
LULCL 2005

Proceedings of the Lesser Used Languages and
Computer Linguistics Conference

Bolzano, 27th-28th October 2005

Isabella Ties (Ed.)

LULCL 2005

Proceedings of the Lesser Used Languages and Computer Linguistics Conference

Isabella Ties (Ed.)

EURAC
research

EUROPÄISCHE
AKADEMIE

ACCADEMIA
EUROPEA

EUROPEAN
ACADEMY

BOZEN - BOLZANO

2006



The proceedings are co-financed by the European Union through the Interreg IIIA Italy-Switzerland Programme

Bestellungen bei:

Europäische Akademie Bozen
Viale Druso, 1
39100 Bozen - Italien
Tel. +39 0471 055055
Fax +39 0471 055099
E-mail: press@eurac.edu

Per ordinazioni:

Accademia Europea Bolzano
Drususallee, 1
39100 Bolzano - Italia
Tel. +39 0471 055055
Fax +39 0471 055099
E-mail: press@eurac.edu

Nachdruck und fotomechanische
Wiedergabe - auch auszugsweise - nur
unter Angabe der Quelle
(Herausgeber und Titel) gestattet.

Riproduzione parziale o totale del
contenuto autorizzata soltanto con la
citazione della fonte
(titolo ed edizione).

Verantwortlicher Direktor: Stephan Ortner
Redaktion: : Isabella Ties
Koordination: : Isabella Ties
Graphik und Umschlag: Marco Polenta
Druck: Fotolito Longo

Direttore responsabile: Stephan Ortner
Redazione: Isabella Ties
Coordinazione: Isabella Ties
Grafica e copertina: Marco Polenta
Stampa: Fotolito Longo

ISBN 88-88906-24-X

Index

Preface	7
Spracherneuerung im Rätoromanischen: Linguistische, soziale und politische Aspekte	11
<i>Clau Solèr</i>	
Implementing NLP-Projects for Small Languages: Instructions for Funding Bodies, Strategies for Developers	29
<i>Oliver Streiter</i>	
Un corpus per il sardo: problemi e prospettive	45
<i>Nicoletta Puddu</i>	
The Relevance of Lesser-Used Languages for Theoretical Linguistics: The Case of Cimbrian and the Suport of the TITUS Corpus.....	77
<i>Ermenegildo Bidese, Cecilia Poletto and Alessandra Tomaselli</i>	
Creating Word Class Tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping	97
<i>Danie J. Prinsloo and Ulrich Heid</i>	
A Comparison of Approaches to Word Class Tagging: Distinctively Versus Conjunctively Written Bantu Languages	117
<i>Elsabé Taljard and Sonja E. Bosch</i>	
Grammar-based Language Technology for the Sámi Languages	133
<i>Trond Trosterud</i>	
The Welsh National Online Terminology Database	149
<i>Dewi Bryn Jones and Delyth Prys</i>	
SpeechCluster: A Speech Data Multitool.....	171
<i>Ivan A. Uemlianin</i>	
XNLRDF: The Open Source Framework for Multilingual Computing	189
<i>Oliver Streiter and Mathias Stuflesser</i>	
Speech-to-Speech Translation for Catalan	209
<i>Victoria Arranz, Elisabet Comelles and David Farwell</i>	

Computing Non-Concatenative Morphology: The Case of Georgian	225
<i>Olga Gurevich</i>	
The Igbo Language and Computer Linguistics: Problems and Prospects.....	247
<i>Chinedu Uchechukwu</i>	
Annotation of Documents for Electronic Editing of Judeo-Spanish Texts: Problems and Solutions	265
<i>Soufiane Roussi and Ana Stulic</i>	
Il ladino fra polinomia e standardizzazione: l’apporto della linguistica computazionale	281
<i>Evelyn Bortolotti, Sabrina Rasom</i>	
Il progetto “Zimbarbort” per il recupero del patrimonio linguistico cimbro	297
<i>Luca Panieri</i>	
Stealth Learning with an Online Dog (Web-based Word Games for Welsh)	307
<i>Gruffudd Prys and Ambrose Choy</i>	
Alphabetical list of authors & titles with keywords	329
Alphabetical list of contributors & contact addresses	335

Preface

On behalf of the programme committee for the 'Lesser Used Languages and Computer Linguistics' conference (LULCL 2005), we are pleased to present the proceedings, which contain the papers accepted for presentation at the Bolzano meeting on 27th-28th October 2005. The contributions published in this volume deal with the main aspects of lesser used languages and their support through computer linguistics, ranging from lexicography to terminology for lesser used languages, and from computational linguistic applications in general to more specific resources such as corpora. Some papers deal specifically with Translation Memory Systems, online dictionaries, Internet Computer Assisted Language Learning (CALL) or Language for Specific Purposes (LSP).

The choice of the conference theme was strongly influenced by the ambition to give lesser used languages an opportunity for visibility without taking into consideration the official number of speakers, but rather the range of technological resources available for each language. Even though some languages do indeed count a considerable number of speakers, technology support may be almost nonexistent. It is therefore remarkable how much has been done in the last decades for languages with few speakers. The Zimbar speakers are the smallest community represented at the conference, which counts about 2230 speakers living in Luserna, Roana, Mezzaselva and Giazza. Despite the small number of native speakers, there are major projects running on this Germanic language. The first project described here (cf. Bidese *et al.*) foresees the storage of Cimbrian textual material in the TITUS Corpus ('Thesaurus of Indo-European Textual and Linguistic Materials'), while the second one provides the guidelines for the Zimbarbort. The latter is a new project on the preservation of the Zimbar language, during which a database of lexical entries will be created (cf. Panieri). Both projects represent a substantial contribution to the preservation of language: through the recovery and storage of textual data they enable researchers to carry out linguistic analyses from several points of view.

Sparseness of data is one of the main characteristics that many lesser used languages share with Zimbar. This influences both the choice of methodology and, of course, the results. Clau Solèrs keynote contribution reflects very well what happens when, for example, specialised terminology has to be elaborated within a small language. The lack of native terminology for many LSPs and the influence of bigger official languages,

such as German in this special case, are just some of the problems Rumantsch and other small languages have to face in order to propose acceptable terminology and preserve language at the same time. The project on the 'Welsh National Terminology Database' reflects the need to find a means between accepted terminology standards used for bigger languages (ISO 704 and ISO 860 norms) and language preservation. This project takes advantage of the similarities between terminology and lexicography, as existing lexicographical resources and applications are used to enrich the terminology database.

Another central topic that lesser used languages have in common is the usability of available data. On the one hand we find the contribution on Judeo-Spanish, where Roussi & Stulic describe how to transliterate and annotate texts written in Hebrew characters and, at the same time, allow users to add their own interpretation and comments. On the other hand, Uchechukwu explains in his contribution the problems related to appropriate font programmes and software compatibility. On the basis of the Igbo language he describes what happens when the amount of data is considerable but not usable (due to the obstacle of accepted format).

Issues of data sparseness and usability determine linguistic research, especially during the phases of data pre-processing, and the amount of time linguists must invest in dealing with linguistic research questions. Uemlianin proposes to use SpeechCluster in order to ensure that linguists can concentrate on linguistic analyses rather than disperse their efforts with formatting or any other time-consuming manual processing.

Trosterud emphasises on the importance of open-source technology for projects on lesser used languages, so as to avoid waste in terms of time and technology, which must be reinvented every single time for every small language. The same point of view is stated by Stuflesser and Streiter as they present their intention to use XNLRDF, a free software package for NLP. Their contribution introduces the existing prototype and outlines future strategies.

A similar aim is pursued by the invited key-note speaker Oliver Streiter, who focuses on this topic, providing a detailed overview on available resources and underlining the importance of mutual support within the research community through data sharing in standard formats, so as to make it usable and accessible to everybody. One of the instruments cited and used most often for data sharing is the Internet, as it allows online storage of data such as dictionaries, language games or terminology data bases (Jones & Prys). This medium is used by Canolfan Bedwyr to publish the web-based word games for Welsh, as well as by the Ladin institutions to disseminate their online

dictionaries (cfr. Bortolotti & Rasom) meant to improve the language skills of native speakers.

Several authors file contributions on tools for the elaboration and storage of language for text analysis and processing of text material with a view to the development of corpora. Puddu points out the importance of corpora for supporting the development of lesser used languages and the main problems connected to corpus design, text collection, storage and annotation for a lesser used language like Sardo (cf. Puddu). Sardinian, like any other lesser used language, has to cope with problems related to retrieval of written text, and in this specific case, also with a second problem: the absence of a standard orthography. The application of a homogeneous tag system, as well as the use of standards on storage, such as the rules elaborated by the EAGLES group (XCES), is suggested.

Prinsloo and Heid describe methodology such as the bootstrapping of resources in order to elaborate language documentation and annotation. They describe the development of different tools to bootstrap tagging resources for Northern Sotho, and resources used to identify verbs and nouns for the disambiguation of closed class items. The Bantu languages and their characteristics are also discussed in the contribution by Taljard and Bosh, who present the problems encountered when dealing with languages with different writing systems – in this special case, Northern Sotho and Zulu. The authors describe the distinct approaches for class tagging according to the different writing systems.

Examples of knowledge extraction and knowledge engineering are discussed in the paper on the FAME project, an Interlingual Speech-to-Speech Machine Translation System for Catalan, English and Spanish developed to assist users in making hotel reservations. The project includes tools for the documentation of data and elaboration of the standard Interchange Format (IF).

It is clear from these contributions that nowadays, a variety of approaches and scientific methodologies are adopted in research on lesser used languages, showing the vitality of research in this specific area.

Thanks to authors who cover a large variety of projects and technologies, an overview of the state of the art in research on lesser used languages can be provided, especially as regards projects on lesser used languages involving computational linguistics in Europe and the world. Central to the conference are both methodological issues, prompted by the described strategies for an efficient support of lesser used languages, and the problems encountered with theoretical approaches developed for major languages but applied to lesser used languages.

The contributions underline the significance of computational linguistics, the methodologies and strategies followed, and their application on lesser used languages. It becomes evident how important decisions on international standards are and which consequences they imply for the standardisation of tools.

This conference would not have been possible without the energy of many people. First of all we would like to thank the authors, who have provided superb contributes. Our gratitude goes also to the reviewers and to the scientific committee for their detailed and inspiring reviews.

Isabella Ties

Spracherneuerung im Rätoromanischen: Linguistische, soziale und politische Aspekte

Clau Solèr

In Graubünden, the minority language Romansch has to assert itself in an environment of bilingualism with German on the one hand, while constantly keeping pace with the changing needs of its speakers on the other. To fulfill this task, terminological precision must continually come to terms with both the spoken language and the existing syntax. Romansch must be able to express a frame of mind that is influenced by the Germanic element, and neologisms must also adapt to the regional varieties for the speakers to be able to identify with them.

Due to the limited political and economic importance of the language, as well as instruction that partly takes place in German only, Romansch is currently lacking the necessary channels for an efficient diffusion of neologisms.

1. Einleitung

Jede Sprache dient im Alltag als Werkzeug und passt sich ihrer Sprachgemeinschaft an; dies im Unterschied zu nur historischen oder kultischen Sprachen. Dabei darf sie sich aber nicht veräußern, nur um modern oder aktuell zu sein. Neben einer spontanen, gelegentlich unerwünschten Erneuerung - die übliche sich langfristig ablaufende Sprachentwicklung steht hier nicht zur Sprache - unterliegt die Sprache Eingriffen aus unterschiedlichsten Richtungen und Kräften und aus verschiedenen Gründen. Wie geschieht das, was entsteht daraus, wem nützt das und wird sie besser oder schlechter? Diese Fragen möchte ich besonders aus der praktischen Erfahrung zu beantworten versuchen und einige Überlegungen dazu anstellen. Vorerst muss ich das Rätoromanische in Graubünden und dessen Stellung im Hinblick auf die Sprachanpassung kurz umreißen. Ich wähle bewusst den Ausdruck Anpassung, um keine Wertung wie Erneuerung, Modernisierung, Einschränkung und Uminterpretation vorweg-zunehmen.

Das Bündnerromanische ist eine eigenständige, neolateinische Sprache auf vorrömischer Grundlage. Seit über 1000 Jahren ist es im vielfältigen Kontakt mit dem Deutschen und während mehreren Jahrhunderten auch mit dem Italienischen (im Engadin besonders wirtschaftlich und in den katholischen Gegenden religiös bedingt). Nach dem Anschluss an die schweizerische Eidgenossenschaft 1803 ist die gelebte und relativ ausgeglichene Dreisprachigkeit der drei Bünde durch das Deutsche als fast

unumschränkte Verkehrs- und Verwaltungssprache ersetzt worden. Die rätoromanischen Ortsdialekte, die in fünf regionalen Schriftidiomen in geografisch und konfessionell mehr oder weniger getrennten Gebieten mit einer ursprünglich traditionellen, heute mehrheitlich touristischen Wirtschaft verwendet werden, haben sich unterdessen zu einer primär gesprochenen Varietät gewandelt. 35.095 Personen nannten bei der Volkszählung 2000 (RÄTOROMANISCH 2004:24) das Romanische als ihre Hauptsprache und insgesamt 60.816 verwenden es im Alltag oder bei der Arbeit, wobei nur zwei Drittel davon in Graubünden leben und der Rest in der schweizerischen Diaspora. Gemäß EUROMOSAIC (1996:34) braucht eine Sprachgruppe mindestens 300.000 Mitglieder für ihre Selbständigkeit und so sind die Aussichten des Romanischen eher düster. In den Gemeinden mit mehr als 20% Romanischsprechern besuchen zwei Drittel der Schüler eine maximal vierjährige romanische Grundschule mit anschließender Einführung ins Deutsche, das in den drei letzten Jahren Unterrichtssprache wird, neben immerhin bis zu 6 Stunden Romanisch als Fach. Die Mittelschulen in Chur und Samedan ermöglichen einen zweisprachigen Maturitätsabschluss. In der Pädagogischen Fachhochschule, die im Unterschied zum bisherigen Lehrerseminar nicht mehr sprachbezogen ist, fehlt eine entsprechende Unterstützung, wie es in den nur deutschen beruflichen Fachschulen auch der Fall ist.

Es fehlt noch der linguistische Zustandsbericht. Die traditionelle Einsprachigkeit mit wenigstens einer Fremdsprache gibt es nur noch bei wenigen, älteren Personen in entlegenen Ortschaften mit geringer Zuwanderung. Sonst leben die Bündnerromanen in einer funktionalen, domänenorientierten und personengesteuerten Mehrsprachigkeit mit jeweils unterschiedlichen Kodes: romanische Ortsmundart gesprochen, teilweise gelesen, aber selten geschrieben, Schweizerdeutsch gesprochen sowie Standarddeutsch als Schriftsprache und teilweise gehört. Man wählt die Sprache relativ wertfrei, und die Phase als Rätoromanisch stigmatisiert war und man daher am Minderwertigkeitskomplex litt, ist heute mehrheitlich überwunden, und zwar in erster Linie wegen der hohen Deutschkompetenz der Romanischsprecher, ihrer besseren Integration in der deutschsprachigen Gesellschaft und letztlich auch wegen der vielen Zuzügler mit noch selteneren Sprachen.

2. Terminologische Anpassung

Ich wähle bewusst den Begriff Terminologie, der die Neologie und die Uminterpretation vorhandener Begriffe einschließt. Dabei hat man sich weniger umfangreiche und weit abgestützte Prozesse vorzustellen, sondern eher zufällige und gelegentlich chaotische Vorschläge, die nach Möglichkeit gesammelt und verbreitet werden. Viele Einträge des Pledari Grond in *Rumantsch grischun* stammen aus allerlei Übersetzungen und

Anfragen von Sprachverwendern, der Rest stammt aus den Regionalwörterbüchern und aus der systematischen Neologie.

3. Das Bedürfnis nach terminologischer Anpassung

Es ist zwar unbestritten, dass keine Sprache von sich aus eine Anpassung braucht, denn Sprachen handeln nun einmal nicht. Trotzdem hat man seit dem Ende des 19. Jh. immer wieder das Romanische als klagende, leidende oder verschufte Sprache anthropomorphologisiert, damit den Rätoromanen ins Gewissen geredet und zugegebenermaßen einiges erreicht. Vieles ist aber auch verdorben worden (Coray 1993). Es ist die Sprachgemeinschaft mit ihren Anwendern, die eine Sprache den Bedürfnissen nach gesicherter und rascher Kommunikation anpasst. Als Sprachverwender gelten grundsätzlich die Sprechenden in ihrem sozialen, wirtschaftlichen und geistigen Umfeld, solange sie sich nicht ausschließlich als Parteivertreter der Sprache verhalten. Puristische oder spracherhaltende Gründe sind politisch und gesellschaftlich begründet und von den Sprachverwendern nur beschränkt getragen. Sie lehnen diese von der Sprachverwaltung vertretene künstliche Erhaltung ab, wie ihr Verhalten u. A. gegenüber dem *Rumantsch grischun* zeigt.

Eine wirkliche Alternative sich sprachlich anzupassen besteht für Sprachverwender von Minderheitensprachen mit einem asymmetrischen Bilinguismus in einem Sprachwechsel, der meistens mehrstufig verläuft, auch wenn dieser Sprachwechsel gerne verschwiegen wird (Solèr 1986:299). Das Englische in bestimmten Wirtschaftsbereichen gilt heute als direkter Weg, wenn es nicht aus Ermangelung einer gemeinsamen Sprache gewählt wird.

4. Methoden der terminologischen Anpassung

In der Vergangenheit hat sich das Romanische den Bedürfnissen mehr schlecht als recht angepasst und ist auch deshalb minorisiert worden. Erst im Zuge der spätromantischen Nationalbewegung, also seit mehr als hundert Jahren, bemüht man sich bewusst und systematisch um eine lexikalische Erneuerung. Heute ist das Romanische terminologisch sehr stark ausgebaut, verglichen mit dem Zustand vor 150 Jahren als „tausenderlei Gegenstände und Tätigkeiten der gebildeten Welt unbekannt oder doch fremd geblieben [waren] (CARISCH 1848:X). Auch die Syntax hat sich erneuert und ist eigenständig(er) geworden. In dieser Zeit veränderten sich die Gesellschaft und Wirtschaft grundlegend. Die obligatorische Volksschule erreichte erstmals eine ganze Bevölkerungsschicht und konnte die Sprache direkt beeinflussen, indem alte örtliche Formen verschwanden, wie Jaberg/Jud (1928) bedauerten.

Nun gilt es zu erklären, wie die Sprache terminologisch angepasst wird. Neben der Verwendung einer phonetisch und morphologisch integrierten fremden Bezeichnung oder gleichzeitig dazu liefert die eigensprachliche Um- und Beschreibung (Periphrase) die wichtigste spontane terminologische Anpassung. Dieses Vorgehen passt auch stilistisch fremde oder unverständliche Terminologie an die Umgangssprache an und steht logischerweise im Widerspruch zur Systematisierung der Fachsprache. Weiterhin gilt die typisch analytische Parataxe einer Volkssprache, wie es das Romanische im Grunde genommen ist. Der Vorteil der hohen Verständlichkeit muss mit der Variabilität erkaufte werden. Hierhin gehören auch die zufälligen, spielerischen Volksbildungen mit üblicherweise nur regionaler und kurzzeitiger Gültigkeit; erwähnt seien vallader: *chasperets* für 'Scheibenwischer', eigentlich 'Kasperlefigur' oder sursilvan: *cutgna* für 'Surfbrett, Snowboard', eigentlich 'Schwarte (vom Holz oder vom Speck)'.

Systemkonform ist auch die professionelle Terminologie oft periphrastisch anstatt derivativ und daraus entstehen, je nach dem Definitionsgrad, linguistische Ungetüme wie *ovs da giaglinas allevadas a terra* für ganz gewöhnliche 'Eier (von Hühnern) aus Bodenhaltung' oder *chapisch da la rullera d'alver da manvella* für 'Kurbelwellenlagerdeckel', das freilich auch nicht verständlicher ist und als einzelner Baustein noch kompliziertere Sätze bilden muss.

Ein typischer und traditioneller Terminologieprozess ist die Analogie. Heute weicht diese endolinguale zugunsten der exolingualen, sich am Deutsch lehrende Bildung zurück wegen ihrer Nähe zur Denkstruktur der Romanischsprecher. Sie verspricht mehr Erfolg als eine Herleitung aus dem Französischen als kaum mehr unterrichtete Fremdsprache oder aus dem Italienischen, das zwar (noch) einen festen Platz in den Bündner Schulen hat, aber nur eine geringe Bedeutung im Alltag genießt.

Analogien zu romanischen Sprachen liegen in den folgenden Beispielen vor¹, (vgl. auch Decurtins 1993, 235-254 passim): Als Alltagsbegriff gilt *schambun* (oit, frz.) 'Schinken'. Der Begriff vl: *levatriza* 'Hebamme' scheiterte als undurchsichtige Bezeichnung für eine einsetzende Professionalisierung und wurde deshalb periphrastisch zu vl: *duonna da part* 'Geburtsfrau' rg: *spendrera* eigentlich 'Rettende', *dunna da part*. Auch *purantina* 'Tragbahre' ist kaum verständlich und konnte *bara* trotz der Homonymie zu 'Leiche' nicht ersetzen. Die Ausdrücke *guid* '(Reise-)Führer' und *guidar* 'führen' sind seltener als *manader* 'Führer, Lenker' und *manar* 'lenken, leiten, führen'. Einsichtig sind *giraplattas* 'Plattenspieler' und modernisiert *giradiscs* 'Diskettenlaufwerk', das

1 Die romanischen Beispiele sind in Rumantsch grischun (rg); die Regionalformen werden bezeichnet als sr = sursilvan, st = sutsilvan, sm = surmiran, pt = puter, vl = vallader; Französisch = frz., Italienisch = it, Oberitalienisch = oit., Rätoromanisch = rtr.

aber schon durch 'CD-Player' internationalisiert wurde.² Die Bezeichnung *telefonin* für 'Funktelefon' konnte sich gegen *natel* als Produktname und besonders *handy* nicht durchsetzen und das westschweizerische *portable* ist geographisch und mental schon zu weit entfernt.

Besonders die ersten grundlegenden Wörterbücher in der ersten Hälfte des 20. Jh. wählten die Analogie. Ein Teil ihrer Vorschläge konnte sich dank der Verbreitung in der damals sprachprägenden Schule sowie dem hohen Ansehen des Französischen und Italienischen durchsetzen und viele Germanismen ersetzen (Solèr 2005).

Wohl immer beeinflusste der Purismus sowohl außer- wie auch innersprachlich die terminologische Anpassung. Zu Beginn des 20. Jh. fielen besonders im Engadin wegen des Irredentismus viele Italianismen trotz ihrer linguistischen Nähe zum Rätoromanischen. Andererseits besteht das Dilemma zwischen neolateinischen Begriffen wie *aspiratur* 'Staubsauger', *mochetta* 'Spannteppich', die aber weniger transparent sind als die transkodischen *tschitschapulvra* 'Staubsauger' und *tarpun stendi* eig. 'gespannter Teppich'. Und genau diese Nähe schafft viele neue Begriffe, die erst rückübersetzt, also deutsch gedacht, verstanden werden: *maisa da mezdi* wörtlich 'Mittagstisch' für 'gemeinsames Mittagessen für ältere Personen' anstatt *gentar cuminaivel*.

Zu den produktiven endolingualen Prozessen gehört die Morphemableitung für die verschiedenen Kategorien. Trotz ihrer grundlegenden Systematik erkennt man zeittypische Vorlieben. Deverbale Agensbegriffe auf *-ader*, *-atur*, *-adur* sind häufig, während Formen auf *-ari*, z.B. sr: *attentari* 'Attentäter', teilweise mit lateinischem *-ARIU*-Formen zusammenfallen; *splanari* 'Hobelbank' ist insofern eine Falschbildung, weil es kein Agens ist und auch nicht zu *-ARIU* gehört. Die *-ist*-Formen wie *schurnalist* 'Journalist' sind nur dann erfolgreich, wenn die Variante *-cher* nicht durch ein deutsches Analogon gestützt wird. Sonst gilt *-ist* als puristisch und High-Variante wie *musicist* 'Musiker', das mit *musicher* eine Low-Variante erzeugt.

Die Prozesse und teilweise deren Resultat werden gebildet mit *-ziun* wie *furniziun* 'Lieferung', *allontanaziun* 'Entfernung' und *exemziun* 'Befreiung, Entbindung' auf ganz unterschiedlicher romanischer Basis oder mit *-ada* wie *zavrada* 'Schafscheide, Aussonderung', *scuntrada* 'Treffen, Zusammenkunft' und, ziemlich heterogen, *auzada* 'Stockwerk', genauer 'Anhebung'.

Auch andere Suffixe sind mehrwertig, so *-al* in *fossal* 'Baugrube, Stadtgraben', *plazzal* 'Baustelle', aber auch *runal* 'Schlepplift' ohne die *-ALIS*-Adjektive zu berücksichtigen. Allgemein bevorzugt das Romanische Periphrasen anstatt der stilistisch markierten

2 Als Abkürzung gilt mehrheitlich „CD“ m/f während *disc cumpact* im romanischen Radio recht geläufig ist.

Adjektive auf *-abel*, *-ibel*, *-aivel*, *-ar*, *-ari*, *-ic* und unterscheidet sich damit stark vom Französischen und Italienischen.³

Mehrdeutig ist auch das Morphem *-et* als Verkleinerung *vegliet* '(kleiner) Alter', als Spezifikation *furtget* 'Gabler', rg: *buffet* 'Blasebalg', sr: *suflet* analog frz. „souflet“, it. „soffietto“, sr: *stizzet* 'Löschhorn', rg: *durmigliet* 'Siebenschläfer' als Lehnbildung bzw. Calque für die kaum verständlichen Formen sr: *glis*, vl: *glira* aus lat. GLIS.

Interessant sind die Bildungen auf *-era*. Während die vom Verb abgeleiteten durchsichtig und verständlich sind, wie *ardera*⁴ 'Verbrennungsanlage', *mulschera* 'Melkstand', *cuera* 'Brutkasten', erweisen sich die vom Nomen gebildeten sehr undurchsichtig wie *balestrera* 'Schießscharte', das primär mit sr: *ballistrar* 'zappeln, störrisch sein, hapern' assoziiert wird, oder sie wirken ambivalent wie *sutgera* 'Sesselbahn', *bobera* 'Bobbahn', *cruschera* sr: 'Drehkreuz, Kreuz Kreuzworträtsel', rg: 'Fadenkreuz' und nicht beispielsweise 'Kreuzung', das *cruschada* heißt und homonymisch ist mit 'Kreuzzug'. Hier erzeugten die verschiedenen Idiome trotz der sogenannten *avischinaziun miaivla* 'sanften Annäherung' der 60er Jahre unterschiedliche Formen, die man zwar gegenseitig verstand, aber nicht zu einer einheitlichen Sprachform beigetragen haben.

Grundsätzlich kann jede Entlehnung als Basiselement dienen, wobei sie mehr an psychologische als an linguistische Grenzen stößt. Anstatt neue Verben direkt mit dem Morphem *-ar* an fremde, meistens deutsche Stämme zu binden wie *bremsar* 'bremsen', *spizzar* 'mit dem Spitzeisen ausschlagen', *clliccar* 'klicken', *checkar* 'merken' (über Deutsch aus dem Englischen), *chiffar* 'kiffen', die die früheren Verben auf *-egiar/-iar* ersetzen, bevorzugt man das analytische *far il* + deutscher Infinitiv *far il clichen* '(den) Klick machen'.

Asyndetische Bildungen sind durchsichtig und treffend wie *tirastruvas* 'Schraubenzieher' und *muntastgala* 'Treppenaufzug', während *tilavent* 'Düse' in Richtung Wetter weist. 'Mutterkuh' *vatga-mamma* drückt auch in der veränderten Abfolge von Bestimmtes-Bestimmendes (Determinat-Determinant) das undefinierte Verhältnis aus. Obgleich analog zu *biomassa* 'Biomasse', ist *biopur* 'Biobauer' gewöhnungsbedürftig aber nötig, weil *pur da bio* wie *pur da latg* 'Milchbauer' zuerst auf das Material oder die Herkunft verweist. Regelmäßige Bildungen wie *telecumandar*, *microcircuit* bleiben elitär.

3 Auf *-ebel* lautet einzig *debel* „schwach“. Formen wie frz. „grippe aviaire“ und it. „influenza aviaria“ für 'Vogelgrippe' sind im Romanischen fast unmöglich und uaulic 'den Wald betreffend', *selvus* 'waldig' wirken exotisch.

4 Dazu „muss“ das Pledari Grond eine Periphrase *implant per arder rument* liefern; die Idiome verwenden zudem sr: *barschera*, vl: *bruschaduoir*.

Auch eine Aktualisierung durch Um- und Neudefinition bestehender, nicht mehr gebrauchter Begriffe ist möglich, trotz der unsicheren Übergangszeit mit Homonymie: Noch heute wird *zavrar* nur auf 'Schafe scheiden' beschränkt, trotz *zavrader* 'Sortwort' und *zavrar* 'sortieren'; man verwendet *sortar* oder das ungenaue *separar* 'trennen'. Eine wirkliche „Herausforderung“ bedeutet eben dessen Bezeichnung: Das surselvische *provocaziun* ist vermutlich zu nahe an die deutsche „Provokation“, so dass man heute vermehrt *sfida*, eine italienische Entlehnung im Engadinischen, verwendet, obwohl *sfida*, und ganz besonders *sfidar* in Rheinschbünden näher an *sfidar*, *disfidar* 'misstrauen' liegt. Eine Erweiterung erfuhr das Verb *sunar* 'musizieren', im Engadin noch 'Glocken läuten', durch die Unterdifferenzierung von 'spielen' und 'abspielen' *sunar ina platta, in(a) CD* anstatt *tadlar, far ir ina platta, in(a) CD*. Dem Biologiebegriff *tessi* 'Gewebe' fehlt das typische Fadenmuster eines gewobenen Tuches, und er ist deshalb nicht alltagstauglich; stattdessen verwendet man konkret *pel* 'Haut', *charn* 'Fleisch' bis zu *material* 'Material'. Auch der Fachausdruck *petroli* für 'Erdöl' wird nur in der engeren Bedeutung von Lampenbrennstoff 'Petrol' wahr genommen und erfordert infolgedessen ein Calque *ieli (mineral)* 'Mineralöl'.

Die gesamtromanische Standardisierung, angestrebt in *Rumantsch grischun*, zeigt im Alltag ihre Grenzen wegen einer hohen Heteronymie. Entweder verwendet man beide Ausdrücke wie *taglia/imposta* 'Steuer', *buis/schluppet* 'Gewehr', *entschaiver/cumenzar* 'beginnen' oder man vereinfacht unzulässig, indem man *glisch* für 'Lampe' anstatt *cazzola* in der Surselva verwendet, wo *glisch* nur 'Licht' bedeutet, weil *lampa* aus puristischen Gründen ausfällt. Manchmal wird der ursprüngliche Begriff missverstanden und das Resultat ist unbrauchbar wie *plimatsch* 'Kissen' in *bischla da plimatsch* 'Lagerhülse' als Umdeutung eines horizontal beweglichen Holzes auf dem Wagen für eine rotierende Drehbewegung, der *rullera* 'Rolllager', *cullanera* 'Kugellager' entsprechen. Heute schmunzelt man über die Pionierbezeichnung sr: *tabla spurteglia* (Gadola 1956:79) für eine 'elektrische Schalt(er)tafel' mit Unterdifferenzierung von 'elektrischer Schalter' und 'Schalterfenster', das inzwischen zu *tavla da distribuziun, cumond* berichtigt wurde. Der ganze Bereich der Elektrizität mit 'Strom', 'Spannung', 'Hochspannung' und 'Starkstrom' usw. wurde erst nach 1990 für das Pledari Grond terminologisch aufgearbeitet; umgesetzt ist es kaum, schließlich ist es ziemlich abstrakt.⁵

5. Auswirkungen der terminologischen Anpassung

Außer in offiziellen Bereichen mit einem vorgeschriebenen Sprachgebrauch wie die dreisprachige Kantonsverwaltung, die Gesetzgebung und die Herstellung von Schulbüchern, ist die terminologische Anpassung ein Zusammenspiel von glücklichen,

5 Die ersten Fachvorschläge wurden 1917 im Chalender ladin veröffentlicht: Davart l'electricited. Terms romauntschs per l'electricited, acceptos dalla Commissiun Linguistica, 70-71.

überzeugenden Vorschlägen auf der einen Seite und einer erfolgreichen Vermarktung auf der anderen Seite. Zuerst zur linguistischen Komponente:

6. Linguistische Identifizierung

Seit der Einführung des *Rumantsch grischun* 1982 bedeutet Terminologie nicht nur eine lexikalische Erweiterung, teilweise in einer Diglossie, sondern auch einen Paradigmawechsel hin zum Einheitsstandard. Neben psychologischen und politischen Hindernissen bestehen auch syntaktisch-semantische Unterschiede. Außer bei Gesprächen in sektoriellen Sprachen zwischen Fachleuten, sind die betroffenen Endanwender Laien, die Romanisch praktisch nur sprechen, und deshalb muss die Fachterminologie folgendes beachten:

- der Begriff muss durchsichtig, transparent sein sowohl elementar (wörtlich) als auch in der Bedeutung (inhaltlich):
sufflafain 'Heugebläse', *tirastapun* 'Zapfenzieher', *pendiculara* 'Seilbahn', *autpledader* 'Lautsprecher', *portasperanza* 'Hoffnungsträger'; problematisch ist sr: *sclausanetg*, rg: *strasarsa* und, trotz des Calques, *cirquit curt* 'Kurzschluss'; *camiun-tractor* erkennt die ländliche Bevölkerung als 'Ackertraktor' und nicht als modernen 'Sattelschlepper', 'LKW';

- er muss sich regional und idiomatisch anpassen:
rg: *tilastruvas*, vl *tirascauvs*, sr: *tilastrubas* konnte in der Lumnezia zu *tre(r)strubas* angepasst werden. Schnell verliert sich aber der Grundbegriff, so für 'Scheibenwischer' mit der Vermischung von 'wischen', 'waschen' und 'trocknen' rg: *fruschavaiders*, *fruschets*, sr: *schubregiaveiders*, *furschets*, st: *furbaveders*, sm: *siaintaveders*, pt: *süjaintavaiders*, *terdschins*, vl: *süaintavaiders*, *terdschins* und die schon erwähnten *chasperets*;
rg: *biancaria* 'Weißwäsche' ist unverständlich im Romanischen mit nur *alv* als Benennung für 'weiß'; üblich sind konkrete Begriffe wie sr: *resti da letg*, vl: *linzöls* 'Bettwäsche', sr: *resti suten* 'Unterwäsche';

- er sollte weder zur Homonymie noch zu Heteronymie führen:
schluppet/buis 'Gewehr' sind regional so verankert, dass keiner davon sich durchsetzen kann; die ungenügende Unterscheidung von *fittar* 'mieten' und *affittar* 'mieten, vermieten' erfordert eine Periphrase *prender/dar a fit* 'in Pacht nehmen/geben';
rg: *taglia* 'Steuer' ist bevorzugt worden, obwohl *imposta* produktiver wäre: **impostabel*, **impostar*, das im PG nur als Part. Perf. *contribuziun imposta* 'auferlegte Leistung' steht und kaum von rg: *impostar* 'aufgeben, einfächern' als Buchwörter bedrängt würde;

vl: *cumischium sindicatoria* 'Geschäftsprüfungskommission' ist neben sr, rg: *cumissium da gestiun* lediglich eine Scheinopposition, weil *gestiun* überall 'Geschäft' bedeutet, aber trotzdem identifiziert sich die Bevölkerung zunehmend mit solchen Schibboleths als Gegenreaktion zu einer drohenden Vereinheitlichung;

- darf im Romanischen dem deutschen Diskurs und Geist⁶ nicht zuwiderlaufen. Begriffe wie *denticulara* 'Zahnradbahn' sind offenbar zu wenig einsichtig und brauchen eine Periphrase *viafier a roda dentada*, um sich von *dentera* 'Zahnspange', *dentadira* 'Gebiss, Zahnung' und *dentigliun* 'Bartenplatte (beim Wal)' abzusetzen, weil viele Morpheme zu schwach und deshalb nicht produktiv sind. Trotzdem vermochten sich auch eigenständige Begriffe durchzusetzen: *runal* 'Schlepplift', *sutgera* 'Sessellift'; *rentier* 'Rentner' ist umstritten wegen des deutschen Synonyms 'Ren, Rentier' und vl: *golier* vermag den üblichen *goli* 'Goali, Torhüter' kaum zu vertreiben.

Fast unüberwindliche Hindernisse für eine Standardisierung stellen die idiomatisch ausgeprägten Bereiche der Speisen und der häuslichen Tierwelt dar. Der Ersatz für sr: *tier* 'Tier' *animal* wird besonders im Engadin pejorativ als 'Viech' verstanden; dessen *bes-cha* wird wiederum mit sr: *bestga* und besonders *bestia* 'Raubtier, Bestie' gleichgesetzt, denn *biestga* gilt dort nur für 'Vieh, Großvieh' und entspricht nicht vl: *besch* 'Schaf'. Die exemplarische Vielfalt belegt die Bezeichnung der Körperteile beim Menschen (PLED RUMANTSCH 3 1984).

Bei bilingualen Sprechern mit einer stark interferierten Sprache betrifft die linguistische Identifizierung nicht nur das definierte Romanisch als postulierte reine Sprache, sondern das gesamte Repertoire (Deutsch und andere Sprachen). Die romanische Form wirkt oft puristisch mit entsprechendem Registerwechsel und verletzt den oft einzigen verfügbaren tieferen Stil des Sprachverwenders; es entsteht ein neues Register. In der geläufigen Jugendsprache wirken *magiel* 'Glas' und *gervosa* 'Bier' stilistisch fremd neben *glas* und *pier*, und *sa participar* 'sich beteiligen' entspricht aus sozialkommunikativen Gründen nicht *far cun* 'mitmachen', das man ersetzen will (Solèr 2002:261).

7. Linguistische Bereicherung und Unsicherheit

Die Terminologie will eine umfassendere Verwendbarkeit der Sprache mit neuen Domänen erreichen, aber sie soll auch die linguistische Ausdrucksmöglichkeit erweitern und so das Romanische als Fachsprache fördern. Wohl sind die derivativen Prozesse linguistisch geeigneter als die analytischen, aber diese werden wegen der

6 Gemäss Ascoli (1880-83:407) „materia romana in spirito tedesco“ und Solèr (2002:261) „mentale Symbiose“.

höheren Transparenz und der Nähe zum Deutschen bevorzugt; auch psychologische Gründe scheinen eine Hürde darzustellen. Spontane und spielerische Bildungen sind Einzelfälle ohne Wirkung, so *idear* 'die Idee haben', *impulsar* 'den Impuls geben' oder *praulistic* 'märchenhaft'. Besonders die Zeitungsleute des 19. Jh. mussten mehr oder weniger eine neue Sprache für die sich stark veränderte Umwelt erschaffen, weil bis anhin nur eine religiöse und juristische Fachsprache bestand und Deutsch keine Alternative war. Noch heute sind die Medien Pioniere, denken wir an 'Seebeben', 'SARS', 'Herdenschutzhund' und 'Vogelgrippe', aber gelegentlich verhindern notdürftige abstrakte Stelzenbegriffe eine genaue und kohärente Terminologie: *chaun da protecziun* 'Herdenschutzhund' anstatt *chaun-pastur*, *chaun pertgirader*; *forzas da segirezza* 'Sicherheitskräfte' anstatt eines konkreten Begriffs *armada*, *polizia*; *bains da provediment* 'Versorgungsgüter' für *victualias*, *provediment* oder *effectiv*, *populaziun da peschs* 'Fischbestand, -population', das romanisch als 'Fischbevölkerung' verstanden wird anstatt (*ils*) *peschs* als Kollektiv.

Einzelelemente lassen sich problemlos austauschen, während mehrgliedrige Begriffe die bestehende Syntax überfordern, indem sie sie verändern oder eine systemfremde Syntax übernehmen:

- Verben mit Präposition im abstrakten Sinn:

metter enturn ideas 'Ideen umsetzen' verstanden als 'Ideen umlegen, töten' anstatt *realisar*, *concretisar ideas*; sr: *fatg en lavur cumina priu ora la lavur da professiun* 'in Gemeinwerk gemacht ausgenommen die Facharbeit' anstatt *auter che*, *cun excepziun da*, *danor* 'anders als, mit Ausnahme von, außer';

- Nominalisierung und Nominalketten:

sm: *La discussiun cun la donna ò savia neir exequeida sainza grond disturbi da canera* 'das Gespräch mit der Frau konnte ohne größere Belästigung durch Lärm durchgeführt werden' anstatt *ins ò savia discorrer cun la donna senza neir disturbo* 'man konnte mit der Frau sprechen, ohne gestört zu werden';

- Stelzensätze und Leerformulierungen:

far adiever dals meds publics da transport en Engiadina Bassa 'Gebrauch machen von den öffentlichen Verkehrsmitteln im Unterengadin' für *ir cun il tren ed auto da posta*; *exequir lavurs da surfatscha* 'Oberflächenarbeiten ausführen' anstatt *far la cuvrida* 'die Abdeckung (der Straße) machen'.

Mit diesen transkodischen Bildungen könnte man sich linguistisch allenfalls abfinden, wenn das Romanische damit nicht noch die Identität verlieren würde. Komplexe Begriffe widersprechen zwar der Sprachgewohnheit, der Tradition der Romanischsprecher, aber die abstrakte, sperrige, styroporartige Syntax hat sich

vom traditionellen Romanisch so weit entfernt, dass man es nur über das Deutsche, versteht, also aus der Rückübersetzung.

8. Sozialpsychologische Aspekte

Das Romanische wird in dörflichen Sprachgemeinschaften und teilweise in den Regionalzentren verwendet; es schafft dort eine lokale Identifikation unter den Romanischsprechern, ganz besonders den Einheimischen, und steht für das Überschaubare gegenüber dem Fremden. Wenn man aber beruflich oder mit einem nichtromanischen Partner eine andere Sprache verwendet, so tut man das emotionslos. Und wenn manche wegen ihrer fehlenden romanischen Fachkompetenz Deutsch verwenden, dann ist das eher ein Reflex der Sprachpolitik, als dass man sich schämt. Es ist zudem eher selten, dass man bewusst neue romanische Ausdrücke sucht, denn allzu oft vergessen besonders die Sprachverwalter und Sprachpfleger, dass das Romanische oft nur informell gesprochen wird und endgültig eine Ko-Sprache des Deutschen ist.

9. Politisch-wirtschaftliche Aspekte

Jede Sprache kann zwar materiell (Terminologie, Neologie) erfolgreich, sozusagen im Labor erneuert werden, aber deren Verbreitung, Implementierung, kann nur die Anwendungsseite (Produkte, Sprachträger usw.) bewirken. Beim Romanischen hingegen sind die anwendungsorientierten Bedingungen überhaupt nicht oder nur schwach erfüllt und auch der technisch-linguistische Bereich ist nicht eindeutig bestimmt (Entscheidungskompetenz, Verbindlichkeit, Verbreitung). Die enge und fast intime Sprachgemeinschaft fordert vom linguistischen Bearbeiter, der zugleich selber betroffen ist, eine technisch-linguistische Spracherneuerung, die einerseits systematisch ist und andererseits auch eine sichere Trivallösung liefert. Diese Individualisierung beeinflusst trotzdem die Spracherneuerung weniger als andere Rahmenbedingungen, nämlich das sprachliche Umfeld, die Nützlichkeit und die Kleinräumigkeit.

Im gemeinsamen Wirtschafts-, Verkehrs-, Ausbildungs- und Kommunikationsraum mit der deutschen Schweiz fehlt dem Romanischen die konkrete, durchgehende Anwendung, die Kommerzialisierung der Sprache, außer in den gesteuerten Bereichen der Verwaltung und Volksschule in denen sie ohne direkte Konkurrenz ist.

10. Verbreitung und Nachhaltigkeit

Im ganzen Anpassungsprozess erweist sich - bei einer minorisierten Sprache nicht unerwartet - ausgerechnet die wichtigste Phase, nämlich die Verbreitung und systematische Anwendung, als schwächstes Glied. Die Anpassung dringt nicht direkt zum Anwender im Berufsalltag, sondern er muss sie bewusst holen und auch bereit sein,

sie zu verwenden; gezwungen wird er kaum und wenn, dann nur in einzelnen Bereichen von befohlener Mehrsprachigkeit. Zudem verstreicht häufig so viel Zeit zwischen dem Vorschlag und der Anwendung beim Endverbraucher, dass der Begriff im technischen Bereich entweder schon veraltet ist oder dass die entlehnte Erstbezeichnung oder ein Trivialbegriff sich eingebürgert hat. Häufig überrumpelt die Entwicklung aber die Sprache regelrecht, so z. B. im Informatikbereich.

Beim Start des *Rumantsch grischun* 1982 war auch die Informatik ein relativ neues und unbekanntes Werkzeug, so dass in dieser Phase auch die romanischen Begriffe dafür geschaffen werden konnten. In der anschließenden rasanten Verbreitung der Informatik sind diese aber durch die internationalen bedrängt oder verdrängt worden, so: *ordinatur* 'Rechner, Computer' > *computer*; *platta fixa* 'Festplatte' > *HD*; *arcunar* 'speichern' > *segirar* 'sichern'; *datas* 'Daten', *datoteca*, 'Datenfile' > *file*; *actualisaziun, cumplettaziun* 'Update' > *update*, *palpader* 'Scanner' > *scanner*. 'Laptop' hat man direkt übernommen ohne *portabel* vorzuschlagen.

Die Zeitungsredaktoren des 19. Jh. konnten ihre neuen, wenig systematischen Begriffe unmittelbar den Lesern konkurrenzlos vermitteln; aber sie verliefen oft im Sand, weil sie nicht systematisch gesammelt und weiter verbreitet wurden. Diese Schwächen versuchte die für das Engadin 1919 begonnene themenorientierte Reihe „S-CHET RUMANTSCH“ in der Zeitung und später in Buchform zu überwinden. Ich möchte es nicht unterlassen, einige phantasievolle Verbreitungsarten wenigstens zu erwähnen:

- mit Metzgereibegriffen bedruckte Papiertüten
- Beschriftungen der Produkte in den Auslagen
- Sportterminologie auf Tafeln in den Turnhallen
- zweisprachige Rechnungsformulare für Autowerkstätten
- Beschreibung und Gebrauchsanweisung auf Produktpackungen⁷

Diese direkten Anwendungen wurden durch sekundäre Listen ergänzt und noch heute veröffentlichen einzelne Zeitungen regelmäßig kleine Wortlisten.

Die wohl erfolgreichste Verbreitung brachte die Schule bis zu den großen Strukturänderungen der 70er Jahre des letzten Jahrhunderts, die eine noch mehrheitlich nur-romanische und ländliche Bevölkerung in eine neue Welt inhaltlich und sprachlich einführte. Diese Periode dauerte so lange, dass eine Schulbuchreihe noch über zwei Schulgenerationen reichte und die gelernten Neuerungen fast lebenslänglich galten. In dieser Zeit fallen auch die ersten systematischen Wörterbücher.

⁷ Eines der wenigen Beispiele ist die „Lataria Engiadinaisa SA, CH-7502 Bever“; in den 90er Jahren waren einige Tierarzneien romanisch beschriftet; die Anschrift Adina Coca Cola blieb ein Werbegag der 90er Jahre.

Wörterbücher und Lexikographie sind unverzichtbare Hilfsmittel für jede Sprache, aber wenig wirkungsvoll für den Sprachverwender, weil er bewusst und außerhalb der Gesprächs, sozusagen metakommunikativ auf sie greifen muss. Sie sind im Romanischen zudem nur referenziell und liefern eine Ersatzbezeichnung für schon bekannte - zwar deutsche - Ausdrücke, aber trotzdem verfügbar im Zeicheninventar der Romanischsprecher (Reiter 1984:289).

Während die hochspezialisierte Terminologie in keiner Sprache zum allgemeinen Wortschatz gehört, sollten die neuen Begriffe des modernen Alltags wie Verkehr, Kommunikation, Unterhaltung, Lifestyle, aber auch der neueren Verwaltung umgesetzt werden. Für eine umfangreichere Durchsetzung, Implementierung, fehlt das romanische Umfeld sowie der Terminologiediskurs. Die bestehenden Medien erfüllen lediglich eine lokale und emotionale Rolle gegenüber einem umfassenden deutschsprachigen Angebot, und so entwickelt sich auch kaum eine Sprachnorm.

Die Verwendung des Romanischen allgemein, und einer offiziellen Sprachform im besonderen, anstatt des Deutschen oder des Englischen ist nur ausnahmsweise bei Kulturtouristen und Heimwehromanen ein kommerzieller Faktor; sonst kann es sogar hinderlich sein, wie die Reaktionen der Bevölkerung auf jegliche Anpassung eindrücklich belegen. Das Romanische besitzt kein geschütztes Sprachgebiet und seine Verwendung kann gesetzlich kaum oder nicht durchgesetzt werden wie beispielsweise in Frankreich.

Die kantonale Verwaltung verwendet die drei offiziellen Kantonsprachen in den Veröffentlichungen und im Internet (Erklärungen, Berichte, Anleitungen, Hinweise, Abstimmungen usw.). In der Verwaltungstätigkeit hingegen ist das Romanische gegenüber dem Deutschen besonders im Fachbereich eingeschränkt: Die romanische Steuererklärung gibt es nicht digital, verschiedene amtliche Formulare können online nur deutsch und gelegentlich italienisch ausgefüllt werden. Offensichtlich trifft folgendes für die regionalen Organisationen, die direkt mit der Bevölkerung arbeiten zu: nur Romanisch selten, zweisprachig ist häufiger, eher plakativ, und mehrheitlich Deutsch. Das ist auch eine Folge des 'polykephalen', sprich regionalisierten Romanisch als Teil einer deutschen Umwelt, und es verunmöglicht eine einheitliche Fachterminologie und ihre einheitliche umgangssprachliche Umsetzung. So bestätigt sich die Feststellung von Haarmann (1993:108) „Hier liegt ein prinzipielles Problem des Minderheitenschutzes. Eine indominante Sprache hat zwar grundsätzlich bessere Chancen zu überleben, wenn ihre Verwendung in Bereichen des öffentlichen Lebens garantiert wird, es besteht aber keine automatische Wechselbeziehung zwischen

einer sprachpolitischen Förderung und der Erhaltung dieses Kommunikationsmediums als Mutter- und Heimsprache“.

Die gewinnorientierte Wirtschaft wählt dementsprechend die beste Sprache. Romanisch verwendet sie identifikatorisch und emotional in den rtr. Regionen, aber nicht als durchgehende Plattform (Banken, Versicherungen). „Unique Selling Proposition“ ist ein Schlagwort und wird bestenfalls im Mäzenatentum eingelöst. Ohne die operative Bedeutung passt sich keine Fachsprache an, oder sie wird nicht systematisch und einheitlich verwendet, sondern als lokale und stilistische (diglossische) Variante, banalisiert als Trivialterminologie. Dann ist auch die domänenspezifische Verwendung des Romanischen und deren Aktualisierung weitgehend illusorisch, und auch die bescheidene berufliche Aus- und Weiterbildung dient bestenfalls für romanische Infrastrukturbetriebe (Lia Rumantscha, Radio, Fernsehen und die Schulunterstufe).

Das Romanische passt sich zwar den neuen Erfordernissen dauernd an, aber weil diese Entwicklung eher spontan als geordnet erfolgt, und weil sie eher die gesprochene Sprache mit einer Trivialterminologie betrifft, fördert sie die zweisprachige Diglossie mit dem Schriftdeutschen in allen Außenbeziehungen und sogar unter Romanischverwendern.

11. Ausblick - aber kaum die Lösung

Das klingt nach einer Bankrotterklärung. Das ist es nicht, aber man muss sich auf die Grundlagen rückbesinnen und in erster Linie die Randbedingungen, die soziolinguistischen, politischen und wirtschaftlichen Voraussetzungen ernst(-er) nehmen.

Zum ersten die Terminologie; anstatt der akademischen und direkt kopierten, sterilen Erneuerungen muss man sich um assoziative - und überschreite sie auch die Einzelsprache - einsichtige oder sogar spielerische, aber praxistaugliche Benennungen bemühen, die lebensnah sind und genaue Inhalte sprachlich sinnvoll und kulturell verträglich umsetzen können.

Die Hauptschwierigkeit ist und bleibt die Verbreitung. Wenn eine Sprache wie das Romanische mehr kulturell, ideell und politisch, als wirtschaftlich begründet ist, erweist sich deren Anpassung (Modernisierung und Standardisierung) umso weniger durchsetzbar.

Psychologischer Druck oder die Drohung eines Sprachniedergangs wirken vielleicht kurzfristig, erwecken Hoffnungen, aber sie wirken niemals nachhaltig.

Dass sich der Riesenaufwand für die Romanisierung des ganzen MS-Office mit der Orthografiekontrolle (Spell-Checker) nicht lohnt, ist leicht vorauszusagen; das

Produkt spricht eine zu kleine Gruppe an und das Bedürfnis nach romanischen Texten kann man nicht künstlich erzeugen.⁸ Mit Sicherheit hilfreich und seit bald 15 Jahren nützlich erwies sich die Terminologearbeit im Pledari Grond der Lia Rumantscha; es ist zwar bescheidener, dafür praxisbezogen und dient zudem als eine Hilfsbrücke zu den Idiomen und sollte somit Spannungen abbauen.

Für eine isolierte Kleinsprache ist es aber unabdingbar, die Sprachverwender schnell, unkompliziert und konkret zu unterstützen. Die privaten und kollektiven Sprachverwalter wie die Lia Rumantscha und der Kanton mit seiner umfassenden Tätigkeit können die Sprachverwender am ehesten überzeugen mit gebrauchsfertigen Vorlagen, schnellen Übersetzungen, gefälligen Texten und mit einem umsichtigen, engen Coaching bei der Sprachverwendung und so wären auch die Empfänger eingebunden.

Für diese Aufgaben braucht es Terminologearbeit. Das ist ein guter Anfang und ist auch zu bewältigen. Die folgenden ebenso notwendigen Schritte müssen zuallererst die Sprachverwender tun.

8 Versuche der LR um 1990 digitales Material für Handwerksbetriebe herzustellen und zu vertreiben scheiterte an den einzelbetrieblichen „Branchenlösungen“, die miteinander unverträglich sind, an der Einheitsform Rumantsch grischun, an der gewohnten deutschen Berufssprache sowie der Einstellung gegenüber der deutschsprachigen Kundschaft.

Bibliographie

Ascoli, G.I. (1880-1883). "Annotazioni sistematiche al Barlaam e Giosafat soprasilvano." *Archivio glottologico italiano*. Roma: Loescher, 7:365-612.

Carisch, O. (1848). *Taschen-Wörterbuch der Rhätoromanischen Sprache in Graubünden*. Chur: Wassali.

Coray, R. (1993). "La mumma romontscha: in mitos." *ISCHI* 77, 4, 146-151.

Decurtins, A. (1993). "Wortschatz und Wortbildung - Beobachtungen im Lichte der bündnerromanischen Zeitungssprache des 19./20. Jahrhunderts." *Rätoromanisch, Aufsätze zur Sprach-, Kulturgeschichte und zur Kulturpolitik. Romanica Rætica* 8, Chur: Società Retorumantscha, 235-254.

EUROMOSAIC (1996). *Produktion und Reproduktion der Minderheitensprachgemeinschaften in der Europäischen Union*. Brüssel/Luxemburg: Amt für amtliche Veröffentlichungen der EG.

Gadola, G. (1956). "Contribuziun alla sligiazium dil problem puril muntagnard." *Igl Ischi*, 42, 33-93.

Haarmann, H. (1993). *Die Sprachenwelt Europas. Geschichte und Zukunft der Sprachnationen zwischen Atlantik und Ural*. Frankfurt: Campus.

Jaberg, K. & Jud, J. (1928). *Der Sprachatlas als Forschungsinstrument*. Halle: Niemeyer.

Pledari Grond (2003) deutsch-rumantsch, rumantsch-deutsch, cun conjugaziuns dals verbs rumantschs. Cuira: Lia rumantscha [CD-ROM].

PLED RUMANTSCH/PLAID ROMONTSCH 3 (1984). *Biologia*. Cuira: Lia rumantscha.

RÄTOROMANISCH (2004). *Facts & Figures*. Cuira: Lia rumantscha.

Reiter, N. (1984). *Gruppe, Sprache, Nation*. Wiesbaden: Harrassowitz.

S-CHET RUMANTSCH (1917-1963). *Fögls per cumbatter la caricatura nella lingua ladina*. Scuol: Uniun dals Grischs.

Solèr, C. (1986). "Ist das Domleschg zweisprachig?" *Bündner Monatsblatt*, 11/12, 283-300.

Solèr, C. (2002). "Spracherhaltung - trotz oder wegen des Purismus. Etappen des Rätoromanischen." *Bündner Monatsblatt*, 4, 251-264.

Solèr, C. (2005). "Co e cura che la scrittira emprenda rumantsch. Cudeschs da scola per la Surselva." *Annalas da la Societad Retorumantscha*. Cuira: Societad retorumantscha, 7-32.

Implementing NLP-Projects for Small Languages: Instructions for Funding Bodies, Strategies for Developers

Oliver Streiter

This research starts from the assumption that the conditions under which ‘Small Language’ Projects (SLPs) and ‘Big Language’ Projects (BLPs) are conducted are different. These differences have far-reaching consequences that go beyond the material conditions of projects. We will therefore try to identify strategies or techniques that aim to handle these problems. A central idea we put forward is pooling the resources to be developed with other similar Open Source resources. We will elaborate the expected advantages of this approach, and suggest that it is of such crucial importance that funding organisations should put it as *condicio sine qua non* into the project contract.

1. Introduction: Small Language & ‘Big Language’ Projects - An Analysis of their Differences

Implementing NLP-projects for Small Languages: Is this an issue that requires special attention? Are Small Language Projects (SLPs) different from ‘Big Language’ Projects (BLPs)? What might happen if SLPs are handled in the same way as BLPs? What are the risks? How can they be reduced? Can we formulate general guidelines so that such projects might be conducted more safely? Although the processing of minority languages and other Small Languages has been subject to a series of workshops, this subject has been barely tackled as such. While most contributions discuss specific achievements (e.g. an implementation or transfer of a technique from Big Languages to Small Languages), only a few articles transcend to higher levels of reflection on how Small Language Projects might be conducted in general.

In this contribution, we will compare SLPs and BLPs at the abstract schematic level. This comparison reveals differences that affect, among other things, the status of the researcher, the research paradigm to be chosen, the attractiveness of the research for young researchers, as well as the persistence and availability of the elaborated data - all to the disadvantage of Small Languages. We will advance one far-reaching solution that overcomes some of these problems inherent to SLPs, that is, *to pool the developed resources with other similar Open Source resources and make*

them freely available. We will discuss, step by step, the possible advantages of this strategy, and suggest that this strategy is so promising and so crucial for the survival of the elaborated data that *funding organisations should put it as condicio sine qua non into their project contract.*

Let us start with the comparison of BLPs and SLPs.

