adapting international standards to fit the needs of less-resourced languages such as Welsh and making frugal use of existing resources by recycling and reusing components as much as possible. Delyth is the Director of SALT Cymru, a Speech and Language Technologies research network for Wales, funded by the Welsh Assembly Government and the European Regional Development Fund.

Luísa Alice Santos Pereira is a high school teacher of Portuguese as a mother tongue who has experience teaching Portuguese as foreign language. She is a collaborator at the Center of Linguistics of the University of Lisbon, where she has been involved in the design and constitution of several corpora, such as the *Reference Corpus of Contemporary Portuguese* and the five corpora of African varieties of Portuguese. Her main interests are Portuguese teaching as foreign language, collocations and idioms, and corpus compilation.

Thomas Schmidt holds a doctoral degree in German linguistics from the University of Dortmund. He has worked as a Language Resource Engineer for a language technology company in Aachen, as a project assistant for the European Language Resource Association in Paris, and as a researcher at the University of Hamburg, the Berlin-Brandenburg Academy of Sciences and the International Computer Science Institute in Berkeley. He is currently conducting a project on 'Computer-assisted methods for creating and analysing multilingual data' at the University of Hamburg. Dr. Schmidt's main research interests are text and corpus technology and computational lexicography.

Rita Veloso is a researcher at the Centre of Linguistics of the University of Lisbon (CLUL) and an invited assistant at the Faculty of Letters of the University of Lisbon (FLUL). She has been working in corpus compilation and annotation since 1995. She has mainly been involved in the compilation of the spoken subcorpus of *the Reference Corpus of Contemporary Portuguese*, and of several other spoken corpora, their PoS annotation and text-to-sound alignment. She is now working on her PhD thesis, a corpus-based description and analysis of relative clause construction strategies.

Paul Videsott is a professor in the Ladin Section of the Faculty of Education at the Free University of Bolzano–Bozen. His research is concerned with Italian, French and Rhaeto-Romance linguistics; alpine dialectology and toponymy; Italian and French scriptology and scriptometry; and corpus linguistics. Currently he is involved with projects on *The Language of Royal Chambers of Paris in the 13th Century* as well as a corpus of literary Ladin.