- **Competition in Big Languages:** Big Languages are processed in more than one research centre. Within one research centre more than one group might work on different aspects of this single language. The different centres or groups compete for funding, and thus strive for scientific reputation (publications, membership in exclusive circles, membership in decision taking bodies) and try to influence the decision-making processes of funding bodies.

- **Niches for Small Languages:** Small Languages are studied by individual persons, small research centres or cultural organisations. Small Languages create a niche that protects the research and the researcher. Direct competition is unusual. This, without doubt, is positive. On the negative side, however, we notice that methodological decisions, approaches and evaluations are not challenged by competitive research. This might lead to a self-protecting attitude that ignores inspiration coming from successful comparable language projects.

- **Big Languages Promise Money:** There is commercial demand for BLPs as can be seen from the funding that companies like Google or Microsoft provide for NLP projects. As these companies try to obtain a relative advantage over their competitors, language data, algorithms, and so forth are kept secret.

- **There is No Money in Small Languages:** Those organisations that fund BLPs are not interested in SLPs. If a Small Language wants to integrate its spellchecker in Microsoft Word, the SLP has to provide the linguistic data with no or little remuneration for Microsoft.

- **Big Languages Hide Data:** Language resources for Big Languages are and have been produced many times in different variants before they find their way into an application, or before they are publicly released. Since research centres for Big Languages compete for funding, recognition and commercialisation, every centre hopes to obtain a relative advantage over their competitors by keeping developed resources inaccessible to others.¹

¹ That this secretiveness might have any advantages at all can be called into question. Compare, for example, the respective advantages Netscape or Sun had from handing over their resources to the Open Source Community. Netscape-based browsers by far outperform their previous competitors such as Internet Explorer or Opera and the data handling in Open Office is going to be copied by the competitor Microsoft Office. As for the scientific reputation, people cited frequently are those who make available their resources including dictionaries and corpora (e.g. Eric Brill, Henry Kucera, W. Nelson Francis,

-
- **Small Languages Shouldn't Do So:** For Small Languages, such a waste of time and energy is unreasonable. Resources that have been built once should be made freely available so that new, related projects can build on top of them, even if they are conducted elsewhere. Without direct competition, a research centre should have no disadvantage by making its resources publicly available. Reasons for not distributing the developed resources are most likely due to the misconception that sharing the data equals to losing the copyright on the data.

However, under the *General Public License* (a license that might be used in SLPs), the distribution of resources requires that copies must contain the appropriate copyright notice (so that the rights remain with the author of the resources). In addition, it has to contain the disclaimer of warranty, so that the author is not liable for any problems others have with the data or programs. Somebody modifying the data or programs cannot sell this modification unless the source code is made freely available, so that everybody, including the author, can take over the resources for further improvements.

The consequence of not sharing the data (i.e., keeping the data on the researcher's hard disk) is that the data will be definitely lost within ten years after its last modification.²

- **BLPs Overlap** in time and create a research continuum. In this research continuum, researchers and resources can develop and adapt to new paradigms (defined as "exemplary instances of scientific research", Kuhn 1996/1962) or new research guidelines. In fact, a large part of a project is concerned with tying the knots between past and future projects. Data is re-worked, re-modelled and thus kept in shape for the future.

- **SLPs are Discontinuous.** There might be temporal gaps between one SLP and the next one. This threatens the continuity of the research, forces researchers to leave the research body, or might endanger the persistence of the elaborated

Huang Chu-ren, Chen Keh-jiann, George A. Miller, Christiane Fellbaum, Throsten Brants, and many others).

2 Reasons for the physical loss of data are: Personal mobility (e.g. after the retirement of a collaborator, nobody knows that the data exists, or how it can be accessed or used). Changes in software formats (e.g. the format of backup programs, or changes in the SCSI controller make the data unreadable). Changes in the physical nature of external memories (punch card, soft floppy disk, hard floppy disk, micro floppy, CD-ROM, magnetic tape, external hard disk, or USB-stick) and the devices that can read them. Hard disk failure (caused by firmware corruption, electronic failure, mechanical failure, logical failure, or bad sectors). The limited lifetime of storage devices is: tapes (2 years), magnetic media (5-10 years) and optic media (10-30 years). This depends very much on the conditions of usage and storage (temperature, light and humidity).

data. The data is unlikely to be re-worked, or ported to new platforms or formats, and thus it risks becoming obsolete or unreadable.

- **BLPs Rely on Specialists:** The bigger the group in a BLP, the more specialists in programming languages, databases, linguistic theories, parsing, and so forth it will integrate. Specialists make the BLP autonomous, since specific solutions can be fabricated when needed.

- **All-rounders at Work:** Specialisation is less likely to be found in SLPs, where one person has to cover a wider range of activities, theories, tools, and so forth. in addition to administrative tasks. Thus, SLP projects cannot operate autonomously. They largely depend on toolkits, integrated software packages, and so forth. Choosing the right toolkit is not an easy task. It not only decides the success or failure of the project, but will also influence the course of the research more than the genius of the researcher. If a standard program is simply chosen because the research group is acquainted with it, a rapid project start might be bought at the price of a troublesome project history, data that is difficult to port or upgrade, or data that does not match the linguistic reality it should describe.

- **BLPs Play with Research Paradigms:** BLPs can freely choose their research paradigm and therefore frequently follow the most recent trends in research. Although different research paradigms offer different solutions and have different constraints, BLPs are not so sensitive to these constraints and can cope successfully with any of them. BLPs must not only be capable of handling new research paradigms; otherwise, the new research paradigms could not survive, BLPs are even expected to explore new research paradigms, as they are the only ones having the gross scientific product that can cope with fruitless attempts and time-consuming explorations. Indeed, we observe that BLPs frequently turn to the latest research paradigm to gain visibility and reputation. Shifts in the research paradigm might make it necessary to recreate language resources in another format or another logical structure.

- **SLPs Depend on the Right Research Paradigm:** SLPs do not dispose of rich and manifold resources (dictionaries, tagged corpora, grammars, tag-sets, and taggers) in the same way as BLPs do. The research paradigm should thus be chosen according to the nature and quality of the available resources, and not according to the latest fashion in research. This might imply the usage of a) example-based methods, since they require less annotated data (cf. Streiter & de Luca [2003]), b) unsupervised learning algorithms, if no annotations are available, or c) hybrid bootstrapping methods (e.g. D. Prinsloo & U. Heid 2006), which are almost

impossible to evaluate. Young researchers may experience these restrictions as a conflict. On one hand, they have to promote their research, ideally in the most fashionable research paradigm; on the other hand, they have to find approaches compatible with the available resources. Fortunately, after the dust of a new research trend has settled³, new research trends are looked at in a less mystified light, and it is perfectly acceptable for SLPs to stick to an older research paradigm, if it conforms to the overall requirements.⁴

- **Model Research in Big Languages:** Research on Big Languages is frequently presented as research on that respective language and, in addition, as research on Language in general. The same piece of research might thus be sold twice. From this, BLPs derive not only a greater reputation and better project funding, but also an increased attractiveness of this research for young researchers. Big Languages, as a consequence, are those languages for which, in virtue of general linguistic accounts, documentary and pedagogic resources are developed. Students are trained in and with these languages in the most fashionable methods, which they learn to consider as superior.

- **SLPs Represent Applied Research - at best!:** SLPs are less likely to sell their research as research on Language in general. In fact, little else but research on English counts as research on Language, and is considered research on a specific language at best.⁵ The less general the scope of research, the less likely it is to be

³ I have taken this term from Harold Somers (1998).

⁴ Although Big Language research centres are free to choose their research paradigm, they more often than not are committed to a specific research paradigm, (i.e., the one they have been following for years or in which they play a leading role. This specialization of research centres to a research paradigm is partially desirable, as only specialists can advance the respective paradigm. However, when they do research on Small Languages, either to extend the scope of the paradigm or to access alternative funding, striking mismatches can be observed between paradigm and resources. Such mismatches are of no concern to the Big Language research centre, which, after all, is doing an academic exercise, but they should be closely watched in SLPs, where such mismatches will cause the complete failure of the project. For example, Machine Translation knows two approaches: rule-based approaches, where linguists write the translation rules; and corpus-based approaches, where the computer derives the translation rules from parallel corpora. Corpus-based approaches can be statistics-based or example-based. Recently, RWTH Aachen University, known for its cutting-edge research in statistical Machine Translation, proposed a statistical approach to sign language translation (Bungeroth & Ney 2004). One year later Morrissey & Way (2005) from Dublin City University, a leading agent in Example-based Machine Translation, proposed "An Example-Based Approach to Translating Sign Languages." The fact, however, that parallel corpora involving at least one sign language are extremely rare and extremely small is done away in both papers, as if it would not affect the research. In other words, the research builds on a type of resource that does not actually exist, just to please the paradigm.

⁵ In a Round Table discussion at the 1st SIGHAN Workshop on Chinese Language Processing, hosted by ACL in Hong Kong, 2000, a leading researcher in Computational Linguistics vehemently expressed his dissatisfaction in being considered only a specialist in Chinese Language Processing, while his colleagues working in English are considered specialists in Language Processing. Nobody would call Chomsky a specialist in American English! Working on a Small Languages thus offers a niche at the price of a stigma that prevents researchers from ascending to the Olympus of science.

taught at university. Students then implicitly learn what valuable research is, that is, research on Big Languages and recent research paradigms.

To sum up, we observed that BLPs are conducted in a competitive and sometimes commercialised environment. Competition is a main factor that shapes the way in which BLPs are conducted. In such an environment, it is quite natural for research to overlap and to repeatedly produce similar resources. Not sharing the developed resources is seen as enhancing the competitiveness of the research centre. It is not considered to be an obstacle to the overall advance of the research field: similar resources are available elsewhere in any case. Different research paradigms can be freely explored in BLPs, with an obvious preference for the latest research paradigm, or for the one to which the research centre is committed. Gaining visibility, funding and eternal fame are not subordinated to the goal of producing working language resources.

The situation of SLPs is much more critical. SLPs have to account for the persistence and portability of their data beyond the lifespan of the project, beyond the involvement of a specific researcher, and beyond the lifespan of a format or specific memory device. As Small Languages are not that much involved in the transition of paradigms, data cannot be reworked, especially if research is discontinuous. The refunding of a project due to a shift in research paradigms or lost or unreadable data is unthinkable. With few or no external competitors, most inspiration for SLPs comes from BLPs. However, the motivation for BLPs to choose a research paradigm and their capacity to handle research paradigms (given per definition) is not that of a SLP. For talented young researchers, SLPs are not attractive. As students, they have been trained in BLPs and share with the research community a system of values according to which other languages and other research paradigms are preferred.

2. Improving the Situation: Free Software Pools

Although most readers might consent with the obvious description of SLPs I have given above, few have turned to the solutions I am about to sketch below. The main reason for this might be possible misconceptions or unsubstantiated fears. Let us start with what seems to be the most puzzling question; that is: how can projects and researchers guarantee the existence of their data beyond the direct influence of the researcher him/herself? To give a hypothetical example: you develop a spellchecker for a language of two hundred speakers, all above the age of seventy (including yourself), and none of them having a computer (except for you). How can you ensure that the data survives? The answer is: Pool your data with other data of the same form and function and let the community take care of YOUR data. If you make your

research results available as free software, other people will take care of your data and upgrade it into new formats, whenever needed. 'But,' you might wonder now, 'why should someone take care of my data on an unimportant and probably dying language?' The answer lies in the pool: even if those people do not care about your data per se, they care about the pool, and once they transform resources for new versions they transform all resources of the pool, well knowing that the attractiveness of the pool comes from the number of different language modules within it. In addition, all language modules have the same format and function and if one module can be transformed automatically, all others can be automatically transformed as well.⁶ But which pools exist that could possibly integrate and maintain your data? Below, you find an overview of some popular and useful pools. This list might also be read as a suggestion for possible and interesting language projects, or as a check-list of components of your language that still need to be developed to be at par with other languages. Frequently, the same linguistic resources are made available to different pools (e.g. in ISPELL, ASPELL and MYSPELL). This enlarges the range of applications for a language resource, increases the visibility, and supports data persistence.

- Spelling, Office, Etc:

ISPELL (lgs. > 50); *spelling dictionary + rules*:

A spellchecker, used standalone or integrated into smaller applications.

(AbiWord, flyspell, WBOSS). (<http://www.gnu.org/software/ispell/>)

ASPELL (lgs. > 70); *spelling dictionary + rules*:

An advanced spellchecker, used standalone or integrated into smaller applications. (emacs, AbiWord, WBOSS)(<http://aspell.sourceforge.net/>)

MYSPELL (lgs. > 40); *spelling dictionaries + rules*:

A spellchecker for Open Office. (<http://linguocomponent.openoffice.org/>)

OpenOffice Grammar Checking (lgs. > 5); *syntax checker*:

A heterogeneous set of grammar checkers for Open Office.

OpenOffice Hyphenation (lgs. > 30); *hyphenation dictionary*:

⁶ I do not know how much of an unmotivated over-generalisation this is. In the Fink project (<http://fink.sourceforge.net>), for example, there is one maintainer for each resource and not for each pool and, as a consequence, not all ispell modules are available. In DEBIAN (<http://www.debian.org>) we find again one maintainer for each resource, but orphaned packages, that is packages without maintainer, are taken over by the DEBIAN QA group.⁸

A hyphenation dictionary for use with Open Office, but used also in LaTeX, GNU Troff, Scribus, and Apache FOP.

OpenOffice Thesaurus (lgs. > 12); *thesaurus*:

A thesaurus for use with Open Office.

(<http://lingucomponent.openoffice.org/>)

STYLE and DICTION (lgs. = 2); *style checking*:

Help to improve wording and readability.

(<http://www.gnu.org/software/diction/diction.html>)

HUNSPELL (lgs. > 10); *spelling dictionary + rules*:

An advanced spellchecker for morphologically rich languages that can be turned into a morphological analyser. (<http://hunspell.sourceforge.net/>).

- Dictionaries:

FREEDICT (lgs. > 50); *translation dictionary*:

Simple, bilingual translation dictionaries, optionally with definitions and API as binary and in XML. (<http://sourceforge.net/projects/freedict/>).

Papillon (lgs. > 8); *multilingual dictionaries*:

Multilingual dictionaries structured according to Mel'čuk's Text <=> Meaning Theory. (<http://www.papillon-dictionary.org/Home.po>)

JMDict (lgs. > 5); *multilingual dictionaries*:

Multilingual translation dictionaries in XML, based on word senses.

(http://www.csse.monash.edu.au/~jwb/j_jmdict.html)

- Corpora:

Universal Declaration of Human Rights (lgs. > 300); *parallel corpus*:

The Universal Declaration of Humans Rights has been translated into many languages and can be easily aligned with other languages. (<http://www.unhchr.ch/udhr/navigate/alpha.htm>)

Multex (lgs. > 9); *corpora and morpho-syntactic dictionaries*:

Parallel corpora of Orwell's 1984 annotated in CES with morpho-syntactic information in ten Middle and Eastern European languages. (<http://nl.ijs.si/ME/V2/>)

- Analysis:

Delphin (lgs. > 5); *HPSG-grammars*:

HPSG-Grammars for NLP-applications, in addition various tools for running and developing HPSG resources. (<http://www.delph-in.net/>)

AGFL (lgs.> 4); *parser and grammars*:

A description of Natural Languages with context-free grammars. (<http://www.cs.ru.nl/agfl/>)

- Generation:

KPML (lgs.> 10); *text generation system*:

Systemic-functional grammars for natural language generation.

(<http://purl.org/net/kpml>)

- Machine Translation:

OpenLogos (lgs. > 4); *Machine Translation software and data*:

An open Source version of the Logos Machine Translation System for new language pairs to be added. (<http://logos-os.dfki.de/>).

3. Strategies and Recommendations for Developers

If there is no pool of free software data that matches your own data, you should try the following: 1) *Convert your data into free software* so that you have a greater chance that others will copy and take care of it; and, 2) *Modify your data so that it can be pooled with other data*. This might imply only a minor change in the format of the data that can be done automatically by a script. Alternatively, *create a community that will, in the longterm, create a pool*. In general, this implies that you separate the procedural components (tagger, spelling checker, parser, etc.) from the static linguistic data; make the procedural components freely available; and, describe the format of the static linguistic data. An example might well be Kevin Scannell's CRUBADAN, a web-crawler for the construction of word lists for ISPELL. The author succeeded in creating a community around his tool that develops spellcheckers for more than thirty Small Languages (cf. <http://borel.slu.edu/crubadan>). Through this split of declarative (linguistic) components on the one hand, and procedural components (programs) on the other, many pools come with adequate tools to create and maintain the data.

Pooling of corpora is not as frequent as, for example, the pooling of dictionaries. The main reason for this may be that corpora are very specific, and document a cultural heritage. Pooling them with corpora of different languages, subject areas, registers, and so forth is only of limited use. Nevertheless, there are some computer-linguistic pools that integrate corpora for computational purposes, and that may, therefore, integrate your corpora and maintain them for you. A description of these mostly very complex pools is beyond the scope of this paper, but the interested reader might check the following projects:

- GATE (<http://gate.ac.uk>);
- Natural Language Toolkit (<http://nltk.sourceforge.net>); and,
- XNL RDF (<http://140.127.211.213/xnlrdf>).

Projects targeting language documentation may also host your corpora, (e.g. the TITUS Project [<http://titus.uni-frankfurt.de/>]). In addition, LDC (<http://www ldc.upenn.edu>) and ELRA (<http://www.elra.info>) are hosting and distributing corpora (and dictionaries) so that your institute might profit financially from sold copies of the corpus you created.

Once you decide to create your own free software (including corpora, dictionaries, etc.), you have to think about the license and the format of the data. From the great number of possible licenses you might use for your project (cf. <http://www.gnu.org/philosophy/license-list.html> for a commented list of licenses) you should consider the GNU General Public License, as this license, through the notion of Copyleft, doesn't give a general advantage to someone who is copying and modifying your software. Copyleft refers to the obligation that:

...anyone who redistributes the software, with or without changes, must pass along the freedom to further copy and change it. (...) Copyleft also provides an incentive for other programmers to add to free software.

(<http://www.gnu.org/copyleft/copyleft.htm>)

With Copyleft, modifications have to be made freely available under the same conditions as you originally distributed your data, and if the modifications are of general concern, you can integrate them back into your software. The quality of your resources improves, as others can find and point out mistakes or shortcomings in the resources. They will report to you as long as you remain the distributor. In addition, you may ask people to cite your publication on the resource whenever using the resource for one of their publications. Without Copyleft, important language

data would already have been lost (e.g. the CEDICT dictionary, after the developer disappeared from the Internet).

After putting your resources with the chosen license onto a webpage, you should try to integrate your resource into larger distributions such as DEBIAN (<http://www.debian.org>) so that, in the long term, these organisations will manage your data. To do this, your resources have to conform to some formal requirements that, although seeming tedious, will certainly contribute to the quality and maintainability of your resources (cf. <http://www.debian.org/doc/debian-policy>, for an example of the requirements of integration in DEBIAN). From DEBIAN, your resources might be migrated without your work into other distributions (REDHAT, SUSE, etc.) and into other modules, perhaps embedded into nice GUIs.

4. Instructions for Funding Organisations

A sponsoring organisation that is not interested in sponsoring a specific researcher or research institute, but which tries to promote a Small Language in electronic applications, should insist on putting the resources to be developed under Copyleft, and make this an explicit condition in the contract. Only this will guarantee that the resources will be continually developed even after the lifetime of the project. 'Copylefting' thus allows for a sustainable development of language resources from punctual or discontinuous research activities. Only Copylefting guarantees that the most advanced version is available to everybody who might need it. In fact, a funding organisation that does not care about the way data can be accessed and distributed after the project's end is, in my eyes, guilty of grossly negligent operation. Too many resources have been created in the past only to be lost on old computers, tapes, or simply forgotten. Adding resources to this invisible pile is of no use.

In addition, the funding organisations may require the sources to be bundled with a pool of Free software resources in order to guarantee the physical preservation of the data and its widest accessibility. Copylefting alone only provides the legal grounds for the survival of the data; handing over the resources to a pool will make them available in many copies worldwide and independent from the survival of one or the other hard disk. Copylefting without providing free data access is like eating without swallowing.

5. Free Software for SLPs: Benefits and Unsolved Problems

Admittedly, it would be naive to assume that releasing project results as free software would solve all problems inherent in SLPs. This step might solve the most important problems of data maintenance and storage, and embed the project into a

scientific context. But can it have more positive effects than this? Which problems remain? Let us return to our original list of critical points in SLPs to see how they are affected by such a step.

- Open Source pools create a platform for research and data maintenance that allows the niche to be assigned to SLPs without having to handle situations of competition;
- Data is made freely available for future modifications and improvements. If the data is useful it will be handed over from generation to generation;
- The physical storage of the data is possible in many of the listed pools, and does not depend on the survival of the researchers hard disk;
- The pools frequently provide specific, sophisticated tools for the production of resources. These tools are a cornerstone of a successful project;
- In addition, through working with these tools, researchers acquire knowledge and skills that are relevant for the entire area of NLP;
- Working with these tools will lead to ideas for improvement. Suggesting such improvements will not only help the SLP to leave the niche, but will finally lead to better tools. For young researchers, this allows them to work on their Small Language, and, at the same time, to be connected with a wider community for which their research might be relevant; and,

Through the generality of the tools (i.e., their usage for many languages) the content of SLPs might become more appropriate for university curricula in computational linguistics, terminology, corpus linguistics, and so forth. Some problems, however, remain, for which other solutions have to be found.

These are:

- Discontinuous research if research depends on project acquisition;
- Dependence on research paradigm. Corpus-based approaches can be used only when corpora are available, rule-based approaches when formally trained linguists participate in the project. To overcome most of these limitations, research centres and funding bodies should continuously work on the improvement of the necessary infrastructure for language technology (cf. Sarasola 2000); and,
- Attracting and binding researchers. As the success of a project depends to a large extent on the researchers' engagement and skills, attracting and binding researchers is a sensitive topic, for which soccer clubs provide an illustrative model. Can a SLP attract top players, or is an SLP just a playground for a talented young researcher who will sooner or later transfer to a BLP? Or can the SLP count

on local players only? A policy for building a home for researchers is thus another sensitive issue for which research centres and funding bodies should try to find a solution.

6. Conclusions

Although the ideas outlined in this paper are very much based on ‘sofa-research’ and intuition, a very schematic and simplistic thinking, informal personal communications, and my personal experience, I hope to have provided clear and convincing evidence that Small Language Projects profited, profit and will profit from joining the Open Source community. For those who want to follow this direction, the first and fundamental step is to study possible licenses (<http://www.gnu.org/philosophy/license-list.html>) and to understand their implications for the problems of SLPs, such as the storage and survival of data, their improvement through a large community, and so forth. This article lists some problems against which the licenses can be checked.

Emotional reactions like “I do not want others to fumble in my data,” or “I do not want others to make money with my work” should be openly pronounced and discussed. What are the advantages of others having my data? What are the disadvantages? How can people make money with Open Source data? As said before, misconceptions, and thus unsubstantiated fears, often lead to a rejection of the Open Source idea than a well-founded argument. This is how humans function, but not how we advance Small Languages.

7. Acknowledgments

This paper would not have been written if I had not met with people like Mathias, Isabella, Christian, Judith, Daniel and Kevin. As a consequence of these encounters, the paper is much more a systematic summary than my original thinking. For mistakes, gaps and flawed arguments, however, the author alone is responsible.

References

Bungeroth, J. & Ney, H. (2004). "Statistical Sign Language Translation." Streiter, O. & Vettori, C. (eds) (2004). *Proceedings of the Workshop on Representation and Processing of Sign Languages*, 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal, May 2004, 105-108.

Kuhn, T.S. (1996/1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 3rd edition.

Morrissey, S. & Way, A. (2005). "An Example-Based Approach to Translating Sign Languages." Way, A. & Carl, M. (eds) (2005). *Proceedings of the "Workshop on Example-Based Machine Translation"*, MT-Summit X, Phuket, Thailand, September 2005, 109-116.

Prinsloo, D. & Heid, U. (this volume). "Creating Word Class Tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping", 97-115.

Sarasola, K. (2000). "Language Engineering Resources for Minority Languages" *Proceedings of the Workshop "Developing Language Resources for Minority Languages: Re-usability and Strategic Priorities."* Second International Conference on Language Resources and Evaluation, Athens, Greece, May 2000.

Scannell, K. (2003). "Automatic Thesaurus Generation for Minority Languages: an Irish Example." *Proceedings of the Workshop "Traitement automatique des langues minoritaires et des petites langues.* 10ème conférence TALN, Batz-sur-Mer, France, June 2003, 203-212.

Somers, H. (1998). "New paradigms." *MT: The State of Play now that the Dust has Settled. Proceedings of the "Workshop on Machine Translation"*, 10th European Summer School in Logic, Language and Information, Saarbrücken, August 1998, 22-33.

Streiter, O. & De Luca, E.W. (2003). "Example-based NLP for Minority Languages: Tasks, Resources and Tools." Streiter, O. (ed) (2003) *Proceedings of the Workshop*

“Traitement automatique des langues minoritaires et des petites langues”, 10ème conférence TALN, Batz-sur-Mer, France, June 2003, 233-242.

Un corpus per il sardo: problemi e prospettive

Nicoletta Puddu

Creating a corpus for minority languages has provided an interesting tool to both study and preserve these languages (see, for example, the DoBeS project at MPI Nijmegen). Sardinian, as an endangered language, could certainly profit from a well-designed corpus. The first digital collection of Sardinian texts was the Sardinian Text Database; however, it cannot be considered as a corpus: it is not normalized and the user can only search for exact matches. In this paper, I discuss the main problems in designing and developing a corpus for Sardinian.

Kennedy (1998) individuates three main stages in compiling a corpus: (1) corpus design; (2) text collection and capture; and, (3) text encoding or mark-up. As for the first stage, I propose that a Sardinian corpus should be mixed, monolingual, synchronic, balanced, and annotated, and I discuss the reasons for these choices throughout the paper. Text collection seems to be a minor problem in the case of Sardinian: both written and spoken texts are available and the number of speakers is still significant enough to collect a sufficient amount of data. The major problems arise at the third stage. Sardinian is fragmented into different varieties, and has not a standard variety (not even a standard orthography). Recently, several proposals for standardization have been made, but without success (see the discussion in Calaresu 2002; Puddu 2003). First of all, I suggest using a standard orthography that allows us to group Sardinian dialects into macro varieties. Then, it will be possible to articulate the corpus into sub-corpora that are representative of each variety. The creation of an adequate morphological tag system will be fundamental. As a matter of fact, with a homogeneous tag system, it will be possible to perform searches throughout the corpus and study linguistic phenomena both in the single macro variety and in the language as a whole.

Finally, I propose a morphological tag system and a first tagged pilot corpus of Sardinian based on written samples according to EAGLES and XCES standards.

1. Perché creare corpora per le lingue minoritarie

La *corpus linguistics* o linguistica dei corpora (da qui LC) risulta di particolare interesse, soprattutto per chi adotti un approccio funzionalista, in quanto “studia la

lingua nel modo in cui essa viene effettivamente utilizzata, da parlanti concreti in reali situazioni comunicative” (Spina 2001:53). L’utilizzo dei corpora, come noto, può essere molteplice: dagli studi sul lessico (creazione di lessici e dizionari di frequenza) a quelli sulla sintassi, fino alla didattica delle lingue e alla traduzione. Per le lingue standardizzate, l’utilizzo di corpora è in grande sviluppo. Tuttavia, anche per le lingue in pericolo di estinzione, la creazione di corpora si può rivelare particolarmente utile. Oltre alle motivazioni comuni alle lingue standardizzate, creare un corpus può essere un valido metodo per conservare la testimonianza della lingua, nella malaugurata ipotesi che essa si estingua. Se un corpus viene infatti definito come “una raccolta strutturata di testi in formato elettronico, che si assumono rappresentativi di una data lingua o di un suo sottoinsieme, mirata ad analisi di tipo linguistico” (Spina 2001:65), è evidente che esso può fungere anche da “specchio” di una lingua in un determinato stato. In questo senso il corpus può porsi come strumento complementare ad atlanti linguistici e indagini mirate, fotografando un campione rappresentativo della lingua.

La presenza di corpora facilita di molto l’analisi di fenomeni linguistici. La presenza di un corpus non elimina certamente i metodi tradizionali di raccolta dati, ma fornisce un valido strumento per testare la validità di una ipotesi anche per studiosi che non possano accedere direttamente ai parlanti. Inoltre, su un corpus è possibile compiere degli studi sulla frequenza, evidentemente molto difficili da realizzare con gli strumenti tradizionali.

Mostrata quindi l’utilità dei corpora anche per le lingue minoritarie, è necessario porre in evidenza le particolari questioni che la linguistica dei corpora si trova ad affrontare nel caso di lingue non standardizzate. In questo studio prenderemo ad esempio il caso del sardo, per evidenziare i possibili problemi (e le eventuali soluzioni da adottare).

Nel caso del sardo, come in molte altre lingue minoritarie in via di estinzione che si aprono solo ora alla linguistica dei corpora, ci troviamo davanti a due questioni fondamentali.

Da un lato è necessario creare quanto prima un corpus per cercare di preservare lo stato di lingua attuale. Nel caso del sardo, sottoposto a massiccia influenza da parte dell’italiano, alcune varietà rischiano una rapida estinzione e sarebbe quanto mai auspicabile raccogliere il prima possibile, con criteri omogenei, dati della lingua parlata che possano essere inseriti in un corpus.

Dall’altro lato, oltre che per la pianificazione del corpus e la raccolta dei dati, è necessario stabilire tutti gli standard di codifica e annotazione e ciò, come vedremo,

crea non pochi problemi nel caso di lingue non standardizzate e frammentarie come il sardo.

2. Un progetto sperimentale: il Sardinian Digital Corpus

Il sardo è, come ben noto, suddiviso in diverse varietà e non standardizzato, nonché in costante regresso. Esistono diverse grammatiche e dizionari e, su internet, è disponibile il *Sardinian Text Database* (<http://www.lingrom.fu-berlin.de/sardu/textos.html>), una raccolta di testi in sardo curata dall'Università di Colonia. Si tratta di un'interessante iniziativa, che però non risponde ai criteri di rappresentatività, campionamento e bilanciamento. I testi vengono infatti inseriti dai vari autori e non vi è uniformità nella codifica.

2.1 La pianificazione

Come noto, la fase di pianificazione è fondamentale in quanto, proprio in questa fase, si prendono decisioni che determinano la fisionomia del corpus e che, da un certo momento in poi, non possono più essere modificate.

Nel descrivere le fasi della progettazione del SDC, seguiremo la fondamentale tassonomia di Ball (anno:pag.).

Una prima distinzione è quella per mezzo. Nel progettare un corpus di una lingua in pericolo di estinzione, ovviamente sarebbe da privilegiare la presenza di campioni di lingua orale. Tuttavia, quando esista una tradizione scritta, la scelta potrebbe ricadere anche su una tipologia di corpus mista, in modo da avere una visione quanto più possibile globale della lingua. Nel caso del sardo, ad esempio, un corpus misto sarebbe possibile, in quanto esiste una certa produzione scritta sia "tradizionale" (romanzi, racconti, testi poetici, articoli di giornale), sia "elettronica" (mailing lists e siti in sardo).

Il SDC dovrebbe essere un corpus monolingue, ma rappresentativo delle diverse varietà. La creazione di corpora multilingui o paralleli potrebbe essere un passo successivo, particolarmente interessante sia dal punto di vista della linguistica comparativa che della glottodidattica.

Per le stesse ragioni enumerate sopra, il SDC si propone come un corpus sincronico, dato che abbiamo mostrato come sia urgente documentare lo stato di lingua attuale. Il confronto con stati di lingua passati e quindi la creazione di un corpus diacronico dovrà essere necessariamente successiva.

Il SDC, data la totale assenza di altri corpora per il sardo dovrebbe quindi essere un corpus di riferimento, basato rigorosamente sui due criteri di campionamento e bilanciamento.

Il corpus si propone, almeno in una fase iniziale, come un corpus aperto, continuamente aggiornabile con nuove acquisizioni, sempre e comunque coerenti con la pianificazione iniziale.

Infine, il corpus dovrà essere annotato. A questo proposito, proponiamo qui anche una prima ipotesi di annotazione del SDC secondo gli standard internazionali.

2.2 *Acquisizione dei dati*

Per quanto riguarda la raccolta dati, non vi sono particolari differenze rispetto a lingue ufficiali e standardizzate. Bisogna pertanto adottare gli accorgimenti tipici della ricerca sul campo.

Problemi ben più importanti si pongono invece per quanto riguarda la codifica dei dati. Bisogna infatti arrivare a una normalizzazione grafica che evidentemente comporta, per i testi non standardizzati, una scelta da parte di chi codifica.

Il sardo rappresenta, sotto questo punto di vista, un caso emblematico. Le differenze tra le diverse varietà sono, infatti, numerose, soprattutto sul piano fonetico. In quale varietà devono essere inseriti i testi nel corpus? E sino a che punto è possibile ridurre le diverse varietà del sardo a una unica macrovarietà?

La questione della standardizzazione della lingua sarda è stata oggetto, negli ultimi anni, di una robusta polemica. Nel 2001 l'Assessorato alla Pubblica Istruzione della Regione Sardegna ha pubblicato una prima proposta di standardizzazione denominata *Limba sarda unificada* (LSU). Tale proposta è stata elaborata da un'apposita commissione e si tratta in sostanza di una varietà che, per ammissione della stessa commissione, per quanto si ponga come obiettivo la mediazione tra le diverse varietà presenti nell'isola, è "rappresentativa di quelle varietà più vicine alle origini storico-evolutive della lingua sarda" (LSU: 5). Di fatto, i tratti scelti per la LSU sono perlopiù logudoresi e sono stati percepiti dai parlanti come tratti "locali" piuttosto che conservativi. Ciò ha portato a una netta opposizione allo standard soprattutto da parte dei parlanti campidanesi: le motivazioni di questa reazione sono analizzate dal punto di vista sociolinguistico in Calaresu (2002) e Puddu (2003, 2005).

Di recente, la Regione Sardegna ha incaricato una nuova commissione di elaborare una lingua standard per usi solo burocratici-amministrativi e, nel contempo, di creare

delle norme ortografiche sarde “per tutte le varietà linguistiche in uso nel territorio regionale”¹.

La soluzione migliore, a mio parere, è quindi di inserire i testi nelle diverse macrovarietà del sardo con una standardizzazione solo ortografica in base alle proposte della commissione. In sostanza, si tratterebbe di operare solo una normalizzazione grafica sui vari testi riconducendoli a macrovarietà e annotando le eventuali differenziazioni fonetiche a parte. Facciamo un esempio: la nasale intervocalica originaria del latino subisce, nelle diverse varietà del sardo campidanese, trattamenti differenti. In alcuni casi viene mantenuta, in alcuni viene raddoppiata, in altri è ridotta a nasalizzazione della vocale precedente, mentre in altri ancora è sostituita dal colpo di glottide. Pertanto, la parola per ‘luna’, può essere pronunciata come [‘luna], [‘lun: a] [‘luâa] o ancora [‘lu/a]. La mia proposta è quindi di trascrivere la forma originaria *luna*, annotando poi la eventuale trascrizione fonetica in un file separato.

Vi è inoltre il problema di inserire nel corpus anche alcune varietà come il gallurese e il sassarese che non sono unanimemente riconosciute come dialetti del sardo. Si potrebbe però operare una distinzione tra ‘nuclear Sardinian’, costituito da campidanese e logudorese e ‘core Sardinian’, con il gallurese e sassarese. La struttura del SDC dovrebbe, in definitiva, essere rappresentata come nella figura 1 in appendice.

2.3 Codifica primaria

La nostra proposta è, come detto, che il SDC sia annotato. Il corpus sarà codificato usando XML e lo standard sarà quello stabilito da XCES (Corpus Encoding Standard for XML). Le linee guida CES raccomandano in particolar modo che l’annotazione sia di tipo stand-off, vale a dire che l’annotazione non sia unita al testo base, ma sia contenuta in altri files XML a esso collegati. Nel caso del SDC l’annotazione stand-off è particolarmente importante, dato che è pensato come un corpus aperto.

- Nella codifica dei corpora, le guidelines CES riconoscono tre principali categorie di informazioni rilevanti: la documentazione, che contiene informazioni globali sul testo, il suo contenuto e la sua codifica; i dati primari, che comprendono sia la “gross structure”, ovvero sia il testo al quale vengono aggiunte informazioni sulla struttura generale (paragrafi, capitoli, titoli, note a piè di pagina, tabelle, figure ecc.) sia elementi che compaiono al livello del sottoparagrafo;

1 “La Commissione dovrà inoltre definire norme ortografiche comuni per tutte le varietà linguistiche in uso nel territorio regionale. In questo modo sarà possibile promuovere la creazione di word processor, correttori ortografici, oltre all’utilizzo e alla diffusione di strumenti elettronici per favorire l’uso corretto della lingua sarda” (<http://www.regionesardegn.it/j/v/25?s=3661&v=2&c=220&t=1>).

-
- l'annotazione linguistica, che può essere di tipo morfologico, sintattico, prosodico ecc.

A titolo esemplificativo, mostrerò di seguito come possa essere codificato un testo preso da un articolo di giornale. Supponiamo di dover codificare il testo seguente pubblicato su un giornale locale.

Su Casteddu hat giogau ariseru in su stadiu Sant'Elia bincendi po quattru a zero contras a sa Juventus. Is festegiamentus funti sighius fintzas a mengianeddu. Cust'annu parit chi ddoi siant bonas possibilidadis de binci su scudetu.

'Il Cagliari ha giocato ieri allo stadio Sant'Elia vincendo per quattro a zero contro la Juventus. I festeggiamenti sono continuati sino all'alba. Quest'anno pare che ci siano buone possibilità di vincere lo scudetto'.

Le informazioni relative alla documentazione saranno contenute in una *header* conforme alle *guidelines* CE. La header minima per questo documento sarà la seguente.

sdc_header.xml

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<cesHeader xmlns="http://www.xml-ces.org/schema" xmlns:xsi="http://www.
w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.xml-ces.org/
schema http://www.cs.vassar.edu/XCES/schema/xcesHeader.xsd" version="1.0">
  <fileDesc>
    <titleStmt>
      <h.title>Sardinian Digital Corpus, demo</h.title>
    </titleStmt>
    <publicationStmt>
      <distributor>University of...</distributor>
      <pubAddress>via...</pubAddress>
      <availability>Free</availability>
      <pubDate>2005</pubDate>
    </publicationStmt>
    <sourceDesc>
    <biblStruct>
```

```

    <monogr>
      <h.title>La Gazzetta di Sardegna</h.title>
      <edition>...</edition>
      <imprint>...</imprint>
      <publisher>...</publisher>
    </monogr>
    <analytic>
      <h.title>Su scudetu in Casteddu</h.title>
      <h.author>Porru</h.author>
    </analytic>
  </biblStruct>
</sourceDesc>
</fileDesc>
<encodingDesc>
  <projectdesc> Il Sardinian Digital Corpus...</projectdesc>
  <editorialDecl Default="n">
    <conformance level="1">CES Level 1</conformance>
    <correction status="medium" method="silent" Default="n" />
    <quotation marks="none" form="std" Default="n">Rendition attribute
      values on Q and QUOTE tags are adapted from ISOpub and ISONum standard
      entity set names</quotation>
    <segmentation Default="n">Marked up to the level of sentence
  </segmentation>
  </editorialDecl>
  <tagsDecl>
    <tagUsage gi="body" occurs="1" />
    <tagUsage gi="p" occurs="2" />
    <tagUsage gi="s" occurs="3" />
  </tagsDecl>
</encodingDesc>
<profileDesc>
  <langUsage>
    <language id="sro" SIL="sro">Campidanese Sardinian</language>

```

```
</langUsage>
</profileDesc>
</cesHeader>
```

La *header* sarà esterna al *CesDoc*. Le varie *headers* saranno salvate in uno *headerbase* esterno al documento e vi si farà riferimento attraverso un'espressione Xpointer nel *CesDoc*. Il *CesDoc*, che contiene il secondo tipo di informazioni si presenterà come segue:

sdc_demo.xml

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<!DOCTYPE cesDoc SYSTEM "http://www.cs.vassar.edu/XCES/dtd/xcesDoc.dtd">
<cesDoc xmlns="http://www.xml-ces.org/schema" xmlns:xlink="http://www.
w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.xml-ces.org/schema http://www.cs.vassar.edu/
XCES/schema/xcesDoc.xsd">
  <xcesHeader xlink:href="sdc_header.xml" />
  <text>
    <body>
      <div type="article" xml:lang="sro">
        <p id="p1"><s id="p1s1">Su Casteddu hat giogau ariseru in su stadium
        Sant&apos; Elia bincendi po quattru a zero contras a sa Juventus.</s>
        <s id="p1s2">Is festegiamentus funti sighius fintzas
        mengianeddu.</s></p>
        <p id="p2"><s id="p2s1">Cust&apos; annu parit chi ddoi siant bonas
        ssibilidadis de binci su scudetu.</s></p>
      </div>
    </body>
  </text>
</cesDoc>
```

2.4 Annotazione

Come detto, il SDC si propone come un corpus annotato. Al XCESDoc faranno quindi riferimento, tramite Xpointer, le annotazioni ai vari livelli. Il più diffuso tipo di annotazione è quello per parti del discorso. Nel caso del sardo un corpus annotato per parti del discorso potrebbe rivelarsi particolarmente utile per la ricerca linguistica, soprattutto nella creazione di una grammatica descrittiva *corpus-based*. L'annotazione per parti del discorso sarà logicamente la prima in ordine di tempo, ma, dato il carattere aperto del corpus, potranno in seguito essere aggiunte annotazioni su altri livelli.

2.5 Il tagset

Nel caso del sardo è necessario creare un apposito tagset, che può essere in parte mutuato dai *tagset* per l'italiano e lo spagnolo creati all'interno del progetto MULTEXT secondo le raccomandazioni EAGLES. Le diverse varietà del sardo non differiscono particolarmente dal punto di vista morfosintattico: questo significa che è possibile definire un unico *tagset* per tutte le varietà. Gli esempi in questo articolo sono in campidanese, ma le etichette si potranno applicare praticamente senza variazioni anche alle altre varietà del sardo.

L'annotazione grammaticale del nostro corpus, compatibile con la CesAna DTD, consisterà di tre livelli:

- la forma base (<base>);
- una descrizione morfosintattica secondo le linee guida EAGLES (<msd>);
- un corpus tag (<ctag>).

In accordo con quanto proposto da EAGLES, abbiamo una descrizione a due livelli:

- la prima, a grana più fine, contiene la descrizione quanto più accurata possibile del token (descrizione lessicale <msd>);
- la seconda invece, "a grana più grossa", è una versione sottodeterminata della prima descrizione (corpus tag <ctag>).

La distinzione a due livelli si mostra particolarmente utile quando si voglia utilizzare un sistema di etichettatura automatica. Alcune categorie sono infatti piuttosto difficili da disambiguare automaticamente ed è pertanto opportuno avere un sistema di etichettatura a grana più grossa. Nel caso del sardo, la creazione o l'implementazione di tagger automatici può essere un passo successivo, ma mi sembra utile, sin d'ora, definire un sistema di etichettatura adatto anche per un futuro utilizzo automatico.

- Il tag <msd> si compone di una stringa di caratteri strutturata nel modo seguente: in posizione 0 il simbolo che codifica la parte del discorso;

-
- nelle posizioni da 1 a n i valori degli attributi relativi a persona, genere, numero, caso ecc.;
 - se un attributo non si può applicare è sostituito da un trattino.

Il tagset qui proposto è simile, come detto, a quello proposto per l'italiano e lo spagnolo all'interno del progetto MULTTEXT (Calzolari & Monachini 1996). Manteniamo innanzi tutto la classificazione in parti del discorso proposta all'interno di MULTTEXT (tab.1 in appendice).

Analizziamo quindi in breve le etichette create per il sardo. Ci soffermeremo solo nel caso in cui vi siano differenze notevoli tra il sardo e le altre due lingue romanze, e nei casi in cui siano state operate scelte differenti.

Nome

Per quanto riguarda la categoria "nome", i tratti presi in considerazione sono esemplificati nella tabella 2 e corrispondono a quelli considerati per l'italiano in Calzolari e Monachesi. La tabella 3 mostra le possibili combinazioni e la traduzione in *ctags*.

Verbo

Nella scelta dei valori del verbo, vi sono alcune modifiche sia rispetto all'italiano che allo spagnolo (tab. 4). Per quanto riguarda il modo, non è inserito tra i codici il condizionale. In sardo, infatti, esso è una forma perifrastica formata tramite un verbo ausiliare (*camp. hai 'avere', log. depi 'dovere'*) più l'infinito del verbo. Pertanto, in conformità con quanto fatto nelle altre lingue per le forme perifrastiche, le due forme saranno etichettate autonomamente.

Il medesimo discorso vale per il futuro, formato nelle diverse varietà dal verbo 'avere' coniugato, la preposizione *a* e l'infinito del verbo.

Il sardo non possiede inoltre forme di passato remoto, ma il passato non durativo è espresso dalla forma perifrastica formata da 'avere' più il participio passato del verbo.

La situazione è piuttosto complessa per quanto riguarda i clitici. Mentre nel caso dell'italiano è previsto un unico codice E per tutti i tipi di clitico, nel caso dello spagnolo ogni tipo di clitico e le possibili combinazioni vengono specificati con diversi codici anche nel *ctag*. Lo spagnolo non presenta però i cosiddetti "clitici avverbiali" (italiano *ne* e *ci*), che in sardo si possono aggiungere al verbo e combinarsi con altri

clitici. Abbiamo quindi mantenuto la convenzione dello spagnolo, aggiungendo però i codici per i clitici avverbiali.

La tabella 5 mostra tutte le possibili combinazioni. Si noti che in sardo non esiste il participio presente e che è possibile aggiungere forme clitiche del pronome solo al gerundio e all'imperativo.

Aggettivo

L'aggettivo in sardo (tabb. 6 e 7) non presenta particolari differenze rispetto all'italiano e allo spagnolo. Il comparativo è normalmente formato con *prus* 'più' seguito dall'aggettivo.

Pronome

Per quanto riguarda i pronomi, in analogia con quanto fatto per lo spagnolo, è stato preso in considerazione anche l'attributo caso (tabb. 8 e 9).

Determinante

Cfr. tabb. 10 e 11 in appendice.

Articolo

Cfr. tabb. 12 e 13 in appendice.

Avverbio

Cfr. tabb. 14 e 15 in appendice.

Determinante

Cfr. tabb. 10 e 11 in appendice.

Articolo

Cfr. tabb. 12 e 13 in appendice.

Avverbio

Cfr. tabb. 14 e 15 in appendice.

Punteggiatura

Cfr. tabella 24 in appendice.

2.4 L'esempio annotato

A questo punto siamo in grado di annotare l'esempio secondo le convenzioni XCES. Per questioni di brevità forniamo l'annotazione solo di una parte del nostro testo.

sdc_annotation.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE cesAna SYSTEM "http://www.cs.vassar.edu/XCES/dtd/xcesAna.dtd">
  <cesAna xmlns="http://www.xml-ces.org/schema"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.xml-ces.org/schema
      http://www.cs.vassar.edu/XCES/schema/xcesAna.xsd"
      version="1.0">
<cesHeader version "2.3">
</cesHeader>
<chunklist xml:base="sdc_demo.xml">
  <chunk type="sentence" xml:base="#p1s1">
    <tok xlink:href="xpointer(string-range('', 0, 2))">
      <orth>su</orth>
      <lex>
        <base>su</base>
        <msd>Tdms-</msd>
        <ctag>RMS</ctag>
      </lex>
    </tok>
    <tok xlink:href="xpointer(string-range('', 3, 10))">
      <orth>Casteddu</orth>
      <lex>
        <base>Casteddu</base>
        <msd>Np..-</msd>
```

```
<ctag>NP</ctag>
</lex>
</tok>
<tok xlink:href="xpointer(string-range('', 11, 13))">
  <orth>hat</orth>
  <lex>
    <base>hai</base>
    <msd>Vaip3s-</msd>
    <ctag>VAS3IP</ctag>
  </lex>
</tok>

<tok xlink:href="xpointer(string-range('', 14, 20))">
  <orth>giogau</orth>
  <lex>
    <base>giogai</base>
    <msd>Vmp--sm-</msd>
    <ctag>VMPSM</ctag>
  </lex>
</tok>

<tok xlink:href="xpointer(string-range('', 20, 27))">
  <orth>ariseru</orth>
  <lex>
    <base>ariseru</base>
    <msd>R-p-</msd>
    <ctag>B</ctag>
  </lex>
</tok>

<tok xlink:href="xpointer(string-range('', 28, 30))">
  <orth>in</orth>
  <lex>
    <base>in</base>
    <msd>Sp</msd>
```

```
<ctag>SP</ctag>
</lex>
</tok>
<tok xlink:href="xpointer(string-range('', 31, 33))">
  <orth>su</orth>
  <lex>
    <base>su</base>
    <msd>Sp-</msd>
    <ctag>SP</ctag>
  </lex>
</tok>
<tok xlink:href="xpointer(string-range('', 34, 39))">
  <orth>stadiu</orth>
  <lex>
    <base>stadiu</base>
    <msd>Ncms-</msd>
    <ctag>NMS</ctag>
  </lex>
</tok>
<tok xlink:href="xpointer(string-range('', 40, 44))">
  <orth>Sant&apos</orth>
  <lex>
    <base>Santu</base>
    <msd>A-pms-</msd>
    <ctag>AMS</ctag>
  </lex>
</tok>
<tok xlink:href="xpointer(string-range('', 45, 48))">
  <orth>Elia</orth>
  <lex>
    <base>Elia</base>
    <msd>Np..-</msd>
    <ctag>NP</ctag>
```

```
</lex>
</tok>
<tok xlink:href="xpointer(string-range('',...))">
</tok>
</chunk>
<chunk type="sentence" xml:base="#p1s1">
  <tok xlink:href="xpointer(string-range('', 0, 2))">
    <orth>is</orth>
    <lex>
      <base>is</base>
      <msd>Tdmp-</msd>
      <ctag>RMP</ctag>
    </lex>
  </tok>
  <tok xlink:href="xpointer(string-range('', 3, 16))">
    <orth>festegiamentus</orth>
    <lex>
      <base>festegiamentu</base>
      <msd>Ncmp-</msd>
      <ctag>NMP</ctag>
    </lex>
  </tok>
</chunk>
.....
</chunklist>
</cesAna>
```

3. Conclusioni

In questo articolo abbiamo mostrato, attraverso un *case study*, le problematiche relative all'applicazione della linguistica dei corpora a lingue minoritarie e non standardizzate. Mi pare quindi che si possano trarre alcune considerazioni più generali applicabili a gran parte delle lingue minoritarie.

Innanzitutto, il lavoro da fare, in casi come quello del sardo, è a più livelli dalla progettazione del corpus, alla raccolta, all'etichettatura dei dati. L'imponenza dell'opera si accompagna alla necessità di accelerare i tempi nel caso di varietà particolarmente a rischio di estinzione.

In secondo luogo, in caso di varietà non standardizzate, si pone, come abbiamo visto, il problema della codifica dei dati. La scelta qui operata, vale a dire di un corpus articolato in sotto-macro-varietà, mi sembra possa essere un buon compromesso tra la necessità di standardizzazione da un lato e il mantenimento delle differenziazioni dall'altro. Ciò che è assolutamente necessario è invece il raggiungimento di una standardizzazione dal punto di vista ortografico.

Infine, l'annotazione per parti del discorso può essere particolarmente utile per la creazione di grammatiche basate sull'uso. Nel caso del sardo, ad esempio, la presenza di una lingua dominante come l'italiano, con una ricca tradizione letteraria e grammaticale, può influenzare i giudizi di grammaticalità dei parlanti, inficiando in alcuni casi i dati raccolti.

Mi pare quindi che, in base a tutte queste riflessioni, la progettazione e lo sviluppo di corpora per le lingue minoritarie debbano assumere un ruolo prioritario in progetti di salvaguardia e politica linguistica.

Ringraziamenti

Ringrazio Andrea Sansò per aver letto con attenzione il manoscritto.

Bibliografia

Ball, C. (1994). *Concordances and corpora for classroom and research*. Online at <http://www.georgetown.edu/cball/corpora/tutorial.html>.

Bel, N. & Aguilar A. (1994). *Proposal for Morphosyntactic encoding: Application to Spanish*, Barcelona.

Blasco Ferrer, E. (1986). *La lingua sarda contemporanea*. Cagliari: Edizioni della Torre.

Calaresu, E. (2002). "Alcune riflessioni sulla LSU (Limba Sarda Unificada)." Orioles, V. (a cura di), *La legislazione nazionale sulle minoranze linguistiche. Problemi, applicazioni, prospettive*. Udine: Forum, 247-266.

Calzolari, N. & Monachini, M. (1996). *Multext. Common Specification and notations for Lexicon Encoding*, Pisa: Istituto di Linguistica Computazionale.

EAGLES (1996) *Recommendations for the morphosyntactic annotation of corpora. EAG-TCWG-MAC/R*, Pisa: Istituto di Linguistica Computazionale.

Ide, N. (1998) "Corpus Encoding Standard. SGML Guidelines for Encoding Linguistic Corpora." *Proceedings of the First International Language Resources and Evaluation Conference*, Paris: European Language Resources Association, 463-70.

Ide, N., Bonhomme, P. & Romary, L. (2000). "XCES: An XML-based Standard for Linguistic Corpora." *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, 825-30.

Ide, N. (2004). "Preparation and Analysis of Linguistic Corpora." Schreibman, S., Siemens, R. & Unsworth, J. (a cura di), *A Companion to Digital Humanities*. London: Blackwell.

Kennedy, G. (1998). *An introduction to Corpus Linguistics*. London: Longman.

Leech, G. & Wilson, A. (1996). *EAGLES recommendations for the morphosyntactic annotation of corpora*. Pisa: Istituto di Linguistica Computazionale.

Regione Autonoma della Sardegna (2001). "*Limba sarda unificada. Sintesi delle norme di base: ortografia, fonetica, morfologia e lessico*", Cagliari.

McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mensching, G. & Grimaldi, L. (2000). *Sardinian Text Database*, <http://www.lingrom.fu-berlin.de/sardu/textos.html>.

Puddu, N. (2003). "In search of the "real Sardinian": truth and representation." Brincat, J., Boeder, W., Stolz, T. (a cura di), *Purism in minor languages, endangered languages, regional languages, mixed languages*. Bochum: Universitätsverlag Dr. N. Brockmeyer, 27-42.

Puddu, N. (2005). "La nozione di purismo nel progetto di standardizzazione della lingua sarda." Carli, A., Calaresu, E. & Guardiano, C. (a cura di), *Lingue, istituzioni, territori. Riflessioni teoriche, proposte metodologiche ed esperienze di politica linguistica*. Bulzoni: Roma, 257-278.

Spina, S. (2001). *Fare i conti con le parole*. Perugia: Guerra Edizioni.

Una commissione tecnico-scientifica per un'indagine socio-linguistica sullo stato della lingua sarda. Online at www.regione.sardegna.it.

Figura 1: La struttura del Sardinian Digital Corpus

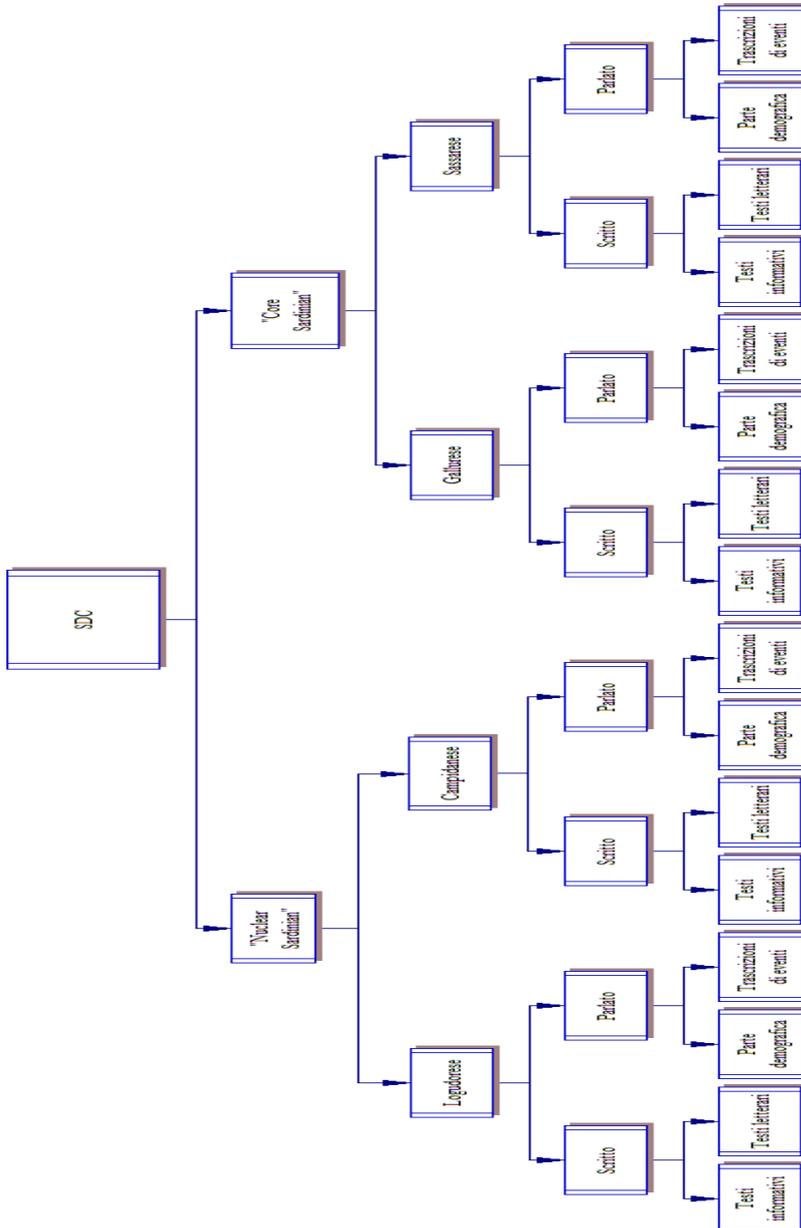


Tabella 1: Codici per le parti del discorso

Parte del discorso	Codice
Nome	N
Verbo	V
Aggettivo	A
Pronome	P
Determinante	D
Articolo	T
Avverbio	R
Apposizione	S
Congiunzioni	C
Numerali	M
Interiezione	I
Unico	U
Residuale	X
Abbreviazione	Y

Tabella 2: Coppie attributo-valore per la categoria "Nome" in sardo

Attributo	Valore	Esempio	Codice
Tipo	comune	libru	c
	proprio	Giuanni	p
Genere	maschile	omini	m
	femminile	femina	f
	comune	meri	c
Numero	singolare	omini	s
	plurale	feminas	p
Caso	///	///	///

Tabella 3: <msd> e <ctag> per la categoria "Nome" in sardo

msd	ctag	esempio
Ncms-	NMS	liburu
Ncmp-	NMP	liburus
Ncmn-	NN	lunis (su/is)
Ncfs-	NFS	domu
Ncfp-	NFP	domus
Nccs-	NNS	meri (su/sa)
Nccp-	NNP	meris (is f.m.),
Np-	NP	Mariu, Maria, Puddu

Tabella 4: Coppie attributo-valore per la categoria "Verbo" in sardo

Attributo	Valore	Esempio	Codice
Status	lessicale	papai	m
	ausiliare	hai/essi	a
	modale	podì	o
Modo	indicativo	papat	l
	congiuntivo	papit	s
	imperative	papa	m
	infinito	papai	n
	participio	papau	p
	gerundio	papendi	g
	Tempo	presente	papu
	imperfetto	papasta	i
persona	prima	seu	1
	seconda	ses	2
	terza	est	3
numero	singolare	papat	s
	plurale	papant	p
genere	maschile	papau	m
	femminile	papada	f
clitico	accusativo	donaddu	a
	dativo	donaddi	d
	avverbiale	donandi	
beninci	r		
	acc+dat	donasiddu	t
	dat+avv	donasindi	u
	avv+dat	donandeddi	v
	dat+avv+acc	mandasinceddu	z

Tabella 5: <msd> e <ctag> per la categoria "Verbo" in sardo

msd	ctag	esempio
Vaip1s-	VAS1IP	hapu/seu
Vaip2s-	VAS2IP	has/ses
Vaip3s-	VAS3IP	hat/est
Vaip1p-	VAP1ICP	eus/seus
Vaip2p-	VAP2IP	eus/seis
Vaip3p-	VAP3IP	hant/funt
Vaii1s-	VAS1II	hia, femu
Vaii2s-	VAS2II	hiast, fiast

Vaii3s-	VAS3II	hiat, fiat
Vaii1p-	VAP1II	emus, femus
Vaii2p-	VAP2II	eis, festis
Vaii3p-	VAP3II	iant, fiant
Vasp1s-	VASXCP	apa, sia
Vasp2s-	VASXCP	apas, sias
Vasp3s-	VASXCP	apas, siat
Vasp1p-	VAP1ICP	apaus, siaus
Vasp2p-	VAP2CMP	apais, siais
Vasp3p-	VAP3CP	apant, siant
Vasi1s-	VAS3CI	hemu, fessi
Vasi2s-	VAS3CI	essist, fessis
Vasi3s-	VAS3CI	essit, fessit
Vasi1p-	VAP1CI	essimus, festus
Vasi2s-	VAP2ICR	essidis, festis
Vasi3p-	VAP3CI	essint, fessint
Vanp---	VAF	hai, essi
Vaps-sm	VAMSPR	apiu, stetiu
Vaps-pm	VAMPPR	stetius
Vaps-sf	VAFSPR	stetia
Vaps-pf	VAFPPR	stetias
Vmip1s-	VMIP1S	papu
Vmip2s-	VMIP2S	papas
Vmip3s-	VMIP3S	papat
Vmip1p-	VMIP1P	papaus
Vmip2p-	VMIP2P	papais
Vmip3p-	VMIP3P	papant
Vmsp1s-	VMSP1S	papi
Vmsp2s-	VMSP2S	papis
Vmsp3s-	VMSP3S	papi
Vmsp1p-	VMSP1P	papeus
Vmsp2p-	VMSP2P	papeis
Vmsp3p-	VMSP3P	papint
Vmii1s-	VMII1S	papemu
Vmii2s-	VMII2S	papást
Vmii3s-	VMII3S	papát
Vmii1p-	VMII1P	papemus
Vmii2p-	VMII2P	papestis
Vmii3p-	VMII3P	papánt

Vmsi1s-	VMSI1S	tenessi
Vmsi2s-	VMSI2S	tenessis
Vmsi3s-	VMSI3S	tenessit
Vmis1s-	VMSI1S	tenessimus
Vmsi2p-	VMSI2P	tenestis
Vmsi3p-	VMSI3P	tenessint
Vmp--pf-	VMPPF	tentas
Vmp--sf-	VMPSF	tenta
Vmp--pm-	VMPPM	tentus
Vmp--sm-	VMPSM	tentu
Vmg----t	VMGT	tzerrienimiddas
Vmg----t	VMGT	tzerriandimiddas
Vmg----t	VMGT	tzerriendimidda
Vmg----t	VMGT	tzerriendimiddus
Vmg----t	VMGT	tzerriendimiddu
Vmg----d	VMGD	tzerriendimì
Vmg----t	VMGT	tzerriendididdas
Vmg----t	VMGT	tzerriendididdas
Vmg----t	VMGT	tzerriendididda
Vmg----t	VMGT	tzerriendididdus
Vmg----t	VMGT	tzerriendididdu
Vmg----d	VMGD	tzerriendidì
Vmg----t	VMGT	tzerriendisiddas
Vmg----t	VMGT	tzerriendisidda
Vmg----t	VMGT	tzerriendisiddus
Vmg----t	VMGT	tzerriendisiddu
Vmg----d	VMGD	tzerriendisì
Vmg----t	VMGT	tzerreindisiddas
Vmg----t	VMGT	tzerriendisidda
Vmg----t	VMGT	tzerriendisiddus
Vmg----t	VMGT	tzerriendisiddu
Vmg----d	VMGD	tzerriendisì
Vmg----t	VMGT	tzerriendisiddas
Vmg----t	VMGT	tzerriendisidda
Vmg----t	VMGT	tzerriendisiddus
Vmg----t	VMGT	tzerriendisiddu
Vmg----d	VMGD	tzerriendisì
Vmg----a	VMGA	tzerriendimì
Vmg----a	VMGA	tzerrienditi
Vmg----a	VMGA	tzerriendiddas

Vmg----a	VMGA	tzerriendidda
Vmg----a	VMGA	tzerriendiddu
Vmg----a	VMGA	tzerriendiddu
Vmg----a	VMGA	tzerriendisì
Vmg----a	VMGA	tzerriendisì
Vmg----u	VMGU	mandendisindi
Vmg----z	VMGZ	mandendisinceddu
Vmg-----	VMG	tzerriendi
Vmmp2sa	VMM2SA	mandaddu
Vmmp2sd	VMM2SD	mandadì
Vmmp2st	VMM2ST	mandadiddu
Vmmp2su	VMM2SU	mandadindi
Vmmp2sv	VMM2SV	mandandeddi
Vmmp2sz	VMM2SZ	mandasinceddu
Vmmp2pa	VMM2PA	mandaiddu
Vmmp2pd	VMM2PD	mandaisì
Vmmp2pt	VMM2PT	mandaisiddu
Vmmp2pu	VMM2PU	mandaisndi
Vmmp2pz	VMM2PZ	mandaisinceddu
Vmmp2s-	VMM2S	manda
Vmmp2p-	VMM2P	mandai
Vmmp2pv	VMM2PV	mandaineddi

Tabella 6: Coppie attributo-valore per la categoria "Aggettivo" in sardo

Attributo	Valore	Esempio	Codice
Tipo	//	//	//
Grado	positivo	bonu	p
	comparativo	mellus	c
	superlativo	mellus	s
Genere	maschile	bonu	m
	femminile	bona	f
	l-spec	druci	c
Numero	singolare	bonu	s
	plurale	bonus	p
Caso	//	//	//

Tabella 7: <msd> e <ctag> per la categoria "Aggettivo" in sardo

msd	ctag	esempio
A-pms-	AMS	bonu
A-pmp-	AMP	bonus
A-pfs-	AFS	bella
A-pfp-	AFP	bellas
A-pcs-	ANS	drucci
A-pcp-	ANP	drucis
A-sms-	AMS	bellissimu
A-smp-	AMP	bellissimus
A-sfs-	AFS	bellissima
A-sfp-	AFP	bellissimas
A-sfs-	AFS	bellissima
A-sfp-	AFP	bellissimas

Tabella 8: Coppie attributo-valore per la categoria "Pronome" in sardo

Attributo	Valore	Esempio	Codice
Tipo	personale	deu	p
	dimostrativo	cuddu	d
	indefinito	calincunu	i
	possessivo	miu	m
	interrogativo	chini	t
	relativo	chi	r
	esclamativo	cantu !	e
	riflessivo	si	x
Persona	prima	deu	1
	seconda	tui	2
	terza	issu	3
Genere	maschile	issu	m
	femminile	issa	f
L-spec	comune	deu	c
Numero	singolare	custu	s
	plurale	custus	p
L-spec	invariante	chini	n
Caso	nominativo	deu	n
	dativo	ddi	d
	accusativo	ddu	a
	obliquo	mei	o

Tabella 9: <msd> e <ctag> per la categoria "Pronome" in sardo

msd	ctag	esempio
Pd-ms--	PDMS	cussu
Pd-mp--	PDMP	cuddus
Pd-fs--	PDFS	cudda
Pd-fp--	PDFP	cuddas
Pi-ms--	PIMS	dognunu
Pi-mp--	PIMP	calincunus
Pi-fs--	PIFS	dognuna
Pi-fp--	PIFP	calincunas
Pi-cs--	PINS	chinechisiat
Ps1ms--	PPMS	miu, nostru
Ps1mp--	PPMP	mius
Ps1fs--	PPFS	mia
Ps1fp--	PPFP	mias
Ps2ms--	PPMS	tuu
Ps2mp--	PPMP	tuus
Ps2fs--	PPFS	tua
Ps2fp--	PPFP	tuas
Ps3ms--	PPMS	suu
Ps3mp--	PPMP	suus
Ps3fs--	PPFS	sua
Ps3fp--	PPFP	suas
Ps3cp--	PPNP	insoru
Pt-cs--	PWNS	chini?
Pt-cn--	PWNN	ita?
Pt-cs--	PWMS	cantu?
Pt-cp--	PWMP	cantus?
Pr-cs--	PWMS	cantu
Pr-cp--	PWNP	cantus
Pr-cs--	PWNS	chini
Pr-cp--	PWNP	calis
Pe-cs--	PWNS	cantu!
Pe-cp--	PWNP	cactus!
Pe-cn--	PWNN	ita!
Pp1csn-	PP1SN	deu
Pp2cs-n	PP2SN	tui
Pp3ms[no]	PP3MS	issu
Pp3fs[no]	PP3FS	issa
Pp1cp[no]	PP1PN	nosus
Pp2cp[no]	PP2PN	bosatrus
Pp3mp[no]	PP3MP	issus
Pp3fp[no]	PP3FP	issas
Pp1cso-	PP1SO	mei
Pp2-so-	PP2SO	ti
P[px]1cs[ad]-	P1S	mi
P[px]2cs[ad]-	P2S	ti

P[px]3cs[ad]-	P3	si
Pp3.pd-	PP3PD	ddis
Pp3.sd-	PP3SD	ddi
Pp3fpa-	PP3FPA	ddas
Pp3fsa-	PP3FSA	dda
Pp3mpa-	PP3MPA	ddus
Pp3msa-	PP3MSA	ddu
P[px]1cp[ad]-	P1P	si
P[px]2cp[ad]-	P2P	si
P..fp---	PFP	mias, custas,
P..fs---	PFS	mia, custa, canta etc.
P..mp---	PMP	mius, custus, cantas etc.
P..ms---	PMS	miu, custu, cantu etc.

Tabella 10: Coppie attributo-valore per la categoria "Determinante" in sardo

Attributo	Valore	Esempio	Codice
Tipo	dimostrativo	cuddu	d
	indefinito	dogna	i
	possessivo	miu	m
	interrogativo	chini	t
	relativo	chi	r
	esclamativo	cantu !	e
Persona	prima	mia	1
	seconda	tua	2
	terza	sua	3
Genere	maschile	custu	m
	femminile	custa	f
L-spec	comune	dogna	c
Numero	singolare	custu	s
	plurale	custus	p
L-spec	invariante	chini	n
	nominativo	deu	n
Possessore	singolare	miu	s
	plurale	nostru	p

Tabella 11: <msd> e <ctag> per la categoria "Determinante" in sardo

msd	ctag	esempio
Dd-ms--	DDMS	cuddu
Dd-mp--	DDMP	cuddus
Dd-fs--	DDFS	cudda
Dd-fp--	DDFP	cuddas
Di-ms--	DIMS	nisciunu
Di-mp--	DIMP	unus cantu
Di-fs--	DIFS	nisciunas
Di-fp--	DIFP	unas cantu
Di-cs--	DINS	chinechisiat
Di-cc--	DINC	dogni
Ds1ms--	DPMS	miu, nostru
Ds1mp--	DPMP	mius
Ds1fs--	DPFS	mia
Ds1fp--	DPFP	mias
Ds2ms--	DPMS	tuu, vostru
Ds2mp--	DPMP	tuus
Ds2fs--	DPFS	tua
Ds2fp--	DPFP	tuas
Ds3ms--	DPMS	suu
Ds3mp--	DPMP	suus
Ds3fs--	DPFS	sua
Ds3fp--	DPFP	suas
Ds3cp--	DPNP	insoru
Dr-cs--	DWNS	cantu
Dr-cp--	DWNP	cantus
Dt-cn--	DWNN	cali
Dt-cs--	DWNS	cantu
Dt-cp--	DWNP	cantus
De-cs--	DWMS	cantu
De-cp--	DWMP	cantus
D..fp--	DFP	mias, custas, cantas ecc.
D..fs---	DFS	mia, custa, canta, ecc.
D..mp---	DMP	mius, custus, cantus, ecc.
D..ms---	DMS	miu, custu, cantu, ecc.

Tabella 12: Coppie attributo-valore per la categoria "Articolo" in sardo

Attributo	Valore	Esempio	Codice
Tipo	definito	su	d
	indefinito	unu	i
Genere	maschile	su	m
	femminile	sa	f
Numero	singolare	su	s
	plurale	is	p
Caso	//	//	//

Tabella 13: <msd> e <ctag> per la categoria "Articolo" in sardo

msd	ctag	esempio
Tdms-	RMS	su
Td[fm]p-	RXP	is
Tdfs-	RFS	sa
Tims-	RIMS	unu
Tifs-	RIFS	una

Tabella 14: Coppie attributo-valore per la categoria "Avverbio" in sardo

Attributo	Valore	Esempio	Codice
Tipo	–	–	–
Grado	positivo	chitzi	p
	superlativo	malissimu	s

Tabella 15: <msd> e <ctag> per la categoria "Avverbio" in sardo

msd	ctag	esempio
R-p	B	mali
R-s	BS	malissimu

Tabella 16: Coppie attributo-valore per la categoria "Preposizione" in sardo

Attributo	Valore	Esempio	Codice
Tipo	preposizione	in, de	p

Tabella 17: <msd> e <ctag> per la categoria "Preposizione" in sardo

msd	Ctag	Esempio
Sp	SP	in

Tabella 18: Coppie attributo-valore per la categoria "Congiunzione" in sardo

msd	ctag	esempio
Cc	CC	ma
Cs	CS	poita

Tabella 19: <msd> e <ctag> per la categoria "Congiunzione" in sardo

msd	ctag	esempio
Cc	CC	ma
Cs	CS	poita

Tabella 20: Coppie attributo-valore per la categoria "Numerale" in sardo

Attributo	Valore	Esempio	Codice
Tipo	cardinale	centu	c
	ordinale	primu	o
Genere	maschile	primu	m
	femminile	prima	f
Numero	singolare	primu	s
	plurale	primus	p
Caso	//	//	//

Tabella 21: <msd> e <ctag> per la categoria "Numerale" in sardo

msd	ctag	esempio
M.ms-	NMS	primu
M.fs-	NFS	prima
M.mp-	NMP	primus
M.fp-	NFP	primas
Mc---	N	unu, centu

Tabella 22: <msd> e <ctag> per la categoria "Interiezione" in sardo

msd	ctag	esempio
l	l	ayo!

Tabella 23: <msd> e <ctag> per la categoria "Residuale" in sardo

ctag	esempio
X	simboli ecc.

Tabella 24:<msd> e <ctag> per la categoria "Punteggiatura" in sardo

ctag	esempio
punct	.,;!?...

The Relevance of Lesser-Used Languages for Theoretical Linguistics: The Case of Cimbrian and the Support of the TITUS Corpus

Ermenegildo Bidese, Cecilia Poletto
and Alessandra Tomaselli

On the basis of the TITUS Project, the following contribution aims at showing the importance of a lesser-used language, such as Cimbrian, for the theory of grammar. In Chapter 1, we present the goals of TITUS and its possibilities in order to analyse old Cimbrian writings. Furthermore, according to these possibilities, the second chapter will summarise some recent results of the linguistic research about relevant aspects of Cimbrian grammar, in particular the syntax of verbal elements, of subject clitics, and of subject nominal phrases. Chapter 3 and 4 discuss which relevance these results can have in the Generative framework, in particular with respect to a generalisation concerning the syntactic change in context of isolation and language contact.*

1. The TITUS Project (<http://titus.uni-frankfurt.de>)

The TITUS Project was conceived in 1987 during the Eighth Conference of Indo-European Studies in Leiden, when some of the participants had the idea to link their work together in order to create a text database for the electronic storage of writings/sources relevant to their discipline.¹ The name of the project was “Thesaurus of Indo-European Textual Materials on Data Media” (Thesaurus indogermanischer Textmaterialien auf Datenträgern). In the first phase, the project aimed at preparing a collection of textual materials from old Indo-European languages, such as Sanskrit, Old Iranian, Old Greek, Latin, as well as Hittite, Old High German and Old English.

In the beginning of the '90s, the rapid increase of electronic storage capacities in data processing led to a second phase of the project in 1994. During the Third Working Conference for the Employment of Data Processing in the Historical and Comparative Linguistics, in Dresden, the newly-founded working group ‘Historisch-Vergleichende Sprachwissenschaft’ (Historic-Comparative Linguistics) of the Society for Computational Linguistics and Language Technology (Gesellschaft für Linguistische

* The present contribution was written by the three authors in complete collaboration. For the formal definition of scholar responsibility, we declare that Ermenegildo Bidese draws up sections 1, 1.1 and 1.2, 2, 2.1, Cecilia Poletto sections 2.2 and 2.3, Alessandra Tomaselli sections 3 and 4. We would like to thank the staff of EURAC for the opportunity to present our research.

¹ Cf. Gippert (1995)

Datenverarbeitung) decided on an extension of the objectives for the 'Thesaurus', including further text corpora from other Indo-European and neighbouring languages, and introduced the new name 'Thesaurus of Indo-European Textual and Linguistic Materials', shortened to the acronym from the German designation: TITUS (Thesaurus indogermanischer Text- und Sprachmaterialien). The addition, 'linguistic materials', emphasizes that TITUS understands itself no longer only as a text database, but also as a 'data pool'.² On the TITUS server, you can find materials and aids for the analysis of the texts as well as, such as, among other things, a currently up-to-date bibliography with the newest publications in Indo-European studies, teaching materials, lexica, glossaries, language maps, audiovisual materials, software and fonts and heaps of helpful links. In fact, since 1995, owing to the above-mentioned conference, TITUS has been present on the World Wide Web with its own site at <http://titus.uni-frankfurt.de>.³ Responsible for the project is the *Institut für Vergleichende Sprachwissenschaft* at the University Johann Wolfgang-Goethe in Frankfurt am Main/Germany (direction: Professor Jost Gippert) in connection with other European universities.

The third phase in the development of the TITUS Project coincides with the explosive expansion of the Internet, and the new possibilities that online communication and Web performance offer. The new target of TITUS is the replacement of static data retrieval by an interactive one.⁴ This means that in order to better comprehend and analyse the texts, further information about the writings are made available to the user, who can then become interactive with the text. Three issues are pursued:

- a graphic documentation of the physical supports of the texts, usually manuscripts and inscriptions;
- an automatic retrieval of word form correspondences in a single text or in an entire language corpus; and,
- an automatic linguistic analysis of occurrences for the morphology of a word or for the basic forms of a verb.⁵

This interactive retrieval system is currently in development.

1.1 *The Cimbrian Texts in the TITUS Project*

The TITUS text database includes two Cimbrian texts provided by Jost Gippert, Oliver Baumann & Ermenegildo Bidese (1999).⁶ They comprise the catechism of 1813

2 Bunz (1998:12)

3 Ibid.

4 Cf. Gippert (2001)

5 Cf. Ibid. Cf. the same for four illustrative examples.

6 The direct links are: <http://titus.uni-frankfurt.de/texte/etcs/germ/zimbr/kat1813d/kat18.htm> and <http://titus.uni-frankfurt.de/texte/etcs/germ/zimbr/kat1842d/kat18.htm>.

(better known as the 'short Cimbrian catechism', written in the Cimbrian variety of the Seven Communities), and a new edition of the same text with slight alterations from 1842.⁷ In fact, this catechism is a Cimbrian translation of the 'Piccolo Catechismo ad uso del regno d'Italia' (Short Catechism for the Italian Kingdom) of 1807. A critical edition of both the original Italian text and the two Cimbrian versions was provided by Wolfgang Meid.⁸ The situation of Cimbrian knowledge at this time (with particular reference to the plateau of the Seven Communities) was very good, even though the use of the local Romance variety - in accordance with what the same text in the introduction testifies - was spreading.⁹ For this reason, and in view of the possibility of comparing this text with the first Cimbrian catechism of 1602, (which represents the oldest Cimbrian writing¹⁰), the 'short catechism' of 1813 and its later version in 1842 are essential sources for studying and analysing the diachronic development of the Cimbrian language.¹¹

On the basis of the above-mentioned critical edition by Professor Meid, we digitised the text in agreement with Meid's linearization of the original version. Moreover, we provided a first linguistic structuring of the text marking, above all, for the prefix of the participle perfect, pronominal clitics, personal pronouns, and the existence particle *-da*.¹²

1.2 The Research of Linguistic Content of the Cimbrian Texts

The first way of accessing the content of the Cimbrian texts is selecting the levels (chapters, paragraphs, verses and lines) into which the text is specifically subdivided in the entry form on the right frame of the text's start page. In this way, you can precisely find any given passage of the Cimbrian text.¹³

7 Cimbrian is a German dialect commonly spoken today in the village of Lusern/Luserna in the region of Trentino in northern Italy. It is also found, albeit in widely dispersed pockets, in the Venetian communities of Mittoballe/Mezzaselva (Seven Communities) and Ljetzan/Giazza (Thirteen Communities), in the northeast of Italy. When the Cimbrian colonies were founded and where the colonists came from are still subjects of controversy, although the accepted historical explanation is that the Cimbrian colonies originated from a migration of people from Tyrol and Bavaria (Lechtal) at the beginning of the second millennium. For a general introduction about the Cimbrian question and this language, cf. Bidese (2004b).

8 Cf. Meid (1985b)

9 Cf. Cat.1813:17-21 in Meid (1985b:35)

10 Cf. Meid (1985a). The first Cimbrian catechism is the translation of Cardinal Bellarmino's 'Dottrina cristiana breve' (cristian short doctrine). In spite of the title, the text is remarkably longer than the 1813's 'short catechism.'

11 Moreover, in TITUS, there is the first part of Remigius Geiser's (1999) self-learning Cimbrian course (cf. <http://titus.fkidg1.uni-frankfurt.de/didact/zimbr/cimbrian.htm>).

12 Cf. for the linguistically analysed texts following links: <http://titus.uni-frankfurt.de/texte/etcs/germ/zimbr/kat1813s/kat18.htm> and <http://titus.uni-frankfurt.de/texte/etcs/germ/zimbr/kat1842s/kat18.htm>.

13 Cf. for a detailed description of all these possibilities Gippert (2002).

A second possibility for content searching is obtained by using TITUS word search engine. By double-clicking on a given word of the Cimbrian text, for example, you can automatically look for its occurrences in the text, for the exact text references, and for the context in which this word is used (including orthographic variants).

A third way of content searching in the Cimbrian texts consists of using a search entry form that you can find when you open the link Switch to Word Index on the right frame of the start page of the text. In the box, you can enter a word and obtain its occurrences in the Cimbrian text.

In conclusion, we can state that the TITUS Project, with all the above-mentioned possibilities (and including the Cimbrian texts with a first linguistic structuring), offer a good starting-point for the research of the diachronic development of Cimbrian's syntax.

2. Some Relevant Aspects of Cimbrian Syntax

In the last decade, three interrelated syntactic aspects of the Cimbrian dialects have become the subject of intensive descriptive studies, from both the diachronic and the synchronic point of view: a) the syntax of verbal elements; b) the syntax of subject clitics; and, c) the syntax of subject NPs. The theoretical relevance of these studies will be discussed in section 4.

2.1 Verb Syntax

As for the syntax of verbal elements, the following descriptive results can be taken for granted:

i) Cimbrian is no longer characterised by the V2 restriction, which requires the second position of the finite verb in the main declarative clause. As the following examples show, the finite verb can be preceded by two or more constituents that are not rigidly ordered, as shown by the fact that both (1) (a and b) and (2) are grammatical. Similar cases of V3 (as in [1a]) or V4 (as in [1b]) are not acceptable, neither in Standard German (cf. 3), or in any other continental Germanic languages:¹⁴

(1a) Gheistar in Giani hat gahakat iz holtz ime balje (/in balt)¹⁵ (Giazza)

Yesterday the G. has cut the wood in the forest

(1b) De muotar gheistar kam Abato hat kost iz mel¹⁶ (Giazza)

The mother yesterday in Abato has bought the flour

¹⁴ Cf. Scardoni (2000), Poletto & Tomaselli (2000), Tomaselli (2004), Bidese & Tomaselli (2005). In the catechism of 1602, there are few examples of V3 constructions, but this is probably due to the fact that there is no relevant context for the topic. Cf. for this problem Bidese and Tomaselli (2005:76ff.)

¹⁵ Scardoni (2000:152)

¹⁶ Ivi:157

-
- (2) Haute die Mome hat gekoaft die öala al mercà¹⁷ (Luserna)
Today the mother has bought the eggs at-the market
- (3) *Gestern die Mutter hat Mehl gekauft
yesterday the mother has flour bought

ii) A correlate of the V2 phenomenology forces the reordering of subject and inflected verb: in the Germanic languages,¹⁸ the subject can be found in main clauses to the right of the inflected verb (but still to the left of a past participle, if the sentence contains one) when another constituent is located in first position, yielding the ordering XP Vinfl Subject ... (Vpast part.). In Cimbrian, the phenomenon of subject - (finite) verb inversion is limited to subject clitics starting from the first written documents (i.e., the Cimbrian catechisms of 1602, here shortened in *Cat. 1602*) (cf. 4), and survived the loss of the V2 word order restriction for quite a long time (cf. 5 and 6). Nowadays, in Giazza, it is only optionally present, and only for some speakers (cf. 7 and 8), while it survives in Luserna (cf. 9 and 10):¹⁹

- (4) [Mitt der Bizzonghe] saibar ghemostert zò bizzan den billen Gottez.²⁰
Through knowledge are-we taught to know the will of God.

- (5) [Benne di andarn drai Lentar habent gahört asó], haben-se-sich manegiart ...²¹

When the other three villages had heard this, had-they taken pains to

...

- (6) [Am boute] [gan ljêtsen] hense getrust gien ...²²
Once in Ljetzan have-they got to go ...

- (7) In sontaghe regatz / In sontaghe iz regat²³ (Giazza)
On Sunday rains-it / On Sunday it rains

- (8) Haute er borkofart de oiar / Haute borkofartar de oiar²⁴ (Giazza)
Today he sells the eggs/today sells-he the eggs

17 Grewendorf & Poletto (2005:117)

18 English has this possibility too, but it is restricted to main interrogatives, while in the other Germanic languages it is found also in declaratives.

19 Bosco (1996) and (1999), Benincà & Renzi (2000), Scardoni (2000), Poletto & Tomaselli (2000), Tomaselli (2004), Bidese & Tomaselli (2005) and Grewendorf & Poletto (2005). That subject clitics continue to invert when nominal subjects cannot is a well-known generalisation confirmed in other language domains, such as Romance.

20 *Cat. 1602:694-5* in Meid (1985a:87)

21 Baragiola 1906:108

22 Schweizer 1939:36

23 Scardoni 2000:144

24 Ivi:155

(9) *Häüte geat dar Giani vort²⁵ (Luserna)

Today goes the Gianni away

(10) Häüte gearar vort (dar Gianni)²⁶ (Luserna)

Today goes-he away (the John)

This seems to indicate that the 'core' of the V2 phenomenon (i.e., the word order restriction) could be lost before one of its main correlates (i.e., pronominal subject inversion).

- Germanic languages can be OV (German and Dutch) or VO (Scandinavian and Yiddish). In Cimbrian, the discontinuity of the verbal complex is limited to the intervention of pronominal elements, negation (cf. 12), monosyllabic adverbs/verbal prefixes,²⁷ and bare quantifiers²⁸ (cf. 13). In fact, from a typological point of view, Cimbrian belongs, without any doubt, to the group of VO languages:

(11a) Häüte die Mome hat gebäschd di Piattn²⁹ (Luserna)

Today the mother has washed the dishes

(11b) *Häüte di Mome hat di Piattn gebäschd³⁰ (Luserna)

(12) Sa hom khött ke dar Gianni hat net geböllt gian pit se³¹ (Luserna)

They have said that the G. has not wanted go with them

(13a) I hon niamat gesek³² (Luserna)

I have nobody seen

(13b) han-ich khoome gaseecht (Roana)

have-I nobody seen

- Residual word order asymmetries between main and subordinate clauses with respect to the position of the finite verb are determined by a) the syntax of some 'light' elements (cf. 14 and 15 for negation and pronominal); b) by the presence of clitics (cf. 14b and 15b versus 16 and 17); and, c) by the type of subordinate clause (cf. 14b and 15b versus 18 and 19):

(14a) Biar zéteren nete³³

We give in not

25 Grewendorf & Poletto 2005:116

26 Ibid.

27 Cf. Bidese 2004a and Bidese & Tomaselli 2005

28 Cf. Grewendorf & Poletto (2005)

29 Ivi:117

30 Ivi:121

31 Ivi:122

32 Ivi:123

33 Baragiola 1906:108

-
- (14b) 'az se nette ghenan vüar³⁴
that they don't put forward
- (15a) Noch in de erste Lichte von deme Tage hevan-se-sich alle³⁵
Even at the break of that day get-they all up
- (15b) 'az se sich legen in Kiete³⁶
that they calm down
- (16) 'az de Consiliere ghen nette auf in de Sala³⁷
that the advisers go not above into the room
- (17) 'az diese Loite richten-sich³⁸
that these people arrange themselves
- (18) umbrume di andar Lentar saint net contente³⁹
because the other villages are not glad
- (19) umbrume dear Afar has-sich gamachet groaz⁴⁰
because the question has got great

2.2 Clitic Syntax

The Cimbrian dialect, contrary to other Germanic languages that only admit weak object pronouns, is characterized by a very structured set of pronominal clitics, like all northern Italian dialects.⁴¹ One important piece of evidence that subject and object pronouns are indeed clitics is the phenomenon of clitic doubling, namely, the possibility to double a full pronoun or an NP with a clitic, already noted in the grammars:

- (20) az sai-der getant diar⁴²
that it will be to you made to you
- (21) Hoite [de muutar] hat-se gakhooft de ojar in merkatn (Roana)
Today the mother has-she bought the eggs at-the market

From a diachronic point of view, this phenomenon already appears for *subject clitics* in *Cat. 1813*, but is limited to interrogative sentences, while in Baragiola (1906) it also appears in declarative sentences. The phenomenon is, nowadays, according to

34 Ivi:111

35 Ivi:109-110

36 Ivi:114

37 Ivi:110

38 Ivi:108

39 Ivi:105

40 Ivi:113

41 For an exhaustive description of the positions of clitics and pronouns in Cimbrian cf. Castagna (2005).

42 Schweizer (1952:27)

the research of Scardoni (2000), no longer productive in Giazza, optional/possible in Luserna,⁴³ but still frequent in Roana.⁴⁴

In *main clauses*, subject clitics are usually found in enclisis to the finite verb (in Giazza, only as a vestige, cf. the above sentences [7] and [8]):⁴⁵

- (22) Bia hoas-to (de) (du)? (Luserna)
How call-you?
- (23) Hasto gi khoaft in ġornal?⁴⁶ (Luserna)
Have-you bought the newspaper?
- (24) Ghestar han-ich ghet an libar ame Pieren (Roana)⁴⁷
Yesterday have-I given a book to P.

In *embedded clauses*, subject clitics occur either in enclitic position to the finite verb or in enclitic position to the conjunction, depending on two main factors: i) the Cimbrian variety under consideration (and the 'degree' of V2 preservation); and, ii) the different types of subordinate clauses. According to what our data suggest, nowadays, enclisis to the finite verb seems to be the rule in Roana (25-8), but Schweizer's grammar (Schweizer 1952) gives evidence for a different distribution of the subject clitics in subordinate clauses. He observes that subject clitics in the variety of Roana usually occur (or occurred) at the Wackernagel's position (WP) in enclisis to the subordinating conjunction (cf. 29-31; cf. the above sentences [14b] and [15b] as well):⁴⁸

- (25) Ist gant zoornig, ambrumme han-ich ghet an libarn ame Pieren (Roana)
(He) has got angry, because have-I given a book P.
- (26) Gianni hatt-ar-mi gaboorsset, benne khimmas-to hoam (Roana)
Gianni has-he-me asked, when come-you home
- (27) Haban-sa-mich gaboorsset, ba ghe-ban haint (Roana)
Have-they-me asked, where go-we today evening
- (28) Haban-sa-mich khött, habat-ar gabunnet Maria nach im beeck (Roana)
Have-they-(to)me said, have-you met M. on the road
- (29) bas-er köt⁴⁹ (Roana)

43 Cf. Vicentini (1993:149-51) and Castagna (2005)

44 Our data suggest that there may be a difference between auxiliaries and main verbs: with the auxiliary 'have', doubling seems mandatory, while this is not the case with main verbs.

45 Some ambiguous forms can also appear in first position; we assume here that when occurring in first position, the pronominal forms are not real clitics, but, at most, weak forms.

46 Vicentini (1993:44)

47 In the variety of Roana, when the subject is definite and preverbal, there is always an enclitic pronoun.

48 Cf. Castagna (2005) as well

49 Schweizer (1952:27)

what-he says

(30) ben-ig-en nox vinne⁵⁰ (Roana)

if-I-him still meet

(31) ad-ix gea au⁵¹ (Roana)

if-EXPL.-I (az-da-ich) go above

All the same, Schweizer (1952) underlines that there are many irregularities in accordance to which subject clitics in embedded clauses can appear in enclisis to the finite verb, or in both positions (clitic doubling). Luserna Schweizer notes that all the pronouns have to be clitized to the complementizer.⁵² But we found evidence for a construction (cf. 32) in which the subject clitic appears in enclisis to the finite verb, probably due to the presence of a constituent between the complementizer and the finite verb (a case of “residual” embedded V2). In this sentence, there is clitic doubling too:

(32) Dar issese darzürnt obrom gestarn honne i get an libar in Peatar⁵³
(Luserna)

He has got angry because yesterday have-I I given a book P.

In main clauses, *object clitics* are always in enclisis to the inflected verb:

(33a) Der Tatta hat-se gekoaft⁵⁴ (Luserna)

The father has-her bought

(33b) Der Tatta *se hat gekoaft⁵⁵ (Luserna)

(34) De muutari hat-sei-se gasecht (Roana)

The mother has-she-her seen

(35) Gianni hatt-an-se gasecht (Roana)

Gianni has-he-her seen

The same is true for embedded declarative clauses:

(36a) I woas ke der Tatta hatse (net) gekoaft⁵⁶ (Luserna)

I know that the father has-her (not) bought

(36b) I woas ke der Tatta *se hat gekoaft⁵⁷ (Luserna)

I know that the father her has bought

50 Ibid.

51 Ibid.

52 Ibid. This analysis is confirmed in the data of Vicentini (1993)

53 Grewendorf & Poletto (2005:121)

54 Ivi:122

55 Ibid.

56 Ivi:123

57 Ibid.

-
- (37) Gianni hatt-ar-mi gaboorsset, bear hat-ar-dich telephonaart (Roana)
Gianni has-he-me asked, who has-he-you called
- (38) kloob-ich Gianni hatt-ar-me ghet nicht ad ander (Roana)
believe-I (that) Gianni has-he-(to)me given nothing else
- (39) biss-i net, Gianni hat-an-en ghakhoofet (Roana)
know-I not, (if) Gianni has-he-him bought

While in Roana, enclisis to the finite verb is the rule in all embedded clauses (including embedded interrogatives), in Luserna, in *relative and embedded interrogative clauses*, subject and object clitics are usually found in a position located to the immediate right of the complementiser (or the *wh*-item).⁵⁸ This corresponds to Wackernagel's position of the Germanic tradition, and is usually hosting weak pronouns in the Germanic languages, which are rigidly ordered (contrary to DPs, which can scramble):

- (40) 's baibe bo-da-r-en hat geet an liber⁵⁹ (Luserna)
the woman who-EXPL.-he-(to) her has given a book
- (41) dar Mann bo dar en (er) hat geet an libar (Luserna)
the man who-EXPL.-he-him (he) has given a book
- (42) Dar Giani hatmar gevorst zega ber (da)de hat o-gerüaft (Luserna)
The G. has-me asked compl. who you has phoned
- (43) I boas net ber-me hat o-gerüaft (Luserna)
I know not who us has phoned
- (44) I vorsmaar zega bar me mage hom o-gerüaf (Luserna)
I wonder COMPL. who me could have phoned

Summarising the data illustrated so far, we can state that:

- Both subject and object clitics are always in *enclisis to the finite verb* in *main clauses* in all varieties;
- Currently in Roana, both subject and object clitics always occur in enclisis to the finite verb in all embedded clauses; and,
- In Luserna, clitics occur in enclisis in embedded declaratives and in WP in relative and embedded interrogatives.

From this we conclude that:

- Luserna displays a split between embedded *wh*-constructions on the one hand and embedded declaratives on the other, while Roana (at least nowadays) does not; and,

⁵⁸ This means that no element can intervene between the element located in CP and the pronoun(s).

⁵⁹ Grewendorf & Poletto (2005:121)

- *No cases of proclisis to the inflected verb are ever found in any Cimbrian variety.*

In general, although Cimbrian, contrary to other Germanic languages, has developed a class of clitic pronouns, it does not seem to have ‘copied’ the syntactic behaviour of subject and object clitics of neighbouring Romance dialects, which realize consistently proclisis to the inflected verb for object clitics in all sentence types, and permit enclisis of subject clitics only in main interrogative clauses, and enclisis of object clitics only with infinitival verbal forms.⁶⁰ On the contrary, enclisis to the inflected verb seems to be the rule in Cimbrian. Proclisis to the inflected verb is not at all attested, and the only other position apart from enclisis is the Germanic WP position in some embedded clause types in the variety of Luserna.

2.3 The Syntax of Subject NPs

As regards the syntax of the subject NPs in Cimbrian, there is evidence of the following aspects:

- Cimbrian is not a *pro-drop* language. As with standard German, English and French, it is characterised by: a) obligatory expression of the subject (cf. 45);
- the use of the expletive pronoun *iz* (cf. 46); and, c) (contrary to standard German) a VO typology and the consequent adjacency of the verbal complex (cf. 47); and, d) a relatively free position of the finite verb:⁶¹

(45) *i han gaarbat (/gaarbatat) ime balt / Haute hani gaarbatat ime balje*⁶²
(Giazza)

Today I have worked in the forest / Today have-I worked in the forest

(46) *Haute iz regat / Haute regatz*⁶³ (Giazza)

Today it rains / Today rains-it

(47) *Gheistar in Giani hat gahakat iz holtz ime balje (/in balt)*⁶⁴ (Giazza)

Yesterday G. has cut the wood in the forest

- Languages requiring a mandatory expression of the subject, such as English or French, see the possibility of putting the subject NPs on the right of the verbal

60 Note that there are Romance dialects that have enclisis to the inflected verb, such as the variety of Borgomanero, studied by Tortora (1997), but this is a Piedmontese dialect, which can not have been in touch with Cimbrian, so we can exclude that enclisis has been developed through language contact with Romance.

61 Cf. Poletto & Tomaselli (2002) and Tomaselli (2004:543). Cf. Castagna (2005) as well.

62 Scardoni (2000:155)

63 Ivi:144

64 Ivi:152

complex only in very limited contexts. From this perspective, it is interesting to note that Cimbrian generally permits it (cf. 48 and 49), similarly to standard Italian (cf. 50), and in opposition to the neighbouring romance dialect, in which the post verbal subject co-occurs with a subject pronoun in a preverbal position (cf. 51 and 52):

(48) Gheistar hat gessat dain manestar iz diarlja⁶⁵ (Giazza)

Yesterday has eaten your soup the girl

(49) Hat gahakat iz holtz dain vatar⁶⁶ (Giazza)

Has cut he wood your father

(50) Lo hanno comprato al mercato i miei genitori

It have bought at the market my parents

(51) Algéri l' à magnà la to minestra la buteleta⁶⁷

Yesterday she has eaten your soup the girl

(52) L' à taià la legna to papà⁶⁸

He has cut the wood your father

3. Cimbrian Data and the Generative Grammar Framework

The results of the syntactic description of some aspects of Cimbrian grammar are relevant for any theoretical framework. In particular, within the Generative Grammar theoretical approach, the data discussed so far is relevant from both a *synchronic* and a *diachronic* point of view.

Cimbrian, having been in a situation of language contact for centuries, offers a privileged point of view for determining how phenomena are lost and acquired. A number of interesting observations can be made concerning language change induced by language contact.

First, Cimbrian shows that the 'correlates' of a given phenomenon (in our case V2) are lost after the loss of the phenomenon itself. More specifically, Cimbrian has maintained the possibility of inverting subject pronouns, while losing the V2 linear restriction. On the other hand, we can also state that the correlates can be acquired before the phenomenon itself: although Cimbrian has not developed a fully-fledged *pro drop* system, it already admits subject free inversion of the Italian type (i.e., the subject inverts with the 'whole' verbal phrase).

65 Ivi:165

66 Ibid.

67 Ibid.

68 Ibid.

Second, syntactic change does not proceed in parallel to the lexicon, where a word is simply borrowed and then adapted to the phonological system of the language.⁶⁹ The syntactic distribution of clitic elements in Cimbrian shows that they have maintained a Germanic syntax, allowing either enclisis to the verb or the complementizer (WP), but never proclisis to the inflected verb, as is the case for Romance. Therefore, even though Cimbrian might have developed (or rather ‘maintained’ / ‘preserved’) a class of clitic elements due to language contact, it has not ‘copied’ the Romance syntax of clitics.

Moreover, the study of Cimbrian also confirms two descriptive generalisations concerning the loss of the V2 phenomenology established on the basis of the evolution of Romance syntax:⁷⁰

- Embedded wh-constructions constitute the sentence type that longer maintains asymmetry with main clauses. This is shown in Cimbrian by the possibility of having clitics in WP only in embedded interrogatives, and relatives in the variety of Luserna; and,
- Inversion of NPs is lost before inversion of subject clitics, which persists for a longer period.

More generally, Cimbrian also confirms the hypothesis first put forth by Lightfoot (1979), and mathematically developed by Clark & Roberts (1993), that the reanalysis made by bilingual speakers goes through ambiguous strings that have two possible structural analyses; the speaker tends to use the more economical one (in terms of movement) that is compatible with the set of data at his/her disposal.

Also, from the *synchronic point of view*, Cimbrian is an interesting study case, at least as far as verb movement is concerned. In V2 languages, it is most probably an Agreement feature located in the C that attracts the finite verb (see Tomaselli 1990 for a detailed discussion of this hypothesis). Cimbrian seems to have lost this property, as neither the linear V2 restriction nor the NP subject inversion are possible at this time. On the other hand, it has not (yet) developed a ‘Romance’ syntax, because clitics are always enclitics in the main clause (both declarative and interrogative). It is a well-known fact (see, among others, Sportiche 1993 and Kayne 1991 & 1994) that in the higher portion of the IP layer, there is a (set of) position(s) for clitic elements, and that subject clitics are always located to the left of object clitics inside the template containing the various clitics.

69 This hypothesis is already been made by Brugmann (1917).

70 See Benincà (2005) for the first generalization, Benincà (1984), Poletto (1998) and Roberts (1993), for the second.

The position of the inflected verb in Cimbrian is neither the one found in V2 language (within the CP domain), nor the lower one found in modern Romance (within the IP domain). The syntax of clitics suggests that, in Cimbrian, the inflected verb moves to a position inside the clitic layer in the high IP (corresponding to the traditional WP), and precisely to the left of clitic elements both in main and embedded declarative clauses.⁷¹ If this theoretical description proves to be tenable, we are now in the position to speculate about a possible explanation.

4. A New Theoretical Correlation 'Visible' in Cimbrian

A further interesting field to explore has to do with the theoretical reason why Cimbrian could not develop a Romance clitic syntax. In other words, there must have been some restriction constraining the speakers to maintain enclisis.

A striking difference between the neighbouring Romance dialects and Cimbrian is the past participle agreement phenomenon. Past participle agreement is mandatory (at least for some object clitics) in Northern Italian dialects (cf. 53), while it is completely absent in Cimbrian. The morphological structure of the Cimbrian past participle has simply preserved the invariant German model, that is, *ge-* ... *-t*, (cf. 54):

(53) (A) so k'el papá li ga visti

I know that the father them-has seen

(54) I woas ke der Tatta hatze (net) gekoaft (Luserna)

I know that the father has-her (not) bought

The existence of past participle agreement is usually analysed in the relevant literature as involving an agreement projection (AgrOP) to which both the object clitic and the verb move; the configuration of spec-head agreement between the two triggers the 'passage' of the number and gender features of the clitic onto the verb yielding agreement on the past participle (see Kayne 1991 and 1993).

We believe that it is the presence of this lower agreement projection that is related to the possibility of having proclisis in Romance, and its absence that constrains Cimbrian to enclisis to the inflected verb. In Cimbrian, the clitic element moves directly to the higher clitic position (within the IP domain), while in Romance, this movement is always in two steps, the first being movement to the lower AgrO projection. In favour of this assumption is the fact that Cimbrian, like all other Germanic varieties, never showed past participle agreement of the Romance type.

71 As we have already noted, the same is true for embedded interrogatives in Roana, while in Luserna, the verb is probably located lower in embedded interrogatives and relative clauses, leaving the clitic in WP alone.

Abbreviations

<i>Cat. 1602</i>	Cimbrian Catechism of 1602 (cf. Meid 1985a)
<i>Cat. 1813</i>	Cimbrian Catechism of 1813 (cf. Meid 1985b)
DP	Determiner Phrase
NP	Nominal Phrase
Vinfl	Inflected Verb
Vpast part.	Participle Past Verb
Wh	(interrogative element)
XP	X-phrase

References

Baragiola, A. (1906). "Il tumulto delle donne di Roana per il ponte (nel dialetto di Camporovere, Sette Comuni)". Padova: Tip, Fratelli Salmin, reprinted in Lobbia, N. & Bonato, S. (eds.) (1998). *Il Ponte di Roana. Dez Dink vo' der Prucka*. Roana: Istituto di Cultura Cimbra.

Benincà, P. (1984). "Un'ipotesi sulla sintassi delle lingue romanze medievali." *Quaderni Patavini di Linguistica* 4, 3-19.

Benincà, P. (2005). "A Detailed Map of the Left Periphery of Medieval Romance." Zanuttini, R. et al. (eds.) (2005). *Negation, Tense and Clausal Architecture: Cross-linguistics Investigations*. Georgetown University Press.

Benincà, P. & Renzi, L. (2000). "La venetizzazione della sintassi nel dialetto cimbro." Marcato, G. (ed.) (2000). *Isole linguistiche? Per un'analisi dei sistemi in contatto. Atti del convegno di Sappada/Plodn (Belluno), 1-4 luglio 1999*. Padova: Unipress, 137-62.

Bidese, E. (2004a). "Tracce di *Nebensatzklammer* nel cimbro settecomunigiano." Marcato, G. (ed.) (2000). *I dialetti e la montagna. Atti del convegno di Sappada/Plodn (Belluno), 2-6 luglio 2003*, Padova: Unipress, 269-74.

Bidese, E. (2004b). "Die Zimbern und ihre Sprache: Geographische, historische und sprachwissenschaftlich relevante Aspekte." Stolz, T. (ed.) (2004). "Alte" Sprachen. *Beiträge zum Bremer Kolloquium über "Alte Sprachen und Sprachstufen" (Bremen, Sommersemester 2003)*. Bochum: Universitätsverlag Dr. N. Brockmeyer, 3-42.

Bidese, E. & Tomaselli, A. (2005). "Formen der ‚Herausstellung‘ und Verlust der V2-Restriktion in der Geschichte der zimbrischen Sprache." Bidese, E., Dow, J.R. & Stolz, T. (eds.) (2005). *Das Zimbrische zwischen Germanisch und Romanisch*. Bochum: Universitätsverlag Dr. N. Brockmeyer, 71-92.

Bosco, I. (1996). 'Christlike unt korze Dottrina': un'analisi sintattica della lingua

cimbra del XVI secolo. Final essay for the degree "Laureat in Modern Languages and Literature." Unpublished Essay, University of Verona.

Bosco, I. (1999). "Christlike unt korze Dottrina': un'analisi sintattica della lingua cimbra del XVI secolo." Thune, E.M. & Tomaselli, A. (eds.) (1999). *Tesi di linguistica tedesca*. Padova: Unipress, 29-39.

Brugmann, K. (1917). "Der Ursprung des Scheinsubjekts 'es' in den germanischen und den romanischen Sprachen." *Berichte über die Verhandlungen der Königl. Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Philologisch-historische Klasse 69/5*. Leipzig: Teubner, 1-57.

Bunz, C.M. (1998). "Der Thesaurus indogermanischer Text- und Sprachmaterialien (TITUS) - ein Pionierprojekt der EDV in der Historisch-Vergleichenden Sprachwissenschaft." *Sprachen und Datenverarbeitung* 1(98), 11-30. <http://titus.uni-frankfurt.de/texte/sdv198.pdf>.

Castagna, A. (2005), "Personalpronomen und Klitika im Zimbrischen." Bidese, E., Dow, J.R. & Stolz, T. (eds) (2005). *Das Zimbrische zwischen Germanisch und Romanisch*. Bochum: Universitätsverlag Dr. N. Brockmeyer, 93-113.

Clark, R. & Roberts, I. (1993), "A Computational Model of Language Learnability and Language Change." *Linguistic Inquiry* 24, 299-345.

Geiser, R. (1999). "Grundkurs in klassischem Zimbrisch." <http://titus.fkidg1.uni-frankfurt.de/didact/zimbr/cimbrian.htm>.

Gippert, J. (1995). "TITUS. Das Projekt eines indogermanistischen Thesaurus." *LDV-Forum* (Forum der Gesellschaft für Linguistische Datenverarbeitung) 12 (2), 35-47. <http://titus.uni-frankfurt.de/texte/titusldv.htm>.

Gippert, J. (2001). *Der TITUS-Server: Grundlagen eines multilingualen Online-Retrieval-Systems* (aus dem Protokoll des 83. Kolloquiums über die Anwendung der Elektronischen Datenverarbeitung in den Geisteswissenschaften an der Universität Tübingen 17. November 2001). <http://www.zdv.uni-tuebingen.de/tustep/prot/>

prot831-titus.html.

Gippert, J. (2002). *The TITUS Text Retrieval Engine*. <http://titus.uni-frankfurt.de/texte/textex.htm>.

Grewendorf, G. & Poletto, C. (2005). "Von OV zu VO: ein Vergleich zwischen Zimbrisch und Plodarisch." Bidese, E, Dow, J.R. & Stolz, T. (eds) (2005). *Das Zimbrische zwischen Germanisch und Romanisch*. Bochum: Universitätsverlag Dr. N. Brockmeyer, 114-128.

Kayne, R.S. (1991). "Romance Clitics, Verb Movement, and PRO." *Linguistic Inquiry* 22, 647-686.

Kayne, R.S. (1993). "Towards a Modular Theory of Auxiliary Selection." *Studia Linguistica* 47, 3-31.

Kayne, R.S. (1994). *The Antisymmetry of Syntax*. Cambridge, Mass.: MIT Press.

Lightfoot, D. (1979). *Principles of Diachronic Syntax*. Cambridge, England: Cambridge University Press.

Meid, W. (1985a). *Der erste zimbrische Katechismus CHRISTLIKE UNT KORZE DOTTRINA*. Die zimbrische Version aus dem Jahre 1602 der DOTTRINA CHRISTIANA BREVE des Kardinals Bellarmin in kritischer Ausgabe. Einleitung, italienischer und zimbrischer Text, Übersetzung, Kommentar, Reproduktionen. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck.

Meid, W. (1985b). *Der zweite zimbrische Katechismus DAR KLÓANE CATECHISMO VOR DEZ BÉLOSELAND*. Die zimbrische Version aus dem Jahre 1813 und 1842 des PICCOLO CATECHISMO AD USO DEL REGNO D'ITALIA von 1807 in kritischer Ausgabe. Einleitung, italienischer und zimbrischer Text, Übersetzung, Kommentar, Reproduktionen. Innsbruck: Institut für Sprachwissenschaft der Universität Innsbruck. <http://titus.uni-frankfurt.de/texte/etcs/germ/zimbr/kat1813d/kat18.htm>.

Poletto, C. (1998). "L'inversione interrogativa come 'verbo secondo residuo': l'analisi

sincronica proiettata nella diacronia.” *Atti del XXX convegno SLI*, Roma: Bulzoni, 311-327.

Poletto, C. & Tomaselli, A. (2000). “L’interazione tra germanico e romanzo in due ‘isole linguistiche’. Cimbri e ladino centrale a confronto.” Marcato, G. (ed.) (2000). *Isole linguistiche? Per un’analisi dei sistemi in contatto. Atti del convegno di Sappada/Plodn (Belluno), 1-4 luglio 1999*. Padova: Unipress, 163-76.

Poletto, C. & Tomaselli, A. (2002). “La sintassi del soggetto nullo nelle isole tedescofone del Veneto: cimbri e sappadino a confronto.” Marcato, G. (ed.) (2002). *La dialettologia oltre il 2001. Atti del convegno di Sappada/Plodn (Belluno), 1-5 Luglio 2001*. Padova: Unipress, 237-52.

Roberts, I. (1993). *Verbs and Diachronic Syntax: A Comparative History of English and French*. Dordrecht: Kluwer.

Scardoni, S. (2000). *La sintassi del soggetto nel cimbro parlato a Giazza*. Final essay for the degree “Laureat in Modern Languages and Literature.” Unpublished Essay, University of Verona.

Schweizer, B. (1939). *Zimbrische Sprachreste. Teil 1: Texte aus Giazza (Dreizehn Gemeinden ob Verona). Nach dem Volksmunde aufgenommen und mit deutscher Übersetzung herausgegeben*. Halle/Saale: Max Niemeyer.

Schweizer, B. (1952). *Zimbrische Gesamtgrammtik. Band V.: Syntax der zimbrischen Dialekte in Oberitalien*. Diessen am Ammersee. Unpublished typescript. Marburg/Lahn, Germany: Institut für die Forschung der Deutschen Sprache.

Sportiche, D. (1993). “Clitic Constructions.” Rooryck, J. & Zaring, L. (eds) (1993). *Phrase Structure and the Lexicon*. Dordrecht: Kluwer, 213-276.

Tomaselli, A. (1990). *La sintassi del verbo finito nelle lingue germaniche*. Padova: Unipress.

Tomaselli, A. (2004). "Il cimbro come laboratorio d'analisi per la variazione linguistica in diacronia e sincronia." *Quaderni di lingue e letterature 28, Supplemento: Variis Linguis: Studi offerti a Elio Mosele in occasione del suo settantesimo compleanno*, 533-549.

Tortora, C.M. (1997). "I Pronomi Interrogativi in Borgomanerese." Benincà, P. & Poletto, C. (eds) (1997). *Quaderni di Lavoro dell'ASIS (Atlante Sintattico Italia Settentrionale): Strutture Interrogative dell'Italia Settentrionale*. Padova: Consiglio Nazionale delle Ricerche, 83-88.

Vicentini, R. (1993). *Il dialetto cimbro di Luserna: analisi di alcuni fenomeni linguistici*. Final essay for the degree "Laureat in Modern Languages and Literature." Unpublished Essay, University of Trento.

Creating Word Class Tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping

Danie J. Prinsloo and Ulrich Heid

To bootstrap tagging resources (tagger lexicon and training corpus) for Northern Sotho, a tagset and a number of modular and reusable corpus processing tools are being developed. This article describes the tagset and routines for identifying verbs and nouns, and for disambiguating closed class items. All of these are based on morphological and morphosyntactic specificities of Northern Sotho.

1. Introduction

In this paper, we report on ongoing work towards the parallel creation of computational linguistic resources for Northern Sotho, on the basis of linguistic knowledge about the language. Northern Sotho is one of the eleven official languages of South Africa, spoken by about 4.2 million speakers in the northeastern part of the country. It belongs to the Sotho family of the Bantu languages (S32), (Guthrie 1971). The three Sotho languages are closely related.

The creation of Natural Language Processing (NLP) resources is part of an effort towards an infrastructure for corpus linguistics and computational lexicography and terminology for Northern Sotho, which is seen as an element of a broader action for the development of Human Language Technology (HLT) and NLP applications for the South African languages.

Parallel resource creation has been attempted as part of our research and development agenda in order to speed up the resource building process, in the sense of rapid prototyping of a part-of-speech (=POS) tagset; a tagger lexicon and (manually corrected) reference corpus; and a statistical tagger. These constitute the first set of corpus linguistic tools to be developed (we report on the first three tools here). At the same time, we intend to verify to what extent 'traditional' corpus linguistic methods and tools (as used for European languages) can be applied to a Bantu language-- an attempt that, to our knowledge, has not been made before.

Two text corpora are used as input to the study. The first is a 43,000 tokens corpus, a selection from the Northern Sotho novel *Tša ka Mafuri* (Matsepe 1974), and the second is the *Pretoria Sepedi Corpus* (PSC) of 6 million tokens, a collection of 327

Northern Sotho books and magazines. These are raw, unannotated corpora, compiled by means of optical character recognition (OCR), commonly known as ‘scanning’, with tokenization done per sentence. The PSC is still in the process of being cleaned from scanning errors. For details regarding the PSC and subsequent applications thereof, see sources such as Prinsloo (1991), De Schryver & Prinsloo (2000, 2000a & 2000b), and Prinsloo & De Schryver (2001).

In this paper, we will discuss our task at both a specific and general level. We report about the specific task of creating resources for Northern Sotho, and our examples and illustrative material will be taken from this language. More generally, we also analyse the exercise in terms of methods and strategies for the joint bootstrapping of different resources for an ‘unresourced’ language, trying to abstract away from language-specific details.

This article is organised as follows: in section 2, we give a brief overview of some of the language-specific phenomena we exploit in resource building; section 3 deals with the component elements of a corpus linguistic infrastructure for Northern Sotho that are presently being constructed, with the steps and procedures used in the process and the characteristics of the resulting resources; section 4 is a methodological conclusion (order of steps in resource creation, role of linguistic knowledge, etc.) and an analysis of the processes in terms of generalisability and portability to other languages such as Sotho, Bantu, and possibly completely different languages.

2. Northern Sotho Linguistics Informing Corpus Technology

A prerequisite to successful interpretation of the criteria for and output of a POS-tagger for Northern Sotho is a brief outline of certain basic linguistic characteristics of the language, especially of nouns and verbs. See Lombard *et al.* (1985), Louwrens (1991), and Poulos & Louwrens (1994) for a detailed grammatical description of this language.

2.1 Noun System: Classifiers and Concords

Nouns in Bantu languages are grouped into different noun classes. Compare Table 1 for Northern Sotho.

Table 1: Noun Classes of Northern Sotho with Examples

Class	Prefix	Example	Translation
1	mo-	monna	man
2	ba-	banna	men
1a	∅	malome	uncle

2b	bo+	bomalome	uncles
3	mo-	monwana	finger
4	me-	menwana	fingers
5	le-	lesogana	young man
6	ma-	masogana	young men
7	se-	selepe	axe
8	di-	dilepe	axes
9	N-/∅	nku	sheep (sg.)
10	di+	dinku	sheep (pl.)
11			
12			
13			
14	bo-	bogobe	porridge
6	ma-	magobe	different kinds of porridge
15	go	go bona	to see
16	fa-	fase	below
17	go-	godimo	above
18	mo-	morago	behind

Nouns are subdivided into different classes, each with its own prefix, and the prefixes of the first ten classes mark singular versus plural forms. Classes 11-13 do not exist in Northern Sotho. The prefixes also generate a number of concords and pronouns that are used to complete phrases and sentences. Consider the following example from Class 1, given in Table 2.

Table 2: Example of a Sentence Consisting of a Noun, Verb, Pronoun and Concords

Monna	yo	o	A	di	rata
noun Cl. 1	demonstrative (pronoun) Cl. 1	subject concord Cl. 1	present tense marker	object concord Class 8/10	verb stem
Man	this	(he)	()	them	loves
This man loves them.					

There are a few hundred closed class items such as the subject concords, object concords, demonstratives (pronouns) and particles. Prime criteria for detecting and tagging nouns will naturally be based on class prefixes and nominal concords and to a limited extent on nominal suffixes such as the locative -ng.

2.2 Verb System: Productivity in Morphology

In the case of verbs, numerous derivations of a single verb stem exist, consisting of the root, plus one or more prefix(es) and/or suffix(es), as is clearly indicated in Table 3, which reflects a subsection (five out of eighteen modules, cf. Prinsloo [1994]) of the suffixes and combinations of suffixes for the verb stem *reka* 'buy.' The complexity of this layout is evident.

Verbal derivations such as those in the rightmost column of Table 3 can all simply be tagged as verbs, or, alternatively, first be morphologically analysed (cf. Taljard & Bosch 2005) and then tagged in terms of their specific verbal suffixes, cf. column 2 versus column 3 in Table 4 with respect to the suffixal cluster 02 ANA in Table 3.

Table 3: Selection of Derivations of the Verb *reka*

MODULE NUMBER AND MARKER	MODULE COMPOSITION	ABBREVIATIONS	STEMS AND DERIVATIONS
01	root + standard modifications	VR	reka
	(Per = Perfect tense)	VRPer	rekile
	(Pas = Passive)	VRPas	rekwa
		VR PerPas	rekilwe
02 ANA	root + reciprocal + standard modifications	VRRec	rekana
		VRRecPer	rekane
		VRRecPas	rekanwa
		VRRecPerPas	rekanwe
03 ANTŠHA	root + reciprocal + causative + standard modifications	VRRecCau	rekantšha
		VRRecCauPer	rekantšhitše
		VRRecCauPas	rekantšhwa
		VRRecCauPerPas	rekantšhitšwe
04 ANYA	root + alternative causative + standard modifications	VRAlt-Cau	rekanya
		VRAlt-CauPer	rekantše
		VRAlt-CauPas	rekanywa
		VRAlt-CauPerPas	rekantšwe
05 EGA	root + neutro passive + standard modifications	VRNeu-Pas	rekega
		VRNeu-PasPer	rekegile
		VRPas	
		VRPerPas	

Table 4: Alternatives in Tagging the Verb *reka*

02 ANA	<i>rekana</i> 'V'	<i>rek</i> 'Vroot' an 'Rec' <i>a</i>
	<i>rekane</i> 'V'	<i>rek</i> 'Vroot' an 'Rec' <i>e</i> 'Per'
	<i>rekanwa</i> 'V'	<i>rek</i> 'Vroot' an 'Rec' <i>w</i> 'Pas' <i>a</i>
	<i>rekanwe</i> 'V'	<i>rek</i> 'Vroot' an 'Rec' <i>w</i> 'Pas' <i>e</i> 'Per'

2.3 Quantitative Aspects of the Lexicon

There are a few marked tendencies in the quantitative distribution of lexical items in Northern Sotho, especially with respect to the relationship between frequency of use and ambiguity.

In our 43,000 word corpus sample, we counted types and tokens, distinguishing nouns, verbs and closed class items. In Northern Sotho, only nouns and verbs allow for productive word formation (i.e., are open word classes), whereas function words, adverbs and adjectives are listed (i.e., belong to closed classes). Note that we did not consider numerals at all; the figures given are to be taken as tendencies. We separately counted forms that can be unambiguously identified as nouns, verbs or elements of one of the closed classes, as opposed to ambiguous forms where more than one word class can be assigned, depending on the context.

All three have many more unambiguous types than ambiguous ones. As is likely in most languages, however, high frequency items are also highly ambiguous (cf. Table 5 below). Nevertheless, if only slightly more than half of the potential verb occurrences in the sample are unambiguous (ca. 5000 tokens), the percentage of unambiguous occurrences of noun candidates is as high as 90% (5800 out of 6300 tokens). Ambiguity with nouns is restricted to rather infrequent items. For closed class items, however, the inverse situation is observed: only little more than 20% of the occurrences of closed class items in our sample are unambiguous, and a small set of closed class item types (88 types), of an average frequency of two hundred or more, constitutes about 40% of the total amount of word forms in the sample. We expect that this distribution will be more or less generalisable to larger data sets of Northern Sotho. It will have an incidence on our approach to the bootstrapping of linguistic resources for this language. Table 5 lists the most frequent (and at the same time most ambiguous) items from the 43,000 word corpus sample with their tags (according to the tagset described in section 3.2) and their absolute frequency in the sample.

Table 5: Most Frequent and Most Ambiguous Items in the Sample

Item	Possible Tags	Freq.
a	CDEM6:CO6:CS1:CS6:CPOSS1:CPOSS6:QUE:PRES	2261
go	CO2psg:CO15:CO17:CS15:CS17:CSindef:PALOC	2075
ka	CS1psg:PAINS:PATEMP:PALOC:POSSPRO1psg	1807
le	CDEM5:CO2ppl:CO5:CS2ppl:CS5:PACON:VCOP	1615
ba	AUX:CDEM2:CO2:CS2:CPOSS2:VCOP	1429
o	CO3:CS1:CS2psg:CS3	1192
ke	AUX:CS1psg:PAAGEN:PACOP	1107

3. Elements of a Scenario for Resource Building for Northern Sotho

3.1 Starting Point and Objectives

For computational lexicography, a sufficiently large corpus is needed, annotated at least at the level of part-of-speech. For the development of automatic tools for syntactic analysis, a more detailed annotation is required. In this paper we concentrate on a step prior to both of these resources, that is, on the creation of smaller, but generic resources to enable part of speech tagging.

Tagset design for Northern Sotho is based on distinctions in traditional Northern Sotho grammar, it is carried out with a view to the kinds of information that would be extracted from a corpus once it has been tagged. As statistical tagging can only be attempted when a sufficiently large training corpus is available, an adaptation of the tagset is likely to be needed when the automatic tagging is tested, since some distinctions from the grammar may not be identifiable in texts without deeper knowledge.

In working towards an annotated training corpus, different procedures are possible in principle: one could manually annotate a significant amount of data, or one could opt for a mixed approach, where certain parts of the corpus would receive manual annotation, and others would be annotated in a semi-automatic fashion, where the results of an automatic pre-classification are manually corrected. Due to the morphological and distributional properties of Northern Sotho discussed in section 2, the following breakdown was chosen:

- Closed class items, as well as other words of very high frequency, were introduced manually to the tagger lexicon, with a disjunctive tag annotation that indicates for each item all its possible tags (Table 5);

- Nouns and verbs can be guessed in the text on the basis of their morphological properties; thus, separate rule-based guessers were developed, and their results were manually corrected in the training corpus; and,
- The disambiguation of closed-class items in context is, to a considerable extent, possible on the basis of rules similar to 'local grammars.' A certain amount of ambiguities in the training corpus have to be dealt with manually.

In the remainder of this section, we report on tagset design (section 3.2); on an architecture for the creation of a tagger lexicon and a training corpus (section 3.3); and, on verb and noun guessing and the disambiguation of closed class items (sections 3.4 to 3.6).

3.2 Tagset Design

The tagset designed for Northern Sotho is organised as a logical tagset (similar to a type hierarchy); this opens up the possibility to formulate underspecified queries to the corpus.

The tagset mirrors some of the linguistic specificities of Northern Sotho, but is also conditioned by considerations of automatic processability with a statistical tagger. The tagset reflects properties of the nominal system of classes and concords: as they are (mostly) lexically distinct, we introduced class-based subtypes for nouns, pronouns and concords, as well as for adjectives: N, ADJ, C (for concord) and PRO (for pronoun) have such subtypes. As concords and pronouns have functionally and/or semantically defined subtypes, we apply the class-based subdivision in fact to the types listed in Table 6:

Table 6: Nominal Categories that have Class-related Subtypes

N	Nouns	CPOSS	possessive concords
ADJ	adjectives	EMPRO	emphatic pronouns
CS	subject concords	POSSPRO	possessive pronouns
CO	object concords	QUANTPRO	quantifying pronouns
CDEM	demonstrative concords		

Given the complexity of the system of verbal derivation (cf. Table 3 above), an attempt to subclassify verbal forms accordingly would have led to an amount of tags (i.e., of distinctions) that would not be manageable with a statistical tagger. Furthermore, as- according to Northern Sotho orthography conventions- concords, adjectives and pronouns are written separately from the nouns and verbs to which they are grammatically related (disjunctive writing), these elements receive their

own tags. Since verbal derivation is written conjunctively (like word formation in European languages), a single ‘verb’ tag (V) proved sufficient (cf. Table 4). As with parts of tense morphology and with word formation in European languages, an analysis of Northern Sotho verbal derivations is left to a separate tool (e.g. to a morphological analyser; see the discussion in Taljard & Bosch 2005).

Other tags cover invariable lexical items:

- adverbs (ADV) and numerals (NUM);
- tense/mood/aspect markers for present tense (PRES), future (FUT), and progressive (PROG);
- auxiliaries (AUX) and copulative verbs (VCOP);
- ideophones (IDEO); and,
- different (semantically defined) kinds of particles that mark a hortative (HORT), questions (QUE), as well as agentive (PAAGEN), connective (PACON), copulative (PACOP), instrumental (PAINS), locative (PALOC) and temporal (PATEMP) constructs.

In principle, our approach to the design of tagsets for nouns and verbs is similar to the one of Van Rooy and Pretorius (2003) for Setswana, but it is much less complex. In the case of verbs we agree on the allocation of a single tag for verb stem plus suffix(es) as well as on separate tags for verbal prefixes:

“[...] verbs are preceded by a number of prefixes, which are regarded as separate tokens for the purposes of tagging. The verb stem, containing the root and a number of suffixes (as well as the reflexive prefix) receives a single tag.” (Van Rooy & Pretorius 2003:211)

Likewise, for nouns, we are in agreement that at this stage in the development of tagsets, certain subclassifications such as the separate identification of deverbatives should be excluded (cf. Van Rooy & Pretorius 2003:210). Our approach differs from Van Rooy and Pretorius among others, in that a much smaller tagset is compiled for both verbs and nouns. In the case of verbs, we do not consider modal categories, and in the case of nouns, we honour subclasses but not divisions in terms of relational nouns and proper names. Consider the following examples illustrating basic differences in terms of the approaches as well as of the complexity of the tags:

(1) Nouns

a) *Mosadi* ‘woman’

Tswana (Van Rooy & Pretorius 2003:217):

Tag category: Common noun, singular; Label: NC1; Intermediate tagset: N101001

Northern Sotho: Noun; Tag: N1

b) *Bomalome* ‘uncles’

Tswana (Van Rooy & Pretorius 2003:217):

Tag category: Relational noun, plural; Label: NR2; Intermediate tagset: N302001

Northern Sotho: Noun; Tag: N2

(2) Verbs

Tswana: *kwala/kwalwa/kwadile*; Northern Sotho: *ngwala/ngwalwa/ngwadile* ‘write/be written/wrote’

Tswana (Van Rooy & Pretorius 2003:219):

Tag category: Lexical verb, indicative, present, active; Label: VIOPA; Intermediate tagset: V000111102000 *kwala*

Tag category: Lexical verb, indicative, present, passive; Label: VIOPP; Intermediate tagset: V0001112102000 *kwalwa*

Tag category: Lexical verb, indicative, past, active; Label: VIODA; Intermediate tagset: V0001141102000 *kwadile*

Northern Sotho: verb; Tag: V

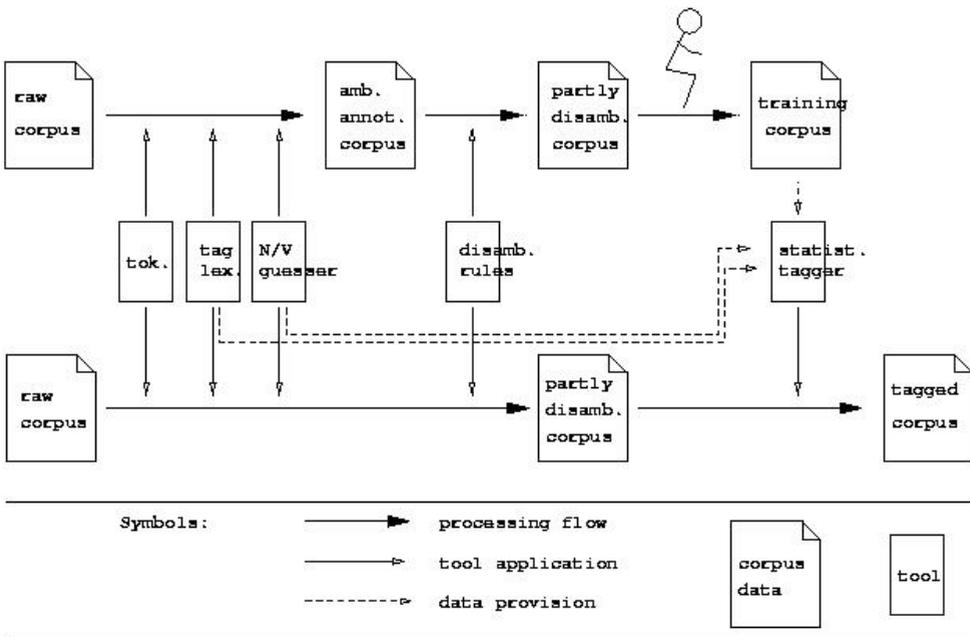
3.3 An Architecture for Parallel Resource Building

Since we opted, as far as POS-tagging is concerned, for an attempt to apply Schmid’s (1994) statistical TreeTagger to Northern Sotho, both a tagger lexicon and a reference corpus for training were needed. Schmid’s TreeTagger was chosen, because it needs much less manually annotated training material than other statistical taggers. For European languages (German, French, English, Dutch, and Italian) training corpora of 40,000 to 100,000 words have proven sufficient to obtain the 96-97% tagging rate that is standard in current applications. Tagging quality of the TreeTagger also depends upon the number of different tags and on the size of the tagger lexicon. It thus seems obvious to bootstrap lexicon and corpus in parallel.

Given the grammatical and distributional properties of Northern Sotho, we opted for the overall approach as sketched above in section 3.1: a list of closed class items and their possible tags is created manually, whereas nouns and verbs are guessed on the basis of morphological rules, and closed class item disambiguation is performed semi-automatically, based on rules, and possibly also on frequency-based heuristics.

Figure 1 shows the strands of corpus annotation, where the (upper) strand leading to the training corpus is meant to be carried out once, whereas the general strand (below) can be repeated for each newly acquired corpus.

Figure 1: Strands of Corpus Annotation



The workflow involves a number of modular tools (developed in the course of the preparation of the training corpus) that can be reused with any additional Northern Sotho corpus. These include a sentence tokenizer; the tagger lexicon and a tool to project its contents (i.e., potentially ambiguous annotations for individual word forms) against the corpus words; guessers for nouns and verbs; and, disambiguation rules for closed class item disambiguation in context.

The procedure sketched here, and depicted in Figure 1, is in fact a combination of rule-based symbolic tagging and statistical tagging, whereby a number of ambiguities are solved by the rule-based component before the statistical tagger is used. This setup is similar to Klatt's (2005) work on a corpus processing suite for German.

3.4 Verb Guesser

In Table 3 above, a few selected examples of derived verb forms of Northern Sotho are given. Except for very frequent forms of a few verbs, most verb forms are marked by unambiguous derivational and inflectional affixes. For example, a word form found in a corpus that ends in *-antšwe* will almost inevitably be a verb form (cf. *rekantšwe* in Table 3).

Consequently, many verb forms can be identified by simple pattern matching. Based on the grammatical system of verb affixation sketched in Prinsloo (1994), we developed a verb form guesser. It compares each candidate form with a list of unambiguous verbal affixes to distinguish verb forms from forms of other categories. Given the productivity of verbal derivation in Northern Sotho (cf. section 2 above), this guesser will be needed on any new corpus of Northern Sotho to be annotated. If required, the grammatical information encoded in the verbal affixes can be made explicit in the annotation (cf. Table 4 above).

3.5 Noun Guesser

Suffixal derivation appears in nouns only to denote locatives, augmentatives/feminins and diminutives. Given the low frequency of these derivations, with the possible exception of the locative, a noun detection strategy based on pattern matching alone, in analogy to that of the verb guesser, will have low recall, even though its precision will be very high.

But nouns are characterised by their class prefixes (cf. Table 1 above); prefixes of classes 1 to 10 indicate singular (classes 1,3,5,7 and 9) versus plural (classes 2,4,6,8 and 10). The prefixes are not, however, unambiguous with respect to classes (*mo-*: class 1,3 and, less relevant, class 18; *di-*: class 8 and 10; etc.). Not all words starting with a syllable that can be a noun prefix are indeed nouns (cf. e.g. the verb form *letetše* 'wait(ed) for' where the first syllable *le-* is not the prefix of class 5).

What is indeed a highly unambiguous indicator of a noun form is its syntagmatic environment, as well as the alternation pattern between singular and plural. Very often, nouns are accompanied by concords or adjectives, as illustrated by the example in Table 2, where the noun *monna* is followed by a demonstrative and a subject concord, both of which show agreement with the noun with respect to the class. Adjectives also show this agreement.

We exploit this regularity in our noun form guesser as follows: to identify items of a given pair of singular/plural classes, we apply word sequence patterns to the corpus data, which rely on the presence of concords, pronouns, adjectives, and so

forth in the neighbourhood of the noun candidates. We check for the existence of such patterns in parallel for singular forms and for their potential plural counterparts. The search is approximative, in so far as it checks the presence of agreement-bearing elements within a window of up to three words left or right of the noun candidate. The rules can, in principle, be triggered either by singular or by plural items (with the exception of class 9 versus class 10, where it is preferable to start from the plural).

Table 7 contains an example of a noun guessing query (simplified, as many potential agreement-bearing indicator items are left out), formulated in the notation of the CQP corpus query language, which underlies the CWB Corpus WorkBench, (Christ *et al.* 1999), used in our experiments as a corpus representation and infrastructure. We indicate (parts of) the queries that extract nouns of classes 7 (and 8).

Table 7: Sample Query for the Identification of Noun Candidates of Classes 7 + 8

(First part of query:
[word = 'sego selo sebatakgomo ...	candidate <i>se-</i> words
setšhaba seatla sello']	as a disjunction;
[][0,2}	followed in distance 0 to 2
[word = 'sa se segolo	by class-7-indicators noted as a
sekhwi sengwe seo sona ... ']	disjunction
)	or (second part of query):
([word = 'sa se segolo	choice of indicators
...]	
[][0,2}	followed in distance 0 to 2
[word =	by candidate words
'sego selo sebatakgomo ...']	
);	
(....)	analogous procedure for
	noun candidates created
	by replacement of <i>se-</i>
	with <i>di-</i> (plus class 8 concords)

When applied to the 43,000 words corpus sample, the query throws up, among others, the results displayed in Table 8.

Table 8: Sample Results of Noun Guessing for Classes 7 and 8

Class 7 cand.	Class 8 cand.	N?	Equivalent(s)
selo	dilo	+	thing, things
setšhaba	ditšhaba	+	nation, nations
sello	dillo	+	(out)cry, outcries
sepetše	*dipetše	–	walked
sekelela	dikelela	–	recommend, disappear

The checking tool is robust towards inexistent forms (cf. **dipetše*) and towards forms that are not nominal (due to the context constraint on agreement-bearing items, (cf. *sekelela* versus *dikelela*).

A first qualitative evaluation of the noun guessing routines on all candidates from the 43,000 word corpus sample seems to suggest that the tool only fails on lexicalized irregular forms (e.g. *mong - beng*, ‘owner(s)’, instead of the hypothetical *mong - *bang*), and on nouns that, mostly due to semantic reasons, do not have both a singular and a plural form (such as *Sepedi* ‘Pedi language and culture’, or *leboa* ‘North’). As for the verb guesser, the noun guesser can be and has to be applied (for quantitative reasons) to any new corpus to be annotated.

3.6 Rules for the Disambiguation of Closed Class Items

Given the high degree of ambiguity in closed class items (see section 2.3), there is a major need for disambiguation strategies for these items. Even though a statistical tagger is designed for this type of disambiguation, a rule-based preprocessing, leading at least to a partial reduction of ambiguity, seems necessary.

We use context-based disambiguation rules, in the spirit of Gross and Silberztein’s local grammars (Silberztein 1993) and of rule-based tagging. As with the noun guessing queries, disambiguation rules are implemented as queries in the format of the CQP language. Some extraction rules exclusively rely on lexical contexts (cf. the topmost part of Table 9), while others involve lexemes and word class tagged items (middle row), or a combination of lexical, categorical and morphological constraints (including, for example, the presence of certain affixes [cf. lower part of Table 9]). The examples in Table 9 all relate to the disambiguation of the form *a*, the most frequent and most ambiguous item in our sample (cf. Table 5).

Table 9: Examples of Disambiguation Queries for the Form *a*

'o O' 'be' 'a'	Sequence of <i>o be a</i> ('he/she was') Hypothesis: <i>a</i> : CS1 Coverage: 109 instances Precision: 109 (100%)
[pos = 'N.{1,2}'] 'a' [pos = 'N.{1,2}'];	<i>a</i> between two nouns ('of': possessive) Hypothesis: <i>a</i> : CPOSS6 Coverage: 42 instances Precision: 42 (100%)
'a'	<i>a</i> preceding a verb form ending in <i>-go</i> (Relative Marker)
[pos = 'V' & word = '.*go'];	Hypothesis: <i>a</i> : CS1 or CS6 or C06 Coverage: 75 instances 63 (80,8%) CS1; 9 (15.4%) CS6; 3 (3.8%) C06.

The examples show that some rules do not fully disambiguate, but leave a set of options. Since we use the rules as a preparatory step to statistical tagging (and to manual disambiguation in the preparation of the training corpus), partial disambiguation is still useful to reduce the effort needed at a subsequent stage (cf. the third example of Table 9, where the choice of eight tags for *a* is reduced to a four-way ambiguity).

4. Methodological Considerations

4.1 Sequencing of Processing Steps

We use semi-automatic procedures to create tagging resources for Northern Sotho. As raw corpora are the only available input, a first step in the project is to define a tagset that underlies all subsequent work (cf. section 3.2).

The creation of the tagger lexicon and the annotation of the training corpus mostly run in parallel. We classify word forms from the corpus, store their (possibly disjunctive) description in the lexicon, and annotate them at the same time in the upcoming training corpus. (We annotate each word form in the corpus with the respective entry from the tagger lexicon.) While the disjunctive annotations remain in the tagger lexicon, context-based rules are used to partly disambiguate the corpus occurrences (cf. section 3.4 and 3.5).

To get the process started efficiently, we first manually annotated the thousand most frequent word forms in the corpus, aiming at a complete coverage of their

potential word class features. This information can be provided easily on the basis of Northern Sotho grammar, as many of them are function words.

Subsequently, we employed semi-automatic procedures (automatic pre-classification of data, followed by manual verification) that focus on high precision, allowing, at the same time, for efficient data production: we capitalised on unambiguous verb and noun forms, covering thereby more than one fourth of all corpus occurrences (tokens), and obtaining in parallel a stock of approximately 2800 additional entries of the tagger lexicon (word form types).

Once nouns and verbs were annotated, disambiguation rules for closed class items were formulated (based on regularities of the Northern Sotho grammar) and applied; many of these contextual constraints involve verbs and nouns. The rules are ordered by specificity: as in many other NLP applications, the most specific cases are handled first; at the end of the cascade, more general rules are applied, which may also be less 'safe' and less effective, that is, have less precision and/or recall.

In conclusion, the strategy may be characterised as 'easy-first' and 'safety-first': for example, as disambiguation rules cannot overwrite (previously verified) lexical data, the overall process is one of monotonic accumulation of information. A bootstrapping procedure proved most efficient, where the validated results of each of the above-mentioned steps are persistently represented in both corpus and lexicon, such that they are available as input for the subsequent steps.

4.2 Reusability of the Created Resources

As mentioned in section 3, our verb and noun guessers can be applied to other Northern Sotho corpora, as can the tool projecting lexical descriptions onto corpus word forms. Given the productivity of verbal derivation and the amount of nouns to be expected in larger corpora, we assume that both tools will prove useful in the preparation of an annotated version of the PSC. Moreover, even though statistical taggers are designed to both disambiguate in context and guess word class values for unknown words (i.e., those not contained in the system's lexicon), reducing the amount of the latter may improve overall output quality.

Obviously, the parallel growth of lexicon and corpus will continue when larger corpora will be treated. At a later stage, we envisage the parallel enhancement of both resources, not only in coverage, but also with respect to the degree of detail covered: some morphological details of nouns (locatives, feminins/augmentatives, and diminutives) and verbs (cf. Tables 3 and 4) can be identified, but are not yet accounted for in our resources. Thus, the tagger lexicon may become part of an NLP-

oriented dictionary that would explicitly store such properties. As far as the corpus is concerned, a multilevel annotation would be more appropriate than the current monodimensional view: without changes to the current annotation, extra layers may be added for the above-mentioned features of nouns and verbs, but also for an appropriate treatment of fused forms (cf. *dirang*, ‘do what?’ from *dira* + *eng*) and of multiword items, for example, idiomatic expressions (cf. *bona kgwedi* ‘see the moon’ i.e., ‘menstruate’). As Northern Sotho orthography is not yet fully standardised, a distinction between standard orthography and observed (possibly deviant) orthography may be introduced through additional layers.

5. Conclusions and Future Work

We reported on an ongoing research and development project for the creation of tagging resources for Northern Sotho. In this context, modular components of a two-layered architecture were created, which are needed in the first place for the preparation of a training corpus for statistical tagging, but which will prove equally useful, we hope, for the later development of larger corpora.

We bootstrap the training corpus and the tagger lexicon in parallel, using semi-automatic procedures consisting of a rule-based automatic pre-classification and subsequent manual validation: the procedures concern the identification of verbal and nominal forms and the disambiguation of closed class items. These procedures are applied one after the other by order of their expected precision (‘easy-first’, ‘safety-first’), leading thereby to a partly disambiguated corpus. For the creation of the training corpus, the remaining ambiguities are removed manually, whereas this task is supposed to be left to the statistical tagger in the later creation of larger corpora.

Linguistic knowledge about the language is extensively used in the definition of the automatic procedures: morphological and morpho-syntactic regularities in the local context provide the starting point for their formulation.

Future work on the tools described in this paper will be devoted to the development of further disambiguation rules, to the finalisation of a fully disambiguated training corpus, and to tagger training and tests. This will allow us to (i) assess tagging quality as obtained by the use of the statistical tagger only in a setup with our rule-based pre-processing, (ii) to stabilise the proposed tagset on the basis of experience with statistical tagging, and (iii) to undertake tagging of the PSC, which could then serve for lexicographic exploration.

A well-designed POS-tagger for Northern Sotho would provide a flying start to the development of similar taggers for the other Sotho languages, the Nguni languages,

and Bantu languages in general. It is expected that only minor adjustments will be required to adapt a POS-tagger for Northern Sotho to the other two Sotho languages (Tswana and Southern Sotho) because these languages are closely related. Pending certain morphological parsing for the Nguni languages (i.e., Zulu, Xhosa, Swazi and Ndebele) the tagger will be equally usable, since these languages do not differ structurally from the Sotho languages. It could finally be extended to other Bantu languages, since Bantu languages in general have a common structure.

6. Acknowledgements

This work was carried out as a joint project between the Department of African languages of the University of Pretoria and the Institut für maschinelle Sprachverarbeitung of Universität Stuttgart. We would like to thank Elsabé Taljard (Pretoria) who contributed to the design of the tagset and who cross-checked a large proportion of the output of our tools. Furthermore, we would like to thank Gertrud Faaß (Stuttgart) for her invaluable help with the implementation of the noun and verb guessers and of the tagging support tools.

References

Christ, O., Schulze, B.M. & König, E. (1999). *Corpus Query Processor (CQP). User's Manual*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart Stuttgart, Germany.

<http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>.

De Schryver, G.M. & Prinsloo, D.J. (2000). "The Compilation of Electronic Corpora, with Special Reference to the African Languages." *Southern African Linguistics and Applied Language Studies* 18(1-4), 89-106.

De Schryver, G.-M. & Prinsloo, D.J. (2000a). "Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1. The *Macrostructure*." *South African Journal of African Languages* 20(4), 291-309.

De Schryver, G.-M. & Prinsloo, D.J. (2000b). "Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 2: The *Microstructure*." *South African Journal of African Languages* 20(4), 310-30.

Guthrie, M. (1971). *Comparative Bantu: an Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages. Vol. 2: The Comparative Linguistics of the Bantu Languages*, London: Gregg Press.

Klatt, S. (2005). *Textanalyseverfahren für die Korpusannotation und Informationsextraktion*. Aachen: Shaker.

Lombard, D.P., Van Wyk, E.B. & Mokgokong, P.C. (1985). *Introduction to the Grammar of Northern Sotho*. Pretoria: J.L. van Schaik.

Louwrens, L.J. (1991). *Aspects of Northern Sotho Grammar*. Pretoria: Via Afrika Limited.

Matsepe, O.K. (1974). *Tša ka mafuri*. Pretoria: Van Schaik.

Poulos, G. & Louwrens, L.J. (1994). *A Linguistic Analysis of Northern Sotho*. Pretoria: Via Afrika Limited.

Prinsloo, D.J. (1994). "Lemmatization of Verbs in Northern Sotho." *SA Journal of African Languages* 14(2), 93-102.

Prinsloo, D.J. (1991). "Towards Computer-assisted Word Frequency Studies in Northern Sotho." *SA Journal of African Languages* 11(2).

Prinsloo, D.J. & de Schryver, G.-M. (2001). "Monitoring the Stability of a Growing Organic Corpus, with Special Reference to Sepedi and Xitsonga." *Dictionaries: Journal of The Dictionary Society of North America* 22, 85-129.

Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees." *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, 44-49.

Taljard, E. & Bosch, S.E. (this volume). "A Comparison of Approaches Towards Word Class Tagging: Disjunctively versus Conjunctively Written Bantu Languages", 117-131.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson.

Van Rooy, B. & Pretorius, R. (2003). "A Word-Class Tagset for Setswana." *Southern African Linguistics and Applied Language Studies* 21(4), 203-222.

A Comparison of Approaches to Word Class Tagging: Distinctively Versus Conjunctively Written Bantu Languages

Elsabé Taljard and Sonja E. Bosch

Northern Sotho and Zulu are two South African Bantu languages that make use of different writing systems, namely, a disjunctive and a conjunctive writing system, respectively. In this paper, it is argued that the different orthographic systems obscure the morphological similarities, and that these systems impact directly on word class tagging for the two languages. It is illustrated that not only different approaches are needed for word class tagging, but also that the sequencing of tasks is, to a large extent, determined by the difference in writing systems.

1. Introduction

The aim of this paper is to draw a comparison of approaches towards word class tagging in two orthographically distinct Bantu languages. The disjunctive versus conjunctive writing systems in the South African Bantu languages have direct implications for word class tagging. For the purposes of this discussion we selected Northern Sotho to represent the disjunctive writing system, and Zulu as an example of a conjunctively written language. These two languages, which belong to the South-Eastern zone of Bantu languages, are two of the eleven official languages of South Africa. Northern Sotho and Zulu are spoken by approximately 4.2 and 10.6 million mother-tongued speakers, respectively. Both these languages belong to a larger grouping of languages, that is, the Sotho and Nguni language groups, respectively. Languages belonging to the same language group are closely related, and to a large extent, mutually intelligible. Furthermore, since all three languages belonging to the Sotho group follow the disjunctive method of writing, the methodology utilised for part-of-speech tagging in Northern Sotho would to a large extent be applicable to the other two Sotho languages (Southern Sotho and Tswana) as well. The same holds true for Zulu with regard to the other Nguni languages (i.e., Xhosa, Swati and Ndebele), which are also conjunctively written languages. The South African Bantu languages are not yet fully standardised with regard to orthography, terminology and spelling rules, and, when compared to European languages, these languages cannot boast a wealth of linguistic resources. A limited number of grammar books and dictionaries

are available for these languages, while computational resources are even scarcer. In terms of natural language processing, the Bantu languages, in general, undoubtedly belong to the lesser-studied languages of the world.

In this paper, a concise overview is first given of the relevant Bantu morphology, and reference is made to the differing orthographical conventions. In the subsequent section, the available linguistic and computational resources for the two languages are compared, followed by a comparison between the approaches towards word class tagging for Northern Sotho and Zulu. In conclusion, future work regarding word class tagging for Bantu languages is discussed.

2. Bantu Morphology and Orthography

According to Poulos & Louwrens (1994:4), "there are numerous similarities that can be seen in the structure (i.e., morphology), as well as the syntax of words and word categories, in the various languages of this family." These languages are basically agglutinating in nature, since prefixes and suffixes are used extensively in word formation.

The focus in this concise discussion on aspects of Bantu morphology is on the two basic morphological systems: the noun class system, and the resulting system of concordial agreement.

2.1 Noun Classes and Concordial Agreement System

The noun class system classifies nouns into a number of noun classes, as signalled by prefixal morphemes also known as noun prefixes. For ease of analysis, these noun prefixes have been divided into classes with numbers by historical Bantu linguists, and represent an internationally accepted numbering system. In general, noun prefixes indicate number, with the uneven class numbers designating singular and the corresponding even class numbers designating plural. The following are examples of Meinhof's (1932:48) numbering system of some of the noun class prefixes:

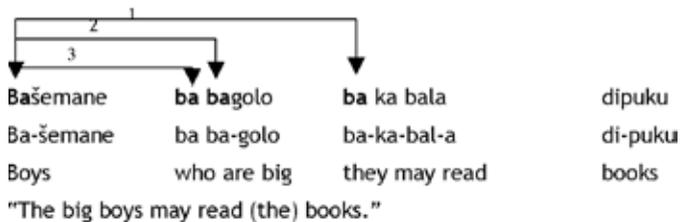
Table 1: Noun Class System: An Illustrative Excerpt

Class #	Northern Sotho	Zulu		
	Prefix	Example	Prefix	Example
1 (sg)	mo-	motho "person"	umu-	umuntu „person"
2 (pl)	ba-	batho "persons"	aba-	abantu "persons"
1a(sg)	Ø-	makgolo "grandmother"	u-	udokotela "doctor"
2b(pl)	bo-	bomakgolo "grandmothers"	o-	odokotela "doctors"
3 (sg)	mo-	mohlare "tree"	umu-	umuthi "tree"
4 (pl)	me-	mehlare "trees"	imi-	imithi "trees"
7 (sg)	se-	setulo "chair"	isi-	isitsha "dish"
8 (pl)	di-	ditulo "chairs"	izi-	izitsha "dishes"
14	bo-	botho "humanity"	ubu-	ubuntu "humanity"

The correspondence between singular and plural classes is not, however, perfectly regular, since some nouns in so-called plural classes do not have a singular form; in Zulu, class 11 nouns take their plurals from class 10, while a class such as 14 is not associated with number.

The significance of noun prefixes is not limited to the role they play in indicating the classes to which the different nouns belong. In fact, noun prefixes play a further important role in the morphological structure of the Bantu languages, in that they link the noun to other words in the sentence. This linking is manifested in a system of concordial agreement, which is the pivotal constituent of the whole sentence structure, and governs grammatical agreement in verbs, adjectives, possessives, pronouns, and so forth. The concordial morphemes are derived from the noun prefixes and usually bear a close resemblance to the noun prefixes, as illustrated by the bold printed morphemes in the following Northern Sotho example:

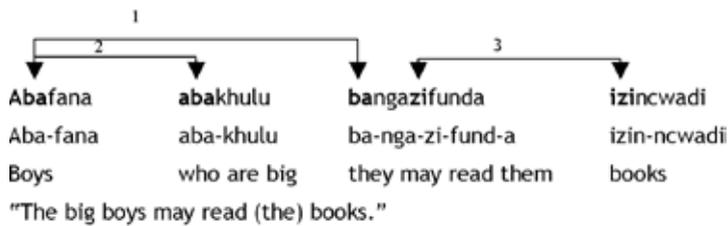
Figure 1: Concordial Agreement - Northern Sotho



In this sentence, three structural relationships can be identified. The class 2 noun *bašemane* 'boys' governs the subject concord *ba-* in the verb *ba ka bala* 'they may read' (1), as well as the class prefix *ba-* in the adjective *bagolo* 'big' (2), and the

demonstrative pronoun *ba*, preceding the adjective. The corresponding Zulu example would be as follows, where (1) indicates subject-verb agreement and (2) is agreement between the noun and the adjective concord *aba-* in the qualificative *abakhulu*. The class 10 noun *izincwadi* 'books' determines concordial agreement of the object concord *-zi-* in the verb (3).

Figure 2: Concordial Agreement - Zulu



The predominantly agglutinating nature of the Bantu languages is clearly illustrated in the above sentences, where each word consists of more than one morpheme. This complex morphological structure will be discussed very briefly by referring to two of the most complex word types, namely nouns and verbs.

2.2 Morphology of Nouns

Nouns as well as verbs in the Bantu languages are constructed by means of the two generally recognised types of morphemes, namely roots and affixes, with the latter subdivided into prefixes and suffixes. The majority of roots are bound morphemes, since they do not constitute words by themselves, but require one or more affixes to complete the word. The root is generally regarded to be "the core element of a word, the part which carries the basic meaning of a word" (Poulos & Msimang, 1996:170). For instance, in the Northern Sotho example *dipuku* 'books', the root that conveys the semantic significance of the word is *-puku* 'book', the morpheme *di-* being the class prefix of class 10. In the Zulu word *izincwadi*, the prefixes are *i-* and *-zin-*, with *-ncwadi* carrying the basic meaning 'book.' By adding the suffixes *-ng* (Northern Sotho) and *-ini* (Zulu), and the prefix *e-* (in the case of Zulu) to the noun, a locative meaning is imparted:

Northern Sotho:	dipukung	di-puku-ng	'in the books'
Zulu:	ezincwadini	e-(i)-zin-ncwadi-ini	'in the books'

2.3 Verbal Morphology

In the case of the verb, the core element that expresses the basic meaning of the word is the verb root. The essential morphemes of a Bantu verb are a subject concord (except in the imperative and infinitive), a verb root, and an inflectional ending. Over and above the subject concord (s.c.), the form of which is determined by the class of the subject noun, a number of other morphemes may be prefixed to a verb root. These include morphemes such as object concords (o.c.), potential and progressive morphemes, as well as negative morphemes. Compare the following example in this regard:

Table 2: Verbal Morphology - Northern Sotho & Zulu

N.S	ba ka di bala	ba	ka	di	bal-	-a
Z	bangazifunda	ba-	-nga-	-zi-	-fund-	-a
	"they can read them"	s.c. cl 2	potential morpheme	o.c. cl 10	Verb	
root	inflectional ending					

It should be noted that whereas object concords also show concordial agreement with the class of the object noun, all other verbal affixes are class independent. Furthermore, verbal affixes have a fixed order in the construction of verb forms, with the object concord prefixed directly to the verb root.

Derivational suffixes may be inserted between the verb root and the inflectional ending. In the following examples, it will be noted that the inflectional ending has changed to the negative *-e/-i* in accordance with the negative prefix *ga-/a-*, for example:

Table 3: Verbal Derivation by Means of Suffixes

N.S.	ga ba rekiše	ga	ba	rek-	-iš-	-e
Z	abathengisi	a-	-ba-	-theng-	-is-	-i
	"they do not sell"	negative morpheme	s.c. cl 2	verb root	suffix	inflectional ending

3. Conjunctive Versus Disjunctive Writing Systems

Following this explanation of the morphological structure of the Bantu languages, a few observations will be made regarding the different writing systems that are followed in the Bantu languages, with specific reference to Northern Sotho and Zulu.

These different writing systems impact directly on POS-tagging, as will be explained below. The following example illustrates the difference in these writing systems:

Table 4: Conjunctivism Versus Disjunctivism

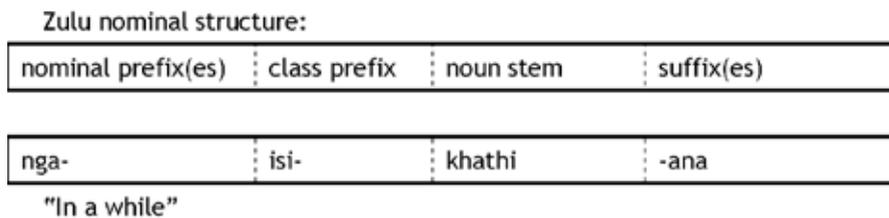
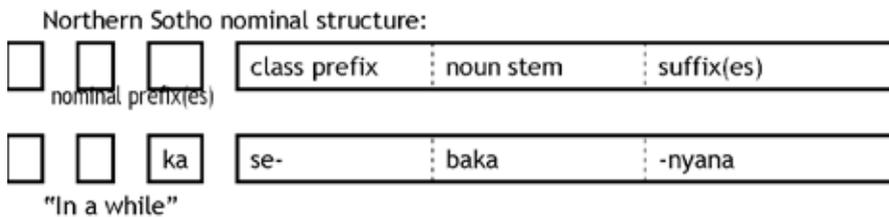
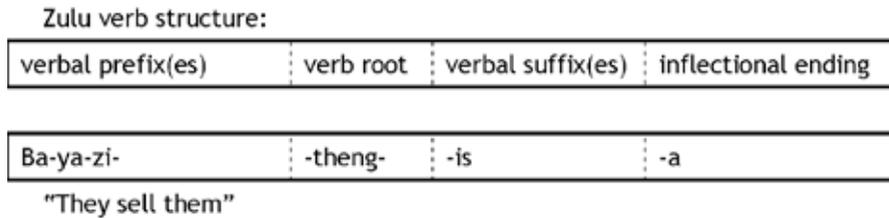
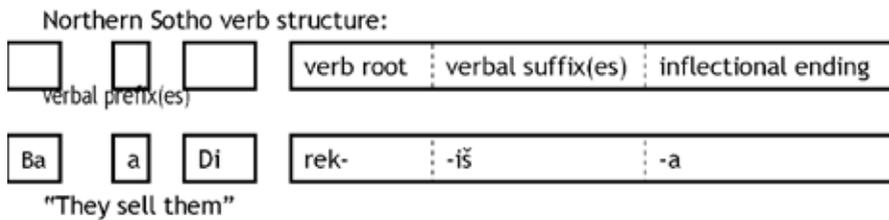
	Orthographical representation	Morphological analysis				
N.S	ke a ba rata	ke	a	ba	rat-	-a
Z	ngiyabathanda	ngi-	-ya-	-ba-	-thand-	-a
	"I like them"	s.c. 1p.sg	PRES	o.c. cl 2	verb root	inflectional ending

The English translation 'I like them' consists of three orthographic words, each of which is also a linguistic word, belonging to a different word category. In the case of the Zulu sentence, where the conjunctive system of writing is adhered to, we observe one orthographic word that corresponds to one linguistic word, which is classified by Zulu linguists as a verb. The orthographic word *ngiyabathanda* is therefore also a linguistic word, belonging to a particular word category. This correspondence between orthographic and linguistic words is a characteristic feature of Zulu that distinguishes it from Northern Sotho. In the disjunctively written Northern Sotho sentence, four orthographic words constitute one linguistic word that is again classified as a verb. In other words, in the latter case, four orthographic elements making up one word category are written as separate orthographic entities.

The reason for the utilisation of different writing systems is based on both historical and phonological considerations. When Northern Sotho and Zulu were first put to writing, mainly by missionaries in the second half of the nineteenth century, they intuitively opted for disjunctivism when writing Northern Sotho, and conjunctivism when writing Zulu. Thus, an orthographic tradition was initiated that prevails even today. Although based on intuition, the decision to adopt either a conjunctive or a disjunctive writing system was probably guided by an underlying realisation that the phonological systems of the two languages necessitated different orthographical systems. As Wilkes (1985:149) points out, the presence of phonological processes such as vowel elision, vowel coalescence and consonantalization in Zulu makes a disjunctive writing system highly impractical: the disjunctive representation of the sentence *Wayesezofika ekhaya* 'He would have arrived at home' as *W a ye s' e zo fika ekhaya* is almost impossible to read and/or to pronounce. In Northern Sotho, these phonological processes are much less prevalent, and, furthermore, most morphemes in this language are syllabic, and therefore pose no problems for disjunctive writing.

What needs to be pointed out at this stage, however, is that there is indeed some overlap with regard to the orthographical systems used by the two languages, and that Northern Sotho and Zulu should rather be viewed as occupying different positions on a continuum ranging from complete conjunctivism to complete disjunctivism. The diagrams below illustrate the degree of overlap between the writing systems of the two languages (dashed lines indicate morphological units, solid lines indicate orthographical units). It can be observed that the disjunctive writing convention in Northern Sotho is mainly applicable to prefixes preceding the class prefix and prefixes preceding the verb root.

Figure 3: Overlap Between Conjunctivism and Disjunctivism



At this stage, it is important to note that the different writing systems utilised by the two languages actually obscure the underlying morphological similarities. These disjunctive versus conjunctive writing systems in the Bantu languages have direct implications for word class tagging, as will be demonstrated later in this paper. In the next section, the available computational resources for the two languages are compared.

4. Computational Linguistic Resources

Existing linguistic and computational resources should be exploited as far as possible in order to facilitate the task of word class tagging. Both languages have unannotated electronic corpora at their disposal - approximately 6.5 million tokens for Northern Sotho, and 5.2 million tokens for Zulu. These corpora were compiled in the Department of African Languages at the University of Pretoria and consist of a mixed genre of texts, including samples of most of the different literary genres, newspaper reports, academic texts, as well as Internet material. Since most of the texts incorporated in the corpora were not available electronically, OCR scanning was done, followed by manual cleaning of scanned material.

The corpora have so far been utilised, among others, for the generation of frequency lists, which are of specific importance for the development and testing of word class tagging, especially in disjunctively written languages. In Northern Sotho, for instance, the top 10,000 types by frequency in the corpus represent approximately 90% of the tokens, whereas in Zulu the top 10,000 types represent only 62% of the tokens. This observation is directly related to the conjunctive versus disjunctive writing systems. Since frequency counts in an unannotated corpus are based on orthographical units, a large orthographic chunk such as *ngiyabathanda* found in Zulu would have a much lower frequency rate than the corresponding units *ke*, *a*, *ba* and *rata* in Northern Sotho. This implies that the correct tagging of the top 10,000 tokens in Northern Sotho, be it manual, automatic, or a combination of both, results in a 90% correctly tagged corpus. The low relation between types versus tokens in Zulu, however, results in a much smaller percentage, that is, only 62% of the corpus being tagged. It furthermore impacts directly on the methodology used for word class tagging in the two languages: the low type/token relationship in Zulu necessitates the use of an additional tool (such as a morphological analyser prototype as described in Pretorius & Bosch 2003) to achieve a higher percentage in the automatic tagging of the Zulu corpus. Let us look at the following examples, which have been analysed by the above-mentioned analyser:

amanzi 'water/that are wet'
a[NPrePre6]ma[BPre6]nzi[NStem]

a[RelConc6]manzi[RelStem]

yimithi 'they are trees'

yi[CopPre]i[NPrePre4]mi[BPre4]thi[NStem]

ngomsebenzi 'with work'

nga[AdvForm]ju[NPrePre3]mu[BPre3]sebenzi[NStem]

bangibona 'they see me'

ba[SC2]ngi[OC1ps]bon[VRoot]a[VerbTerm]

abathunjwa '(they) who are taken captive/they are not taken captive'

aba[RelConc2]thumb[VRoot]w[PassExt]a[VerbTerm4]

a[NegPre]ba[SC2]thumb[VRoot]w[PassExt]a[VerbTerm4]

Examples with more than one analysis exhibit morphological ambiguity that, in most cases, can only be resolved by contextual information. Nevertheless, a morphologically analysed corpus provides useful clues for determining word class tags, since the output of the morphological analysis is a rich source of significant information that facilitates the identification of word classes. For example, the above morphologically analysed words lead to the following information regarding further processing on word class level:

Table 5: Zulu Morphological Analysis and Word Classes

Output of morpho-logical analysis	Word class	Examples
[NPrePre] and/or [BPre] + [NStem] + ...	NOUN	
	amanzi	
[CopPre] + [NStem] + ...	COPULATIVE	yimithi
[SC] + [VRoot] + ... OR [NegPre] + [SC] + [VRoot] + ...	VERB	bangibona abathunjwa
[RelConc] + ...	QUALIFICATIVE	abathunjwa; amanzi
[AdvForm] + ...	ADVERB	ngomsebenzi

Concerning the tags used in the above morphological analysis, it should be noted that “tags were devised that consist of intuitive mnemonic character strings that abbreviate the features they are associated with.” (Pretorius & Bosch 2003:208).

The word class tagset for Zulu is based on the classification by Poulos and Msimang (1996:26). More will be said about this tagset further on in the paper. The features and tags concerned are as follows:

Table 6: Zulu Tags - An Illustrative Excerpt

Tag	Feature
[AdvForm]	Adverbial formative
[BPre6]	Basic prefix class 6
[CopPre]	Copulative prefix
[NegPre]	Negative prefix
[NPrePre6]	Noun preprefix class 6
[NStem]	Noun stem
[OC1ps]	Object concord 1st pers singular
[PassExt]	Passive extension
[RelStem]	Relative stem
[SC2]	Subject concord class 2
[VRoot]	Verb root
[VerbTerm]	Verb terminative

In this paper, it is argued that the difference in writing systems dictates the need for different architectures, specifically for a different sequencing of tasks for POS-tagging in Northern Sotho and Zulu. The approaches followed to implement word class taggers for Northern Sotho and Zulu will be presented in the following section.

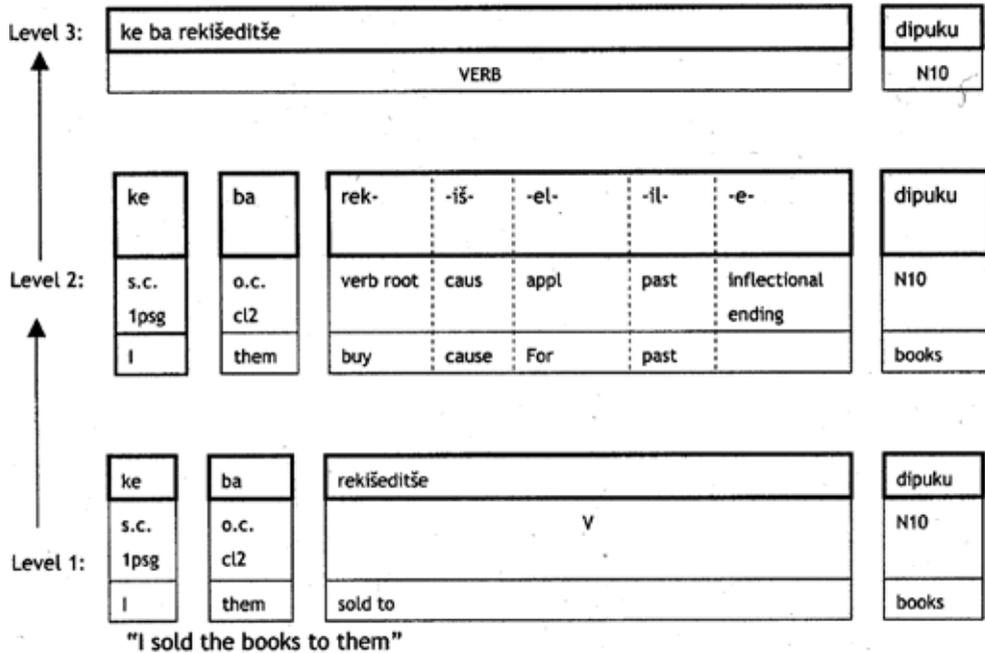
5. Comparison of Approaches Towards Word Class Tagging for Northern Sotho and Zulu

With regard to Northern Sotho, the term POS-tagging is used in a slightly wider sense, following Voutilainen (Mitkov 2003:220) who states that POS-taggers usually produce more information than simply parts of speech. He indicates that the term ‘POS-tagger’ is often regarded as being synonymous with ‘morphological tagger’, ‘word class tagger’ or even ‘lexical tagger.’ POS-tagging for Northern Sotho results in a hybrid system, containing information on both morphological and syntactic aspects, although biased towards morphology. This approach is dictated, at least in part, by the disjunctive method of writing, in which bound morphemes such as verbal prefixes

show up as orthographically distinct units. As a result, in Northern Sotho, orthographic words do not always correspond to linguistic words, which traditionally constitute word classes or parts of speech. Rather than to see this as a disadvantage, it was decided to make use of the morphological information already implicit in the orthography, thus doing morphological tagging in parallel to a more syntactically-oriented word class tagging. It is, therefore, not necessary to develop a tool for the separation of morphemes, since this is largely catered for by the disjunctive orthography of Northern Sotho. As a result, all verbal prefixes can, for example, be tagged by making use of standard tagging technology, even though they are actually bound morphemes belonging to a complex verb form. A further motivation for the tagging of these bound morphemes is the fact that they are grammatical words or function words belonging to closed classes that normally make up a large percentage of any Northern Sotho corpus. Tagging of these forms would therefore result in a large proportion of the corpus being tagged. The decision to tag all orthographically distinct surface forms, regardless of whether these are free or bound morphemes, resulted in a tagset that is somewhat larger than normal: even though only nine word classes are traditionally distinguished for Northern Sotho, the proposed tagset contains thirty-three tag types. This number is further increased by the distinction of class-based subtypes for some of these tag types: the category *EMPRO* (emphatic pronoun) for example, has seventeen subtypes in order to account for the pronouns of the first and second person, as well as those of the different noun classes. The total number of tags comes to 155. (For a full discussion of the tagset design, see Prinsloo & Heid in this volume.)

However, the existence of complex morphological units whose parts are not realised as surface forms necessitates a multi-level annotation. A separate tool such as a morphological analyser would be needed for the analysis of *inter alia* verbal derivations of Northern Sotho. Typical examples that would need to be analysed by such a tool would be verbal suffixes. Such a multi-level approach could be represented as follows:

Figure 4: Multi-level Approach Towards Word Class Tagging



It should be noted that there are cases where the object concord appears within the verbal structure, notably the object concord of the first person singular. This particular object concord distinguishes itself from other object concords in that it is phonologically and orthographically fused to the verbal root. All other object concords are written separately from the verbal root and are thus easily identifiable, except for the object concord of class 1 before verb stems commencing with *b-*, for example, *mo + bona > mmona* 'see him/her.' A procedure similar to the one illustrated above would be needed for these cases.

In the case of Zulu, morphological aspects need not be included in the word class tagging, since these are already accounted for in the morphological analysis. This difference in approach to the tagsets can be mainly ascribed to the different writing systems. The word class tagset for Zulu used for purposes of illustration above is based on the classification by Poulos & Msimang (1996:26) according to which "words which have similar functions, structures and meanings (or significances) would tend to be classified together as members of the same word category [...]" The tagset comprises the following: Noun, Pronoun, Demonstrative, Qualificative, Verb, Copulative, Adverb, Ideophone, Interjection, Conjunction, and Interrogative. It is well known that the

degree of granularity of a tagset should be appropriate to the purposes of the tagged corpus (Allwood *et al.* 2003:230).

The following diagram is a summary of the distinct approaches towards word class tagging as exemplified in the two Bantu languages, Northern Sotho and Zulu. The tasks that need to be performed are similar, but the approaches and sequencing of tasks differ significantly. It is noticeable that, in Northern Sotho, no dedicated tool is needed for the separation of morphemes, since this is already implicit in the disjunctive writing system. The tagger caters to a certain extent for morphophonological rules, but is especially significant for the second level, where morphosyntactic classification of morphemes takes place. Analysis of word formation rules would only need to be done on level II, for which a morphological analyser is needed.

In the case of Zulu, the morphological analyser plays a significant role in levels I and II, where constituent roots and affixes are separated and identified by means of the modelling of two general linguistic components. The morphotactics component contains the word formation rules, which determine the construction of words from the inventory of morphemes (roots and affixes). This component includes the classification of morpheme sequences. The morphophonological alternations component describes the morphophonological changes between lexical and surface levels (cf. Pretorius & Bosch 2003:273-274). Finally, Northern Sotho and Zulu are on a par in level III, where the identification of word classes, associated with the assigning of tags, takes place.

Figure 5: Task Sequencing in Northern Sotho and Zulu

Tasks		Northern Sotho	Zulu
LEVEL III Identification of word classes / categories			
LEVEL II Classification of morpheme sequences		Grammar	
Classification of morphemes	word formation rules	Morphol. analyser	Morphological analyser
	morphosyntactic classification of morphemes	Tagger	
LEVEL II			
LEVEL I Morphophonological rules			
LEVEL I Separation of morphemes		∅	

6. Conclusion and Future Work

In this paper, a comparison of approaches towards word class tagging in two orthographically distinct Bantu languages, namely Northern Sotho and Zulu, was drawn. The disjunctive versus conjunctive writing systems in these two South African Bantu languages have direct implications for word class tagging. Northern Sotho on the one hand resorts to a hybrid system, which contains information on both morphological and syntactic aspects, although biased towards morphology. In the case of Zulu, on the other hand, morphological aspects need not be included in the word class tagging, since these are already accounted for in the morphological analysis. Word class tags for Zulu are associated with syntactic information. The work described in this paper is of crucial importance for pre-processing purposes, not only for automatic word class taggers of Northern Sotho and Zulu, but also for the other languages belonging to the Sotho and Nguni language groups.

Regarding future work, two significant issues have been identified. First, cases of ambiguous annotation require the application of disambiguation rules based mainly on surrounding contexts. A typical example of ambiguity is that of class membership, due to the agreement system prevalent in these languages. For instance, in Northern Sotho as well as Zulu, the class prefix of class 1 nouns is morphologically similar to that of class 3 nouns, that is, *mo-* (N.S) and *umu-* (Z). This similarity makes it impossible to correctly assign class membership of words such as adjectives, which are in concordial agreement with nouns, without taking the context into account. Secondly, the standardisation of tagsets for use in automatic word class taggers of the Bantu languages needs serious attention. A word class tagset based on standards proposed by the Expert Advisory Group on Language Engineering Standards (EAGLES) was recently proposed for Tswana, a Bantu language belonging to the Sotho language group, by Van Rooy & Pretorius (2003). Similarly, Allwood *et al.* (2003) propose a tagset to be used on a corpus of spoken Xhosa, a member of the Nguni language group. In order to ensure standardisation, and therefore achieve reuseability of linguistic resources such as word class tagsets, this initial research on the standardisation of tagsets needs to be extended to all the Bantu languages.

7. Acknowledgements

We would like to thank Uli Heid for unselfishly sharing his knowledge and expertise with us. His comments on an earlier version of this paper added immeasurable value to our effort.

References

Allwood, J., Grönqvist, L. & Hendrikse, A.P. (2003). "Developing a Tagset and Tagger for the African Languages of South Africa with Special Reference to Xhosa." *Southern African Linguistics and Applied Language Studies* 21(4), 223-237.

"Eagles." Online at: <http://www.ilc.cnr.it/EAGLES/home.html>.

Meinhof, C. (1932). *Introduction to the Phonology of the Bantu Languages*. (trad. van Warmelo, N). Berlin: Dietrich Reimer/Ernst Vohsen.

Poulos, G. & Louwrens, L.J. (1994). *A Linguistic Analysis of Northern Sotho*. Pretoria: Via Afrika Limited.

Poulos, G. & Msimang, T. (1996). *A Linguistic Analysis of Zulu*. Pretoria: Via Afrika Limited.

Pretorius, L. & Bosch, S.E. (2003). "Computational Aids for Zulu Natural Language Processing." *Southern African Linguistics and Applied Language Studies* 21(4), 267-82.

Prinsloo, D.J. & Heid, U. (this volume). "Creating Word Class Tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping", 97-115.

Van Rooy, B. & Pretorius, R. (2003). "A Word-class Tagset for Setswana." *Southern African Linguistics and Applied Language Studies* 21(4), 203-222.

Voutilainen, A. (2003). "Part-of-Speech Tagging." Mitkov, R. (ed.)(2003). *The Oxford Handbook of Computational Linguistics*. Oxford University Press: Oxford, 219-232.

Wilkes, A. (1985). "Words and Word Division: A Study of Some Orthographical Problems in the Writing Systems of the Nguni and Sotho Languages." *South African Journal of African Languages* 5(4), 148-153.

Grammar-based Language Technology for the Sámi Languages

Trond Trosterud

Language technology projects are often either commercial (and hence closed for inspection), or small projects that run with no explicit infrastructure. The present article presents the Sámi language technology project in some detail and is our contribution to a concrete discussion on how to run medium-scale, decentralised, open-source language technology projects for minority languages.

1. Introduction

This article presents a practical framework for grammar-based language technologies for minority languages. Such matters are seldom the topic of discussion; one usually goes directly to the scientific results. In order to obtain these results, however, a good project infrastructure is needed. Moreover, for minority languages, the bottleneck is often represented by the lack of human expertise, that is people with a knowledge of the language, linguistics, and language technology. In such situations, we need to organise work in order to facilitate cooperation and avoid duplication of effort. Although the model presented here can hardly be considered the ultimate one, it is the result of accumulated experience gained from different types of projects, commercial, academic and grass-roots Open source, and we hereby present it as a possible source of inspiration.

The Sámi languages make up one of the seven subbranches of the Uralic language family, Finnish and Hungarian being the most well-known members of two of the other sub-branches. From the point of view of typology, the Sámi languages have many properties in common with the other Uralic languages, but several non-segmental morphological processes have entered the languages as well. There are six Sámi literary languages: North, Lule, South, Kildin, Skolt and Inari Sámi. All of them are written with the Latin alphabet (including several additional letters), except Kildin Sámi, which uses the Cyrillic alphabet.

Prior to our project, the main focus within Sámi computing was the localisation issue. Four of the six Sámi languages have letters that are not to be found in the Latin 1 (or Latin 2) code table. At present, this issue is more or less resolved and North Sámi is the language with fewest speakers that at the same time is localised – out of the

box – on all three major operating systems. No other language technology application existed prior to our work.

2. Project Status Quo, Goals and Resources

The work is organised in two projects, with slightly different goals. It started out as a university-based project, with the goal of building a morphological parser and disambiguator for North, Lule and South Sámi, in order to use it for scientific purposes (i.e., creating a tagged corpus with a Web-based graphical interface and using it for syntactic, morphological and lexical research, publishing reverse dictionaries, etc.). In 2003, the Norwegian Sámi parliament asked for advice on how to build a Sámi spellchecker. They considered the construction of this tool as vital for the use of North Sámi as an administrative language. As a result of this, there are now three people working on the University project and four a half people working on the Sámi parliament project. These projects will run with the present financing for another 2 years.

The status quo is that we have a parser with a recall of 80 - 93% on the grammatical analysis of words in running text (modulo genre) and we disambiguate the morphological output with a recall of 99% and a precision of 93%; the outcome is slightly worse for syntactic analysis. The parsers behind these results contain approximately 100 morphophonological rules, 600 continuation lexica and 2000 disambiguation rules.

The figures below show the output for the morphological parser for the sentence *Mii háliidit muitalit dan birra* "We would like to tell about it".

3. Choice of approach

3.1 A Grammatical Versus Statistical Approach

We use a grammar-based, rather than a statistical approach (proponents of the statistical approach often refer to this dichotomy as a choice between a 'symbolic' and a 'stochastic' approach), which means that our parsers rely on a set of grammar-based, manually written rules, that can be inspected and edited by the user. There are several reasons for our choice:

- We think some of the prerequisites for good results with the statistical approach are not present in the Sámi case;
- We want our work to produce grammatical insight, not only functioning programs, and,
- On the whole, we think the grammatical approach is better for our purposes.

Addendum to (1): Good achievements with a statistical approach require both large corpora and a relatively simple morphological structure (low wordform/lemma ratio), as is the case for English. Sámi, as in the case with many other languages, has a rich morphological structure and a paucity of corpus resources, whereas the basic grammatical structure of the language is reasonably well understood.

Addendum to (2): Our work is a joint academic and practical project. Work on minority languages will typically be carried out as cooperative projects between research institutions and in-group individuals or organisations devoted to the strengthening of the languages in question. Whereas private companies will look at the ratio of income to development cost and care less about the developmental philosophy, it is important for research institutions to work with systems that are not 'black boxes', but that are able to give insight into the language beyond merely producing a tagger or a synthetic voice.

Addendum to (3): We are convinced that grammar-based approaches to both parsing and machine translation are superior to the statistical ones. Studies comparing the two approaches, such as Chanod & Tapanainen (1994), support this conclusion.

This does not mean that we rule out statistical approaches. In many cases, the best results will be achieved by combining grammatical and statistical approaches. A particularly promising approach is the use of weighted automata, where frequency considerations are incorporated into the arcs of the transducers. We plan to apply standalone statistical methods after plan the grammatical analysis gives in. In other words, the cooperation should be ruled by the motto: *'Don't guess if you know.'*

3.1 Choosing Between a 'Top-down' and a 'Bottom-up Approach'

Within grammatical approaches to parsing there are two main approaches, which we may brand 'top-down' and 'bottom-up'. The top-down approach tries to map a possible sentence structure upon the sentence, as a possible outcome of applying generative rules on an initial S node. If successful, the result is a syntactic tree displaying the hierarchical structure of the sentence in question.

The bottom-up approach, on the other hand, takes the incoming wordforms and the set of their possible readings as input. Then they disambiguate multiple readings based upon context and build structures.

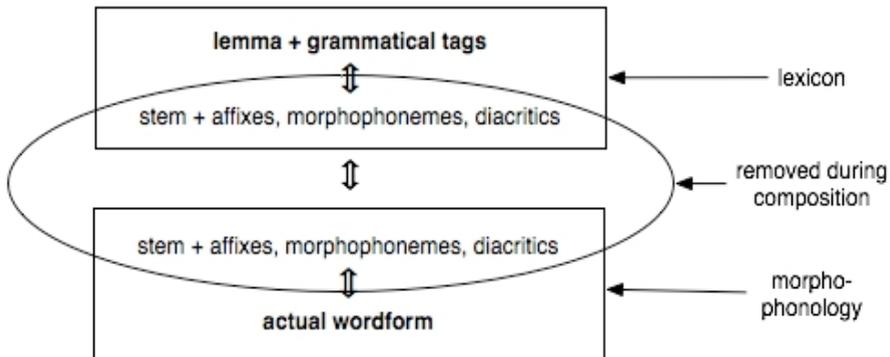
We chose a bottom-up approach, because it proved robust, was able to analyse any input and gave good results.

4. Linguistic Tools

4.1 The Tools Behind our Morphological Analyser

For our morphological analyser, we build finite-state transducers and use the finite-state tools provided by Xerox (documented in Beesley & Karttunen 2003). For morphophonological analysis, we have the choice of using the parallel, two-level morphology model (dating back to Koskenniemi 1983) with *twolc* or the sequential model (presented in Karttunen *et al.*1992) with *xfst*. Xerox' advice is to use the latter; we use the former, but we see this mainly as a matter of taste. The morphophonological and lexical tools are composed into a single transducer during compilation, as described in the literature (cf. the figure below).

Figure 2: A Schematic Overview of the Lexicon and Morphophonology of the Parser.



A more serious question is the choice of Xerox tools versus Open Source tools. In our project, we have no wish to modify the source code of the rule compilers themselves, but we notice that all binary files compiled by the *xfst*, *lexc* and *twolc* compilers are copyrighted property of the Xerox Corporation. It is as if you were to write your own 'C' program, but the compiled version of your program was copyright-owned by Kernighan and Ritchie, the authors of the C compiler. That said, it has been a pleasure working with Xerox: they have been very helpful, and as they see no commercial potential in Sámi, we notice no practical consequences of using proprietary compilers.

4.2 The Tools Behind our Disambiguator

For disambiguating the output of the morphological transducer we use constraint grammar. This is a framework dating back to Karlsson (1990), and the leading idea being that, for each wordform of the output, the disambiguator looks at the context and removes any reading that does not fit the context. The last reading can never be removed and, in the successful case, only the appropriate reading is left. The Brill tagger can be seen as a machine-learning variety of the constraint grammar parser.

There are several versions of the constraint grammar compilers. The original one was written in Lisp by Fred Karlsson. Later, Pasi Tapanainen wrote a compiler in 'C', called CG-2; this version may be licensed from <http://www.connexor.com>. We use an open source version of this compiler, made by Eckhard Bick. It must be stressed that the debugging facility of the Connexor compiler is superior to its competitors.

The optimal implementation would probably be to write the constraint grammar as a finite state transducer, as suggested in the Finite State Intersection Grammar framework. So far, nothing has come out of this work.

4.3 One-base, Multi-purpose Parsers

Working with minority languages, the lack of human resources is often as hard a problem as the lack of financial ones. With this in mind, avoiding duplicating work becomes crucial. The most time-consuming task in any linguistic project is building and maintaining a lexicon, be it in the form of a paper dictionary, a semantic wordnet, or the lexicon for a parser. The optimal solution is to keep only one version of the lexicon and extract relevant information from it, in order to automatically build paper and electronic dictionaries, orthographical wordlists or parsers. In our project, this has not yet been implemented, but for new languages we try out prototype models in order to make this work for new languages. Our plan is to use XML as text storage, and various scripts to extract the relevant lexicon versions.

It goes without saying that we use only source for a morphological transducer for linguistic analysis, pedagogical programs, spellers, and so forth. These applications often need slightly different transducers, in which case we mark the source code so that it is possible to compile different transducers from the same source code. For the academic project we make a tolerant parser that analyses as much of the attested variation as possible. The spellchecker has a totally different goal: here we build a stricter version that only accepts the forms codified in the accepted standard. This approach is even more appropriate, as we are the only language technology project working on Sámi. Any further application will build upon our work, and our goal is to make it flexible enough to facilitate this.

5. Infrastructure

5.1 Computer Platform

Our project is run on Linux and Mac OS X (Unix). The Xerox tools come in a Windows version as well, but the lack of a decent command-line environment and automatic routines for compiling makes it impractical. The cvs base is set up on a central Linux machine. We also use portable Macintoshes, both because they have a nice interface and because they offer programs that make it easier to work from different locations, such as the SubEthaEdit program mentioned below.

5.2 Character Set and Encoding

Most commercially interesting languages are covered by one of the 8-bit ISO standards. Many minority languages fall outside of this domain. It is our experience that it is both possible and desirable to use UTF-8 (multi-byte Unicode) in our source code (i.e., build the parser around the actual orthography of the language in question, rather than to construct some auxiliary ASCII representation). With the latest versions of the Linux and Unix operative systems and shells, we have access to tools that are UTF-8 aware and, although it takes some extra effort to tune the development tools to multi-byte input, the advantage is a more readable source code (with correct letters instead of digraphs) and an easier input/output interface, as UTF-8 now is the de facto standard for digital publishing.

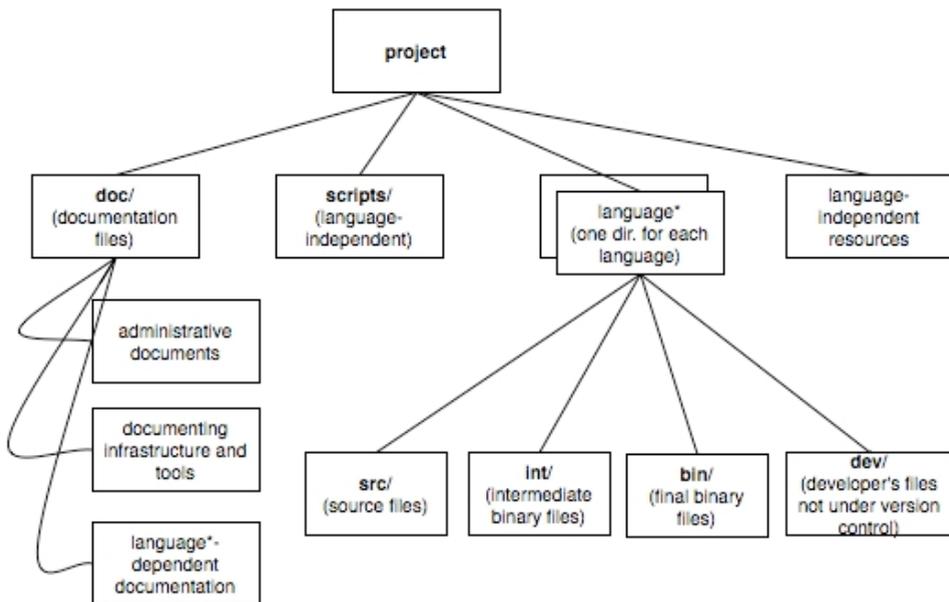
There is one setting where one could consider using a transliteration, namely, for languages using syllabic scripts, such as Inuktitut and Cherokee. In a rule stating that a final vowel is changed in a certain environment, a syllabic script will not give any single vowel symbol to change; rather than changing, for instance, *a* to *e* in a certain context, the rule must change the syllabic symbol BA to BE, DA to DE, FA to FE, GA to

GE, and so forth. It still may be better, however, to use the original orthography; each case requires its own evaluative process.

5.3 Directory Structure

We have put some effort in finding a good directory structure for our files. The philosophy is as follows: different types of files are kept separate. (The source files have their own directory, and binary and developer files are kept separate.)

Figure 3: Directory Structure



All our source and documentation files are under version control using cvs. This means that the original files are stored on our central computer (with backup routines), and that each co-worker 'checks out' a local copy that becomes his or her version to work on. After editing, the changed files are then copied back, or 'checked in' to the central repository. For each check-in, a short note on what has been done is written. We also have set up a forwarding routine, so that all co-workers get a copy of all cvs log messages via email.

Figure 4: Quote from cvs Log

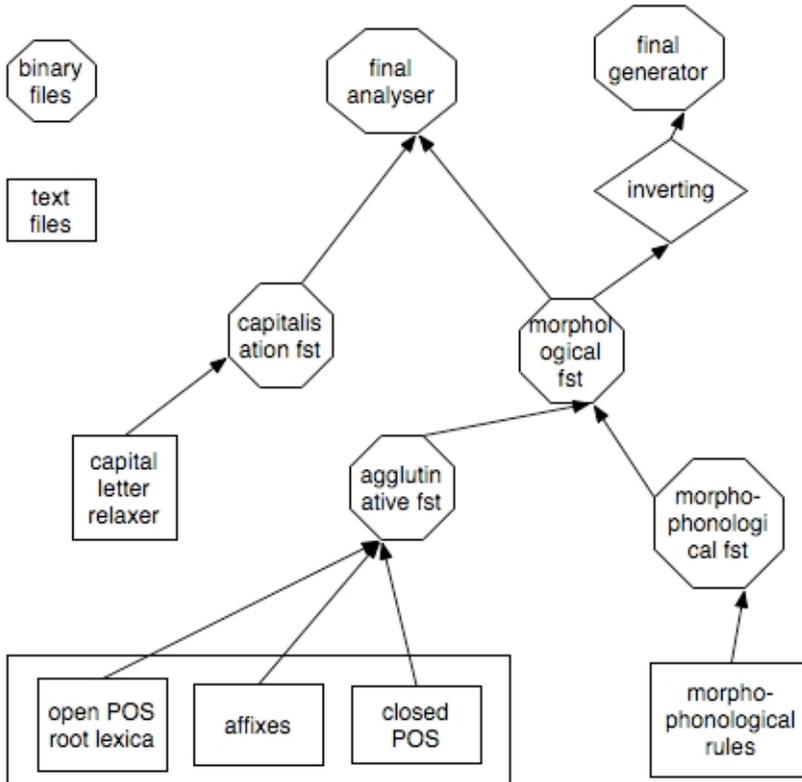
```
-----
revision 1.63
date: 2005/05/24 14:23:14; author: maren; state: Exp; lines: +2 -1
Extended the rule "Optional Vowel Shortening after Short 1st Syllable" to work a
lso for sequences where j:i is the first consonant in the consonant cluster.
-----
revision 1.62
date: 2005/04/13 06:43:21; author: sjur; state: Exp; lines: +2 -2
Minor editorial change.
-----
revision 1.61
date: 2005/04/01 13:40:01; author: trond; state: Exp; lines: +2 -2
Added some instructive Y1 symbols to some of the examples following the rules.
-----
revision 1.60
date: 2005/03/07 14:20:11; author: trond; state: Exp; lines: +5 -0
Changed rule "Stem vowel in Past Tense...", by adding reference to j:, and a ref
erence to Y6 as well as to Y7.
-----
revision 1.59
date: 2005/03/07 09:10:21; author: thomas; state: Exp; lines: +1 -1
Have done nothing.
-----
```

Using cvs (or some other version control system) is self-evident to any programmer, and it may be seen as quite embarrassing that such a trivial fact should be even mentioned here. It is our experience that the use of version control systems is by no means standard within academic projects, and we strongly urge anyone not using such tools to consider doing so. Backup routines become easier, and, when expanding from one-person projects to large projects, it is a prerequisite for when several co-workers collaborate on the same source files. We even recommend cvs for one-person projects. Using cvs, it is easier to document what has been done earlier, and to go back to previous versions to find out when a particular error may have crept in.

5.5 Building with 'Make'

Another self-evident programmer's tool is the use of makefiles, via the program *make*. In its basic form, *make* functions like a script and saves the work of typing the same long set of compilation commands again and again. With several source files, it becomes important to know whether one should compile or not. The *make* program keeps track of the age of the different files, and compiles a new set of binary files only when the source files are newer than the target binary files. The picture shows the dependencies between the different source and binary files.

Figure 5: Dependencies in the Project's Makefile



5.6 Tracing Bugs

As the project grows, so do the number of people debugging it, and hence the number of bugs and errors. We have designed an error database, in our case *Bugzilla*, which keeps track of errors. The database can be found at the address <http://giellatekno.uit.no/bugzilla/>. Interested persons may visit the URL. There is a requirement that you log in with an e-mail account and (preferably) a name, but otherwise the bug database is open for inspection.

5.7 Internal Communication in a Decentralised Project

We have co-workers in Tromsø, Kautokeino and Helsinki. Crucial for the project's progress is the possibility of coordinating our work. For that, we have the following means:

- We have made a project-internal newsgroup. Discussion is carried out in this environment rather than in personal emails, since more than one person may have something to say on the issue, and since it is easier to go back to earlier discussions using the newsgroup format.

- For simultaneous editing of the same document, be it source code or a meeting memo, we use a program called *SubEthaEdit* (<http://www.codingmonkeys.de/subethaedit/> - [only for Mac OS X]). This program makes it possible for several users to edit the same file at the same time. Combined with the use of the telephone (or voice chat!), we may discuss complicated matters on a common rule set while editing together, even though we sit in different countries.

- For informal discussions, we use chat programs. The built-in Mac OS X chat application iChat also facilitates audio and video chats with decent to high quality of the video and sound (mainly restricted by the available bandwidth).

- We have meetings over the phone; although we planned to conduct them using iChat (with up to ten participants in the same audio chat), technical problems with a firewall has stopped us from doing this.

- The cvs version control and *Bugzilla* error database also facilitate working in several locations.

5.8 Source Code and Documentation

In our experience, a systematic approach to documentation is also required also when the project engages only one worker, and it is indispensable when the number of workers grows beyond two. Working on the only Sámi language technology project in the world, we acknowledge that all future work will take our work as a starting point. We thus work in a one hundred-years perspective, and write documentation so that those who follow us will be able to read what we have done.

We document:

- The external tools we use (with links to the documentation provided by the manufacturer);
- The infrastructure of our project; and,
- Our source files and the linguistic decisions we have made.

In an initial phase, we used to write documentation in HTML, which was available only internally on the project machines. We now write documentation in XML, and convert it to HTML via the XML publishing framework (Forrest, <http://forrest.apache.org/>). Documentation can be published in many ways, but it is our experience that it is most convenient to read the documentation in a hypertext format such as HTML .

Since the documentation has grown, we also use a search engine (provided by Forrest) to find what we have written on a given topic.

The internal documentation of our project is open for inspection at the website <http://divvun.no/> (the proofing tools project), as well as <http://giellatekno.uit.no> (the academic disambiguator project). The technical documentation is in English, and can be found under the tab *TechDoc*.

Our source code is open as well, it is downloadable via anonymous cvs via our technical documentation. We believe that publishing the source code of projects like this will lead to progress within the field, not only in general, but especially for minority language projects.

By publishing the documentation and the source code, we make it easy to explain what we do; we hope that it will inspire others to perhaps give us some constructive feedback as well. The only possible drawback of this openness is that it exposes our weaknesses to the whole world. So far, we have not noticed any negative effects in this regard.

6. Costs

Except for the computers themselves and the operating system and applications that come with them, we have mostly used free or open-source software for all our tasks. In the few cases where we have paid for software, there are free or open-source alternatives. The notable exception is the set of compilers for morphophonological automata. For analysing running text and generating stray forms, the Xerox tools can be used in the versions found in Beesley & Karttunen (2003). For our academic project, these tools have proven good enough, but in order to generate larger paradigms, the commercial version of the tools is needed.

As for the computers, the only really demanding task is compiling the transducers. If one is willing to wait five minutes for the compilation, any recent computer can do fine, but the top models cut compilation time to less than half of what the cheapest models can do. Macs turned out to be a good choice for our projects, and the cheapest Mac can be bought for roughly 500 USD/EUR. One good investment, though, is to buy more RAM than what can be found on the standard configuration.

7. Summary

When doing language technology for minority languages, we are constantly faced with the fact that there are few people working with each language, and that different language projects set off in different directions, often due to mere coincidence. Our answer to this challenge is to share both our experience and our infrastructure with

others. By doing this, we hope that people will borrow from us, and comment upon what we do and how we do it. We also look forward to being confronted with other solutions and to borrowing improvements back.

References

Beesley, K.R. & Karttunen, L. (2003). *Finite State Morphology*. Stanford: CSLI Publications. <http://www.fsmbook.com/>.

Bick, E. (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Dr. phil. thesis, Aarhus University Press.

Brill, E. (1992). "A Simple Rule-based Part of Speech Tagger." *Proceedings of the Third Conference on Applied Natural Language Processing*. ACL, Trento, Italy, 1992.

"Bugzilla." Online at <http://giellatekno.uit.no/bugzilla/>.

"Bures boahhtin sámi giellateknologiiija prošektii." Online at <http://giellatekno.uit.no>.

Chanod, J.P. & Tapanainen, P. (1994). "Tagging French: Comparing a Statistical and a Constraint-based Method." *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, 149-156.

"Connexor." Online at <http://www.connexor.com>.

"Divvun - sámi korrektuvrareaiddut." Online at <http://divvun.no/>.

"Forrest." Online at <http://forrest.apache.org/>.

Jelinek, F. (2004). "Some of my best friends are linguists." *LREC 2004*. <http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf>.

Karlsson, F. (1990). "Constraint Grammar as a Framework For Parsing Running Text." Karlgren, H. (ed.) (1990). *Papers presented to the 13th International Conference on Computational Linguistics*, 3, 168-173, Helsinki, Finland, August. ICCL, Yliopistopaino, Helsinki.

Karttunen, L., Kaplan, R.M. & Zaenen, A. (1992). "Two-level morphology with composition." *COLING'92*, Nantes, France, August 23-28, 141-148.

Koskenniemi, K. (1983). *Two-level Morphology: A General Computational Model for Word-form Production and Generation*. Publications of the Department of General Linguistics, University of Helsinki.

Samuelsson, C. & Voutilainen, A. (1997). "Comparing a linguistic and a stochastic tagger." *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997.

"SubEthaEdit." Online at <http://www.codingmonkeys.de/subethaedit/>.

Voutilainen, A. Heikkilä, J. & Anttila, A. (1992). *Constraint Grammar of English, A performance-Oriented Introduction*, 21. Helsinki: Department of General Linguistics, University of Helsinki.

The Welsh National Online Terminology Database

Dewi Bryn Jones and Delyth Prys

Terminology standardisation work has been on-going for the Welsh language for many years. At an early date, a decision was taken to adopt international standards such as ISO 704 and ISO 860 for this work. It was also decided to store the terminologies in a standard format in electronic databases, even though the demand in the early years was for traditional paper-based dictionaries. Welsh is now reaping the benefits of those far-seeing early decisions. In 2004, work began on compiling a national database of bilingual (Welsh/English) standardised terminology. Funded by the Welsh Language Board, it will be made freely available on the World Wide Web. Electronic databases already in existence have been revisited and reused for this project, with a view to updating them to conform to an ISO terminology mark-up framework (TMF) standard. An additional requirement of this project is that the term lists should be packaged and made available in a compatible format for downloading into popular Termbase systems found in translation tool suites such as Trados, Déjà Vu and Wordfast. As far as we know, this is the first time that a terminology database has been developed to provide a freely available Termbase download utility at the same time as providing an online searchable facility. Parallel work of utilising an ISO lexical mark-up framework (LMF) compliant standard for another project, namely, the LEXICELT Welsh/Irish dictionary, has provided the opportunity to research similarities and differences between a terminological concept-based approach and a lexicographical lexeme-based one. Direct comparisons between TMF and LMF have been made, and both projects have gained from new insights into their strengths and weaknesses. This paper will present an overview of a simple implementation for the online database, and attempt to show how frugal reuse of existing resources and adherence to international standards both help to maximise sparse resources in a minority language situation.

1. Introduction

Terminology for Welsh has seen increased activity over the last ten years, in particular in government administration and the public sector, following the passing of the Welsh Language Act 1994 and the establishment of the National Assembly for Wales. Many bilingual Welsh/English dictionaries have been published by various organisations operating in Wales covering subject fields within secondary education,

academia, health, social services and public administration. Welsh terminology is generally perceived as merely an aid to standardised translation for English terms (cf. Prys 2003).

Depending on the organisation responsible, for the commissioning of a dictionary, dissemination to translators and the public at large is done via paper-based and/or electronically-based means. As a result, however, provision of standardised terminology is fragmented and dispersed in nature. Translators have to keep and maintain their own collection of paper-based dictionaries, and/or keep track of where and how to access electronic versions. Finding a Welsh translation may involve laboriously looking in more than one dictionary.

Meanwhile, the public at large, who would not have such a collection of terminology dictionaries, would not be part of a determinologization process, where specialised terms become incorporated into general language, thereby safeguarding or increasing the presence of Welsh in the commercial, printed media and popular culture sectors.

Thus, the Welsh Language Board commissioned the e-Welsh team to develop the Welsh National Online Terminology Database in order to centralise and facilitate efficient terminology dictionary dissemination.

2. Requirements for the Welsh National Online Terminology Database

The Welsh National Online Terminology Database project requirements were comprised of two parts. First, previously published dictionaries of standardised terminology, in particular those commissioned by the Welsh Language Board, were compiled and stored into a new online terminology database system. This new online terminology database would constitute the second part of the requirements, wherein its purpose would be to provide a freely available central Web-based resource supporting the dissemination of Welsh standardised terminology via the greatest number of formats, channels and mechanisms.

The system would support:

- searching and presenting term translations across one or more dictionaries stored in the system;
- downloading of complete dictionaries in various formats for offline use and integration into the translators' own Translation Memory environments' termbase tools such as *Trados Multiterm*, *Deja Vu*, *WordFast* and a dictionary product developed by e-Welsh called *Cysgeir*; and,

-
- presentation of its data as XML for possible incorporation with other online terminology systems.

Since the database system would be a repository of published standardised terminology there were no requirements for wider-scoped terminology management functionalities such as editing and standardisation process support.

3. Standards in Welsh Terminology

Since an early date, ISO international standards have been adopted in Welsh terminology. In 1998, members of the e-Welsh team compiled a guideline document on the use of ISO standards (Prys & Jones 1998) for all terminology standardisation providers in Wales.

The document mandated the use of principles and methods from ISO 704 and 860. The guidelines helped to raise the discipline of terminology standardisation for Welsh above what might otherwise be typical of a lesser-used and resourced language where:

- the work may be led by linguists with insufficient subject specialist knowledge, or subject specialists with insufficient linguistic expertise;
- less technically competent subject specialists experts would independently develop a paper-based dictionary in a word processing application; and,
- new words and terms are too easily coined along with spelling reforms in a misguided attempt to widen the appeal of the language.

The guidelines mandated the use of databases in accordance with ISO/TR 12618:1994 for terminology. The guideline document advised on the structure of such databases, as well as on the fields (or data categories) to be populated for any Welsh terminology dictionaries.

The development of dictionaries would be conducted in tabular format with columns to store fields such as terms, term plurals, disambiguators for homonyms and Welsh grammatical information such as parts-of-speech. Crucially, each row represented the concept level.

Thus, by employing a single table in a simple database application, no special terminological tools are required. A consistent bidirectional dictionary is easily created, whilst report and macro functionality found in Office productivity software can be used to create printed versions. A single table is also simple to host on a website.

Below is a typical example of a Welsh/English terminology dictionary entry:

home help (*of person*) cynorthwydd cartref **eb** cynorthwywyr cartref;

home help (*of service*) **cymorth cartref** **eg**

cynorthwydd cartref **eg** **cynorthwywyr cartref** home help;

In effect, this simple adoption of recommendations from the ISO standards made all previous terminology dictionaries commissioned in Wales 'future proof' and available to the Welsh National Online Terminology Database project electronically and already in a database format.

Over the course of numerous dictionary commissions, the guidelines have become outdated compared to the latest ISO standards. Further needs have been identified, including those to improve the interoperability of data and the reuse of software components. Weaknesses were identified in the guidelines both in the structure of terminological data and in the specification of data category selection. More specifically:

Structure - a single database table is too rigid a structure. Columns were sometimes duplicated as a means of overcoming this inflexibility in order to store multiple terms for a single concept. This is bad database design and practice.

Data Category Selection - although a data category selection was specified, their actual names for use in database tables were not. Therefore, across many dictionary database tables, the field for containing, for example, the English term, has different names such as [English] and [Saesneg] and [Term Saesneg].

Thus, the Welsh National Online Terminology Database project presented an opportunity to update and extend our adoption of standards.

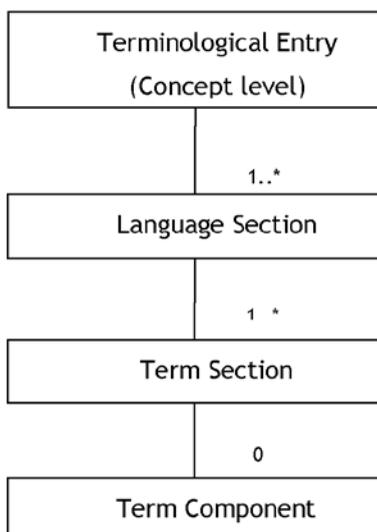
4. Welsh National Online Terminology Database

4.1 Use of Standards for the Welsh National Online Terminology Database

The Welsh National Online Terminology Database would need an improved structure in order to scale up to the number of terms and richness of data that terminological entry records may be expected to store in the future. The improved structure would come from ISO 16642, a.k.a. Terminological Mark-up Framework (TMF).

The TMF simply defines an abstract or meta-model for terminological entries. From the meta-model, Terminological Mark-up Languages (TML) can be derived to facilitate the representation and transfer of terminology data. Thus, adoption of the TML data model, as illustrated in the following figure, can be seen to provide a much-improved representation for all terminology entries in the Welsh National Online Terminology Database.

Figure 1: ISO 16642 / TMF Meta-Model Structure for Terminological Entries



The structure is concept-based, with a hierarchical structure containing multiple language sections, each containing one or more terms in the language that represents the concept in question. At the conceptual level, conceptual or classification system data can be added to increase the granularity of terms classification up and beyond the containing dictionary and the subject field implied by the dictionary title.

Language sections provide the means for storing terms from one, two or any number of languages. This gives the potential for multilingual Welsh terminology dictionaries and incorporating with languages that are related or widely used by Welsh speakers such as Spanish and other Celtic languages.

The term section provides an efficient means for associating many terms that represent a concept in a particular language. The TMF also specifies the mechanism by which its structure is populated with data categories selected from a data category register or catalogue as defined in ISO 12620:1999.

The catalogue contains well-defined data categories and pick list values for use within a TML structural representation. Essentially, ISO12620:1999 provides standardised names for data categories.

4.2 Implementing TMF: a Simple First Implementation

a) TMF Structure Compliance

A very simple implementation was completed that allowed us to quite easily use and benefit from aspects of the TMF. We did not intend to derive our own terminology representation from the TMF, preferring instead to use an already existing XML format. The TMF XML implementation chosen was TBX¹. TBX complies with the requirements of the TMF meta-model. It is a flexible format that allows users to specify their own data categories selection from ISO 12620 and specification of their own data categories via its eXtensible Constraint Specification (XCS).

A number of resources from the TBX home page at <http://www.lisa.org> are available to aid in its adoption, in particular, documentation to describe the standard further. To describe the structure, an XML Schema definition is provided. Also, a collection of ISO 12620 data categories are described and provided in a default eXtensible Constraints Specification file (TBXDv04.xml).

There is limited tool support for TBX, but with a freely available XML Schema Definition tool within the Microsoft .NET framework - xsd.exe², we were able to generate serializable C# classes. The generated C# code would be available to any other code written for constructing TMF compliant representations of terminological entries for inclusion in the Welsh National Online Terminology Database system. Construction would simply involve constructing object instances of the generated TBX classes and assigning values to member variables. When such objects instances are serialized with via the .NET framework's XML serializer, the resulting XML conforms to the original TBX schema. The following shows the generated C# class for the top level TBX TermEntry element.

```
[System.Xml.Serialization.XmlRootAttribute(Namespace="", IsNullable=false)]
public class termEntry
}

    /// <remarks/>

    public noteText_impIDLangTypTgtDtyp descrip;

    /// <remarks/>
    ///
    [System.Xml.Serialization.XmlElement]
```

1 <http://www.lisa.org/tbx>

2 <http://msdn.microsoft.com/library/en-us/cptools/html/cpconXMLSchemaDefinitionToolXsdexe.asp>

```

public descripGrp[] descripGrp;

/// <remarks/>
public noteText_impIDLangTypTgtDtyp admin;
/// <remarks/>
public adminGrp adminGrp;

/// <remarks/>
public transacGrp transacGrp;

/// <remarks/>
public noteText_impIDLang note;

    /// <remarks/>
public @ref @ref;

/// <remarks/>
public xref xref;

/// <remarks/>
[System.Xml.Serialization.XmlElementAttribute(«langSet»)]
public langSet[] Items;

    /// <remarks/>
[System.Xml.Serialization.XmlAttributeAttribute(DataType="ID")]
public string id;
}

```

b) Data Category Selection

Data categories, in particular TBX's eXtensible Constraints Specification support, also need to be available for the construction of terminological entries. However, we decided, since there aren't yet a great number of data categories used by Welsh dictionaries, that it was simpler to hardcode the placement and setting of data

categories into the TBX structure with wrapper code for the generated C# TBX code. Thus, the wrapper code provides easy access to the superset of all fields or data categories from all imported dictionaries.

The default selection of data categories given by TBX correspond to most fields used in previous dictionaries. Newly utilised categories from ISO 12620 would aid in the machine processing of terms. For example, the SortKey is used for implementation of Welsh sort order for all destined lists or dictionaries exports. Some data categories were created in order to store efficiently data for the typical Welsh/English dictionary entry given earlier.

- Welsh Part-Of-Speech (WelshPartOfSpeech)

The standard picklist of values for representing part-of-speech for Welsh terms could be maintained with the addition of the WelshPartOfSpeech data category.

- Welsh Plural (WelshPlural)

Further grammatical information for a term such as the plural can be stored under this data category.

- Disambiguator (concept-disambiguator)

A simple disambiguating field containing a brief explanation for the context of term had been mandated by previous guidelines in cases of Welsh or English homonyms, for example:

primate (=bishop)	achesgob
primate (=monkey)	pmat

The default language for the concept-disambiguator data category is English. However, when a Welsh language disambiguator needs to be included, this would be contained within TBX's <foreign lang="cy"> XML tag.

- Dictionary (originatingDictionary)

The dictionaries from which a term originates can be noted in the TBX representation via the use of this new category.

- Responsible Organisation (responsibleOrganisation)

The organisation responsible for the standardisation of the term can be noted in this new category. The additional data categories specification is given in the extract below from TBXDv04_CY.xml:

```
<!-- data categories added for Welsh terminology -->
<descripSpec name="concept-disambiguator" >
```

```

<contents datatype="noteText" targetType="none"/>
    <levels>termEntry </levels>
</descripSpec>

<termNoteSpec name="WelshPartOfSpeech">
<contents forTermComp="yes" datatype="picklist" targetType="none">
    adf ans be eb eg eg/b ell adj n v
</contents>
</termNoteSpec>

<termNoteSpec name="WelshPlural">
<contents datatype="plainText" targetType="none" forTermComp="yes"/>
</termNoteSpec>

<termNoteSpec name="originatingDictionary" >
<contents datatype="noteText" targetType="bibl"/>
    <levels>langSet termEntry term </levels>
</termNoteSpec>

<termNoteSpec name="responsibleOrganisation" >
<contents datatype="noteText" targetType="respOrg"/>
    <levels>termEntry term </levels>
</termNoteSpec>

```

An example of a TBX XML string with Welsh data categories in action:

```

<termEntry id="tid-tg-31">
<descripGrp>
    <descrip type="concept-disambiguator">(of person)</descrip>
</descripGrp>
<admin type="elementWorkingStatus">consolidatedElement</admin>

<langSet lang="en">
<ntig>

```

```

<termGrp id="tid-tg-31-en-1">
  <term>home help</term>
  <termNoteGrp><termNote type="termType">entryTerm</termNote></termNoteGrp>
  <termNoteGrp>
    <termNote type="grammaticalNumber">singular</termNote></termNoteGrp>
  </termGrp>
  <adminGrp>
    <admin type="SortKey">home help$sf$(of person)$sf$</admin>
  </adminGrp>
</ntig>
</langSet>

<langSet lang="cy">
  <ntig>
    <termGrp id="tid-tg-31-cy-1">
      <term>cynorthwydd cartref</term>
      <termNoteGrp><termNote type="termType">entryTerm</termNote></termNoteGrp>
      <termNoteGrp>
        <termNote type="WelshPartOfSpeech">eg</termNote>
      </termNoteGrp>
      <termNoteGrp>
        <termNote type="WelshPlural">cynorthwywyr cartref</termNote>
      </termNoteGrp>
      <termNoteGrp>
        <termNote type="grammaticalNumber">singular</termNote>
      </termNoteGrp>
    </termGrp>
    <adminGrp>
      <admin type="SortKey">cynorthwydd cartref$sf$seg</admin>
    </adminGrp>
  </ntig>
</langSet>
</termEntry>

```

c) Database Design

The TBX XML representation needs to be stored in a relational database. Database schemas can also usually be generated from XML Schemas. Since the Welsh National Online Terminology Database has the simple purpose of being a storage facility for the dissemination and presentation of standardised and thus fixed or published terminology data in various formats, a sufficient yet effective solution would be to store the entire TBX representation in the database as a string. Thus, the database design contains a single table for containing all terminological entries.

Table 1: Database Table Design for Storing the Terminological Entries

	TermEntries
PK	TermEntry_ID
	TermEntry

Term data within XML strings cannot be accessible in a relational manner for search and querying, and so forth. Therefore, an index table is added, consisting of each term's XPath's that point to its locations in the containing TBX strings stored under TermEntry_ID.

Table 2: Database Table Design for Index to Terminological Entries

	TermIndex
PK	TermIndex_ID
PK	TermEntry_ID
PK	Language_ID
	Term
	TermEntry_XPath
	SortKey

Table 3: An Example Entry in the TermIndex Table

TermIndex	Value
TermIndex_ID	2614
TermEntry_ID	1328
Language_ID	cy
Term	golwg grŵp
TermEntry_XPath	//termGrp[@id='tid-tg-2554-cy-1']
SortKey	golwg grwp\$sf\$sf\$eb

4.3 TBX transformations

Presenting and transforming TBX strings in the database in various other formats such as HTML, CSV, and so forth would involve using simple XSLT transformations.

- CSV

A review of the import process of commercial termbase products used by Welsh translators showed that, in one way or another, all of them support importing terminology data from CSV formatted files.

The following program creates an English to Welsh CSV format data for the term 'group view' (XPath = termGrp id="tid-tg-2554-en-1"):

```
<?xml version="1.0" encoding="UTF-8" ?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="text"/>
  <xsl:template match="/">

    <!-- loop through each term in the Welsh language section -->
    <xsl:for-each select="//langSet[@lang='cy']/termGrp">
      <!-- transform English XPathed term details -->
      „<xsl:value-of select="//termEntry/@id"/>“,
      „<xsl:value-of select="//termGrp[@id='tid-tg-2554-en-1']/term\"/>“,
      “<xsl:value-of select="//descrip[@type='concept-disambiguator']\"/>“,
      “<xsl:value-of select="//termGrp[@id='tid-tg-2554-en-1']
        //termNote[@type='partOfSpeech']\"/>“,
      <!-- transform the only/first or next Welsh term's details -->
```

```

"<xsl:value-of select="term"/>\",
"<xsl:value-of select="termNoteGroup/termNote[@type='WelshPartOfSpeech']"/>"
"<xsl:value-of select="termNoteGroup/termNote[@type='WelshPlural']"/>"

<!-- insert newline -->
<xsl:text>&#10;</xsl:text>

</xsl:for-each>
</xsl:template>
</xsl:stylesheet>

```

The sample outputs CSV lines that provide each Welsh term translation for a specific English term for a specific concept (and concept-disambiguator value).

A specific XPath (and therefore XSLT) is required for each terminological entry transformation. This has made a negligible difference to the performance of transforming entire dictionary contents and/or multiple entries.

- Trados

Trados' termbase product MultiTerm imports terms marked up in its own MultiTerm XML format, as well as CSV. Importing terms from CSV files in most translation memory systems (including Trados) is quite a manual and complex task for some translators, despite both having wizards to help in mapping CSV fields to their termbase fields and structures. Fortunately, Trados MultiTerm has its own XML mark-up language to facilitate importing terminology. Although not entirely compliant to TMF or TBX, MultiTerm XML does follow the TMF meta structure. Data category selection and definition is described in XML Termbase Definition files (with extension .xdt).

Thus, the Welsh National Online Terminology Database employs a simple XSLT to construct Multiterm XML from TBX by mapping data categories in identical structures, and provides a fixed accompanying .xdt file.

```

<?xml version="1.0" encoding="UTF-8" ?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="xml"/>
  <xsl:template match="/">

```

```

<!-- Copy concept level data -->
<conceptGrp>
  <descripGrp>
    <descrip type="TermId">
      <xsl:value-of select="//termEntry/@id"/></descrip>
    </descripGrp>"

<!-- Copy English language section and terms -->
<xsl:for-each select="//langSet[@lang='en']/termGrp">
  <languageGrp>
    <language type="English" lang="EN-GB"/>
    <xsl:if test="//descrip[@type='concept-disambiguator']">
      <descripGrp><descrip type="Disambiguator">
        <xsl:value-of select="//descrip[@type='concept-disambiguator']"/>
      </descrip></descripGrp>
    </xsl:if>
    <xsl:if test="//termNote[@type='partOfSpeech']">
      <descripGrp><descrip type="Part Of Speech">
        <xsl:value-of select="//termNote[@type='partOfSpeech']"/>
      </descrip></descripGrp>
    </xsl:if>
    <termGrp><term><xsl:value-of select="term"/></term></termGrp>
  </languageGrp>
</xsl:for-each>

<!-- Copy Welsh language section and terms -->
<xsl:for-each select="//langSet[@lang='cy']/termGrp">
  <languageGrp>
    <language type="Cymraeg" lang="GA-WA"/>
    <descripGrp><descrip type="Rhan Ymadrodd">
      <xsl:value-of select="//termNote[@type='WelshPartOfSpeech']"/>
    </descrip></descripGrp>
    <xsl:if test="//termNote[@type='WelshPlural']">

```

```

<descripGrp>
  <descrip type="Lluosog">
    <xsl:value-of select="//termNote[@type='WelshPlural']/>
  </descrip>
</descripGrp>
</xsl:if>
<termGrp><term><xsl:value-of select="term"/></term></termGrp>
</languageGrp>
</xsl:for-each>
</conceptGrp>

</xsl:template>
</xsl:stylesheet>

```

- Deja Vu

Deja Vu's Termbase can import terminology from a number of simple formats for example, CSV text, Excel, and Access. It does not support import of terminological data in any flavour of XML.

Its import process contains target templates that are similar in structure and data categories to other common termbases such as Eurodictautom and TBX. However, the user is left with the manual task of mapping the CSV/Excel flat structure to the destination template. Simplifying the import process is under continued investigation.

- WordFast

WordFast does not support CSV, but its terminology support is a simple bilingual glossary text file with a source column and a target column that is easily loaded into the Wordfast toolbar. The XSLT transformation for exporting to this format is trivial as it is a subset of CSV.

- Cysgeir

Cysgeir is a Welsh/English dictionary software product produced by the e-Welsh team that incorporate many Welsh NLP features such as a lemmatizer for aiding in the search for Welsh words and terms. Cysgeir installs onto Windows PCs and integrates with popular Office productivity software and some translation memory packages, such

as WordFast. Cysgeir supports loading terminology and browsing multiple dictionaries described in Cysgeir's own proprietary formatted dictionary files. Integrating functionality for exporting Cysgeir dictionary files from TBX descriptions stored in the Welsh National Online Terminology Database system is under development.

- Raw TBX

Simple ASP.NET code can expose the raw TBX XML strings to other online terminology systems from specially crafted URLs. An example URL supported by the Welsh National Online Terminology Database is: <http://www.e-gymraeg.co.uk/bwrdd-yr-iaith/termau/tbx.aspx?en=view>.

The result returned is merely the concatenation of all TBX strings found in the database containing the English term 'view'. The system adds header information along with a link to the data category selection described in our eXtensible Constraints Specification file - TBXv04_CY.xml.

4.4 Current State

The national online database has been online since August 2005 and was launched with standardised terminology for Information Technology. The URL is <http://www.e-gymraeg.org/bwrdd-yr-iaith/termau>. Despite being a very simple use and implementation of the TMF on a Microsoft .NET/ ASP.NET/ SQL Server platform, it has already proven to be effective in its support of Welsh translators and software localisers. The website provides a simple search panel and presents results to the right of the screen, as seen in the screenshot:

Figure 2: Screenshot of Search Results from the Welsh National Online Terminology Database



During a period of ten weeks between August and mid-October 2005, over six hundred queries were made to this one dictionary, and it has been downloaded in its entirety as a CSV file over thirty times with twenty downloads requesting Welsh as the source language. The work of importing other pre-existing electronic dictionaries is about to begin and two new dictionaries covering ecology and retail will soon be added.

5. Perspectives and Conclusion

This paper has illustrated, for the interest of other lesser-resourced languages, a very simple implementation of online dissemination of terminology and adoption of the TMF standards. International standards have helped to steer terminology standardisation for Welsh (despite being a lesser spoken and resourced language) on a productive and sound course. Adoption of these standards in the past facilitated the creation of many dictionaries that, in the future, would prove easy to include in the Welsh National Online Terminology Database. Frugal reuse of existing resources is key for the development of language technology for lesser-resourced languages such as Welsh.

In the meantime, the scope of ISO TC37 has expanded to cover the principles, methods and applications related to terminology, language resources and knowledge organisation for the multilingual information society. A family of standards are being developed with common principles that deal with lexicons, terminology, morpho-syntactic annotation, word segmentation and data category management. Therefore, we consider that continued gradual adoption of the ISO standards on terminology and lexicography will maximise reuse levels of linguistic data and software components even further.

Parallel work also conducted by the e-Welsh team within the LEXICELT Welsh/Irish dictionary project, where the ISO lexicography standard (the Lexical Mark-up Framework ISO/CD 24613) has been used, identified similarities with the ISO TMF, and opportunities for reusability and merging of language resources.

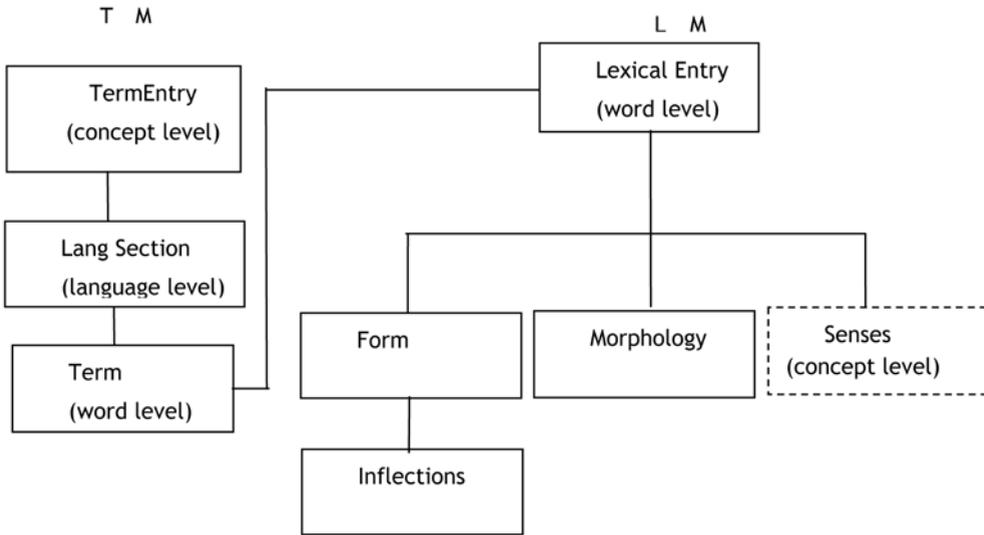
LMF aims to be a standard for MRD and NLP lexicon standards, and therefore define, in a similar way to the TMF, a meta structure along with mechanisms for data category selections from meta data registries. Nevertheless, whereas TMF is concept oriented, LMF is lexeme-based. Most existing lexical standards are said to fit with the LMF, such as the Text Coding Initiative (TEI), OLIF, CLIPS, WordNet, and FrameNet.

Ultimately, the same structure and content can be used for a number of different purposes, from speech technology and printed dictionaries, to machine translation, where, at the moment, individual standalone application specific standards exist.

Data from interoperable terminology and lexicography resources would not only enrich each other, but also use the same software components and systems.

A bridge identified by many between terminological and lexical meta-models links a term entry in the TMF to the lexeme entry in the LMF. Thus, either of the two database implementations could be linked, or a super structure could be constructed containing a superset of term entry and sense entry data categories organised on concept or lexeme basis.

Figure 3: Bridging ISO's TMF and LMF Meta_Model Structures



Linking lexical data and terminological data can improve the richness of a term's grammatical information beyond the capabilities of the TMF. Already noted above, such a link would be an aid in the Welsh terminology standardisation process. In adhering to ISO 860 recommendations, term candidates are evaluated for inflected forms. Such inflected forms may exist already under the corresponding lexeme in its Form->Inflections section, or may be simply added to improve the lexical database.

Larger opportunities for reuse exist if we realise that the adoption of standards in the Welsh terminology could go beyond merely standardised translations, in particular, adopting advanced aspects of the standards such as concept modelling.

Classification of terms and their finer-grained organisation into a conceptual organisation may offer users of the Welsh National Online Terminology Database system better discovery and comprehension of terms in Welsh than in English. It may even contribute to improving English terminology standardisation, since many organisations and/or terminology dictionaries have different terms representing identical concepts that hamper and frustrate cooperation.

LMF supports translation by linking senses in the same way that TMF's term entries are linked as subordinates to the same concept entity. Simple, direct and bidirectional links between senses in respective LMF entries may be sufficient for simple bilingual dictionaries, but more complicated and multilingual lexical dictionaries require an interlingual concept system to handle the various levels of concept precision. Thus, further opportunities for the reuse of software components for concept modelling exists.

Such observations and gradual expansion of international standards adoption for Welsh language technology can only further 'future-proof' Welsh against all future developments and applications in language technology such as semantic web, knowledge bases, and machine translation.

6. Acknowledgements

The Welsh National Online Terminology Database project is funded by the Welsh Language Board.

References

“Cronfa Genedlaethol o Dermau.” Online at <http://www.e-gymraeg.org/bwrdd-yr-iaith/termau>.

“Déjà Vu.” Online at <http://www.atril.com/default.asp>.

“Microsoft .NET framework - xsd.exe”. On line at:
<http://msdn.microsoft.com/library/en-us/cptools/html/cpconXMLSchemaDefinitionToolXsdexe.asp>

ISO/TR 12618:1994 *Computational aids in terminology - Creation and use of terminological databases and text corpora*. (TC37/SC3).

ISO 860:1996 *Terminology work - Harmonization of Concepts and Terms*, (TC37/SC1).

ISO 12620:1999 *Computer Applications in Terminology - Data Categories* (TC37/SC3).

ISO 704:2000 *Terminology Work - Principles and Methods* (TC37/SC1).

ISO 16642:2003 *Computer Applications in Terminology - Terminological Markup Framework* (TC37/SC3).

ISO/CD 24613 *Language Resource Management - Linguistic Annotation Framework* (TC37/SC4).

Prys, D. (2003). “Setting the Standards: Ten Years of Welsh Terminology Work.” *Proceedings of the Terminology, Computing and Translation Conference*, Swansea University, March 27-28, 2004. Swansea: Elsavier.

Prys, D. & Jones, J.P.M. (1998). *The Standardization of Terms Project*. Report prepared for the Welsh Language Board.

"TBX TermBase eXchange." Online at <http://www.lisa.org/tbx>.

"Trados." Online at <http://www.trados.com/>.

"Wordfast." Online at <http://www.wordfast.net/>.

SpeechCluster: A Speech Data Multitool

Ivan A. Uemlianin

When collecting and annotating speech data, to build a database, for example, speech researchers face a number of obstacles. The most obvious of these is the sparseness of data, at least in a usable form. A less obvious obstacle, but one that is surely familiar to most researchers, is the plethora of available tools with which to record and process the raw data. Some example packages include: EMU, Praat, SFS, JSpeechRecorder, Festival, HTK, and Sphinx. Although, prima facie, an embarrassment of riches, each of these tools proves to address a slightly different set of problems, to be slightly (or completely) incompatible with the other tools, and to demand a different area of expertise from the researcher. At best, this is a minor annoyance; at worst, a project must expend significant resources to ensure that the necessary tools can interoperate. As this work is no doubt repeated in unrelated projects around the world, an apparently minor problem becomes a possibly major - and undocumented - drag on progress in the field. This danger is especially extreme in research on minority and lesser-spoken languages, where a lack of resources or expertise may completely preclude research. Researchers need some way of abstracting from all these differences, so they can conduct their research. The simplest approach would be to provide an interface that can read and write the existing formats, and provide other facilities as required.

On the WISPR project-- developing speech processing resources for Welsh and Irish- - we have adopted this approach in developing SpeechCluster. The intention behind SpeechCluster is to enable researchers to focus on research rather than file conversion and other low-level, but necessary preprocessing. SpeechCluster is a freely available software package, released and maintained under an open-source license.

In this paper, we present SpeechCluster (reviewing the requirements it addresses and its overall design), we demonstrate SpeechCluster in use, and finally, we evaluate its impact on our research and outline some future plans.

1. The Context

Lesser-used languages (LULs) are often lesser-resourced languages. Majority languages have wealthy patron states, with the money and the labour force to develop language resources as required. For example, Microsoft alone has over 6000 hours of US English speech data at its disposal (Huang 2005). Patron organisations of lesser-

used languages often do not have access to such power, and they must use their resources wisely.

Research and development in language technology brings many stimulating challenges. With LULs especially, these challenges may include considerations about the status and use of the language (users and patrons of the language are likely to take an active interest, and often language technology research can become part of the life of the language itself).

Research and development in language technology also brings a great deal of tedious labour. Data must be collected and archived, and there are several layers of processing that need to be done before any 'interesting' R&D can begin. Often, the physical forms of the storage and the processing tools- the file formats and software implementations- provide obstacles of their own.

Since these obstacles are contingent upon the machinery rather than the research problem, they are often categorised as 'chores' and tackled quite differently to other tasks on the project. At worst, these obstacles will be tackled manually; at best, scripts will be written *ad hoc* as the need arises, to be discarded (or 'archived') at the end of the project. These approaches are inefficient, especially when compared to the conscious and investigative approach taken with other parts of the project. Resources are wasted, and specialists can spend significant portions of their time involved with inappropriate (and more importantly, unpleasant) activities.

In the speech research department of a large corporation, the costs associated with this waste can be passed on to the customer; in smaller research establishments, these costs may preclude speech research altogether.

2. Our Problem

Our goals on the WISPR project are to develop speech technology resources for Welsh and Irish that we can make freely available to commercial companies. There are currently no such resources at all for Irish, and very limited resources for Welsh (language resources available for Welsh include two text resources: CEG, a 1 million word balanced and tagged corpus (Ellis *et al.* 2001) and a large collection of webpages (Scannell 2004], both of which are for non-commercial use only; a telephone speech corpus (Jones *et al.* 1998); and a small, experimental speech database (Williams 1999).

The project must therefore begin from the bottom, starting with data collection and annotation, moving on to developing necessary speech databases, acoustic models (AMs) and language models (LMs), and, finally, developing packaged software artefacts

that can be used by external developers. With limited resources of time, money and labour, every administrative chore added to the workflow reduces resources available for more delivery-oriented tasks.

The first decision to the problem of 'chores' was that the solution should be a Speech Processing Resource in its own right. The solution should consist of a set of reusable, extensible, shareable tools to be made available to (a) ourselves on future projects, and (b) other teams working on speech processing projects around the world.

2.1 Requirements

The main design goals of our solution are as follows:

- researchers should be able to work independently of data format restrictions;
- necessary, complicated, but uninteresting tasks should be automated;
- interesting, but complicated tasks should be made simple;
- researchers should be able to address linguistic problems with linguistic solutions;
- the toolkit should be increasingly simple to maintain and develop; and,
- the toolkit should encourage its own use and development.
- Researchers should be able to work independently of data format restrictions.

Data can be collected, transcribed and stored in a range of formats. Each of the range of available tools for language technology research accepts or generates data in its own format, or in a limited range of standard formats. Researchers should not have to worry about which format works with which application: they should be able to pick the application necessary (or preferred) for the research problem, and the data should be readily accessible in the correct format.

- Necessary, complicated, but uninteresting tasks should be automated.

This applies to life in general, of course.

- Interesting but complicated tasks should be made uncomplicated.

Due to the nature of the field, where it is often necessary to process large sets of data, many of the more interesting problems (e.g. building AMs for speech recognition) involve procedures that are repetitive (e.g. those that have to be applied to every item in a corpus) or complicated (e.g. initialising a system). Researchers learning about a new area are hampered when these tasks dominate learning time.

-
- Researchers should be able to address linguistic problems with linguistic solutions.

Often, a linguistic problem, or a problem initially described in language terms (e.g. retranscribing the data using a different phoneset) has to be redescribed in programming terms before it can be addressed. Problems should be addressable in the terms in which they occur.

- The toolkit should be increasingly simple to maintain and develop.

Over its lifetime, any toolkit increases in functionality: new problems occur and new tasks become possible. If extensions are increasingly difficult to implement, the toolkit eventually disintegrates (e.g. into a library of loosely-related scripts), becomes impossible to maintain, and falls into disuse. A well-designed toolkit can avoid this fate.

- The toolkit should encourage its own use and development.

It should be preferable to use the toolkit than to revert to the bad old ways. Nevertheless, further use of the toolkit should stimulate researchers to confront it with new problems, and to think of new areas in which the toolkit might be used.

If possible, the toolkit should be extensible by the researchers themselves, rather than having to rely on a separate maintainer. In this case, the design of the toolkit should promote the writing of readable, reusable code.

3. A Solution

3.1 Introduction

Our first (and current) attempt at a software artefact that meets these requirements is the SpeechCluster package (Uemlianin 2005a). SpeechCluster is a collection of small programs written in a programming language called Python. Python has a very clear, readable syntax, and is especially suited for projects with several programmers, or with novice programmers. As such, it suited our aim of encouraging non-programmers to write their own tools.

The SpeechCluster package consists of a main SpeechCluster module with the basic API, and a number of modules that can be used as command line tools. The tools are intended to be used as such, but they can also be used as 'examples', or as a basis for customisation or further programming with SpeechCluster.

Table 1 shows a list of the tools currently available as part of SpeechCluster: Below, we look at two of these in more detail before exploring the use of SpeechCluster as an API. Finally, we look at SpeechCluster in a larger system.

Table 1: SpeechCluster command-line tools

Tool	Function
segFake	'Fake autosegmentation' of a speech audio file
segInter	Interpolates labels into a segmented but unlabelled segment tier
segMerge	Merges separate label files
segReplace	Converts labels in a label file
segSwitch	Converts label file format
splitAll	Splits audio/label file pairs

3.2 Using SpeechCluster I: The Tools

a) SegSwitch

SegSwitch is a label file format converter. It converts label files between any of the formats supported by SpeechCluster (currently, Praat TextGrid, *esps* and the various HTK formats [i.e., the simple *.lab* format and the multi-file *.mlf* format]). This kind of format conversion is a very common task. For example, HTK requires files to be in its own *esps*-like format, but our team prefers to handlabel files in Praat, which outputs its own TextGrid format. Festival uses an *esps*-like format that is slightly different from HTK's.

SegSwitch has a simple command-line interface (see Table 2), in which single files or whole directories can be converted easily and perfectly.

Table 2: segSwitch usage

Usage:	<pre>segSwitch -i <infilename> -o <outfilename></pre>
Examples:	<pre>segSwitch -d <dirname> -o <outFormat></pre> <pre>segSwitch -i example.lab -o example.TextGrid</pre> <pre>segSwitch -d labDir -o textGrid</pre>

A simple facility like this has a remarkable effect on the efficiency of a team. The team no longer has to worry about in what file format they have to work. They can concentrate on the research task converting files in and out of particular formats as needed. In a sense, the two parts of the work- the research and the bookkeeping- have been separated, and the bookkeeping is done by the tools. This division of labour is repeated between the tools and the SpeechCluster module itself. As much of the low-level data manipulation as possible is carried out by SpeechCluster, so that the tools themselves can be written in simple, task-oriented terms.

Table 3 shows the main code for *segSwitch* (excluding the command-line parsing and the loop over files in a directory): all of the work of file format conversion is done by

the code shown. Looking past the Python syntax, this code is a direct implementation of an intuitive statement of the task (see Table 4).

Table 3: Simplified python code for segSwitch

Line	Code
1	from speechCluster import *
2	
3	def segSwitch(inFn, outFn):
4	"""
5	Args: string inFn: input filename
6	string outFn: output filename
7	Returns: None
8	Uses filename extensions to determine input
9	& output formats.
10	"""
11	spcl = SpeechCluster(inFn)
12	ofext = os.path.splitext(outFn)[1]
13	outFormat = SpeechCluster.formatDict[ofext.lower()]
14	out = spcl.write_format(outFormat)
15	open(outFn, 'w').write(out)

Table 4: segSwitch task statement

Line(s)	Task
11	read in an input file
12-13	work out from the output filename what the output format should be
14	generate the output format data
15	write the data out to a file, using the output filename given.

All of the hard programming is hidden in the SpeechCluster module, which is imported in line 1, and which provides useful facilities like formatDict(ionary) and write_format(format).

b) *SplitAll*

One of the special features of SpeechCluster is that it will treat a sound file (i.e., recorded speech) and its associated label file as a pair, and can manipulate them together. *SplitAll* shows this in action.

SplitAll addresses the problem of the researcher who requires a long speech file to be split into smaller units along with its associated label file; for example, one may require a long utterance containing pauses to be split into its constituent phrases. Of course, data can be recorded or segmented into shorter units before it is labelled, but when data is re-used, its requirements often change.

This task is a minor inconvenience if you just have one or two files, but if you have five hundred (or even just fifty) it becomes important to automate it. Furthermore, it would be better psychologically if the researcher could envisage this as a single task, rather than two related tasks (i.e., splitting the *wav* file; then splitting the *label* file to match). The best option is to delegate the task to a machine.

As with *segSwitch*, *splitAll* has a simple command-line interface (see Table 5).

Table 5: splitAll usage

Usage		
<code>splitAll -n <integer> -t <tierName> [-l <label>]</code>		
		inDir outDir
Example		Splits into
<code>splitAll -n 5 -t Phone inDir outDir</code>		5 phone chunks
<code>splitAll -n 1 -t Word inDir outDir</code>		single words
<code>splitAll -n 5 -t Second inDir outDir</code>		5s chunks
<code>splitAll -n 1 -t Phone -l sil inDir outDir</code>		split by silence

SplitAll is intended to be used on directories of files and can process hundreds of speech/label file pairs in moments. Again, the effect is to separate the researcher from the drudgery of looking after files.

Apart from a function that parses the command-line parameters into the variable *splitCriteria*, the code for *splitAll* is just as simple as that for *segSwitch* (see Table 6). The excerpt seen here loops through the filestems in a directory, a filestem being a filename without its extension (e.g. *example.wav* and *example.lab* have the same filestem 'example'). Line 8 generates a *speechCluster* from a filestem: this means that all files with the same filestem- (e.g. a *wav* file and a *lab* file) are read into the one *speechCluster*. Line 9 then calls *split*, saving the results into the given output directory.

Table 6: Simplified python code for splitAll

Line	Code
1	<code>from speechCluster import *</code>
2	
3	<code>def splitAll(splitCriteria, inDir, outDir):</code>
4	<code> stems = getStems(inDir)</code>
5	<code> for stem in stems:</code>
6	<code> fullstem = '%s%s%s' % (inDir, os.path.sep, stem)</code>
7	<code> print 'Splitting %s.*' % fullstem</code>
8	<code> spcl = SpeechCluster(fullstem)</code>
9	<code> spcl.split(splitCriteria, outDir)</code>

This codewalk tells you nothing about how `SpeechCluster.split(splitCriteria)` works, but that's the point. The `SpeechCluster` module provides facilities like `split()` that allow the researcher to phrase problems and solutions in task-oriented terms rather than programming-oriented terms.

3.3 *SpeechCluster as an API*

The two main design features of `SpeechCluster` are:

- 7 it stores segmentation details internally in an abstract format; and,
- 8 it can treat an associated pair of sound and label files as a unit.

In terms of the facilities `SpeechCluster` provides, these features translate into the methods shown in Table 7.

Table 7: `SpeechCluster` methods

Interface (i.e. read/write) methods	
<code>read_format(fn)</code>	<code>write_format(fn)</code>
<code>read_ESPS(fn)</code>	<code>write_ESPS(fn)</code>
<code>read_HTK_lab(fn)</code>	<code>write_HTK_lab(fn)</code>
<code>read_HTK_mlf(fn)</code>	<code>write_HTK_mlf(fn)</code>
<code>read_HTK_grm(fn)</code>	<code>write_HTK_grm(fn)</code>
<code>read_stt(fn)</code>	<code>write_stt(fn)</code>
<code>read_TextGrid(fn)</code>	<code>write_TextGrid(fn)</code>
<code>read_wav(fn)</code>	<code>write_wav(fn)</code>
Methods for manipulating label and sound files (and label/sound file pairs)	
<code>merge(other)</code>	
<code>replaceLabs(replaceDict)</code>	
<code>setStartEnd(newStart, newEnd)</code>	
<code>split(splitCriteria, saveDir, saveSegFormat)</code>	

When programming using `SpeechCluster` as a library, the researcher/developer can program using the linguistic terms of the problem, not the programming terms of the programming language.

There is documentation available (Uemlianin 2005a), and the Python `pydoc` facility allows the researcher/developer to access documentation 'interactively' (see Figure 1).

3.4 *Using SpeechCluster II: Making a New Script*

Although there are tools provided as part of the `SpeechCluster` package, the `SpeechCluster` module itself presents an accessible face, and it is hoped that

researcher/developers are able to use SpeechCluster to build new tools for new problems.

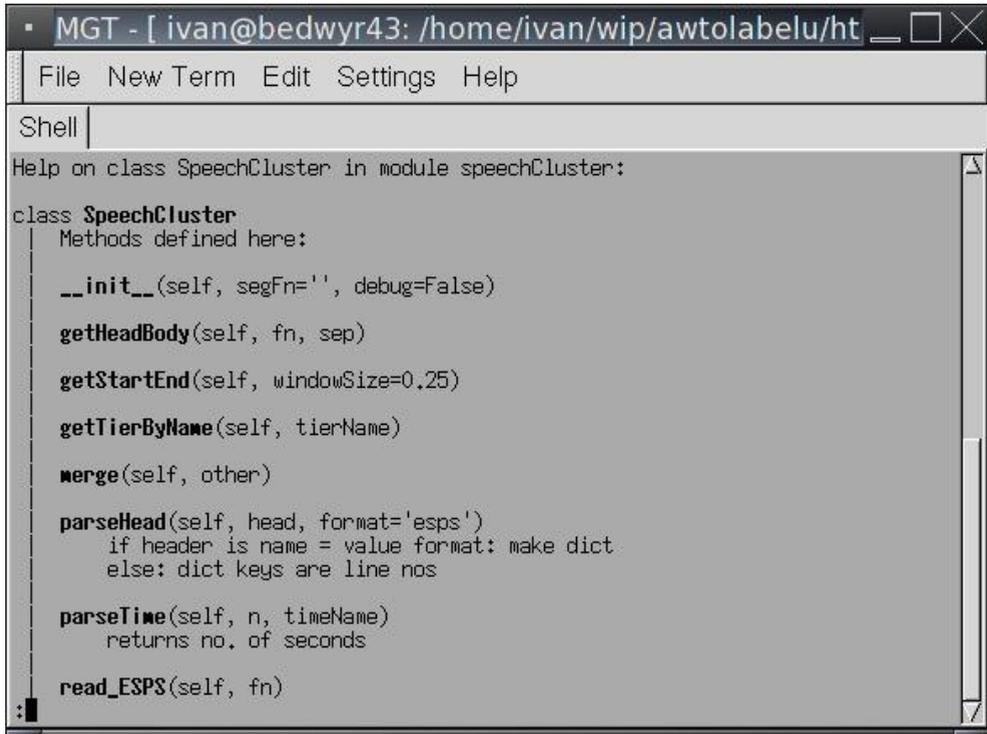
a) SegFake

SegFake provides an example of using SpeechCluster to help write a script to address a specific problem. Handlabelling is passé. It is laborious, tedious and error-prone; but sometimes researchers in LULs have to do it. If there are no AMs to do time-alignment, there seems to be no alternative to labelling the files by hand.

When labelling prompted speech (e.g. recited text), the phone labels are more-or-less given (i.e., from a phonological transcription of the text). The labeller is not really providing the labels, only the boundary points. It would be helpful if the task could be reduced to specifying phone boundaries in a given label file. In other words, if the task could be divided between SpeechCluster and the human: SpeechCluster generates a label file in a requested format with approximate times, and the human corrects it.

This was the idea behind segFake. SegFake detects the end-points of the speech in the wav file (currently it assumes a single continuous utterance) and evenly spreads a string of labels across the intervening time. A resulting TextGrid is shown in Figure 1. Clearly, the probability of any boundary being correct approaches zero, but the task facing the human labeller has been substantially simplified.

Fig. 1



```
MGT - [ ivan@bedwyr43: /home/ivan/wip/awtolabelu/ht _ □ ×
File New Term Edit Settings Help
Shell
Help on class SpeechCluster in module speechCluster:
class SpeechCluster
  Methods defined here:
  __init__(self, segFn='', debug=False)
  getHeadBody(self, fn, sep)
  getStartEnd(self, windowSize=0.25)
  getTierByName(self, tierName)
  merge(self, other)
  parseHead(self, head, format='esps')
    if header is name = value format: make dict
    else: dict keys are line nos
  parseTime(self, n, timeName)
    returns no. of seconds
  read_ESPS(self, fn)
```

We can phrase a more explicit description of the problem (see Table 8); once the problem has been thus specified, translating it into Python is simple (see Table 9), and then this tool can be used from the command line (see Table 10). `segFake` results, viewed in Praat, are shown in Figure 2.

Table 8: Pseudocode representation of `fakeSeg` problem

GIVEN:	wav file list of N labels
in the wav file, identify endpoints of speech: START, END	
$T = \text{END} - \text{START}$	
$L = T / N$	
Specify label boundaries, starting at START and incrementing by L	

Fig. 2



Table 9: segfake in python

Line	Code
1	def fakeLabel(fn, phoneList, tierName='Phone', outFormat='TextGrid'):
2	seg = SpeechCluster(fn)
3	segStart, segEnd, fileEnd = seg.getStartEnd()
4	width = (segEnd - segStart)*1.0 / len(phoneList)
5	tier = SegmentationTier()
6	# start with silence
7	x = Segment()
8	x.min = 0
9	x.max = segStart
10	x.label = SILENCE_LABEL
11	tier.append(x)
12	for i in range(len(phoneList)):
13	x = Segment()
14	x.label = phoneList[i]
15	x.min = tier[-1].max
16	x.max = x.min + width
17	tier.append(x)
18	# end with silence
19	x = Segment()
20	x.min = tier[-1].max
21	x.max = fileEnd

22	x.label = SILENCE_LABEL
23	tier.append(x)
24	tier.setName(tierName)
25	seg.updateTiers(tier)
26	outFormat = SpeechCluster.formatDict['.%s' \\ % outFormat.lower()]
27	return seg.write_format(outFormat)

Table 10: segFake usage

Usage
<pre>segFake.py -f <filename> -o (TextGrid esps htklab) <phones> segFake.py -d <dirname> -t <transcription filename> -o <format></pre>
Example
<pre>segFake.py -f amser012.wav -o TextGrid m ai hh ii n y n j o n b y m m y n y d w e d i y y n y b o r e segFake.py -d wav -t trans.txt -o TextGrid</pre>

3.5 Using SpeechCluster III: A Bigger Example: PyHTK

So far, SpeechCluster has been shown in fairly limited contexts, essentially as a file management tool to protect researchers from administrative drudgery. This was one of the key goals of SpeechCluster. We have seen this producing a quantitative effect: giving the researcher a bit more time, but not really changing the kind of work a researcher can do. The next example shows that SpeechCluster can have a qualitative effect too.

The Hidden Markov Model Toolkit (HTK) (Woodland *et al.* 1994) is a toolkit to build HMMs, primarily for Automatic Speech Recognition (ASR), but it is also beginning to be used for research in speech synthesis or Text-to-Speech (TTS). HTK also provides facilities for language modelling used in ASR, but is increasingly being applied to problems in other domains. Like DNA sequencing (e.g. Kin 2003). It is a *de facto* standard in academic speech technology research, and no doubt has similar penetration into commercial research and development, particularly with small and medium-sized enterprises (SMEs). Although it is not open-source software, it is available free of

charge, and the models generated can be used commercially with no license costs. Compared with other such toolkits (e.g. Sphinx and ISIP) it is usable, powerful, and accurate. Nevertheless, it is still not easy to use. HTK is:

- Difficult technically: the ideal HTK user is a computer scientist who understands HMM internals, is comfortable with the command-line, and can write supporting scripts as required; and,
- Complicated and time-consuming: use of HTK involves writing long chains of heavily parametrised commands, tests, adjustments and iterations.

This is no criticism of HTK, of course (HMM building is complicated), but the consequence is that its use is limited to computer scientists already working on speech technology research projects (mostly ASR). This is normal (all part of the academic way of institutionalising specialisation); however, it acts as a limit on the usability of language resources (i.e. corpora), and on the potential of language researchers.

PyHTK (Uemlianin 2005b) aims to change all that. *PpyHTK* is a Python wrapper around HTK, hiding the complexities of building and using HTK models behind a very simple command-line interface. A selection of commands from an HTK session is shown in Figure 3.

These commands are roughly equivalent to the command `pyhtk.py -a hmm4`.

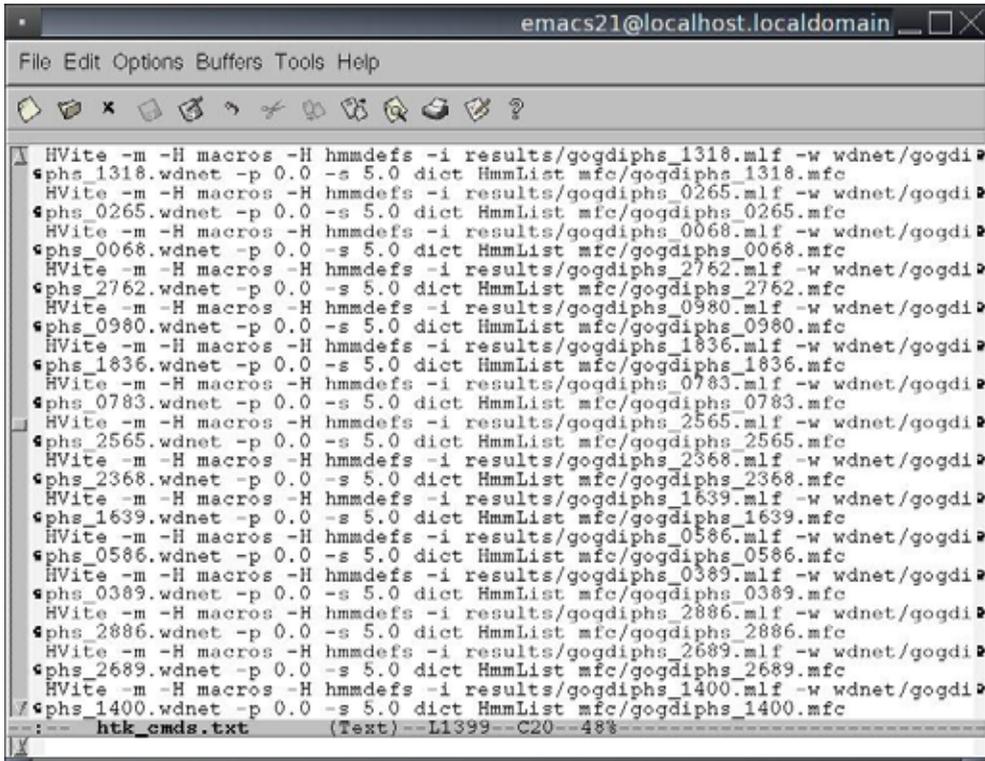
Nobody would type out all those HTK commands longhand. As in the case of some of the functions of *SpeechCluster*, each project will write their own little scripts to generate the commands. As with *SpeechCluster*, this reduplication of code is a huge and invisible waste of effort; and more so here than with *SpeechCluster*, writing scripts to run HTK requires a familiarity with HTK and at least some familiarity with the ins and outs of HMMs. HTK is very far from being 'plug-n-play'.

With *pyHTK* all that is required to get started is a speech corpus (i.e., a set of *wav* files) and some level of transcription. *PyHTK* uses *SpeechCluster* to put everything into the formats required by HTK, and then runs the necessary HTK commands to build a model and/or conduct recognition. In other words, *SpeechCluster* acts as an interface between HTK and your data, and *pyHTK* acts as an interface between you and HTK.

With *pyHTK* as an interface, HTK can be used with no knowledge or understanding at all of the underlying technology. It is perhaps true that 'a little knowledge is a dangerous thing,' but *pyHTK* should not be seen as promoting a lack of understanding. Rather, with *pyHTK* you can:

- 'Try out' ASR research and get more seriously involved if it looks worthwhile;

Fig. 3



```
emacs21@localhost.localdomain
File Edit Options Buffers Tools Help
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_1318.mlf -w wdnet/gogdi
sphs_1318.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_1318.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_0265.mlf -w wdnet/gogdi
sphs_0265.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_0265.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_0068.mlf -w wdnet/gogdi
sphs_0068.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_0068.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_2762.mlf -w wdnet/gogdi
sphs_2762.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_2762.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_0980.mlf -w wdnet/gogdi
sphs_0980.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_0980.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_1836.mlf -w wdnet/gogdi
sphs_1836.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_1836.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_0783.mlf -w wdnet/gogdi
sphs_0783.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_0783.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_2565.mlf -w wdnet/gogdi
sphs_2565.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_2565.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_2368.mlf -w wdnet/gogdi
sphs_2368.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_2368.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_1639.mlf -w wdnet/gogdi
sphs_1639.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_1639.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_0586.mlf -w wdnet/gogdi
sphs_0586.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_0586.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_0389.mlf -w wdnet/gogdi
sphs_0389.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_0389.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_2886.mlf -w wdnet/gogdi
sphs_2886.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_2886.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_2689.mlf -w wdnet/gogdi
sphs_2689.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_2689.mfc
Hvite -m -H macros -H hmmdefs -i results/gogdiphs_1400.mlf -w wdnet/gogdi
sphs_1400.wdnet -p 0.0 -s 5.0 dict Hmmlist mfc/gogdiphs_1400.mfc
:-- htk_cmds.txt (Text)--L1399-C20-48%
```

- Learn about the technology at your own pace, while you work, instead of having to cram up-front; and,
- Start working in a new area without having to hire a new team.

As part of *pyHTK*, *SpeechCluster* can have a qualitative effect on a team's potential: new areas of research and development (ASR, TTS and language modelling) become accessible. For example, we have built a diphone voice with Festival (Black *et al.* 1999). We have gathered the data (recording a phonetically balanced corpus of around 3000 nonsense utterances). Before we can build the voice, we must label the data (i.e., provide time-stamped phonological transcriptions). Labelling all the data by hand would have taken around 100 person-hours. In a small team, this kind of labour-time is not available.

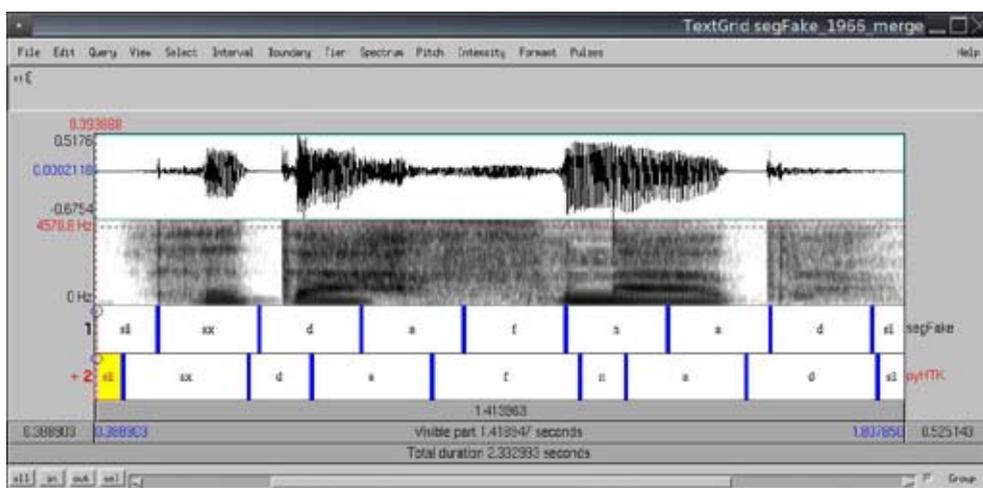
Instead, using *SpeechCluster* and *pyHTK*, we have been able to do the following:

- Use *segFake* to generate a starter segmentation of the data (we also manually transcribed just under 100 of the files, i.e., about 3% of the data).
- Iterate *pyHTK* twelve times overnight on the given segmentation. This involved: building an AM based on the given segmentation; re-labelling the

segFake'd training data with the AM; and, saving the generated segmentation for the next iteration.

The resulting segmentation would not satisfy a linguist- Figure 4 compares a segFake segmentation with an execution of this process- but the boundaries are sufficiently accurate to build a voice with Festival. Although we manually labelled a very small proportion of the data, we hypothesise that this had little effect on the quality of the final voice. In other words, SpeechCluster and pyHTK have enabled an almost fully automated build of a synthetic voice.

Fig. 4



4. Conclusion

Using SpeechCluster can save time and avoid a lot of stress. Developing SpeechCluster (using resources that would have been spent on repetitive chores) has resulted in a deliverable artefact: a reusable, shareable and extensible software package for manipulating speech data.

SpeechCluster has been developed as part of the WISPR project, and the facilities it offers reflect the tasks we have faced on WISPR. In future, SpeechCluster will accompany our work in a similar way; consequently it is not possible to predict entirely the development of SpeechCluster. However, two directions can be indicated:

- Corpora: It would be useful if SpeechCluster could treat a speech corpus with the same abstraction as it treats a sound/label file pair. In this case, the corpus, as a unit, could be described (e.g. counts of and relations between various

units) and manipulated (e.g. subset selection). This direction will include a layer for compatibility with EMU.

- Festival: We are developing a Python wrapper for Festival, similar to pyHTK for HTK. This development may have implications for changes in SpeechCluster.

SpeechCluster can be downloaded from the address given in the documentation (Uemlianin 2005a). We are making it available under an open-source (BSD) license. We take seriously the proposition that SpeechCluster should be a usable, shareable resource. We encourage researchers and developers in the field to use SpeechCluster, and we shall as far as possible maintain and update SpeechCluster in line with users' requests.

5. Acknowledgements

This work is being carried out as part of the project 'Welsh and Irish Speech Processing Resources' (WISPR) (Williams *et al.* 2005). WISPR is funded by the Interreg IIIA European Union Programme and the Welsh Language Board. I would also like to acknowledge support and feedback from other members of the WISPR team, in particular Briony Williams and Aine Ni Bhrian.

References

Black, A. W., Taylor, P. & Caley, R. (1999). *The Festival Speech Synthesis System*. <http://www.cstr.ed.ac.uk/projects/festival/>.

Ellis, N. C. et al. (2001). *Cronfa Electroneg o Gymraeg (CEG): A 1 Million Word Lexical Database and Frequency Count for Welsh*.

http://www.bangor.ac.uk/ar/cb/ceg/ceg_eng.html.

Huang, X. (2005). *Challenges in Adopting Speech Technologies*. CSTR-21. Edinburgh, September 2005.

Jones, R.J. et al. (1998). "SpeechDat Cymru: A Large- Scale Welsh Telephony Database." *Proceedings of the Workshop on "Language Resources for European Minority Languages*, May 27th 1998, Granada, Spain.

Kin, T. (2003) "Designing Kernels for Biological Sequence Data Analysis." Doctoral thesis. School of Knowledge Science, Japan Advanced Institute of Science and Technology.

Scannell, K.P. (2004). *Corpus Building for Minority Languages*. Online at <http://borel.slu.edu/crubadan/index.html>

Uemlianin, I. (2005a). *SpeechCluster Documentation*. Online at <http://www.bangor.ac.uk/~cbs007/speechCluster/README.html>

Uemlianin, I. (2005b). *PyHTK Documentation*. Online at <http://www.bangor.ac.uk/~cbs007/pyhtk/README.html>

Williams, B. (1999). "A Welsh Speech Database: Preliminary Results." *Proceedings of Eurospeech 99*, September 1999, Budapest, Hungary.

Williams, B., Prys, D. & Ni Chasaide, A. (2005). "Creating an Ongoing Research Capability in Speech Technology for two Minority Languages: Experiences from the

WISPR Project.” *Interspeech 2005*. Lisbon.

<http://www.bangor.ac.uk/ar/cb/wispr.php>

Woodland, P.C. et al. (1994). “Large Vocabulary Continuous Speech Recognition Using HTK.” *Acoustics, Speech, and Signal Processing*, ii, 125-128.

<http://htk.eng.cam.ac.uk/>

XNLRDF: The Open Source Framework for Multilingual Computing

Oliver Streiter and Mathias Stuflesser

XNLRDF (Natural Language Resource Description Framework) attempts to collect, formalise and formally describe NLP resources for the world's writing systems so that these resources can be automated in language-related applications like Web-browsers, mail-tools, Web-crawlers, information retrieval systems, or computer-assisted language learning systems. XNLRDF is a free software, distributed in XML-RDF or as database dump. It proposes to replace idiosyncratic ad-hoc solutions for Natural Language Processing (NLP) tasks within the aforementioned applications with a standard interface to XNLRDF. The linguistic information in XNLRDF extends the information offered by Unicode so that basic NLP tasks like language recognition, tokenization, stemming, tagging, term-extraction, and so forth can be performed without additional resources. With more than 1000 languages at use on the Internet (and their number continually rising), the design and development of such software has become a pressing need. In this paper, we describe the basic design of XNLRDF, the metadata, the type of information the first prototype already provides, and our strategies to further develop the resources.

1. XNLRDF as a Natural Extension of Unicode

1.1 *The Advantages of Unicode*

Unicode has simplified the completion of NLP tasks for many of the world's writing systems. Whereas, in the past, specific implementations were required, nowadays programming languages like Java, C++, C or Perl provide an interface to Unicode properties and operations. Unicode not only describes 'code elements'¹ of scripts by assigning the code element a unique code point, but it also assigns properties like uppercase, lowercase, decimal digit, mark, punctuation, hyphen, separator, or the script to the code elements. In addition, it defines operations on the code elements such as uppercasing, lowercasing and sorting. Thus, computer applications, especially those operating in multilingual contexts, are better off when processing texts in Unicode than in traditional encodings such as LATIN1, BIG5 or KOI-r.

¹ Similar, but not identical to characters (cf. our discussion in section 2.1).

1.2. *The Inadequacy of the Notions of Unicode for NLP Metadata*

On the other hand, the conceptual framework of Unicode is limited. Its principal notions are code elements and scripts. Important notions such as character, language or writing system have, astonishingly, no place in Unicode. As Unicode describes mainly scripts, two languages that use the same script (e.g., Italian and French) are essentially the same! The fact that French uses with 'ç' (the cedille) a character unknown in Italian is not formally described in Unicode. For this reason, additional knowledge (e.g., about languages, regions or legacy encodings) has been integrated into Unicode/Internationalisation programming libraries for a limited number of languages (e.g., ICU, Basis Technology Rosette, Lextek Language Identifier).²

As for the notion of language, it is not only absent from the formal framework of Unicode, but to our knowledge, nobody has attempted, except for limited purposes, a large-scale mapping between Unicode scripts and the world's most important language identification standards (i.e., ISO 639 and the SIL-codes of Ethnologue). This is astonishing, as neither the language code, nor the code of the locality of a document, nor the script taken in isolation are sufficiently rich to serve as metadata. Metadata in language-related applications have the function to map a document to be processed to the adequate NLP resources. In XNLRDF, the notion of writing system is used for this purpose. It represents a first large-scale attempt to map scripts onto language identification standards.

1.3 The Writing System in XNLRDF as Metadata

In XNLRDF, the writing system of a text document is an n-tuple of metadata categories, which include the language, the locality and the script as the most distinguishing ones. In Belgium, for example, text documents are produced (at least) in Dutch, French and German. The locality, therefore, is not enough as a single discriminative feature of these documents. Neither is the category language taken by itself, since Dutch, French and German are written in other countries as well. Furthermore, even the tuple language_locality, as it is frequently used (e.g., FR_be, NL_be), is not sufficient for all text documents and NLP resources. There exist localities that have two or more alternative scripts for the same language. For example, Serbian in Serbia and Montenegro is written with the Latin or Cyrillic scripts, and Hausa in Nigeria in the Latin or Ajami scripts.

An extended analysis of the world's writing systems reveals that at least four more categories are required for an unambiguous specification of the writing system. These

² For a detailed survey, see Unicode Inc. (2006).

categories are: the orthography, the writing standard, the time period of the writing system, and (for transliterations), a reference to another writing system.

Supporting evidence for the necessity of these categories comes, for example, from Abkhaz. Not only has Abkhaz been written with two different Cyrillic alphabets, but also with two different Latin alphabets, one between 1926 and 1928, and another between 1928 and 1937. One might want to distinguish these writing systems by their name (the standard) or by the time period. In such cases, we do not exclude the first solution, although there is frequently no standard name for the standard. If possible, we prefer to use the time period, as it offers the possibility to calculate intersections with other time constraints (e.g., the production date of the document).

The writing standard is best explained by the different, concurring, isochronic writing standards for Norwegian: Nynorsk, Bokmål, Riksmål and Høgnorsk are different contemporaneous conventions essentially representing the same language (<http://en.wikipedia.org/wiki/Norwegian>).

The orthography is best illustrated by the spelling reform of German, where the new orthography came into force in different localities at different times, and overlapped with the old spelling for a different number of years. Again, use of the time period is a nice feature, but it does not allow dispensing with the category of orthography. Unfortunately, orthographies, also frequently lacking a standard name, are referred to at the time of their introduction as 'new' in opposition to 'old.' This denomination of orthographies, however, becomes meaningless in a diachronic perspective where each 'new' orthography will eventually grow 'old.'

1.4 The Case of Braille and other Transliterations

Reference is necessary to correctly represent transliterations, that is, transliterations in the sense of one-to-one mappings, but also as one-to-many or many-to-many mappings. Reference introduces recursiveness into the metadata of the writing system, a complexity that is hard to avoid. Braille is a good example of a transliteration system that changes with the standards and spelling reforms of the referenced writing system. There exists a Norwegian Braille derived from Nynorsk, and a Norwegian Braille derived from Bokmål. By the same principle, Braille of the new German orthography is different from Braille based on the old German orthography.

Similarly, Braille changes with respect to the *locality of the Braille documents* that might differ in origin from the *locality of the referenced writing system*. For example, Spanish Braille in a Spanish-speaking country is different from the Spanish of a Spanish-speaking country represented as Braille in the USA. We can only handle

this complexity precisely when we allow writing systems to refer to each other recursively. Thus Braille, as with other transliteration systems, is represented as a writing system with its own independent locality, script, standard (e.g., contracted and non-contracted), and time period. The language of the transliteration and the referenced writing system are nevertheless the same, although XNLRDF would allow this to change for the transliteration as well.

A transliteration is thus marked by a reference to another writing system, and, in the descriptive data, mappings between these two systems in the form of a mapping table, (e.g., between Bokmål and Bokmål Braille). Mappings between writing systems are a natural component in the description of all writing systems, even if they do not represent transliterations of each other, (e.g., mappings between hànyǔ pīnyīn, wade-giles, gwoyeu romatzyh and bopomofo/zhùyīn fúhào). The Braille of Mandarin in the People's Republic, incidentally, is a transliteration of hànyǔ pīnyīn.

To sum up, the metadata needed to identify the appropriate or best NLP resources for the processing of a text-document are much more complex than what current standards have defined. In other words, relying on only one part of the metadata, such as the Unicode scripts or the language codes combined with locality codes, is not always accurate and thus not completely reliable for automated NLP-tasks. If NLP-technologies on the Web have, until now, not suffered from this important misconception (e.g., in the metadata specification in the HTTP or XML header), it is because they either target about two dozen common languages (applying default assumptions that prevent less frequently used writing systems from being correctly processed), or because a linguistically informed human mediates between documents and resources.

2. XNLRDF and Information Needs Beyond Unicode

Let us assume, for expository purposes, that an NLP-application can correctly identify the writing system of a document to be processed, and that this writing system contains references to Unicode scripts or code points. In effect, little follows from this, as Unicode defines only a very limited amount of operations, and defines them only for a script and not a writing system. The task of XNLRDF is thus to reformulate the operations defined in Unicode in the terms of a writing system, and, secondly, to enlarge the linguistic material so that more operations than those defined in Unicode become possible.

2.1 Unicode and Characters: Uppercasing, Lowercasing and Sorting

Contrary to a common sense understanding of Unicode, the conceptual design of Unicode avoids the notion of character, since this is a language-specific notion, and languages are not covered by Unicode. Unicode refers instead to code elements (), which frequently coincide with characters, but also contain combining characters such as diacritics. Characters and code elements further differ, if ligatures (Dutch 'ij', Spanish 'll', 'ch', Belorussian Lacinka 'dz') are to be treated as one character in a language. Uppercasing of ligatures is thus essentially undefined, and will produce from 'xy' uniformly either 'Xy' or 'XY', without knowing the requirements of the writing system. It is thus obvious that specifying the character set of writing systems and describing the mapping between the characters (e.g., for uppercasing and lowercasing) is one principle task in XNLRDF, just as lowercasing, for example, is an important step in the normalisation of a string (e.g., for a lexicon lookup or information retrieval).

Similarly, the sorting of characters, the second operation defined in Unicode (e.g., for the purpose of presenting dictionary entries or creating indices), depends on the writing system, and can only be approximately defined on the basis of the script. Thus, Unicode might successfully sort 'a' before 'b', but already the position of 'á' after 'a' or after 'z' is specific to each writing system. Another example is the Spanish 'll.' Although it is no longer considered a character, it maintains its specific position between 'l' and 'm' in a sorted list. Thus, sorting requires basic writing system-specific information, which XNLRDF sets out to describe. What this example also shows is that the definition of collating sequences for the characters of a writing system is independent from the status of the character (base character, composed character, contracted character, contracted non-character, context-sensitive character, foreign character, swap character, etc.).

2.2 Linguistic Information: What Else?

The operations covered by Unicode are limited, and most NLP-applications would require much more linguistic knowledge when processing documents in potentially unknown writing systems. First, an application might need to identify the encoding (e.g. KOI-R), the script (Cyrillic), the language (Russian), the standard (civil script), and orthography (after 1917) of a document. Part of this information might be drawn from the metadata available in the document, from the Unicode range, or the URL of a document (in our example, <http://xyz.xyz.ru>), but filling in the remaining gaps, (e.g., mapping from the encoding KOI-R to the language Russian, from the language to potential scripts, or from a script to a language) requires background information about the legacy encodings and writing systems. This background information is

available in XNLRDF. Information supporting the automatic identification of writing systems with no or incomplete metadata will also be supported by XNLRDF in the form of character n-grams. These n-grams are compiled from classified text samples or corpora within or without XNLRDF. Thus, for each writing system, XNLRDF allows to give information on URLs of other documents (of the same writing system), to raw text collections, and to elaborated corpora.

From the identified writing system, the application can start to retrieve additional resources that support the segmentation, stemming, hyphenation, and so forth of the document. A Web-crawler, for example, would try to find those text units (words and phrases) that are suitable for indexing. In most cases, the document will be segmented into words using a limited number of writing-specific word-separating characters (e.g., empty space, comma, hyphen, etc.). Although Unicode should provide this information, writing systems also differ as to which characters are unambiguous word separators, ambiguous ones, or not word separators at all. Thus, within those languages using Latin script, some integrate an empty space into a word, for example, Northern Sotho (Prinsloo & Heid 2006), while others like Lojban integrate ‘,’ and ‘.’ in the middle of a word. Unconditionally splitting a text in these languages with the empty space character (), a comma (‘,’), or a period (‘.’) would cut words into undefined chunks.

For writing systems that do not mark word boundaries (e.g., Han-characters or Kanji), the Web-crawler should index either each character individually (this is what Google does), or identify words through word lists and/or rules. Spelling variants (humour, humor), writing variants (Häuser, Haeuser, Häuser or □,□, Wan), inflected word forms (come, came), abbreviations (European Union, EU) should be mapped onto their base forms to improve the quality of document retrieval. All these are basic operations XNLRDF sets out to cover.

3. Difficulties in Obtaining Information beyond Unicode

The need for a linguistic extension of Unicode has long been recognised, and most of the information that applications, as the one sketched above, require is available from online resources. Thus, NLP-applications, at least theoretically, could get them automatically from the Web. If this were without problems, XNLRDF would be a redundant copy of other online information. However, for several reasons, the resources on the Web or the information contained within cannot be accessed, extracted and integrated by these applications (and by humans only with difficulty). First, there might exist difficulties to find and access the resources:

- Resources can not be found because metadata are not available; or,

-
- The resource is not directly accessible for applications: for example, accessing it requires transactions like registering, submitting passwords, entering the credit card number, etc.

Then, once a resource is found and accessed, there might be difficulties to extract or understand the necessary information, such as:

- The resource is not formally structured;
- The information within the resource is formally structured, but the syntax of the structure is not defined: for example, fields are separated by a special character, but the character used is not specified;
- The information is ambiguous, as in the following example: “Abkhaz is a North West Caucasian language with about 105,000 speakers in Georgia, Turkey and Ukraine (...) The current Cyrillic-based system” (<http://www.omniglot.com/writing/abkhaz.htm>), which does not specify which region actually is using or not using the Cyrillic-based script at present;
- The syntax is defined, but the semantics of the units are not as defined as it could be through using XML namespace, and so forth. With a namespace, a NOUN-tag can be linked to the URL containing the definition of the tag. Thus, different NOUN-tags could be used without confusion;
- The information in the different resources is not compatible, that is, the notion of language varies greatly between resources. To give one example, what the Office of the High Commissioner for Human Rights (Universal Declaration of Human Rights) describes as Zapoteco is not covered in Omniglot, and is split into more than fifty languages by Ethnologue and the Rosetta Project ; or
- Most resources are *language-centred* and do not put the *writing system* into the centre of the description. To understand how serious this misconception is, imagine you search a Chinese document and get Chinese in Braille, which is Chinese to the same degree as what you expected to get.

In view of all this, there is an enormous need to bring the available resources together and make them compatible, available and parsable; otherwise, the information will be barely usable for NLP-applications. This compiling work necessarily involves a combination of the linguists’ careful classification, description, and automated approaches to knowledge acquisition. Both techniques will first exploit other resources relevant for XNLRDF.

4. Related Activities and Available Resources

Fortunately, XNLRDF is embedded in a wide field of research activities that create, document and make accessible natural language resources. What makes XNLRDF stand out in its field is its focus on Natural Language Processing resources on the one hand, and fully-automated access to the data by an application on the other. Nevertheless, XNLRDF will try to profit from related projects and to comply with available standards.

Repositories of the world's languages are available online. Figuring most prominently among them are: Omniglot, Ethnologue, The Rosetta Project, TITUS, and the Language Museum (<http://www.language-museum.com>). Although these resources offer rich information on scripts and languages, they are almost unusable for computer applications, as they are designed for human users. The difficulties in using Ethnologue, for example, derive from its focus on spoken languages and its tendency to introduce new languages where others just see regional variants of the same language. This problem has been inherited by the Rosetta Project and the World Atlas of Language Structures (Haspelmath *et al.* 2005). In addition, some sites (e.g., the Language Museum) use scanned images of characters, words and texts that of course are almost impossible to integrate into NLP resources. Still other sites (e.g., TITUS) use mainly transcriptions or transliterations that are equally worthless without a formal definition of the mappings applied. Currently, the information available on these sites is checked and integrated manually into the XNLRDF data structure.

OLAC, the Open Language Archives Community project, is setting up a network of interoperating repositories and services for hosting and accessing NLP resources. The project's aims and approaches are very close to those of XNLRDF, and we foresee a considerable potential for synergy. The metadata and their definition is what will be most relevant to XNLRDF. However, the OLAC user scenario assumes a human user looking for resources and tools, whereas XNLRDF is designed to allow applications to find resources autonomously given a text document to be processed and a task to be achieved.

Closely related to OLAC is the E-MELD project, which supports the creation of persistent and reusable language resources. In addition, queries over disparate resources are envisaged. To which extent XNLRDF can profit from E-MELD has yet to be investigated in detail.

Data consortia like ELRA or LDC host NLP resources that can be identified through the machine-readable metadata in OLAC. However, resources are not freely accessible. Commercial transactions are required between the identification of the

resource and the access to the resource. For this reason, these resources will remain unexplored, even if prices are modest. Although ELRA and LDC have their merits, for small languages, better solutions are available for the hosting of data (cf. Streiter 2005).

Project Gutenberg provides structured access to its 16,000 documents (comprising about thirty languages) through an XML-RDF. Unfortunately, information characterising text T1 as translation of T2 is still not provided, that is, although parallel corpora are implicitly present, they are not identifiable as such. In theory, the documents of Project Gutenberg could be used to build up corpora in XNLRDF. Such a copying of resources, however, might only be justifiable for writing systems for which little corpus material is available. More important might be a mapping from the writing system of XNLRDF to the documents of Project Gutenberg, thus translating the available XML-RDF in terms of XNLRDF.

Free monolingual and parallel corpora are available at a great number of sites, most prominently at <http://www.unhchr.ch/udhr/navigate/alpha.htm> (Universal Declaration of Human Rights in 330 languages), <http://www.translatum.gr/bible/download.htm> (The Bible), and The European Parliament Proceedings Parallel Corpus (<http://people.csail.mit.edu/koehn/publications/europarl>), among others. Those documents that support otherwise underrepresented writing systems will be integrated into XNLRDF in the form of corpora.

The Wikipedia project is interesting for XNLRDF for a number of reasons. First, it provides documents that can be used to build corpora without infringing upon copyrights. Second, as the Wikipedia is available in more than one hundred languages, thousands of quasi-parallel texts become accessible. Third, the model of cooperation in Wiki projects, and the underlying software, will indicate the way XNLRDF will go. Thus, XNLRDF will gradually enlarge the community of researchers involved to the point that the world's linguists will be able to collect the data they need for their writing systems. This issue will be further discussed below.

5. Conceptual Design of XNLRDF

The purpose of XNLRDF is to find adequate NLP resources to process a text document. To this end, the metadata of the document and the resource are matched. The better the match, the more suitable the resource is for the processing of the document. The metadata matched are those categories that make up the writing system.

5.1 Finding Resources via the Writing System

The writing system has a function similar to SUBJECT.LANGUAGE in the OLAC-metadata, defined in Simon & Bird (2001) as “[...] a language which the content of the resource describes or discusses.” A writing system in XNLRDF is defined by the n-tuple of the category’s language, locality, script, orthography, standard, time period, and reference to another writing system. The writing system is a property of the text document and the resource. In XNLRDF, for each writing system there is a more abstract writing system (e.g., without constraints in locality) as a fallback solution to fill in empty categories with default assumptions. In general, for each language there is one writing system without a locality that provides a default locality in the event that no locality can be derived from the document. (Cf. Plate 1: Different Writing Systems for Mandarin Chinese. The first row is the fallback with the default locality.) These underspecified writing systems are currently also used (and perhaps incorrectly) for supranational writing systems, (e.g., English-based writing in the UN).

Plate 1: Writing Systems for Mandarin Chinese. Note the first row showing Chinese without locality as a super-regional language. In case of doubt, the application has to assume China as the locality where the text-document originated.

id	lg id	loc id	orth id	script id	default lg	default loc	default enc	valid from	valid to	default to 1	source
9105	->chinese, mandarin	0	0	->chinese simplified		->china	->gb2312	->1949-01-01	->3000-01-01	->228	->en.wikipedia.org
41396	->chinese, mandarin	->china	0	->chinese traditional			->utf8	->0500-01-01	->3000-01-01	->9105	->en.wikipedia.org
31121	->chinese, mandarin	->china	0	->chinese simplified			->gb2312	->1949-01-01	->3000-01-01	->9105	->en.wikipedia.org
41333	->chinese, mandarin	->singapore	0	->chinese traditional			->utf8	->1965-01-01	->3000-01-01	->9105	->en.wikipedia.org
41270	->chinese, mandarin	->singapore	0	->chinese simplified			->gb2312	->1980-01-01	->3000-01-01	->9105	->en.wikipedia.org
37520	->chinese, mandarin	->taiwan	0	->chinese traditional			->big5	->1950-01-01	->3000-01-01	->9105	->en.wikipedia.org

The inclusion of a writing system into XNLRDF is pragmatically handled. Included are all writing systems for which text documents with yet uncovered combinations of language, locality, and so forth can be found. The same pragmatic approach is used to (or not to) distinguish languages and dialects. Thus, dialects are treated identically to languages, whenever documents of that variant are found (e.g., Akan Akuapem, Akan Asante and Akan Fante). A writing that claims to be representing a language family is registered with a writing system of this language family. The same goes for localities; whenever a document is reasonably associated with one region - even if that region is not a recognised geographical, administrative or economical body - the region will be included as locality.

5.2 The Names of Metadata Categories

All this leads to the overall problem that, for the main categories of the writing system, no standardised identifiers are available. We already discussed the lack of standard names for the standard and orthography of a writing system. But in addition, languages, localities and scripts do not necessarily have standard names or standard codes, albeit XNLRDF tries to integrate the ISO 339 codes for languages (ISO 639 2006) (the 2-letter code for languages ISO-639-1 and the 3-letter code for languages ISO-639-2), the SIL-codesVersion 14 of Ethnologue, the Unicode-naming of scripts, and ISO-3166 (ISO 3166 2006) encoding of localities (countries, regions, and islands).

A number of limitations, however, make these codes difficult to use: ISO-639-1 covers only a few languages; ISO-639-2 assigns more than one code to one language; both ISO norms assign one code to sets of languages, language families, and so on; and, SIL-codes change from version to version (about every four years), and do not cover historic languages, artificial languages, language groups or languages that exist only as written standard.

The situation for the encoding of languages will improve with the adoption of the draft ISO/DIS 639-3 as a standard (presumably in 2006), as it will combine the respective advantages of the SIL-codes and the ISO-codes. Until then, applications will continue to use the RFC 3066 standard for HTTP headers, HTML metadata and in the XML lang attribute. 2 and 3-letter codes are interpreted as ISO-639-1 or ISO-639-2 respectively. ISO-639-1 can be mapped on ISO-639-3, and ISO-639-2 is identical to ISO-639-3, so that, in the future, only ISO-639-1 (transitional) and ISO-639-3 will be needed (for more information on this development, consult the webpages http://en.wikipedia.org/wiki/ISO_639-3, <http://www.ietf.org/rfc/rfc3066.txt> and <http://www.ethnologue.com/codes/default.asp>). SIL-codes will then become superfluous in XNLRDF, and languages that are not written can be removed from XNLRDF. The advantage of ISO-639-3 is that it can group together individual spoken languages (such as two dozen spoken Arabic languages) to 'macro languages' (Arabic), thus preventing writing systems from being fragmented due to the fragmentation of languages.

Most reliably, however, the categories of the writing system can be accessed with their natural language name in one the world's major writing systems, for which XNLRDF guarantees an unambiguous match. As a consequence of this recursion, as outlined in Gödel's 'Incompleteness Theorems', neither the names nor the categories can be formally defined; they can only be explained by the use they are put to (e.g., the material that is attached to a name). Fortunately, this problem is not inherent

to XNLRDF, but is also shared by other classification standards like ISO norms and SIL codes.

5.3 Linguistic Information for Writing Systems

A writing system is associated via a resource type with the corresponding resources. Writing systems stand in a many-to-many relation to encoding (Plate 2), numerals (Plate 3), and function words (Plate 4); characters; sentence separators; word separators; URLs (classified according to genres); dictionaries; monolingual and parallel corpora; and, n-gram statistics.

Plate 2: A writing system (Mandarin Chinese in Taiwan) related to ENCODING.

id	wr id	enc id	source
227	->37520	->utf8	-> www.basistech.com
1759	->37520	->utf16	-> www.unicode.org/onlinedat/languages-scripts.html
1760	->37520	->utf32	-> www.unicode.org/onlinedat/languages-scripts.html
228	->37520	->big5	-> www.basistech.com
1761	->37520	->utf-7	-> www.unicode.org/onlinedat/languages-scripts.html

Plate 3: A writing system (Thai) related to NUMERALS.

id	wr id	number	arabic numeral	source
377	->27286	->๙	->9	-> www.alanwood.net/unicode/thai.html
375	->27286	->๘	->8	-> www.alanwood.net/unicode/thai.html
373	->27286	->๗	->7	-> www.alanwood.net/unicode/thai.html
371	->27286	->๖	->6	-> www.alanwood.net/unicode/thai.html
369	->27286	->๕	->5	-> www.alanwood.net/unicode/thai.html
367	->27286	->๔	->4	-> www.alanwood.net/unicode/thai.html
365	->27286	->๓	->3	-> www.alanwood.net/unicode/thai.html
363	->27286	->๒	->2	-> www.alanwood.net/unicode/thai.html
362	->27286	->๑	->1	-> www.alanwood.net/unicode/thai.html
331	->27286	->๐	0	-> www.alanwood.net/unicode/thai.html

Plate 4: A writing system (Thai) related to FUNCTION_WORDS.

id	wr id	function word	determiner	article	quantifier	marker	question marker	imperative marker	request marker	topic marker	contrastive topic marker	focus marker	case marker	negation marker	preposition
246	>27286	->แล้ว	0	0	0	0	0	0	0	0	0	0	0	0	0
273	>27286	->จน	0	0	0	0	0	0	0	0	0	0	0	0	0
278	>27286	->จนกว่า	0	0	0	0	0	0	0	0	0	0	0	0	0
237	>27286	->อยู่	0	0	0	0	0	0	0	0	0	0	0	0	0
272	>27286	->แล้ว	0	0	0	0	0	0	0	0	0	0	0	0	0
277	>27286	->แล้ว	0	0	0	0	0	0	0	0	0	0	0	0	0
219	>27286	->การ	0	0	0	0	0	0	0	0	0	0	0	0	0
279	>27286	->ที่	0	0	0	0	0	0	0	0	0	0	0	0	0
233	>27286	->การ	0	0	0	0	0	0	0	0	0	0	0	0	0
282	>27286	->มา	0	0	0	0	0	0	->1	0	0	0	0	0	0
241	>27286	->ไป	0	0	0	0	0	0	0	0	0	0	0	0	0
286	>27286	->ไป	0	0	0	->1	0	0	0	0	0	0	0	0	0
284	>27286	->ไป	0	0	0	->1	0	0	0	0	0	0	0	0	0
280	>27286	->ไป	0	0	0	0	0	0	0	0	0	0	0	0	0
294	>27286	->ไป	0	0	0	->1	0	0	0	0	0	0	0	0	0

5.4 Methods and Implementation

The data-model is implemented in a relational database, which provides all means to control the coherence of the data, create backups, and allow people from different parts of the world to work on the same set of data. For applications working with relational databases, this data can be downloaded under the GNU Public License as database dump (PostgreSQL). An Interface to the database has been created as a working tool for the creation and maintenance of data.

An additional goal is to make XNLRDF available in XML-RDF. RDF, a framework for the description of resources, has been designed by the W3C to describe resources with their necessary metadata for applications rather than for people (Manola & Miller 2004). Whereas, in the relational database, the defaulting mechanism is programmed as a procedure, in XML-RDF defaults are compiled out. In this way, the information in XNLRDF can be accessed through a simple look-up with a search string such as 'Thai', 'Thailand', 'Thai;Thailand', and so forth.

6. Envisaged Usage and Impact

In order to give a word-to-word translation, for example, within a Web-browser, the Web-browser has to know where to find a dictionary and how to use it. With only one such resource, a special function within the Web-browser might handle this (e.g., a number of FIREFOX add-ons do exactly this). But with hundreds of language resources, a more general approach is required that not only involves adequate resources, but also metadata with an NLP-specific metadata dictionary and metadata syntax. NLP-operations like tagging or meaning disambiguation for annotated reading have then to be defined recursively in the metadata syntax: in this way, a tagger can call a tokenizer if it can't perform tokenization itself.

The substantiation of the concept of XNLRDF will thus consist of compiling XNLRDF into an Mozilla-compatible RDF and integrating it into an experimental Mozilla module. Not only is Mozilla a base for a great number of very popular applications (e.g., Firefox, Thunderbird, Bugzilla, Netscape, Mozilla Browser, and Mozilla e-mail), but it also disposes of an RDF-Machine that can be accessed via JavaScript and XPConnect (Boswell *et al.* 2002). A minor test-application of XNLRDF in Mozilla might thus have a tremendous impact.

Less spectacular than the still pending integration into Mozilla is the testbed where XNLRDF is currently used. It serves as a linguistic database for Gymn@zilla, a CALL system that handles about twenty languages, with new languages added on a regular basis (Streiter *et al.* 2005). In general, CALL systems are very likely to be the first applications to profit from XNLRDF. They are frequently applicable to many languages and require relatively uncomplicated operations. In fact, many CALL modules are freely available, and, to some extent, language independent (e.g., Hot Potatoes). In practice, however, they are often only suited for an undefined group of languages (e.g., they require a blank to separate words). With the linguistic intelligence of XNLRDF, such modules could not only extend the range of languages, but also generate better exercises and provide better feedback.

Web-crawlers and IR systems are other candidates that will certainly profit from XNLRDF. While most IRs may be tuned to one or a few languages, they generally lack the capacity to process a wide range of languages. The large amount of NLP systems integrated in Google shows the importance of linguistic knowledge in IR.

To sum up, we not only hope to bring many more languages to text document processing applications, but hope to do this in a standard format that can be easily processed by XML or XML-RDF-enabled applications.

7. Status of the Project and Future Developments

The project is still an unfunded garage project. In the previous project phase, we defined the base and implemented the first model in a relational database. An interface to that database has been created to allow new data to be entered via the WWW. After inserting more than 1000 writing systems and getting a better understanding of the framework necessary to describe a writing system, we are currently adding linguistic information to describe the writing systems. The data structures for characters, corpora, dictionaries, and so forth are still changing when new requirements or linguistic complexities are encountered. URLs and corpora are collected to support the description of the writing system and as useful material to be integrated in XNLRDF for NLP-applications (e.g., for the creations of word lists).

In the meantime, we hope to attract more researchers to collaborate in the project. It is impossible to answer now, whether or not the project will be as open as the Wikipedia. It is certain, however, that this endeavour will require the collaboration of a wide range of researchers around the globe. Very likely, small tools will be created around XNLRDF that will illustrate the use the resource can be put to, and motivate linguists to enter data for their language (writing system). Such tools will have the additional advantage to check the accuracy and completeness of the data.

8. Glossary

Language

Language is one of the discriminating features that defines a writing system in XNLRDF. XNLRDF uses language identification standards such as ISO-639-1 and ISO-639-2 to map language names to unambiguous language codes.

Locality

Locality is one of the discriminating features that defines a writing system. ISO 3166 is the standard that defines locality codes. However, XNLRDF pragmatically includes a region as locality, whenever there is a document that is reasonably associated with the region. This applies also even if the region is not a recognised geographical, administrative or economic body.

Metadata Categories

Metadata in language-related applications have the function to map a document to be processed to the appropriate NLP (natural language processing) resources. XNLRDF uses the categories of language, locality, script, orthography, standard, time period,

and reference to another writing system. Together, these NLP metadata categories in XNLRDF define a writing system.

Natural Language Resource

A natural language resource in XNLRDF refers to structured linguistic information and/or NLP applications that are accessible to machines via a clearly defined writing system. Types of resources include, for example, encoding, numerals, function words, characters, sentence separators, word separators, URLs, dictionaries, corpora, and n-gram statistics, as well as applications for basic NLP tasks such as language recognition, tokenization, stemming, tagging, segmentation, hyphenation, and indexing of complex NLP implementations such as term extraction, document retrieval, meaning disambiguation, and ComputerAssisted Language Learning tools.

Orthography

Orthographies can sometimes be tracked using the time period category. However, different orthographies might coexist for a certain time span (e.g., in German, after the latest orthography reform). Therefore, locality is one of the discriminating features that defines a writing system.

Reference

Reference is used in XNLRDF to describe transliterations. The transliteration is a writing system on its own, but can only be understood and correctly processed when referring to another underlying writing system. For example, a text written in Braille can only be understood and processed when referred to the underlying writing system (e.g., Braille referring to standard German in Austria in 'new orthography'). Reference is a recursive category. It is one of the discriminating features that defines a writing system.

Script

In Unicode, legacy scripts are named (e.g., Latin, Arabic and Cyrillic). XNLRDF uses these script names as a discriminating feature to define writing systems.

Time Period

The time period offers the possibility to calculate intersections with other time constraints (e.g., between the validity of an orthography and the production date of the document). Therefore, time period is one of the discriminating features that defines a writing system.

Writing Standard

Sometimes the same language can be written in different, concurring, isochronic writing standards. For example, Nynorsk, Bokmål, Riksmål and Høgnorsk are different contemporaneous conventions that represent Norwegian. Therefore, writing standard is one of the discriminating features that defines a writing system.

Writing System

The writing system helps to map a document to the adequate NLP resources necessary to process the document. A writing system in XNLRDF is defined by the n-tuple of language, locality, script, orthography, standard, time period and reference to another writing system. In XNLRDF, for each writing system there are also more abstract writing systems -(e.g., those without constraints in locality) as a fallback to fill in the missing information with default assumptions.

XNLRDF

XNLRDF stands for 'Natural Language Resource Description Framework.' It is an Open Source Framework for Multilingual Computing, designed to allow applications to find language resources autonomously, given a text document to be processed and a task to be achieved. XNLRDF is distributed either in XML-RDF or as database dump.

References

Boswell, D. *et al.* (2002). *Creating Applications with Mozilla*. Sebastopol: O'Reilly.

"E-MELD". Online at <http://emeld.org>.

"Ethnologue". Online at <http://www.ethnologue.com>.

"European Parliament Proceedings Parallel Corpus 1996-2003". Online at <http://people.csail.mit.edu/koehn/publications/europarl>.

Haspelmath, M. *et al.* (eds) (2005). *The World Atlas of Language Structures*. Oxford: Oxford University Press.

"Hot Potatoes". Online at <http://web.uvic.ca/hrd/halfbaked>. ISO 639 (1989). Code for the representation of the names of languages.

ISO 3166-1 (1997). Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes.

ISO 3166-2 (1998). Codes for the representation of names of countries and their subdivisions -- Part 2: Country subdivision code.

ISO 3166-3 (1999). Codes for the representation of names of countries and their subdivisions -- Part 3: Code for formerly used names of countries.

Manola, F. & Miller, E. (eds) (2004). "W3C Recommendation 10". *RDF Primer*, February 2004. <http://www.w3.org/TR/rdf-primer/>.

Norwegian (2006, March 7). In Wikipedia, The Free Encyclopedia. Retrieved March 7, 2006. Online at <http://en.wikipedia.org/wiki/Norwegian>.

"OLAC, the Open Language Archives Community project". Online at <http://www>.

language-archives.org/documents/overview.html.

“Omniglot”. Online at <http://www.omniglot.com>.

Prinsloo, D. & Heid, U. (this volume). “Creating Word Class Tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping”, 97-115.

“Rosetta Project”. Online at <http://www.rosettaproject.org>.

Simons, G. & Bird S. (eds) (2001). *OLAC Metadata Set*. <http://www.language-archives.org/OLAC/olacms.html>.

Streiter, O. (this volume). *Implementing NLP-Projects for Small Languages: Instructions for Funding Bodies, Strategies for Developers*, 29-43.

Streiter, O. *et al.* (2005). “Dynamic Processing of Texts and Images for Contextualized Language Learning”. *Proceedings of the 22nd International Conference on English Teaching and Learning in the Republic of China (ROC-TEFL)*, Taipei, June 4-5, 278-98.

“TITUS”. Online at <http://titus.uni-frankfurt.de>.

“Translatum”. Online at <http://www.translatum.gr/bible/download.htm>.

“Unicode Enabled Products”. Online at <http://www.unicode.org/onlinedat/products.html>.

“Universal Declaration of Human Rights” Online at <http://www.unhchr.ch/udhr/index.htm>.

Speech-to-Speech Translation for Catalan

Victoria Arranz, Elisabet Comelles and David Farwell

This paper focuses on a number of issues related to adapting an existing interlingual representation system to the ideosyncracies of Catalan in the context of the FAME Interlingual Speech-to-Speech Machine Translation System for Catalan, English and Spanish. The FAME translation system is intended to assist users in making hotel reservations when calling or visiting from abroad. Following a brief presentation of the Catalan language, we describe the system and review the results of a major user-centered evaluation. We then introduce Interchange Format (IF), the interlingual representation system underlying the translation process, and discuss six types of language-dependent problems that arose in extending IF to the treatment of Catalan, along with our approach to dealing with these problems. They include the lack of dialog-level structural relationships, conceptual gaps, the lack of register distinctions (e.g. specifically and formality), the treatment of proper names, the lack of a method for dealing with partitives and conceptual overgranularity. Finally, we summarise the contents and suggest some future directions for research and development.

1. Introduction

The goal of this paper is to review a number of problems that arose in adapting an existing Interlingua, Interchange Format (IF), to the treatment of Catalan, and to describe our approach to dealing with them. As classes, these problems are not peculiar to Catalan per se, but the language presents an interesting case study in terms of their particular manifestation and how they might be dealt with. They include the need for representing dialogue-level structural relations, dealing with conceptual gaps, the need for representing register distinctions, a semi-productive method for dealing with proper names, the need for representing partitive references, and dealing with conceptual overgranularity. This effort was part of the development of the FAME Interlingual Speech-to-Speech Machine Translation System for Catalan, English and Spanish, which was carried out between 2001 and 2004.

In section 2, we provide a background for the discussion, giving some information on Catalan language, and describing the project and the translation system. In section 3, we briefly describe an evaluation procedure, and present some results from a major user-centred evaluation. This section proves the feasibility of the adaptation and the success of the system, which was publicly demonstrated at the 2004 Forum

of Cultures in Barcelona (with a very positive outcome). In section 4, we discuss Interchange Format (IF), the interlingua underlying the translation process, the inadequacies encountered and the modifications made while adapting the framework to Catalan and Spanish. Finally, in Section 5, we summarize the results and conclude with a discussion of future directions.

2. Background

The Catalan language, with all its variants, is spoken in the Païssos Catalans which include the Spanish regions of Catalonia, Valencia and Balearic Islands, the French department of the Pyrénées Orientales, and the Italian area of Alghero. Inside the Spanish territory, Catalan is also spoken in some parts of Aragon and Murcia. Catalan is a Romance language and shows similarities with other languages belonging to the Romance family, in particular with Spanish, Galician and Portuguese. Nowadays, Catalan is understood by 9 million people and spoken by 7 million people.

The FAME Interlingual Speech-to-Speech Translation System for Catalan, English and Spanish was developed at the Universitat Politècnica de Catalunya (UPC), Spain, as part of the recently completed European Union-funded FAME project (Facilitating Agents for Multicultural Exchange) that focused on the development of multi-modal technologies to support multilingual interactions (see <http://isl.ira.uka.de/fame> for details). The FAME translation system is an extension of the existing NESPOLE! translation system (Metze *et al.* 2002; Taddei *et al.* 2003) to Catalan and Spanish in the domain of hotel reservations. At its core is a robust, scalable, interlingual speech-to-speech translation system having cross-domain portability that allows for effective translingual communication in a multi-modal setting. Although the system architecture was initially based on NESPOLE!, all the modules have now been integrated on an Open Agent platform (Holzapfel *et al.* 2003; for details see <http://www.ai.sri.com/~aaa>). This type of multi-agent framework offers a number of technical features for a multi-modal environment that are highly advantageous for both system developers and users.

Broadly speaking, the FAME translation system consists of an analysis component and a generation component. The analysis component automatically transcribes spoken source language utterances and then maps that transcription into an interlingual representation. The generation component maps from interlingua into target language text that, in turn, is passed to a speech synthesiser that produces a spoken version of the text. The central advantage of this interlingua-based architecture is that, in adding additional languages to the system (such as Catalan and Spanish), it is only necessary to develop new analysis and generation components for each new language

in order to be able to translate into and out of all of the other existing languages in the system. In other words, no source-language-to-target-language specific transfer modules are required, as would be the case for transfer systems, with the result that the development task is considerably simplified.

For both Catalan and Spanish speech recognition, we used the JANUS Recognition Toolkit (JRTk) developed at Universität Karlsruhe and Carnegie Mellon University (Woszczyzna *et al.* 1993). For the text-to-text component, the analysis side utilises the top-down, chart-based SOUP parser (Gavaldà 2000) with full domain action level rules to parse input utterances. Natural language generation is done with GenKit, a pseudo-unification-based generation tool (Tomita *et al.* 1988). For both Catalan and Spanish, we use a Text-to-Speech (TTS) system fully developed at UPC, which uses a unit-selection-based, concatenative approach to speech synthesis.

The Interchange Format (Levin *et al.* 2002), the interlingua used by the C-STAR Consortium (see <http://www.c-star.org> for details), has been adapted for this effort. Its central advantage in representing dialogue interactions such as those typical of speech-to-speech translation systems is that it focuses on identifying the speech acts and the various types of requests and responses typical of a given domain. Thus, rather than capturing the detailed semantic and stylistic distinctions, it characterises the intended conversational goal of the interlocutor. Even so, in mapping into or out of IF, it is necessary to take into account a wide range of structural and lexical properties related to Catalan and Spanish.

For the initial development of the Spanish analysis grammar, the already existing NESPOLE! English and German analysis grammars were used as a reference point. Despite using these grammars, great efforts had to be made to overcome important differences between English, German and the Romance languages in focus. The Catalan analysis grammar, in turn, was adapted from the Spanish analysis grammar, and, in this case, the process was rather straightforward. The generation grammars for Catalan and Spanish were mostly developed from scratch, although some of the underlying structure was adapted from that of the NESPOLE! English generation grammar. Language-dependent properties such as word order, gender and number agreement, and so forth needed to be dealt with representationally, but on the whole, starting with existing structural descriptions proved to be useful. On the other hand, the generation lexica play a significant role in the generation process and these had to be developed from scratch. As for the generation grammars, however, a considerable amount of work took place in parallel for both Romance languages, which contributed to a more efficient development of both the Catalan and Spanish generation lexica.

3. Evaluation

The evaluation performed was done on real users of the speech-to-speech translation system, in order to both:

- examine the performance of the system in as real a situation as possible, as if it were to be used by a real tourist trying to book accommodation in Barcelona; and,
- study the influence of using speech input, and thus Automatic Speech Recognition (ASR), in translation.

3.1 Task-Oriented Evaluation Metrics

A task-oriented methodology was developed to evaluate both the end-to-end system (with ASR and TTS) and the source language transcription to target language text subcomponent. An initial version of this evaluation method had already proven useful during system development, since it allowed us to analyse content and form independently, and thus contributed towards practical system improvements.

The evaluation metric used recognises three main categories (Perfect, Ok and Unacceptable), where the second was further subdivided into Ok+, Ok and Ok-. During the evaluation, this metric was independently applied to two separate parameters, form and content. In order to evaluate form, only the generated output (text or speech) was considered by the evaluators. To evaluate content, evaluators took into account both the input utterance or text and the output text or spoken utterance. Accordingly, the meaning of the metrics varies depending on whether they are being used to judge form or to judge content:

- Perfect: well-formed output (form) or communication of all the information the speaker intended (content).
- Ok+/Ok/Ok-: acceptable output, grading from only some minor error of form (e.g. missing determiner) or some minor uncommunicated information (Ok+) to some more serious problem of form or uncommunicated information content (Ok-).
- Unacceptable: unacceptable output, either essentially unintelligible (form) or information unrelated to the input (content).

3.2 Evaluation Results

The results obtained from the evaluation of the end-to-end translation system for the different language pairs are shown in Tables 1, 2, 3 and 4. The results obtained from the translation of clean audio-transcriptions are summarised in Tables 5, 6, 7 and 8. From the results, we can conclude that many of the errors are caused by the

ASR component. This is particularly so when translating from English¹ into Catalan or Spanish. For instance, if we consider the form parameter, Tables 7 and 8 show that there are no unacceptable translations when using the text-to-text interlingual translation system for the English-Catalan and English-Spanish, while Tables 3 and 4 show that performance drops 5.99% and 9.6%, respectively, when using the speech-to-speech system.

In fact, the interlingual translation component performs very well when used on text input and degrades when using speech input. However, it should be pointed out that, even so, results remain rather good for the end-to-end system. For the worst of our language pairs (English-Spanish), a total of 62.4% of the utterances were judged acceptable in regard to content. This is comparable to evaluation results of other state-of-the-art systems such as NESPOLE! (Lavie *et al.* 2002), which obtained slightly lower results and was performed on Semantic Dialog Units (see below) instead of utterances (UTT), thus simplifying the translation task. The Catalan-English and English-Catalan pairs were both quite good with 73.1% and 73.5% of the utterances being judged acceptable, respectively, and the Spanish-English pair performs very well with 96.4% of the utterances being acceptable.

Table 1: Evaluation of End-to-End Translation (with ASR) for the Catalan-English Pair. Based on 119 UTTs.

SCORES	FORM	CONTENT
Perfect	70.59%	31.93%
OK+	5.04%	15.12%
OK	6.72%	9.25%
OK-	9.25%	16.80%
Unacceptable	8.40%	26.90%

Table 2: Evaluation of End-to-End Translation (with ASR) for the Spanish-English Pair. Based on 84 UTTs.

SCORES	FORM	CONTENT
Perfect	92.85%	71.42%
OK+	4.77%	11.90%

¹ It should be pointed out that the efforts to develop the ASR systems were focused on the Catalan and Spanish language models. The language model for the English ASR was used as is, when provided by the NESPOLE! partners. As a result, the English ASR was not as domain sensitive and, consequently, more error prone. The only work done was to enlarge its lexicon.

OK	1.19%	7.14%
OK-	0%	5.96%
Unacceptable	1.19%	3.58%

Table 3: Evaluation of End-to-End Translation (with ASR) for the English-Catalan Pair. Based on 117 UTTs.

SCORES	FORM	CONTENT
Perfect	64.96%	34.19%
OK+	15.39%	11.97%
OK	8.54%	14.52%
OK-	5.12%	12.82%
Unacceptable	5.99%	26.50%

Table 4: Evaluation of End-to-End Translation (with ASR) for the English-Spanish Pair. Based on 125 UTTs.

SCORES	FORM	CONTENT
Perfect	64.80%	17.60%
OK+	4.80%	10.40%
OK	12.00%	18.40%
OK-	8.80%	16.00%
Unacceptable	9.60%	37.60%

Table 5: Evaluation of Translation for Audio Transcription of the Catalan-English Pair. Based on 119 UTTs.

SCORES	FORM	CONTENT
Perfect	85.72%	73.10%
OK+	5.89%	13.45%
OK	2.52%	4.20%
OK-	4.20%	6.73%
Unacceptable	1.69%	2.52%

Table 6: Evaluation of Translation for Audio Transcription of the Spanish-English Pair. Based on 84 UTTs.

SCORES	FORM	CONTENT
Perfect	96.42%	91.66%
OK+	2.38%	3.57%
OK	0%	0%
OK-	0%	3.57%
Unacceptable	1.20%	1.20%

Table 7: Evaluation of Translation for Audio Transcription of the English-Catalan Pair. Based on 117 UTTs.

SCORES	FORM	CONTENT
Perfect	89.75%	88.89%
OK+	8.55%	1.70%
OK	1.70%	0.85%
OK-	0%	4.28%
Unacceptable	0%	4.28%

Table 8: Evaluation of Translation for Audio Transcription of the English-Spanish Pair. Based on 125 UTTs.

SCORES	FORM	CONTENT
Perfect	95.2%	82.4%
OK+	4%	7.2%
OK	0.8%	3.2%
OK-	0%	5.6%
Unacceptable	0%	1.6%

4. Interchange Format

In this section, we discuss the use of the Interchange Format (IF) for Machine Translation, and then we examine some problems of applying IF to new languages such as Catalan and Spanish.

4.1 Introduction to Interchange Format and Discussion

IF is based on Searle's Theory of Speech Acts (Searle 1969). It tries to represent the speaker's intention rather than the meaning of the sentence per se. In the hotel reservation domain, there are several speech acts, such as giving information about

a price, asking for information about a room type, verifying a reservation, and so on. Since domain concepts such as prices, room type and reservation are included in the representation of the act, in our interlingua, such speech acts are referred to as Domain Actions (DAs), and they are the type of actions that are discussed. These DAs are formed by different combinatory elements expressing the semantic information that needs to be communicated.

Generally speaking, an IF representation has the following elements:

Speaker's Tag + DA + Arguments

The Speaker's Tag may be *a* for the agent's contributions, or *c* for the client's.

Inside the DA we find the following elements:

- Speech Act: a compulsory element that can appear alone or followed by other elements. Examples of Speech-Acts include: give-information, negate, request-information, etc.
- Attitude: an optional element that represents the attitude of the speaker when explicitly present. Some examples are: +disposition, +obligation, and so on.
- Main Predication: a compulsory element that represents what is talked about. Examples of these elements are: +contain, +reservation, and so on; and,
- Predication Participant: optional elements that represent the objects talked about, for instance, +room, +accommodation, and so on.

The DA is followed by a list of arguments. These elements are expressed by argument-value pairs positioned inside a list and separated by a “,”.

By way of example, an IF representation of the sentence in (1) contains all the elements mentioned above:

(1) *Would you like me to reserve a room for you?*

IF: a: request-information+disposition+ reservation+room (for-whom=you, who=i, disposition=(who=you, desire), room spec=(quantity=1, room))

From this representation we know that the speaker is the agent and that he is asking for some information. The attitude expressed here is a desire of a second person singular, that is, the client. The main predication is to make a reservation and the predication participant is one room.

Interchange Format is heavily influenced by English, and this may cause problems when using it to represent Romance languages such as Catalan or Spanish. Most of these problems are solvable, however, and in general, IF works rather well to represent both languages. The following subsections describe six different issues that have been encountered when adapting the IF to Catalan and Spanish.

4.2 Dialogue Context Ambiguity

The meaning of an expression sometimes changes depending on the dialog context. That is to say, a unique expression can have different meanings depending on its place in the conversation. This is the case, for instance, of the Catalan expression *digui'm*. In Catalan, it has a different meaning depending on whether it is used when answering a telephone call or when responding to a suggestion. This difference in meaning is seen here in examples (2) and (3):

(2) 9-ENG-CLIENT: *Shall I give you my Visa number then?*

10-CAT-AGENT: *Digui'm*.

Go ahead.

10-IF:*a: request-action+proceed (who=you, communication-mode=phone)*

(3) CAT-AGENT: *Viatges Fame, digui'm?*

Fame Travel. **Hello?**

IF: a: introduce-self (who=name-viajes_fame)

a: dialog-greet (who=you, to-whom=I, communication-mode=phone)

In example (2), the agent uses the expression to indicate to the client that he is ready and that the client may proceed to give his visa number. However, in (3), the expression appears at the beginning of a conversation, as a kind of greeting indicating to the client that the agent is already listening to him.

There is currently no way to represent dialog structure information within the interlingual formalism, and so only one of the translations (go ahead) is used as a default; but a solution would not be difficult to implement. The first step would be to represent various types of conversational contexts (opening, response-to-offer, etc.), and then to modify the analysis grammars to parse differently according to context. In this case, the analyser recognises that it is parsing (and thus interpreting) a dialogue-opening segment (indicating one meaning, i.e., hello) or a post offer-information segment (indicating a different meaning, i.e., go ahead).

4.3 Formality Feature

In Catalan and Spanish, there is a distinction between formal and informal personal pronouns, especially for second person singular and plural. However, as the IF is influenced by English, this distinction is not reflected in this interlingua. In example (4), the verbal form *ajudar-lo* (to help you) implies a formal relationship between the speaker and singular addressee, while in (5) *ajudar-te* (to help you), the implied relationship is familiar.

(4) CAT-AGENT: *¿En què puc ajudar-lo?* IF: a: offer+help (help=(who=i, to-whom= you))

(5) CAT-AGENT: *¿En què puc ajudar-te?*

IF: a: offer+help (help=(who=i, to-whom= you))

But if we inspect the IF representations for both examples, we see that they are the same. This is due to the lack of a formality feature in this interlingua. This does not imply any problem when translating from Catalan/Spanish into English, as the latter does not have any formal register; but it could cause a loss of meaning when translating from Catalan into Spanish or vice versa, for instance, or from either of these two languages into French, for example, which also makes a second person register distinction.

To solve this problem of representing register, we can add a new argument-value pair to the IF with the argument [formal=] and the values (yes) or (no). When implementing this new feature, the IF representation for examples (4) and (5) would be (6) and (7), respectively.

(6) IF: a: offer+help (help=(who=i, (to-whom= you, formal=yes)))

(7) IF: a: offer+help (help=(who=i, (to-whom= you, formal=no)))

Through the use of these new argument-value pairs, we would be able to communicate the feature of formality and have it available in the target language, if applicable.

4.4 Conceptual Gaps in Catalan and Spanish

Another problem we had to overcome when developing the Catalan and Spanish grammars had to do with the lexicons. Since IF was developed with English as point of reference, there are IF values that refer to lexical items that do not exist in Catalan or Spanish per se. In essence, the semantic field is not divided equivalently between the languages. Sometimes, it is a word or an expression, such as Christmas crackers, that does not exist either in the Catalan or Spanish culture. When facing this problem we maintain the same English word, as there is no cultural equivalent.

Sometimes the solution is not that straightforward, however, given that the word without equivalent in Catalan or Spanish is an important word in the dialogue. This is the case of king-size bed and queen-size bed, as shown in example (8). Both words are rather important within the hotel reservation domain we work in, and what's more, the client is supposed to be an English speaker, so he would most definitely use it. As a consequence, we could not adopt the solution proposed in the previous example, and we had to introduce phrasal equivalents based on already existing Catalan/Spanish words referring to bed types to cover those two values. The Catalan and Spanish equivalents would be un llit *extragran* and un llit *gran*, for Catalan, and una cama *extragrande* and una cama *grande*, for Spanish.

- (8) ENG-CLIENT: I would like a room with a **king-size** bed.
IF: c: give-information+disposition+bed (disposition=(who=I, desire),
room-spec=(quantity=1, room, contain=(quantity=1, **king-bed**)))

4.5 Proper Nouns

Currently, all proper names are included in the IF by the use of values. That is to say, each proper name is represented by a different value. In our domain, proper names are mainly person names, street names, city names, hotel names, names of monuments, museums and other attractions, and so on. For instance, the proper name Hotel Duc de la Victoria is represented in the IF under the class **barcelona-hotel-names** by the value [name-hotel_duc_de_la_victoria]. Although this is a good way to represent proper names when they are well known to the interlocutors, we should also point out that it implies a great effort on the developer's part. Whenever a new proper name is added, it should be first included in the IF Specification files, and then both analysis and generation grammars of all languages have to be updated to include this new proper name. As a consequence, all developers have to be aware of the new values included in the IF specifications, especially those working on the analysis side. Otherwise, this proper name will not be analysed.

Moreover, when developing our analysis grammars, we had to deal with the phenomena of bilingualism in Barcelona, and in Catalonia in general. In Barcelona, both Catalan and Spanish are spoken, and when using a proper name there is a certain degree of code mixing. As a result the name may be in Catalan, in Spanish, or in a mixture of both languages, as shown in (9). When including proper names in Catalan analysis grammars, we took into account all those forms: the Catalan name, the Spanish name and the hybrid version were added under the IF value representing the proper name, as shown in (10):

- (9) CAT: *Carrer Pelai* (Pelai Street)

SPA: *Calle Pelayo*

SPA/CAT: *Calle Pelai*

(10) [name-carrer_pelai]

(carrer pelai)

(calle pelayo)

(calle pelai)

In any case, while this way of treating proper names is adequate, a good way to avoid the significant effort it entails would be to generate proper name representations automatically, or scrap proper name representations altogether and pass proper name forms directly to the target language as strings. Either way, the key is to deal with proper name translation independently from translation of other expressions.

4.6 Catalan 'de' Partitive

In Catalan there is a phenomenon called *de* partitive. This construction is used when a qualifying adjective has an elided head, or when it is used in construction with the impersonal pronoun *en*, as shown in example (11a). In the sentence (11b), there is a noun phrase *llit extragran* (king-size bed) formed by a head noun (*llit* - bed) and a qualifying adjective (*extragran* - king-size). In Catalan, this noun phrase can be transformed by eliding the head noun *llit*, inserting the pronoun *en* in its place and introducing the adjective *extragran* by the preposition *de* (in this case *d'*).

(11a) CAT: *En tenim un d'extragran*

ENG: *We have one in king-size.* b) CAT: *Tenim un llit extragran*

ENG: *We have a king-size bed*

At first sight, if we want to have an interlingua representation for (11a), it would be example (12). In this representation, we have an [object-spec=] argument that contains a subargument [size=] with the value (king-bed-size). The focus of the representation is on the size of the object without explicitly mentioning the type *per se*.

(12) give-information+existence+object (provider=we, object-spec=
(quantity=1, size=king-bed-size))

However, (11a) actually continues to refer to a king-size bed. Furthermore, ideally, the interlingua should be language independent. It would not be fair to create a new value such as *king-bed-size* only for Catalan especially, since we could represent this sentence through already existing values and arguments, as in example (13).

-
- (13) give-information+existence+object (provider=we, object-spec=
(quantity=1, bed-spec=king-bed))

In this representation, the segment *un d'extragran* is represented by the subargument [bed-spec=], used to represent the types of beds.

4.7 Excess of Conceptual Granularity

The Interchange Format is a formalism intended to express the meanings of different parts of a sentence. In some cases, however, the representation of this meaning is too specific with respect to one or another of the languages in the system. This is especially common in regard to representing modifiers such as adjectives. For example, (14) shows two different IF values that, with respect to Catalan and Spanish, can both be taken to mean the same thing: 'old.' The corresponding terms for both values is *vell* in Catalan and *viejo* in Spanish. The difference between the English lexical counterparts has less to do with semantics as such, but rather with their distributional properties.

- (14) [ancient]
[antique]

In this case then, the IF values are too specific, since the meaning they convey could be included under one value. The solution is to introduce an IF class value for values [ancient] or [antique], and then to map the Catalan or Spanish lexeme to or from this class value. In translating from English to Catalan or Spanish, one simply moves from the particular value to the class value, since there are no possible equivalents associated with the particular values. When translating from Catalan or Spanish into English, the English appropriate lexeme is selected on the basis of class of head element modified.

5. Conclusions

This presentation began with a brief description of the FAME Speech-to-Speech Machine Translation System, and the results of a user-oriented evaluation of the system for both voice (with ASR) and clean audio-transcription inputs. It was observed that the interlingual translation component performs very well when used on clean input and that, as expected, worsens in performance when used on spoken input. Nonetheless, we are satisfied with system performance, although we acknowledge that further work should be done, especially in order to improve the ASR throughput.

Next, Interchange Format (IF) was introduced, and we examined a number of language-particular problems that arose while applying IF to the representation of

Catalan. In each case, we described the solutions we used, or propose to use, to overcome these problems, including improvements to IF that should widen its coverage and make it easier to be used by developers.

In the future, we hope to continue to develop the system along three general lines:

- We would like to implement the changes and improvements proposed for IF, and see how they work and in which way they help to widen the coverage of the interlingua;
- We would like to improve the ASR component of our translation system, and try to find solutions to overcome possible problems due to spontaneous speech and disfluencies; and,
- We also expect to extend the coverage of our grammars and lexica, not only to other areas of the travel domain, but also to other domains such as medicine.

6. Acknowledgments

This research has been partially financed by the FAME (IST-2001-28323) and ALIADO (TIC2002-04447-C02) projects. We would especially like to thank Climent Nadeu and Jaume Padrell, for all their help and support in numerous aspects of the project. We are also grateful to other UPC colleagues, such as Josè B. Mariño and Adrià de Gispert, and to our colleagues at CMU, Dorcas Alexander, Donna Gates, Lori Levin, Kay Peterson and Alex Waibel, for all their feedback and assistance.

References

“C-STAR.” Online at <http://www.c-star.org>.

“FAME.” Online at <http://isl.ira.uka.de/fame>.

Gavaldà, M. (2000). “SOUP: A Parser for Real-world Spontaneous Speech.” *Proceedings of the 6th International Workshop on Parsing Technologies (IWPT-2000)*, Trento, Italy.

Holzapfel, H. *et al.* (2003). *FAME Deliverable D3.1: Testbed Software, Middleware and Communication Architecture*.

Lavie, A. *et al.* (2002). “A Multi-Perspective Evaluation of the NESPOLE! Speech-to-Speech Translation System.” *Proceedings of ACL-2002 Workshop on Speech-to-Speech Translation: Algorithms and Systems*. Philadelphia, PA, 121-128.

Levin, L. *et al.* (2002). “Balancing Expressiveness and Simplicity in an Interlingua for Task based Dialogue.” *Proceedings of ACL-2002 Workshop on Speech-to-Speech Translation: Algorithms and Systems*. Philadelphia, PA, 53-60.

Metze, F. *et al.* (2002). “The NESPOLE! Speech-to-Speech Translation System.” *Proceedings of HLT-2002*, San Diego, California.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, UK: Cambridge University Press.

Taddei, L. *et al.* (2003). *NESPOLE! Deliverable D17: Second Showcase Documentation*. <http://nespole.itc.it>.

“The Open Agent Architecture™.” Online at <http://www.ai.sri.com/~oaa>.

Tomita, M. & Nyberg, E.H. (1988). “Generation Kit and Transformation Kit, Version 3.2,

User's Manual." Technical Report CMU-CMT-88-MEMO, Center for Machine Translation, Carnegie Mellon University, Pittsburgh, PA.

Woszczyna, M. *et al.* (1993). "Recent Advances in JANUS: A Speech Translation System." *Proceedings of Eurospeech-1993*, Berlin.

Computing Non-Concatenative Morphology: The Case of Georgian¹

Olga Gurevich

Georgian (Kartvelian) is a less commonly studied language, with a complex, non-concatenative verbal morphology. This paper examines characteristics of Georgian that make it a challenge for language learners and for current approaches to computational morphology. We present a computational model for generation and recognition of Georgian verb conjugations, and describe one practical application of the model to help language learners.

1. Introduction

Georgian (Kartvelian) is the official language of the Republic of Georgia and claims about 4 million native speakers. Georgian morphology is largely synthetic, with complex verb forms that can often express the meaning of a whole sentence. Georgian has sometimes been called agglutinative (Hewitt 1995), but such classification does not fully describe the complexity of the language.

Descriptions of Georgian verbal morphology emphasise the large number of inflectional categories, the large number of elements that a verb form can contain, the dependencies between the occurrence of various elements, and the large number of regular, semi-regular, and irregular patterns of formation of verb inflections. All of these factors make computational modeling of Georgian morphology a rather daunting task. To date, no successful large-scale models of parsing or generation of Georgian are available.

In this paper, I propose a computational model for parsing and generation of a subset of Georgian verbal morphology that relies on a templatic, word-based analysis of the verbal system, rather than assuming compositional rules for combining individual morphemes. I argue that such a model is viable, extensible, and capable of capturing the generalisations inherent in Georgian verbal morphology at various levels of regularity.

¹ This research was in part supported by the Berkeley Language Center. Thanks to Mark Kaiser, Claire Kramsch, Lisa Little, Nikolas Euba, David Malinowski, and Sarah Roberts for many hours of productive discussion and wonderful suggestions, and to Aaron Siegel for technical support. I am eternally grateful to Vakhtang Chikovani and Shorena Kurtsikidze for help in creating the website, and for introducing me to Georgian. I alone am to blame for any errors and omissions.

I begin with a brief overview of Georgian verbal morphology, emphasising the factors that complicate its computational modelling. I present an analysis grounded in word-based approaches to morphology and Construction Grammar, and suggest that this type of analysis lends itself more easily to computational implementations than analyses that assume morpheme-based compositionality. Following a brief overview of existing approaches to computational morphology, I propose a model for Georgian and describe it in detail. The model is currently implemented as a cascade of finite-state transducers (Beesley & Karttunen 2003), but probabilistic and connectionist extensions or alternative implementations are plausible. Finally, I describe a practical application of this model for language learning: an online database of Georgian verb conjugations.

2. An Overview of Georgian Verbal Morphology

The morphosyntax of Georgian verbs is characterised by a variety of lexical (irregular), semi-regular, and completely regular patterns. The verb forms themselves are made up of several kinds of morphological elements that recur in different formations. These elements can be formally identified in a fairly straightforward fashion; however, their function and distribution defy a simple compositional analysis, but instead are determined by the larger morphosyntactic and semantic contexts in which the verbs appear (usually tense, aspect, and mood) and by the lexical properties of the verbs themselves. The combination of morphosyntactic and lexical factors also determines the case marking on the verb's arguments.

The specific types of morphological elements and peculiarities in their function and distribution are described below. The main point of this section is that a language learner and a computational model are faced with patterns in which formal elements (morphs) do not have identifiable, context-independent meanings that can be combined compositionally to form whole words. Rather, they must contend with a variety of patterns at various degrees of regularity. In computational terms, this amounts to a series of rules of varying specificity, backed up by defaults.

The linguistic analysis at the core of the computational model splits Georgian verbs into several lexical classes. The lexical classes are described on the basis of *example paradigms*, using frequent verbs belonging to each class. This is in contrast to a more rule-oriented description in which lexical classes may be identified by some morphological or syntactic feature. In the rest of this section, I argue that an example-

based description is the only one plausible for learners of Georgian, and provides a good basis for computational modeling as well.

2.1 Series and Screeves

Georgian verbs inflect in tense / mood / aspect (TAM) paradigms called *screeves* (from *mck'rivi* 'row'). There are a total of eleven screeves in Modern Georgian, although only ten are actively used. Screeves can be grouped into three *series* based on morphological and syntactic commonalities, as in Table 1:

Table 1 - Series and Screeves

Series I		Series II	Series III
Present sub-series	Future sub-series	(aorist)	(perfect)
Present	Future	Aorist	Perfect
Imperfect	Conditional		Pluperfect
Present subjunctive	Future subjunctive	Aorist subjunctive	(Perf. subj.) [*]

Knowing the series and screeve of a verb form is essential for being able to conjugate it. Screeve formation exhibits a number of lexical, semi-regular, and regular patterns, some of which are examined below.

Georgian verbs are often divided into four conjugation classes, based mostly on valency (cf. Harris 1981). For now, I will concentrate on transitive verbs; it will be necessary to mention the other classes (unergative, unaccusative, and indirect) in the discussion of case-marking below. The structure of a verb form can be described using the following (simplified) template:

(Preverb₁)-(Pron1₂)-(PRV₃)-root₄-(TS₅)-(Scr₆)-(Pron2₇)²

The approximate function of each element is as follows:

- Preverb - marks aspectual distinctions, lexically associated with each verb (similar to verbal prefixes in Slavic or German).
- Pron1 - Prefixal pronominal agreement slot.
- PRV - pre-radical vowel slot, serves a variety of functions in different contexts.
- Root - the only required part of the verb form.
- TS - Thematic Suffix. Participates in the formation of several tenses, predicts certain inflectional properties of the verb.

* The Perfect Subjunctive is almost never used in contemporary Georgian.

2 Cf. Hewitt 1995.

- Scr - Screeve marker. This is a screeve (tense) ending which may depend on verb class and agreement properties.
- Pron2 - suffixal agreement slot.

The preverb, root, and thematic suffix must be lexically specified in all cases, although their distribution follows a somewhat regular pattern described in the next section. Other elements in the template are distributed according to more or less regular principles, although some lexical exceptions do exist.

The templatic composition of the Georgian verb forms suggests, at first blush, an agglutinative structure. However, a closer examination of the morphological elements in the verbal template and their function provides evidence against such an analysis. In particular, the morphological elements do not have identifiable meanings independent of context, and their meanings do not compositionally comprise the meanings of the words in which they participate. As argued in Gurevich (2003), the morphological elements of Georgian cannot be thought of as *morphemes*, or smallest meaningful elements of form. Rather, word-level constructions determine both the meaning of the whole word, and the collection of morphological elements that comprise the word. This combination of templatic morphological structure and non-compositional meaning construction makes Georgian inflectional morphology look *non-concatenative*.

As an illustration, let us examine the formation of the verb *xat'va* 'paint' in Table 2. The screeves (and, more generally, series) govern the distribution of the morphological elements.

Table 2: Screeves of *xat'va* 'paint'

Series		Screeve	2SgSubj, 3Obj form
I	Pres. subseries	Present	<i>xat'-av</i> 'You paint'
		Imperfect	<i>xat'-av-di</i> 'You were painting'
		Pres. Subj.	<i>xat'-av-de</i> 'You should paint'
	Fut. subseries	Future	<i>da-xat'-av</i> 'You will paint'
		Conditional	<i>da-xat'-av-di</i> 'You would paint'
		Fut. Subj.	<i>da-xat'-av-de</i> 'If you could paint'
II	Aorist	<i>da-xat'-e</i> 'You painted'	
	Aor. Subj.	<i>da-xat'-o</i> 'You have to paint'	
III	Perfect	<i>da-g-i-xat'-avs</i> 'You have painted'	
	Pluperfect	<i>da-g-e-xat'-a</i> 'You should have painted'	

In addition to the multitude of morphological elements in any given verb form, the distribution and lexical dependency of the elements makes a learner's task difficult. Preverbs, thematic suffixes and screeve endings present particular difficulties.

The preverbs form a closed class of about eight. A preverb (*da-* for the verb ‘paint’) appears on forms from the Future subgroup of series I, and on all forms of series II and III in transitive verbs. The preverbs are by origin spatial prefixes that now mark perfective aspect. However, the presence of a preverb on a verb form signals more than just a change in aspect. For example, the preverb differentiates the Conditional from the Imperfect, and the meaning of the two screeves differs in more than aspect. An additional difficulty is in the lexical connection between prefixes and verb roots, similar to the verbal prefixes in Slavic or German. Table 3 demonstrates some of the lexically-dependent morphological elements, including several different preverbs (row ‘Future’).

Similarly, thematic suffixes (otherwise known as screeve suffixes or screeve formants) form a closed class and are lexically associated with verb roots. In general, thematic suffixes do not appear to have independent meaning. Rather, they serve to mark the inflectional class of the verb, because they determine certain patterns of inflectional behavior in different screeves.

On transitive verbs, thematic suffixes appear in all series I forms. Their behavior in other series differs by individual suffix: in series II, most suffixes disappear, though some seem to leave partial ‘traces’. In series III, all suffixes except *-av/-am* disappear in the Perfect screeve; and in Pluperfect, all suffixes disappear, but the inflectional ending that takes their place does depend on the original suffix (rows ‘Present’ and ‘Perfect’ in Table 3).

The next source of semi-regular patterns comes from the inflectional endings in the individual screeves and the corresponding changes in some verb roots (row ‘Aorist’ in Table 3).

Finally, another verb form relevant for learners is the *masdar*, or verbal noun, which is the closest substitute of the infinitive in Georgian. The *masdar* may or may not include the preverb and/or some variation of the thematic suffix (last row in Table 3). The formation of the *masdar* is particularly important, as it is the reference form listed in most Georgian dictionaries, even though it might not even start with the same letter as an inflected verb form.

Table 3: Lexical Variation

	‘Bring’	‘Paint’	‘Eat’
Present	<i>igh-eb-s</i>	<i>xat'-av-s</i>	<i>ch'-am-s</i>
Future	<i>c'amo-ighebs</i>	<i>da-xat'avs</i>	<i>she-ch'ams</i>
Aorist, 3Sg Subject	<i>c'amoigh-o</i>	<i>daxat'-a</i>	<i>shech'am-a</i>
Perfect	<i>c'amough-ia</i>	<i>dauxat'-avs</i>	<i>sheuch'am-ia</i>
Masdar (verbal noun)	<i>c'amo-gh-eba</i>	<i>da-xat'-va</i>	<i>ch'-am-a</i>

In many cases, the inflectional endings and root changes can be determined if we know the thematic suffix of the verb (cf. the painstakingly detailed description of such patterns in Hewitt 1995). However, there are exceptions to most such connections, and learning the patterns based on explicit rules seems virtually impossible.

On the other hand, screeve formation in some instances presents amazing regularity. Thus, the Imperfect and First Subjunctive screeves are regularly formed from the Present. Similarly, the Conditional and Future Subjunctive are formed from the Future. And for most (though not all) transitive verbs, the Future is formed from the Present via the addition of a preverb.

Additionally, the number of possible combinations of inflectional endings, root changes and other irregularities is also finite, and some choices tend to predict other choices in the paradigm of a given verb (e.g. the selection of thematic suffix or Aorist 2Sg Subj ending often predicts the Aorist Subjunctive ending). Although the rule-based analysis is unproductive, Georgian verbs *can* be classified according to several example paradigms, or inflectional (lexical) classes. This is similar to the inflectional class distinctions made in Standard European languages; the major difference is that the number of classes is much greater in Georgian than in other languages. One such classification is presented in Melikishvili (2001), distinguishing seventeen inflectional classes for transitive verbs alone, and over sixty classes overall. While the exact number of inflectional classes is still in question (see the discussion in section 4.4), the general example-based approach seems the only one viable for Georgian.

The next section deals with subject and object agreement, a completely regular yet non-concatenative phenomenon.

2.2 Subject and Object Agreement

A Georgian verb can mark agreement with both its subject and its object via a combination of prefixal and suffixal agreement markers, as in Table 4:

Table 4: Agreement in Present

Subj	OBJECT				
	1SG	1PL	2SG	2PL	3
1SG	--	--	g -xat'av	g -xat'av-t	v -xat'av
1PL	--	--	g -xat'av-t	g -xat'av-t	v -xat'av-t
2SG	m -xat'av	gv -xat'av	--	--	xat'av
2PL	m -xat'av-t	gv -xat'av-t	--	--	xat'av-t
3SG	m -xat'av-s	gv -xat'av-s	g -xat'av-s	g -xat'av-t	xat'av-s

3PL	m-xat'av-en	gv-xat'av-en	g-xat'av-en	g-xat'av-en	xat'av-en
-----	-------------	--------------	-------------	-------------	-----------

The distribution and order of attachment of agreement affixes has been the subject of much discussion in theoretical morphological literature (Anderson 1992; Halle & Marantz 1994; and Stump 2001). To simplify matters for the computational model, I assume here that the prefixal and suffixal markers attach to the verb stem at the same time, and indicate the combined subject and object properties of a paradigm cell.

While the prefixal markers and the suffix *-t* appear in all screeves, the suffixes in 3Sg and 3Pl Subject forms are screeve-dependent (cf. row 'Aorist' in Table 3). These suffixes therefore belong to the semi-regular patterns, while the rest of the agreement system is completely regular.

Another difficulty arises in series III for transitive verbs. Here, the subject and object agreement appears to be the inverse of that in series I and II (Table 5; notice the different designation of rows and columns). This phenomenon, called *inversion*, corresponds to a reverse case marking of the nominal arguments (see next section). Several analyses have been proposed suggesting that, in inversion, the semantic subject corresponds to a 'surface' indirect object, and the semantic object corresponds to a 'surface' subject (Harris 1981). However, a simple difference in linking does not fully explain the paradigm composition. In inverted paradigms, plural number agreement is still sensitive to the semantic arguments (namely, the semantic subject / agent triggers plural agreement regardless of other agreement or case-marking facts).

Table 5: Agreement in Perfect

Object	Subject				
	1SG	1PL	2SG	2PL	3
1SG	--	--	g-xat'av	g-xat'av-t	v-xat'av
1PL	--	--	g-xat'av-t	g-xat'av-t	v-xat'av-t
2SG	m-xat'av	gv-xat'av	--	--	xat'av
2PL	m-xat'av-t	gv-xat'av-t	--	--	xat'av-t
3SG	m-xat'av-s	gv-xat'av-s	g-xat'av-s	g-xat'av-t	xat'av-s
3PL	m-xat'av-en	gv-xat'av-en	g-xat'av-en	g-xat'av-en	xat'av-en

2.3 Subject and Object Case Marking

Case marking of nominal arguments in Georgian is not constant, but depends on the conjugation (valency) class of the verb and the series / screeve of the verb forms. Transitive verbs can follow one of three patterns, depending on series:

- (1) k'ac-i dzayl-s xat'av-s
 man-NOM dog-DAT paint.Pres.3SgSubj

“The man paints / is painting the dog.” (Series I, Present - **Pattern A**)

- (2) k’ac-**ma** dzayl-i daxat’a
 man-**ERG** dog-**NOM** paint.Aor.3SgSubj

“The man painted the dog.” (Series II, Aorist - **Pattern B**)

- (3) k’ac-**s** dzayl-i t’urme dauxat’avs
 man-**DAT** dog-**NOM** apparently paint.Perf.3SgSubj

“The man has painted the dog.” (Series III, Perfect - **Pattern C**)

Table 6 demonstrates the case-marking patterns by series for all four conjugation classes. Only transitive and unergative (active intransitive) verbs show variability by series. Unaccusative verbs always follow Pattern A (similar to the standard nominative/accusative pattern in European languages), and indirect verbs always follow Pattern C (the inverse pattern). In order to assign correct case marking, a learner of Georgian must recognise the conjugation class of each verb, as well as the series / screeve for some of the verb classes.

Table 6 - Case-Marking Patterns

Series	Transitive	Unaccusative	Unergative	Indirect
I	A	A	A	C
II	B	A	B	C
III	C	A	C	C

2.4 Summary

The formation of the screeves exhibits several irregular, semi-regular, and regular patterns. The morphological elements in the Georgian verb template are easy to identify, suggesting an agglutinative structure. However, closer inspection reveals that the morphological elements may not have easily identifiable meanings or functions (cf. preverbs, thematic suffixes, and screeve endings). Moreover, even if we manage to find meanings for these elements, the meanings will not predict the distribution of such elements across different verbs, verb types, and screeves. Such non-compositionality in meaning makes Georgian more similar to morphologically non-concatenative languages such as Arabic and Hebrew.

On the basis of the data above, it is argued in (Gurevich 2003) and (Blevins forthcoming 2006) that a word-based morphological theory is more appropriate for Georgian. In such a theory, word formation is determined by whole-word patterns, such that the whole word carries morphosyntactic properties, and they need not be assigned to individual morphemes. Gurevich (forthcoming 2006) suggests that such

patterns may be represented as *constructions*, or form-meaning pairings in which the elements of form need not match the elements of meaning one-to-one. The analysis is based on insights of Construction Grammar (Fillmore 1988; Goldberg 1995). It is argued that the main organising unit and the best level for morphosyntactic constructions in Georgian is the series. The series provides a base for expressing the more or less regular patterns of Georgian morphosyntax. The less regular and more lexicalised information, on the other hand, is best expressed using inflectional (lexical) classes of verbs.

3. Approaches to Computational Morphology

3.1 *Standard Assumptions and Difficulties Presented by Georgian*

Many contemporary approaches to computational morphology are based on, or can be easily translated into, finite-state networks (FSN). In such approaches, an arc in the FSN often corresponds to a phoneme or morpheme, and the recognition or generation of each arc advances the state in the network. Many approaches, including Beesley & Karttunen (2003), are implemented as two-way finite-state transducers (FST) in which each arc corresponds to a mapping of two elements, for example, a phoneme and its phonetic realisation, or a morpheme and its meaning. As a result, FST morphology very often assumes morpheme-level compositionality, the idea that the meaning of a word is compositionally made up from the meanings of its constituent morphemes. FST morphology has, for the most part, been applied to concatenative morphological systems like Finnish, although there have been some recent applications to templatic morphology such as Arabic (Beesley & Karttunen 2003).

As demonstrated in the previous section, assumptions of morphemic compositionality do not serve well to describe the verbal morphology of Georgian. The Georgian verb forms are made up of identifiable morphological elements (i.e., elements of form), but the meaning of these elements is not easily identifiable, and does not stay constant in different morphosyntactic contexts.

A computational system appropriate for Georgian should be able to accommodate the templatic nature of Georgian verb forms and its patterns of regularity and sub-regularity. Overall it should be able to describe the following:

- Meaning carried by a whole word form rather than by individual morphemes;
- Lexical root alternations and suppletion;

-
- Lexical class-dependent screeve formation (e.g. the endings in the Aorist);
 - The dependency between the formation of some screeves from that of others (e.g. the Imperfect from the Present); and
 - The multiple exponence of agreement, that is, the use of suffixes and prefixes simultaneously, and the simultaneous expression of subject and object agreement.

The linguistic analysis of Georgian verbal morphology suggested in the previous section relies on insights from Construction Grammar. Unfortunately, there are currently no computational implementations of CG capable of handling complex morphological systems. Bryant (2003) describes a constructional syntactic parser, based on general principles of chart parsing. However, this parser cannot yet handle morphological segmentation, and adapting it for Georgian would require substantial revision.

Fortunately, FST tools for computational morphology have advanced to the point where they can handle some aspects of non-concatenative morphology. The next section briefly describes the approach in Beesley & Karttunen (2003) and what makes it a possible candidate for modelling at least a subset of Georgian verbal morphology.

3.2 Xerox Finite-State Morphology Tools

Beesley & Karttunen (2003) present the state-of-the-art set of tools for creating finite-state morphological models. The book is accompanied by implementations of the two Xerox languages: *xfst* (designed for general finite-state manipulations) and *lexc* (designed more specifically for defining lexicons). Since our goal was to reproduce morphotactic rules of word formation rather than the structure of the lexicon, *xfst* was used.

Xfst provides all of the basic commands for building up single or two-level finite-state networks (i.e., transducers), such as concatenation, intersection, and so forth. In addition, *xfst* has several built-in shortcuts that make network manipulation easier, such as various substitution commands. *Xfst* distinguishes between words of a natural language (composed of single characters) and multi-character symbols, used in our model to indicate morphosyntactic properties such as person or number. Each completed arc in a finite-state network compiled using *xfst* represents a mapping between a set of morphosyntactic and semantic properties (on the upper side) and a full word form that realises those properties (on the lower side).

Another very useful feature of *xfst* is the ability to create scripts with several commands in a sequence. The later commands can operate on the output of earlier commands, and can thus create a *cascade* of finite-state transducers. *Xfst* also provides convenient ways of outputting all the words recognised by a given transducer, which proved very useful in the creation of the online reference (see section 5). An updated version of *xfst* (Beesley & Karttunen forthcoming 2006) also includes support for utf-8.

While finite-state technology is very good at generating and recognising regular expressions, it has a harder time capturing other features of natural language such as non-concatenative morphological structure. The next section describes some adaptations that allow FST to handle many of the non-concatenative patterns in Georgian.

In addition, FST is not designed to represent a dynamic, living mental lexicon of an actual speaker. It does not provide any mechanisms for probabilistic decisions, or for recognition and generation of novel inflectional forms. The concluding section discusses some possible future developments in this area.

4. Computational Model of the Georgian Verb

4.1 General Idea

As argued above, Georgian verb morphology can be described as a series of patterns at various levels of regularity. Most of the patterns specify particular morphosyntactic or semantic properties of verb forms and the corresponding combinations of elements in the morphological templates. In the model proposed here, screeve formation is viewed as lexical or semi-regular, and pronominal agreement is viewed as completely regular.

Screeve formation for different conjugation classes (transitive, unergative, unaccusative, and inverse) is fairly different in Georgian, and so each conjugation class is implemented as a separate network. Nevertheless, the principles for composing each network are the same.

The model is implemented as a cascade of finite-state transducers, that is, as several levels of FST networks such that the result of composing a lower-level network serves as input to a higher-level network. The levels correspond to the division of templatic patterns into completely lexical (Level 1) and semi-regular (Level 2). Level 3 contains completely regular patterns that apply to the results of both Level 1 and Level 2. The result of compiling Level 3 patterns is the full set of conjugations for the

verbs whose lexical information is included in Level 1. The FST model can be used both for the generation of verbal inflections and for recognition of complete forms.

In general, the most specific or irregular information is contained at the lower levels. The higher levels, by contrast, contain defaults that apply if there is no more specific information. The verbs explicitly mentioned in the lexical level (Level 1) are representative examples of lexical classes, as posited by the linguistic analysis in section 2. Through the use of diacritics and replacement algorithms, other verbs are matched to their lexical classes and are included in the resulting network.

The main advantage of this implementation is in the separation of lexical, or irregular, verb formation patterns from the semi-regular or completely regular patterns. The initial input to the FST cascade includes only the necessary lexical information about each verb and verb class; the computational model does the rest of the work.

The model described here served as the basis for an online reference on Georgian verb conjugation, described in section 5. This practical application underlies some of the specific choices in implementing the model.

The current implementation of the model focuses on transitive verbs; however, there are obvious ways of extending the model to apply to other verb classes.

4.2 Level 1: The Lexicon

The first level of the FST model contains lexically specific information. There are two separate networks. The first network contains information about the gloss and masdar or the verb stem.

The second network contains several complete word forms for each verb stem, providing all the lexically-specific information needed to infer the rest of the inflections. For the most regular verbs, these are:

- Present screeve, no overt agreement (corresponds to 2Sg Subject, 3Sg Object);
- Future screeve, no overt agreement;
- Aorist screeve, no overt agreement;
- Aorist, 3Sg Subject, no overt object agreement; and
- Aorist Subjunctive.

Some verbs need additional forms in order to describe their paradigms:

- Present screeve, 3Pl Subject (most verbs have the ending *-en*, but some end in *-ian*); and
- Perfect screeve.

The inflected forms are represented as two-level finite-state arcs, with the verb stem and morphosyntactic properties on the upper side, and the inflected word on the lower side, as in Figure 1. The purpose of the stem is to uniquely identify each verb. Verb roots in Georgian are often very short and ambiguous; therefore a combination of the verb root plus thematic suffix was used. In some cases, even this combination is insufficient to identify the verb uniquely; in such cases, the preverb may be necessary as well. It is only important that the verb stem can be uniquely matched in the network containing glosses; thus, the stem has no theoretical significance in this model.

Another challenge is posed by the non-concatenative nature of verb agreement. Recall from section 2 that verb agreement is realised by a pre-stem affix and a final suffix. Since many of the word forms in Level 1 contain preverbs, the agreement affix would need to be infixated into the verb form at a later level. Beesley & Karttunen provide some fairly complex mechanisms for doing infixation in FST; however, the fixed position of the agreement affixes in the Georgian verb template allows for a much simpler solution. The forms on Level 1 contain a place holder “+Agr1” for the prefixal agreement marker (Figure 1), which is replaced by the appropriate marker in the later levels.

The Level 1 network is produced via scripts from a table of verb forms containing only the necessary lexical information. Redundancy in human input is thus minimised.

Figure 1 - Simplified FST Script

```

#          Level          1          --          Lexicon
# +Agr1 is the place holder for agreement

[[[xat' +Pres .x.      +Agr1 xat' ] |
[xat' +Fut  .x.      da +Agr1 xat' ] |
[xat' +Aor  .x.      da +Agr1 xat'e ] |
[xat' +Aor +3SgSubj .x.      da +Agr1 xat'a ] |
[xat' +Aor +3PlSubj .x.      da +Agr1 xat'es ] |
[xat' +Perf .x.      da +Agr1 xat'avs]]

# Level 2 -- Derivative forms
# Conditional
regex [Level1.i .o. [Level1.u & [?* +Fut ?*]]].i [+Cond: di ];
substitute symbol 0 for +Fut

```

4.3 Level 2: Semi-regular Patterns

The purpose of Level 2 is to compile inflectional forms that are dependent on other forms (introduced in Level 1), and to provide default inflections for regular screeve formation patterns.

An example of the first case is the Conditional screeve, formed predictably from the Future screeve. The FST algorithm is as follows:

- Compile a network consisting of Future forms;
- Add the appropriate inflectional suffixes (-di for 1st and 2nd person subject, -da for 3rd person subject);
- Replace the screeve property "+Fut" with "+Cond"; and
- Add the inflectional properties where needed.

The replacement of screeve properties is done using the 'substitute symbol' command in *xfst*; other operations are performed using simple concatenation commands.

An example of the latter is the addition of 3Pl Subject forms of the Present screeve. The default suffix is *-en*, which is added to all verbs unless an exception is specified at Level 1. The basic algorithm is as follows:

- Compile a network of Present forms, excluding the forms for which both 3Pl Subject forms are already specified;

-
- Add the suffix *-en*; and
 - Add the morphosyntactic property “+3PlSubj”.

All of the patterns defined at Level 2 are then compiled into a single network, which serves as input to Level 3.

4.4 Level 3: Regular Patterns

The purpose of Level 3 is to affix regular inflection, namely, subject and object agreement. As described in section 2, agreement in Georgian is expressed via a combination of a prefix and a suffix that are best thought of as attaching simultaneously and working in tandem to express both subject and object agreement. Thus, the compilation of Level 3 consists of several steps, each of which corresponds to a paradigm cell.

In each step, all of the word forms from Level 2 are taken as input. The place holder for the pre-stem agreement affix is then replaced by the appropriate affix (in some cases, this is null), and the appropriate suffix is attached at the end, as in Figure 1. The resulting networks are then compiled into a single network.

The only difficulty at this level arises when dealing with the ‘inverted’ screeves (Perfect and Pluperfect). As demonstrated in section 2, the morphological agreement in these screeves is sensitive to the case-marking of the nominal arguments, which is the reverse of the regular pattern. However, the composition of the agreement paradigm is sensitive to the semantic roles played by the arguments: plural number agreement is still triggered by the semantic agent. In this case, the computational implementation was motivated by the practical application of the model to the online reference. A separate set of paradigm cells was created for the inverted tenses, interpreting the properties ‘Subject’ and ‘Object’ as semantic. The resulting FST network thus shows no relation between inverted and non-inverted forms (i.e., it does not capture the generalisation behind inversion). Such an interpretation was sufficient for the purposes of the conjugation reference. However, the model could easily be amended to incorporate a different analysis of inversion that relies on the distinction between semantic and morphological arguments.

4.5 Treatment of Lexical Classes

The input to Level 1 contains a representative for each lexical class, supplied with a diacritic feature indicating the class number. Other verbs that belong to those classes could, in principle, be inputted along with the class number, and the FST model could

substitute the appropriate roots in the process of compiling the networks. There are, however, several challenges to this straightforward implementation:

- Verbs belonging to the same class may have different preverbs as well as different roots, thus complicating the substitution;
- For many verbs, screeve formation involves stem alternations such as syncope or vowel epenthesis, again complicating straightforward substitution; and
- Suppletion is also quite common in Georgian, requiring completely different stems for different screeves.

As a result, even for a verb whose lexical class is known, several pieces of information must be supplied to infer the complete inflectional paradigm. The FST substitution mechanisms are fairly restricted, and so the compilation of new verbs is currently done using Java scripts performing simple string manipulations. Such an implementation still makes use of the division into lexical classes. The scripts make non-example verbs look like example verbs in Level 1 of the FST network by creating the necessary inflected forms, but the human input to the scripts need only include the information necessary to identify the lexical class of the verb. Future improvements to the computational model may include a more efficient method of identifying lexical classes within FST itself.

The exact number of lexical classes has not been established with full certainty. Melikishvili (2001) relies entirely on morphological characteristics of verb inflection and categorises verb forms into sixty-three different classes; seventeen of those are for transitive verbs. This classification, however, makes some distinctions that can be merged in the computational model; for example, certain types of non-productive stem extensions can be considered part of the lexically specified verb stem.

Another issue is the psychological reality of the lexical classes. A pilot survey of morphological productivity, conducted with adult speakers of Georgian, suggests that speakers conjugating nonce verbs rely more on frequent inflectional patterns than on a rule-based comparison with existing verbs based on morpho-phonological similarities with the nonce verbs (Gurevich forthcoming 2006). Such a reliance on frequency is not reflected in Melikishvili's classification. The computational model proposed here takes a small step in this direction by relying on frequent verbs as example paradigms; however, the model does not have any built-in way to accommodate the relative frequency of different inflectional patterns. The concluding section suggests some possible improvements for the future.

4.6 Case Frames

As described in Section 2, another complicating feature of the Georgian verb is the variability of case-marking patterns for the verb's arguments. For the purposes of the online conjugation reference, it was necessary to present the case-marking information with each verb. Fortunately, the case marking patterns depend almost entirely on the conjugation class and TAM series of the verb³. Since the goal of the online reference is to *describe* the morphosyntactic patterns of Georgian, it was sufficient to simply mention the case-marking pattern for each verb type.

If the purpose of the morphological transducer is to supplement a syntactic parser, the case-marking information could be represented as a feature structure and associated with each verb type.

4.7 Summary

The computational model presented here accommodates many properties of Georgian verbal conjugation that make it challenging: the templatic structure of the verb forms; the non-concatenative nature of word meaning construction; the large number of irregular and semi-regular word formation patterns; and the interaction between word formation and case marking on the verb's arguments. The model crucially relies on classification of verbs into lexical classes with example paradigms for each class. The two-level mappings inherent in FST mean that the model can be used for generation as well as recognition.

5. Practical Application: An Online Reference

5.1 Purpose

The computational model described here serves as the basis for an online reference on Georgian verb conjugation. The goal of the online reference is to aid the learners of Georgian in a number of ways:

- It provides complete conjugation tables for two hundred frequently-used verbs;
- The verb database can be searched using any verb form or its English translation; and,
- For many verb forms, real-life examples from the Internet, as well as audio and video sources, are provided (along with translations).

3 One of the exceptions is the verb *ic*'is 'he/she knows.' Although this verb is transitive, its subject must always be in the Ergative and its object, in the Nominative.

-
- Several types of exercises are available on the website; answers are automatically checked for correctness.
 - The online reference is meant as an addition to the classroom or self-study using a textbook, such as (Kurtsikidze forthcoming 2006).

5.2 Website Design

The website is divided into four sections: 'Verb Conjugation', 'Examples', 'Exercises', and 'Resources'.

The section on verb conjugation is the core of the reference tool. It provides complete tables of verb conjugations, accessible through browsing by individual verb (in Georgian or in English), or by searching. The conjugated forms are produced using the FST model described in the previous section; the forms are then automatically inputted into a MySQL database and displayed on the website using PHP. In addition to displaying verb forms, the site allows the user to search for a given verb form, using the recognition capabilities of the FST network. This search capacity demonstrates a major advantage of online resources over print.

Many of the verb forms are accompanied by handpicked examples of usage from print sources (mainly online newspapers and chat rooms), audio (from recorded naturalistic dialogues), and movie clips. The examples are provided as complete sentences and short paragraphs; translations are available for all examples. Audio and video examples are likewise accompanied by transcriptions and translations. I am very grateful to Vakhtang Chikovani for finding and translating the examples.

The 'Examples' section of the website provides a different way to access the print, audio, and video examples. This can be done through browsing by verb, or by searching (again, in Georgian or in English).

The 'Exercises' section contains several different types of exercises to provide additional practice for using and conjugating verbs. Many of the exercises are generated based on the conjugated forms or the handpicked examples, and so the correctness of the answers can be checked automatically.

Finally, the 'Resources' section contains links to various online and bibliographical resources about Georgian, as well as technical suggestions for using Georgian fonts.

The website will be operational in spring 2006; anyone interested in using it should contact this author.

6. Conclusions and Further Work

This paper represents a first attempt at modelling Georgian verbal morphology using easily available, off-the-shelf technology such as FST. Using some adaptations to accommodate the templatic and non-compositional structure of the Georgian verbs, we were able to make significant progress and produce one practical application of the computational model for language learners. In short, the model provides a convenient method for representing the existing lexicon for computational applications such as parsing or generation.

Naturally, each technology has its drawbacks. FST provides no way to incorporate frequency information about the Georgian lexicon, and, in general, is not an accurate model for how verbs are learned. Unfortunately, creating a statistically sensitive model of the Georgian lexicon is not currently an easy proposition, as there are no available corpora of Georgian, and no immediate ways of obtaining statistical distributions.

This project will develop in several ways in the future. First, the existing model will be enriched with more verb types and more inflectional parameters (such as the use of pre-radical vowels and productive passivization and causativization processes). Second, I plan to explore ways to incorporate statistical information into the model, either through the use of connectionist networks or by putting numerical transition probabilities on the different arcs in the FST transducers. The eventual goal would be to create a model that can be used for learning Georgian verb conjugations, which could produce a finite-state network of complete word forms. We also hope that this model and the online reference and collection of examples can serve as the basis for the creation of a corpus of spoken Georgian. Information collected in the corpus can then be used to inform and improve future computational models.

References

Anderson, S.R. (1992). *A-Morphous Morphology*. Cambridge, England; New York: Cambridge University Press.

Beesley, K. & Karttunen L. (forthcoming 2006). *Finite-State Morphology*. Second Edition. Cambridge/ New York: Cambridge University Press.

Beesley, K. & Karttunen L. (2003). *Finite-State Morphology*. Cambridge / New York: Cambridge University Press.

Blevins, J.P. (forthcoming 2006). *Word-Based Morphology*. *Journal of Linguistics*.

Boeder, W. (1969) "Über die Versionen des georgischen Verbs." *Folia Linguistica* 2, 82-152.

Fillmore, C.J. (1988) "The Mechanisms of 'Construction Grammar'." *BLS* 14, 35-55.

Goldberg, A.E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Gurevich, O. (2003). "On the Status of the Morpheme in Georgian Verbal Morphology." *Berkeley Linguistic Society* 29, 161-172.

Gurevich, O. (forthcoming 2006). *Constructional Morphology: The Georgian Version*. PhD Dissertation, UC Berkeley.

Halle, M. & Marantz, A. (1994). "Distributed Morphology and the Pieces of Inflection." Hale K. and Keyser S.J.(eds) (1994). *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*. Where:, MIT Press.

Kurtsikidze, S. (forthcoming 2006). *Essentials of Georgian Grammar*. München: LINCOM Europa.

Melikishvili, D. (2001). *Kartuli zmnis ughlebis sist'ema* [Conjugation system of the Georgian verb]. Tbilisi: Logos presi.

Stump, G.T. (2001). *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge, New York: Cambridge University Press.

The Igbo Language and Computer Linguistics: Problems and Prospects

Chinedu Uchechukwu

Computer Linguistics is a wholly undeveloped and an almost unknown area of research in the study of Nigerian languages. Two major reasons can be given for this state of affairs. The first is the lack of training of Nigerian linguists in this discipline, and the second is the general newness of computer technology in the country as a whole. This situation, however, is most likely to change as a result of the increasing introduction of computer technology in the country, and in the institutions of higher learning in particular. Such a change is highly promising and most welcome, but it also makes obvious three main aspects of computer technology that have to be properly addressed before one can speak with confidence of computer linguistics in connection with any Nigerian language. These three aspects are: appropriate font programs, good input systems, and compatible software. This paper looks at the Igbo language in the light of these points. Section 1, which serves as an introduction, presents the major problems confronting the language with regard to its realisation in the new technology. Section 2 presents the strategies adopted to take care of these problems. Section 3 examines the benefits of such strategies on the development of an Igbo corpus and lexicography, as well as the issue of computer linguistic tools (like spell checkers) for the language. Finally, section 4, the conclusion, examines the prospects of full-fledged computer linguistics in the Nigerian setting.

1. Introduction

There are several issues that constitute a major hinderance to the development of computer linguistics in Nigeria. These range from the implementation problems that confront the information technology policies of the different Nigerian governments, to the lack of harmony between such policies and the Nigerian educational systems, as well as the effects of all these on the different Nigerian languages.

With regard to the government policy, Nigeria has already had two different computer-related policies. The first was the Nigerian Computer Policy of 1988, and the second is the newly enacted Nigerian National Policy for Information Technology (IT) of 2001. As regards the Nigerian National Computer Policy, a comparison of its goals with actual practice has revealed that the policy itself has not been fully

implemented at all. While Yusuf (1998) sees its lack of success as resulting from teachers' incompetence, Jegede & Owolabi (2003) have come to the following conclusions in their own survey: the policy's software and hardware stipulations are completely outdated and not maintained, its teachers' in-service training provision has never been practiced, and the stipulated number of computers per secondary school is rarely to be found. All these can only but confirm the conclusion that "the current pronouncements are obsolete and need to be updated within the dynamic world of computers" (Jegede & Owolabi 2003: 9). The recent Nigerian National Policy for Information Technology has not fared any better. With the assumption that information technology can be said to have started more intensively in the country with the return of democracy in 1999 (Ajayi 2003), the failure of the previous national computer policy is often overlooked, so that the failings of the older policy are simply being repeated. But some of what is being presented as the 'achievements' of the new policy actually diverts attention from simple core issues that need to be addressed before such achievements can be effective nationwide. The first example of such an achievement is a concentration of energy on the readily visible Internet access and on IT workshops for high officials of the government at the federal level (while leaving the average civil servants of the lower cadre to find the means of helping themselves). However, it is especially those of the lower cadre who are surely to be involved in the actual implementation/execution of the policies. The second example is the establishment of the National Information Technology Development Agency (NITDA) with the sum of about \$10 million (Nigerian National Policy for Information Technology 2001:vii). One of the agency's achievements is its 'Nigerian Keyboard' project, which only led to the production of a downloadable single keyboard dll file for the Microsoft operating system (NITDA: <http://www.nitda.org/projects/kbd/index.php>). It shall be shown below how the effort that is being made by the private sector, with little or no financial support, is yielding more benefit. Finally, just like the computer policy of 1988, the more recent policy has also not contributed much to the educational sector. In his examination of the impact of the most recent policy on Nigeria's educational system, Yusuf (2005) sees it as not providing any specific provision for (or application to) education, being market driven, dependent on imported software, and without any specific direction at the institutional levels. His conclusions are that:

The need for integration in teaching and learning, the need for quality professional development programs for pre-service and serving teachers, research, evaluation and development, and the development of local context software are not addressed (Yusuf 2005:320).

The overall conclusion from this overview is that the two policies, as well as their implementation, have not contributed much to the Nigerian educational system.

While agreeing with the analysis of the authors cited above, I would add, however, that one should also bear in mind that “nobody can give what he does not have.” For example, one should not expect the bureaucrats, who have always had their secretaries type their letters, to suddenly understand and fully implement IT policies that they did not encounter in the course of their training. The same also applies to the institutions of higher learning. Here one should not, for example, expect lecturers in the Departments of Linguistics, who did not do any computer-based research work in the course of their studies, to suddenly start supervising PhD projects in computer linguistics. In other words, two groups are involved: the civil service and the teaching force. The civil service is a system that has operated over many decades without the help of computer technology and that consequently can neither fully appreciate nor implement the computer technology-related policies as they affect the educational sector. That also explains why much of the input into making the computer technology immediately relevant to the Nigerian educational system is coming from other channels than the federal civil service system itself. The second side is the teaching force within the Nigerian educational institutions. The majority of Nigerian linguists of the past and present generation were not trained in the area of computer-based research for the simple reason that the ordinary typewriter usually was the best machine available to them at the time they were trained. It is obvious that these scholars would train their successors in line with what they knew. For this simple reason, it is unjustifiable to expect them, as well as those they trained, to suddenly start teaching computational linguistics. The conclusion is that, just like within the civil service, the proper use of computer technology within the educational sector, especially with regard to the Nigerian languages, has to come through other channels than those established by the government, and would involve the inevitable but voluntary contribution of both private institutions and individuals. This is the simple reality that confronts most Nigerian languages today.

Finally, the above state of affairs in both the civil service and the educational sector can be described as human resources related. It is also from this angle that most analysts of the computer and IT policies of the Nigerian National Computer Policy (1988) and the Nigerian Policy on Information Technology (2001) have looked at it. A further confirmation of this is the drive of one single state of the federation, Jigawa State, to simply ignore the slowly grinding federal structure and seriously

invest in information and communication technology hardware, as well in the training of its civil servants and teachers. This exemplary position, which other states of the federation are now imitating¹, was facilitated through an agreement between the Jigawa state government and Informatics Holdings of Singapore in 2001². A further boost to the effort of this single state is the recent W 21 (Wireless Internet) award assigned to the State Governor, Alhaji Ibrahim Saminu Turaki, by the Wireless Internet Institute of the United States of America, which is an international recognition of Jigawa State's investment in ICT and human resources development. However, taking care of the human resources issue does not also simultaneously take care of the technical needs of the Nigerian languages; on the contrary, it makes these technical needs ever more apparent. Thus, the limited increase in computer literacy is enough to make apparent that the Nigerian languages are confronted with two main problems: (1) an appropriate input system in the form of keyboards; and, (2) the fonts for a satisfactory combination of diacritics for the scripts of the individual languages.

In the next section, I shall give an overview of the effort made so far to take care of these two problems for the Igbo language.

2. Computer-Related Problems and their Solutions

The input system and the appropriate font to display the inputted characters are so intertwined that progress in one cannot take place without a similar progress in the other. For the Igbo language (as well as other Nigerian languages), this has meant a constant pendulum movement between fixing the input system and fixing the font. But generally, the effort to solve the input problem for the Igbo language and other Nigerian languages had two main phases: the typewriter phase and the computer phase.

2.1 The Typewriter Phase

The typewriter phase was engineered by Kay Williamson and her colleagues, first at the University of Ibadan and later at the University of Port Harcourt. This involved removing certain foreign symbols on the standard typewriter keyboard and replacing them with special symbols used in Nigerian languages, like the hooked letters of Hausa, or by diacritics like tone marks and sub-dots. The diacritic keys were changed to become 'dead' keys, so that the diacritic (tone mark or subdot or both) was typed before the letter which bears it. With the start of the National Computer Policy in 1988 this effort no longer had any further support and consequently came to an end.

1 <http://www.onlinenigeria.com/articles/ad.asp?blurb=117>

2 <http://www.e-lo-go.de/html/modules.php?name=News&file=article&sid=7512>

2.2 The Computer Phase

The computer phase, on the other hand, witnessed different stages. The first stage, from 1985 onwards, was not concerned with a physical input system (i.e. a keyboard), but mainly with the development of the appropriate font that could be inputted with the available English keyboard. This effort was led by Victor Manfredi on behalf of the Journal of the Linguistic Association of Nigeria (JOLAN), supported by Edward *Ogwejiofor*, a Macintosh programmer in Boston. The first version of the font was called *JolanPanNigerian*; it was expanded to include symbols for other major languages of West Africa, and was consequently renamed *PanKwa*. The main drawback with *PanKwa* was its restriction to Macintosh computers, which were not used in Nigeria; it has also not been possible to adapt it to other operating systems such as DOS, Windows or Linux (for further details see Uchechukwu 2004). This situation remained unchanged until 2000.

The next stage has been aided through a convergence of some favourable factors during the 21st century, including the availability of virtual keyboards, the founding of Unicode, as well as the drive towards the development of a physical keyboard for Nigerian languages. All these, however, have led to four major lines of effort: the aforementioned Nigerian Keyboard Project of the National Information Technology Development Agency (NITDA), the independent endeavours of KQYIN/Lancor, *Alt-1*, and my collaboration with Andrew Cunningham (<http://www.openroad.net.au>).

2.2.1 The Nigerian Keyboard Project

The keyboard layout produced by NITDA is messy. It is not fully Unicode compatible and does not provide the means for adding some diacritics, like a macron. It is not surprising that not much work has gone into the project since the release of its downloadable *dll* file for the Microsoft OS. Finally, the effort mentioned in the sections below can only buttress the point that NITDA's Nigerian Keyboard Project has become another white elephant.

2.2.2 The KQYIN Keyboard

The KQYIN keyboard is not a Nigerian keyboard layout project, but a business venture of LANCOR Technologies of Boston, MA in the United States. The LANCOR Multilingual Keyboard Technology (LMKT) (<http://www.lancorltd.com/konyin.html>) has been used by the company to create multilingual keyboards for different languages. The

KQYIN keyboard involved the use of the LANCOR Multilingual Keyboard Technology (LMKT) to create a physical multilingual keyboard for the Nigerian languages.

Some observable changes that the keyboard has undergone can be summarised as follows. First of all, for the characters (especially the vowels) that are combined with diacritics (in the form of definite symbols placed under the vowels), the company initially used the Yoruba Ife System, which involved the use of a short vertical line under the appropriate character. This was demonstrated on the company's website with an instruction on how to key-in the Yoruba name of the president that contained such characters. Later, the vertical line was replaced with a dot, which is more widespread in the scripts of other Nigerian languages, including Igbo. This could be seen as an improvement, as it would also increase the marketability of the company's keyboard in Nigeria.

2.2.3 *The ALT-I Keyboard*

The African Languages Technology Initiative (Alt-I) can be described as an organisation whose aim is to appropriate modern ICTs for use in African Languages. The company hopes to achieve this through advocating ICT and also delivering ICT-related services. But more relevant here is the organisation's effort to produce a Yoruba keyboard. The organisation listed some of its achievements in this regard on its website. These include:

- The production of an installable keyboard driver, which it hoped to start marketing by 2004;
- Demonstrations (including installation) of its keyboard driver at the following seven universities in the Western part of Nigeria: University of Ibadan; Olabisi Onabanjo University, Ago-Iwoye; University of Ilorin; Lagos State University, Ojo; University of Lagos; Adekunle Ajasin University, Akungba; and Obafemi Awolowo University, Ile Ife;
- The endorsement of its keyboard by the *Yoruba Studies Association of Nigeria* (YSAN) at the 2003 Annual Conference of the Yoruba Studies Association of Nigeria between the 4th and 8th November 2003; and,
- The 2003 IICD award of the African Information Society Initiative (AISI) on Local Content Applications.

Finally, the organisation has started to reach out to other Nigerian languages outside the Western and predominantly Yoruba-speaking part of the country. There is no doubt that the aim is a 'Nigerian Keyboard.' A hint in this direction is the comparison of the KQYIN keyboard with the ALT-I keyboard: "the Alt-I keyboard is superior to the Lancor

product on the grounds that Alt-I considered a lot of human factor engineering and other social issues, which Lancor seems to have overlooked in their keyboard design” (Adegbola 2003). I do not know the details of the issues between the two efforts, but with a population of about 120 million the Nigerian market is large enough for more keyboards. I now turn to the development of the Igbo keyboard.

2.2.4 The Igbo Keyboard

This is simply an effort that arose through my collaboration with Andrew Cunningham. The effort is not supported by any business or charity organisation. From the outset, the focus was to find a solution that would exploit the already available keyboard layouts and adapt them for the Igbo language without building a physical keyboard from scratch.

There are many virtual keyboards on the net that could be altered to that effect, but Tavultesoft’s (www.tavultesoft.com) ‘Keyman’ program was found to be the best. Two possible physical keyboards came into consideration at the initial stage: the German keyboard and the English keyboard. The drawback of the English keyboard is the requirement to hold down or combine not less than three different keys in order to realise a single subdotted character. Such a method is tedious and not particularly appealing. That is why I chose the German keyboard. The special German characters can thus be replaced with specific Igbo characters as shown in Table 1 below.

The third column of Table 1 shows a further combination of the subdotted characters with tone marks. Through the collaboration with Andrew Cunningham, all these and many other changes (especially with regard to the consonants) were incorporated and used to build an Igbo keyboard layout that can freely be downloaded from the Tavultesoft website. Later a similar keyboard map was also made for the English keyboard for people who have access only to the English keyboard. But as has already been pointed out, the users of the English keyboard simply have to cope with the tedious key combinations. I have therefore donated the Tavultesoft keyboard program, together with physical German keyboards, to the Department of Nigerian Languages at the University of Nigeria, Nsukka, as well as to some other Igbo scholars; since then I have been receiving feedback that was further incorporated to refine the program for both the average user as well as for the linguist’s most complicated needs. This has led to the development of the second version of the program. Due to the use of the English language in Nigeria, the third version of the program has now been made QWERTY-based like the English and Danish keyboards, thus replacing the

QWERTZ layout of the German keyboard. In addition, it also has a much better display of the characters than is shown in the third column of Table 1. Like the previous versions, the program shall also be freely available.

Table 1: Some Igbo Characters

German Keys	Simple Igbo Replacements	Tone Marked			
Ü ü	Ụ ụ	Ú ú	Ù ù	Ū ū	ṁ
Ö ö	Ọ ọ	Ó ó	Ò ò	Õ õ	ṅ
Ä ä	Ị ị	Í í	Ì ì	Ī ī	
ß	Ñ ñ				

Finally, while the Igbo Keyman keyboard is a virtual keyboard, its transformation into a physical Igbo keyboard shall be taken up at the appropriate time. For the time being, it has contributed to taking care of the language’s input system. The problem of the appropriate font programs to go with the keyboard has also been taken care of through the increasing number of Unicode-based font programs. Thus, the two aspects of (1) an appropriate input system and (2) the fonts mentioned at the end of section 1 of this paper have been addressed. The next step is to use these facilities to tackle specific linguistic problems of the language. In the next section, I shall present my efforts in this direction, especially with regard to the development of a corpus model for the Igbo language.

3. Computer Technology and the Igbo Corpus Model

Of all the different activities involved in developing the Igbo Corpus Model³ obtaining the proper OCR software was indeed difficult, but more difficult was (and still is) finding an adequate corpus development, manipulation or query system and the use of the software to process Igbo texts.

Some of the programs I initially experimented with were either theory dependent, required some manipulation of the system, or required personally writing the internal components needed; all this would involve an initial preoccupation with more theoretical issues than with the practical development of the corpus itself. It is for this reason that I chose the following three pieces of software: WordSmith, Corpus Presenter, and the Word Sketch Engine (WSE). Some factors influenced my choice.

3 Partly supported through the Laurence Urdang Award 2002 of the European Association for Lexicography.

These programs are relatively theory neutral, have friendly GUI, and are explicit in their claim to be Unicode-based.

With regard to the Igbo texts, one can differentiate between two types. The first type is made up of texts without tone marks:

! ka nọriri Obinna? Obinna tughariri hu Ogbenyeanu...

The second type is tone marked like the text below:

! ka nọriri Óbìnnà? Óbìnnà tughàriri hú Ógbènyèánu

A typical Igbo text written by native speakers for fellow native speakers is usually not tone-marked, simply because many find it tedious, although tone marking Igbo texts would make a great deal of difference. However, for any serious linguistic work or research, the Igbo texts are usually tone-marked. The tone-marked Igbo text above was produced with version 2.0 of the Igbo Keyman program. The rendition in version 3, which is soon to be released, is much better. However, I used mainly Igbo texts without tone marks in my work with the above-named three corpus programs. I examined them based on (1) how they handled the text input; and, (2) how they handled Igbo words, with and without diacritics. I shall present the programs individually.

3.1 WordSmith

The basic problem encountered with WordSmith is that not ALL components of the software are able to handle the Igbo texts appropriately.

For text input, it is not possible to add words with diacritics, either directly through the Igbo Keyman keyboard or simply by copying and pasting. For example, a keyboard input of the word *ọgụ* results in 'üß' (see the WordSmith-Concordance screenshot), while pasting the word only yields a blank in the entry box. In both cases, activating 'Search' does not yield any results.

Figure 1: Text Input in WordSmith

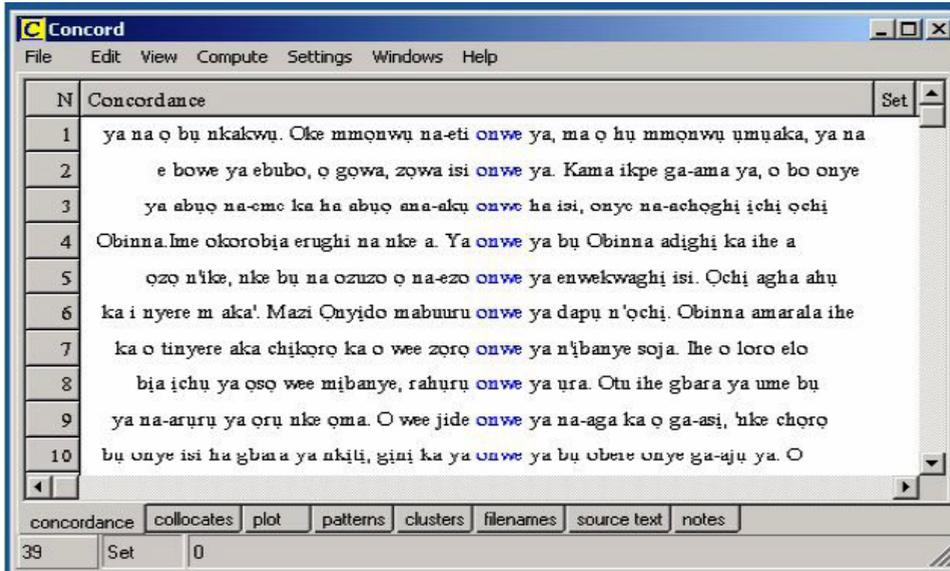


In terms of processing an entry, the concordance component of the software does not function properly. As long as a word to be searched does not bear any diacritics

(like a subdot), the program sorts it appropriately, as can be seen in the concordance of ONWE.

Generally, WordSmith 4.0 is much better than its previous version, but it still has the problem of not being fully Unicode compatible at all levels. Its use for the Igbo language is therefore extremely limited.

Figure 2: Concordance Search for 'onwe' in *WordSmith*



3.2 Corpus Presenter

The problems encountered here are the same as in WordSmith: not all components are fully Unicode-based.

Inputting the Igbo word *ndi* with the Keyman program, or simply by copying and pasting, yields *nd?*, the same result as in WordSmith (This can be seen in the Image of the Corpus Presenter Search):

Figure 3: Text Input in *Corpus Presenter*

Left flank	Keyword	Right flank
ya ghar??, maka na n'ime	[nd?]	?gb? ya no n'ogbe ha n'Ach?..
nan? mmad? iri banyere soja,	[nd?]	?z? ka n'r? na-epioghar? n'?...
akwa, puta n'?z? j?wa	[nd?]	o dugara ?! akwa?be ha si!
lechaa anya, hap? ah?h?a	[nd?]	?z? h?r? ?kaz?, o nwere ihe ?
zamaram. egbe elu igwe zuo g...	[nd?]	mmad? wee malite kwabawa...

In its Dataset component, the program processes an Igbo text in a different manner. The text is usually well displayed as a dataset, as can be seen from the screenshot of the Corpus Presenter Dataset.

Figure 4: Corpus Presenter Dataset

'I ka n'oriri, Obinna?' Obinna tughariri hu Ogbenyeanu. Mgbe Ogbenyeanu ji ruo ya nso tukwasia ya aka n'ubu, Obinna amaghi. Nke a na-eme ya ugboro an? mmadu na-aju ya ka n'oriri, n'ime otu izu a. Gini mere o ji buru nani ya ka ha na-ahu na-aju udi ajuju a kwa oge. Um? okorobia juru obodo na-erughari, o buru nani Obinna ka ha na-abia aju ma o ka n'oriri. Obinna echechaala echiche maka nke a, o gba ya ghariri, maka na n'ime ndi ogbo

But switching to the 'Text Retrieval Level' for the manipulation of the displayed text simply turns the characters with subdots into question marks. This can be seen from the screenshot Text Retrieval component.

Figure 5: Corpus Presenter Text Retrieval

'? ka n?r?r?, Obinna?' Obinna t?ghar?r? h? Ogbenyeanu?. Mgbe Ogbenyeanu? ji ruo ya nso t?kwas? ya aka n'ubu, Obinna amagh?. Nke a na-eme ya ugboro an? mmad? na-aj? ya ? ka n?r?r?, n'ime otu izu a. G?n? mere o ji b?r? nan? ya ka ha na-ah? na-aj? ?d? aj??? a kwa oge. ?m? okorob?a juru obodo na-erughar?, ? b?r? nan? Obinna ka ha na-ab?a aj? ma ? ka n?r?r?. Obinna echechaala echiche maka nke a, ? gba ya ghar??, maka na n'ime nd? ?gb? ya no n'ogbe ha n'Ach?, nan? mmad? iri banyere soja, nd? ??z? ka n?r? na-epioghar? n'?n?

Finally, one major point of difference between WordSmith and Corpus Presenter is in the making of word lists. While WordSmith can do it without loss of data, Corpus

Presenter simply leaves out ALL the characters with diacritics. This can be seen in the screenshot of the Corpus Presenter and WordSmith wordlists.

Figure 6: Corpus Presenter Word List

Word	Frequency
?g??	00001
?G??!. '!	00001
?g?d?	00002
?g?d?.	00001
?garanya	00001
?gb?	00002
?gba	00005
?gbagbu	00001

Figure 7: WordSmith Word List

N	Word	Freq.
3	○	637
4	KA	535
5	○	400
6	A	384
7	HA	325
8	IHB	297
9	BU	281

Generally, the two programs are very good for manipulating texts of European languages, with Corpus Presenter also having the further advantage of the capacity for POS-tagging. But with regard to the Unicode scripts of an African language like Igbo, they both have their limitations.

I shall now turn to the next program, which is the most promising, the most user-friendly, and the most suitable for use in teaching corpus linguistics at an elementary or advanced stage.

3.3 Word Sketch Engine (WSE)

WSE is so far the only relatively theory-neutral program that has been able to handle Igbo texts without tone marks. The different components of the program are also reliable. For example, words could be keyed in or copied into the Concordance component of the program without any loss or distortion. This applies to both words without subdots and those with subdots, as can be seen from the two screenshots:

Figure 8: A Word Without Subdots in WSE

Home	Concordance	Word Sketch	Thesaurus	Sketch-Diff	Frequency	Collocation
KWIC/Sentence	View options	Sample	Filter	Sort		

#278 o dugara ulo akwa ebe ha si | E jiri **okwu** ndi a kowaa akuko maka unu anumanu
#342 anumanu , Obinna buru mmadu na - ekwu **okwu** . Nke oso , ma nke mbe mere , ma nke
#375 ime ya , e bo ya mbe . Mbe anaghi ekwu **okwu** . O naghi aru uka , maka na ebe
#745 onyezinyere gi ka i bia gwawa m udi **okwu** a n ' ututu a , ka o bu ihe si
#953 gi na Ogbenyeanu ka unu bia gwawa m **okwu** n ' ututu a ? Unu che . . . ? '
#980 bia gbafe ha ato ebe ha no na - ekwu **okwu** a , Obinna wee loo aso , na - ele anya
#2375 na - agba oso egbe . Otu nwoke kwuru **okwu** si , na i kpasuo mmadu iwe oku ,
#3042 uka kwadoro na ha ga - eje tikpoo **okwu** mmuo obodo . O bu Obinna so na ndi
#3989 ya onu , egwu anaghi ekwe ya asa ya **okwu** , maka na o maghi nke o ga - eme ,
#4025 mmadu na soja . Ndi Igbo na - ekwu n ' **okwu** si , na a gachi eji maka mgbagbu ghara
#5704 o bu na ha ejideghi ya . Nke a bu **okwu** aka ya . Nani ihe di ya mkpa bu ka
#7865 oso . Ma o maghi na o bu ezi **okwu** na icheku oku na - adasa ndi mmadu
#8390 Eneka ga - acho iji ya wee gwa Obinna **okwu** ma onu bie . Ma , ndi Igbo si na

Figure 9: A Word With Subdots in WSE

Home	Concordance	Word Sketch	Thesaurus	Sketch-Diff	Frequency	Collocation
KWIC/Sentence	View options	Sample	Filter	Sort		

#2754 aku ? ' Nna anyi ukwu , i si m mee **gini** ? Abiara m ka i nyere m aka . Mani Onyido
#4639 nke bu onye isi ha gbara ya nkiti , **gini** ka ya onwe ya bu obere onye ga - aju
#5086 Onyido na - adi ire nke ukwu , ma **gini** mere o ji ruo na nke ya , ghara idi
#6401 idi oku ahụ i manyere na - eme **gini** ugbu a ? ' ' Eji m ya agu akwukwo
#8504 Chineke—e | Chin e ke—e || ' ' Obinna , **gini** ka i na - emere mkpotu ? Ya bu na

The collocation analysis of WSE does not present any difficulties; neither does it distort the characters of the language. This can be seen in the collocation screenshot:

Figure 10: Collocation in WSE

The screenshot shows the WSE interface with the following elements:

- Navigation tabs: Home, Concordance, Word Sketch, Thesaurus, Sketch-Diff, **Frequency**, Collocation.
- Sub-navigation tabs: KWIC/Sentence, View options, **Sample**, Filter, Sort, and three menu icons.
- Results list:
 - #6657 ihi na e jidela **onye** gworo ya . Nke **ozo** , o no na o jere gworo ogwu ,
 - #5487 amaghi ka a ga - esi kororo **onye ozo** . Olee ka a ga - esi kororo umu uwa
 - #1432 osoro soja chuwa . O kweghi na **onye ozo** ga - ama ebe o na - eje ezo , maka na
 - #8161 ya adila nso , maka na otu **onye** dibia **ozo** gwara ya na o baa soja , na o ga - abu
 - #439 ya . Kama ikpe ga - ama ya , o bo **onye ozo** ihe o mere ; nsi nkita riri eje rewe
 - #7380 ihe o bua onye Igbo maara , **onye** Igbo **ozo** maara ya , maka na ihe di n ' etiti onye

Finally, the texts processed with the three different programs are Igbo texts without tone marks. The way each program handles such a text determines the extent to which the program can be taken into consideration for texts with greater character combination and complication. This simply means that, at the level of texts without tone marks, WordSmith and Corpus Presenter are of limited use. For WSE, however, an investigation is still to be made into its further use for processing Igbo texts that have been complicated through the combination of more diacritics with the subdotted words. In the next section, I shall briefly discuss the effect of the above problems on Igbo lexicography, as well as the efforts to develop a spellchecker for the language.

3.4 Igbo Lexicography and the Igbo Spellchecker

The three lexicographic works of the language to be examined here are Williamson's Igbo-English Dictionary (1972), Echeruo's Igbo-English Dictionary (1998), and Igwe's Igbo-English Dictionary (1999). The spellchecker is a project that I am currently working on with Kevin Skannel (<http://borel.slu.edu/nlp.html>).

Each of the three dictionaries have imprints of the technological stages of the time when they were written. Williamson's dictionary was produced with the typewriter. Its legacy is the imprint it has left on Igbo orthography. A particular tone of the language known as the 'downstep' was marked in her dictionary through placing the dash '-' on the sound segments that incorporate the downstep. This means the following forms for the vowels without subdots: ā ē ō ī. The same was also done for the subdotted vowels. But as was pointed out in section 2 above, this was achieved through the physical adjustment of the typewriter. Through such a method, Igbo texts can be properly tone marked with the old typewriter.

The presentation of the Igbo characters in Echeruo's dictionary has been strongly influenced by the available fonts within the Microsoft operating system. The author

simply used the German umlauted vowels (shown in the first column of Table 1 above) instead of the Igbo subdotted vowels. But, with his method, an Igbo word like ùtò ‘sweet, sweetness’, whose tone is indicated on the word itself, is written as ütö [LH]. This simply means marking the tone extra. Such a method, however, becomes irrelevant when one tries to use it for the representation of a fully tone-marked Igbo text. The method has not found much acceptance (Uchechukwu 2001), but the author has also agreed to change it in line with the existing orthography (Anyanwu 2000). There is no doubt that this would become easier for him as a result of the growing improvements in computer technology, including the freely available Igbo keyboard and Unicode-based font programs.

Igwe’s dictionary was written in line with the existing orthography; however, the author’s effort involved the use of the ‘Symbols’ windows within Microsoft’s *Word* to painfully click on the individual Igbo characters of his 850-page dictionary! The only mark the method left on his work can be seen in the combination of the vowel <i> with a tone mark. Within the Microsoft 95/98 system used by the author, this lower case vowel is automatically changed into an upper case vowel through such a combination. Thus, a combination of the vowel <i> with the high tone symbol (accent acute) <´ > yields <I>; and a combination with the low tone symbol (grave accent) <` > is realized as <I>. This is regardless of whether the letter occurs at the beginning of a word, in the middle, or at the end. But with the current improvements in the different operating systems, as well as the available Igbo keyboard, such problems can now be completely addressed.

The spellchecker is still an on-going project between Andrew Scannel and myself. For the time being, this is restricted to Aspell and Igbo texts without tone marks. The development of a spellchecker for fully tone-marked Igbo texts shall be taken into consideration at the appropriate time.

4. Prospects for Computer Linguistics

The above situation of the Igbo language has not only highlighted the stages involved in the struggle of the language with modern technologies, but also how this development can be enhanced.

The texts of the Igbo language without tone marks can be processed to some extent by some corpus processing or development systems. Such texts are of little research significance, however, since they cannot graphically represent the very phenomena that are of interest to both the ordinary linguist and the computational linguist. Compared with the situation of such matters just a few years back, it is already a great advancement to have some software that can handle Igbo texts without tone

marks. But a further step in the direction of laying a good foundation for future computer linguistics within the Nigerian setting requires the different sophisticated corpus software, acoustic phonological system, spellchecker and so on to be in a position to handle the scripts of the language, whether with or without tone marks.

The conclusion is that through developing the Igbo keyboard, as well as through the availability of freely downloadable Unicode-based fonts, the problem confronting the average Nigerian language is now solely a software problem and no longer the problem of a physical keyboard or an operating system. Two additional points support this conclusion. First of all, the same Keyman program can be used to write more keyboard maps for other Nigerian languages, thus making it unnecessary to invest in building a physical keyboard from scratch. All that the users with different native languages within the Nigerian setting need to do is simply click on their language keyboard map. This solution is not likely to change, because the English keyboard has become part and parcel of the computer hardware within the Nigerian setting. Thus, the production of a physical keyboard for a Nigerian language would definitely involve an expansion of the physical English keyboard. The second point is the present effort by Andrew Cunningham and Tavultesoft to further port the Keyman program into the Linux operating system. This should make available to Linux users the same keyboard facility that the Keyman program has provided for the Windows operating system. The effect would be to have the appropriate software also within the Linux OS. Both developments can only but further the point that the keyboard and font stages have been addressed. Computer linguistics within the Nigerian setting now faces the problem of developing the necessary programs that make use of the facilities presently available.

Finally, the picture of the Igbo language presented here shows that the current excitement about the new technology should not make us overlook the simple fact that the three necessary elements for the development of computational linguistics for any African language are: (1) an appropriate font program; (2) a good input system; and (3) compatible computer programs. Thus, the development of computational linguistics for an average African language depends on the extent to which these three aspects have been taken care of for the respective language.

References

Adegbola, T. (2003). *2003 Annual report on the activities of African Languages Technology Initiative (Alt-I)*. <http://alt-i.org/2003Report.doc>.

Ajayi, G.O. (2003). "NITDA and ICT in Nigeria." *Paper presented at the 2003 Round Table Talk on Developing Countries Access to Scientific Knowledge*. (The Abdus Salam ICTP, Trieste, Italy, 23 October 2003).

Anyanwu, R.J. (2000). "Echeruo, Micheal J.C. 1998. Igbo-English Dictionary: A Comprehensive Dictionary of the Igbo Language, with an English-Igbo Index." *Frankfurter Afrikanistische Arbeitspapier*. 12: 147-150.

Echeruo, M.J.C. (1998). "Igbo-English Dictionary: A Comprehensive Dictionary of the Igbo Language, with an English-Igbo Index." Yale: Yale University Press.

Egbokhare, F.O. (2004). *Breaking Barriers: ICT-Language Policy and Development*. Dugbe, Ibadan: ODU'A Printing & Publishing Company Ltd.

Jegede, P.O. & Owolabi, J.A. (2003). "Computer Education in Nigerian Secondary Schools: Gaps Between Policy and Practice." *Meridian: A Middle School Computer Technologies Journal*. Raleigh, NC: NC State University, 6(2).
<http://www.ncsu.edu/meridian/sum2003/nigeria/print.html>.

Nigeria National Computer Policy (1988). Lagos: Federal Ministry of Education.

Nigerian National Policy for Information Technology (IT) (2001).
<http://www.nitda.gov.ng/docs/policy/ngitpolicy.pdf>

Uchechukwu, C. (2001). "Echeruo na Eceruo, kedu nke ka mma...?" *KWENU*, 1(8), 16-22.

Uchechukwu, C. (2004). "The Representation of Igbo with the Appropriate Keyboard."

Paper presented at the International Workshop on Igbo Meta-Language. (University of Nigeria, Nsukka, 18 April 2004).

Yusuf, M.O. (1998). "An Investigation into Teachers' Competence in Implementing Computer Education in Nigerian Secondary Schools." *Journal of Science Teaching and Learning*, 3(1 & 2), 54-63.

Yusuf, M.O. (2005). "Information and Communication Technology and Education: Analysing the Nigerian National Policy for Information Technology." *International Education Journal*, 6(3), 316-321.

Annotation of Documents for Electronic Editing of Judeo-Spanish Texts: Problems and Solutions

Soufiane Roussi and Ana Stulic

The result of an interdisciplinary process comprising Linguistics, Information and Computer Sciences, this contribution consists of modelling the annotated electronic editing of Judeo-Spanish¹ texts written in Hebrew characters, following the principle of document generation in a collaborative work environment. Our approach is based on the concept of digital document annotation that places mark-up at any text level, starting with the character resulting from the transcription. Our point of view is that the annotations of a 'translated/interpreted' document can have two different purposes: to interpret (to add new mark-up in order to propose a different interpretation from the one formulated at the beginning); and, to comment (make a comment on the interpretation done by another author). Our aim is to make it possible for the reader/user to interact with the document by adding his own interpretation (translation) and/or comments on an interpretation made by another author. We present a model for the description of annotation in response to our problem.

1. Introduction

In this paper, we will explore the problem of digital document annotations in application to the very specific problem of building a Judeo-Spanish corpus. We will briefly present the interest of such an enterprise, together with some difficulties related to building corpora in general, as well as those specific to the Judeo-Spanish case. Considering the recent developments in information technology (IT), we will take into account the concepts of digital documents, automatic generation of documents, and production of digital documents in collaborative mode, and then, apply them to our problem. Finally, we will propose a prototype model for Judeo-Spanish corpus building in the context of a collaborative environment. This proposal offers some conceptual and methodological solutions based on existing technologies, but leaves open the question of technical realisation.

¹ The language of the Sephardic Jews, who, after being expelled from Spain at the end of 15th century, settled in the greater Mediterranean area.

2. Building a Judeo-Spanish Corpus

2.1 *The Research Value of a Judeo-Spanish Corpus*

Judeo-Spanish is the language spoken by the Sephardic Jews, who, after being expelled from Spain at the end of 15th century, settled in the greater Mediterranean area. It represents a variety of Spanish that has followed its own development path since late 15th century (though not without any contact with the Iberian Peninsula), and is relatively well documented. Many original documents in Judeo-Spanish, such as manuscripts, books and other printed material have been conserved. Linguistic fieldwork from the beginning of 20th century also yielded a certain number of oral transcriptions.

From a linguistic point of view, Judeo-Spanish is very interesting because it offers numerous possibilities for comparative and historical linguistic analysis, in that peninsular Spanish is itself very well documented in terms of the pre-expulsion period. Equally, the original sources are of great value for historical and cultural research.

Unfortunately, in many countries where it was kept alive for centuries, Judeo-Spanish has been in progressive decline since the beginning of the 20th century, and, at present, it is no longer spoken in many cities of Balkan Peninsula (formerly the centres of Judeo-Spanish culture). Therefore, the editing of original Judeo-Spanish sources can also contribute to the preservation of knowledge about this language.

The approach adopted here in treatment of Judeo-Spanish documents has been primarily oriented to their usage as a corpus for linguistic research, but it can be extended to other uses as well.

2.1 *General Problems of Linguistic Corpora Editing*

In the humanities (philology, history, literature and linguistics), the word 'corpus' has traditionally referred to any body of texts that are used for research. In modern linguistics, it refers most often to relatively large collections of texts that represent a sample of a particular variety or use of language(s). Language corpora can be in the form of manuscripts, paper-printed, sound recordings (spoken corpora), or machine-readable. Nowadays, it has become common to think of it especially in this latter form- and not without reason. The development of computer-readable corpora has enlarged the possibilities of linguistic research by simplifying search tasks, and making possible the use of large portions of texts.

Compilations of electronic corpora, especially those with historical dimensions, rely upon existing written documents, which are frequently philological editions. In

electronic corpora, regardless of whether the electronic text is made from a source document (such as a manuscript or original edition) or a philological edition, the authors of the corpus must develop annotation strategies in order to represent the source document from which the text is derived. The information commonly provided in annotations concerns metadata about the digital document itself (as well as the source document), but it can also deal with the linguistic properties of the text, such as parts of speech, lexemes, prosody, phonetic transcription, semantics, discourse organisation, co-reference, and so forth. Designed to be global (extending to the entire corpus) and universal in their validity, most types of linguistic annotations represent complex tasks that are economically very expensive and inconvenient for lesser-used language corpora that are designed principally for scientific research. On the other hand, de Haan's proposal of problem-oriented tagging offers another point of view (de Haan 1984). In this approach, the users take a corpus, either annotated or unannotated, and add to it their own form of annotation, oriented to their own research goal. This type of annotation seems very promising in the context of specific research needs, provided that the annotational system in question is supplied with a dynamic and interactive dimension.

Although in language corpora building, emphasis is frequently placed on developing software search possibilities and linguistic annotations, one crucial question remains: Where do the texts come from? Though reproducing the raw text of a source document may seem like a simple task, it often isn't, especially when dealing with ancient texts, or texts in writing systems other than Latin character set. In current corpora building practices, the old philological problems are still of current interest. In the traditional philological paper editions, the text is determined by the editor (on the basis of the source document[s]), together with the critical apparatus and the writing system, and the reader/researcher can only accept the editor's interpretation. In electronic corpora, the approach is similar: while the source document remains inaccessible for practical reasons, whatever the editorial choices of the corpus author are (including critical apparatus, writing system and annotations), the user cannot intervene or adapt them to his/her own purposes.

On the other hand, source documents have never been more accessible in technical terms, albeit on condition that they are available in digital form, as an image or as a sound file. The possibility of consulting the digital image of a source document in parallel to its electronic transcription would enable the researcher to be critical of the editor, and would resolve some of the philological problems caused by the necessarily arbitrary choices one is forced to make in a philological edition.

2.2 Specific Problems in the Judeo-Spanish Context

The most salient specificity of Judeo-Spanish texts is the writing system in which they were composed, and it represents, at the same time, the most important difficulty in their editing and computer processing.

The Judeo-Spanish documents produced in the post-expulsion period were commonly written in an adaptation of Hebrew script (in this context its distinctive Rashi version is frequently used, but the square Hebrew script is equally found; the difference between the two is only in the form of the letters). The practice of using Hebrew script for texts in Romance languages was already very common before the expulsion (see Minervini 1992).

In the history of writing systems, the adaptation of a script originally designed for one language into the writing system of another language is a cultural phenomenon that has been frequently repeated; it leads to the development of new conventions adapted for the target language that involve the use of graphemes coming from the source language's writing system.

In its original form, the Hebrew script made no use of vocalic graphemes, because in most of the cases, the realised linguistic contrast was of grammatical and not lexical character. In order to avoid certain ambiguities, some letters progressively acquired vocalic meaning in certain contexts in Biblical Hebrew (*yod*, *waw*, *he* and *aleph*). The fully vocalised writing system of Hebrew was designed much later, and has been mostly reserved to the texts of the Bible (for more details see Sampson 1997:123-129).

Similar to the Hebrew writing tradition, in Spanish texts, the fully vocalised script was reserved only to translations of the Bible and sacred texts, but in other texts the usage of letters with vocalic meaning was extended to all contexts, with the particularity that two letters, *yod* and *waw*, could denote two vowels each, /e/ and /i/, and /o/ and /u/ respectively. Also, a diacritic sign (of different shape in different times and traditions) has been introduced above certain letters to create new graphemes for consonants that had no counterpart in Hebrew writing system, or that lacked phonological value in Hebrew (Sampson 1997:123-129).

Nevertheless, the history of this adaptation shows many variations in application of conventions. One of the sources of variation comes from the possibility to use different letters for the same phoneme (this kind of variation is found even within the same text). Although some of the basic principles of adaption of Hebrew script for the Spanish language have probably been transmitted over generations, the reading and

writing of Hebrew represented the only constant knowledge, so conventions applied to Spanish could be updated at any time. On the other hand, Judeo-Spanish evolved phonologically, and this was also reflected in the writing system (Pascual Recuero 1988).

The main, but not the only problem, in the editing Judeo-Spanish texts, is the underspecified use of vowel graphemes. This becomes even more complex if we consider the fact that the vowel system has suffered some modifications and that reconstruction on the basis of 15th century Spanish can not be completely reliable (especially if we take into account the fact that 15th century Spanish is known through the different variations it presented, including in the vowel system).

Two approaches are possible: (1) conserving the original script conventions in every way, by *transliteration* of original documents, which means replacing each grapheme by another one; and, (2) interpreting vowel graphemes, by *transcription*, which means specifying the vowels where their presence is indicated. If carried out in the traditional sense of a philological edition, both have their advantages and drawbacks. The transliteration conforms to the source, so that the researcher can rely on its fidelity, but it doesn't make the text more accessible in terms of intelligibility. The transcription is certainly more intelligible, but the choices of interpretation of vowels done on the basis of reconstruction are determined (and fixed forever) by the transcriber, and fidelity to the source is lost². The need for both transliteration and transcription as research tools has been recognised in the study of Judeo-Spanish texts; they are both used in different contexts, and sometimes even the parallel versions of texts are proposed, as in the edition of Jewish medieval texts from Castilla and Aragon by Laura Minervini (1992); also, a similar solution is proposed independently in Stulic (2002) for the editing of a 19th century Judeo-Spanish newspaper *El amigo del pueblo*.

In this paper, we wanted to model a solution for the electronic editing of such texts that could encompass both approaches, and maybe even offer something more as a research tool. Although we took as a starting point a very concrete Judeo-Spanish text from the 19th century, the problems we are endeavouring to solve could derive from any text edition where a choice between the intelligibility of the text and the fidelity to the source is imposed.

2 The text transcribed for syntactic analysis wouldn't be useful, for example, for research on phonological issues or writing system history.

3. The Digital Document and Annotation

3.1 What is a Digital Document?

The beginning of the year 2000 showed an increasing spread of online environments, which has been facilitated by the use of databases for the storage of content, the automatic generation of digital documents, and the use of interactive 'fill-in' forms. It is this favorable situation, together with numerous developments in Web technologies, that leads us to the creation of a Judeo-Spanish corpus in a digital environment.

First of all, what is a digital document? Following the definitions in use among French scholars, it can be defined on three levels: as an *object* (material or immaterial), as a *sign* (meaningful element) and as a *relation* (communication vector). As an object, a digital document can be defined as *"a data set organised in a stable structure associated with formatting rules to allow it to be read both by its designer and its readers"* (Pédaugue 2003). As a sign, a digital document is *"a text whose elements can potentially be analysed by a knowledge system in view of its exploitation by a competent reader"* (Pédaugue 2003). From the social perspective, a digital document can be viewed as *"a trace of social relations reconstructed by computer systems."* (Pédaugue 2003).

For our approach, the perspective of a digital document as an object (material or immaterial) seems the most relevant. As such, the digital document opens three principal issues: storage, plasticity, and its programmability (Rouissi 2004). It is this latter characteristic that captures our attention here. Dealing with the annotation of electronic documents, our aim is to make use of solutions made possible by automatic generation.

In a functional approach, a digital document can be considered as any other document. In documentalist theory, the definition of a document emphasised the function (something that serves as an evidence) more than the actual physical form (paper, stone or antilope) (see Schürmeyer 1935; Briet 1951; Otlet 1990). The irrelevance of support became only more evident in the case of the digital document. Buckland writes *"The algorithm for generating logarithms, like a mechanical educational toy, can be seen as a dynamic kind of document unlike ordinary paper documents, but still consistent with the etymological origins of "docu-ment", a means of teaching - or, in effect, evidence, something from which one learns"* (Buckland 1998). In accordance with this point of view, and considering the digital document as a research tool, we would like to explore its programmable possibilities.

3.2 Automatic Generation of Electronic Documents

Current Web technology offers the possibility of creating a document upon request. In this context, the result of an execution of a computer program is a document: a Web page obtained on the basis of one or multiple resources (program, database, cascading style sheets, etc.). The automatic generation of Web documents is based on the use of scripting programming languages whose execution takes place on the server. These technologies accelerate the treatment, making it possible to surpass the limits of HTML (Hypertext Markup Language), which remains static, and only permits the handling of the document's layout. Equally, a connection is established with the databases where the information to be furnished in a given context is stored. Thanks to these principles of functioning, and with an appropriate editing program, individual appropriation by Information Technology non-professionals has developed. The ease with which it is possible to reuse documentary resources existing on the Web only contributes to an even greater development of digital document production. The correction, modification and adding possibilities facilitate digital document production in an autonomous mode. Production in autonomous mode is defined by a user, who elaborates the content and defines its layout for his personal use, or for that of other users. This production is realised with the help of suitable IT material and programs. The autonomous mode means that the user has all the creative freedom (choice of layout, colours, fonts, format, file names, diffusion and storage hosting, etc.), but also presupposes that he has all the technical competence needed. On the other hand, by conserving the autonomy of production while facilitating the exchange, the semi-autonomous mode can be applied in a collaborative environment where production rules need to be followed. It is in this context that we wish to situate our work on digital document annotation. Our aim is to put at the user's disposal (in this case, scholars working on Judeo-Spanish documents) a tool modelled on the principle of semi-autonomous mode production of digital documents.

In this context, the rules are defined at technical level (layout rules and data structuration) but the decision to produce and to publish is made by the user (Rouissi 2005). This implies of course that he can be identified by the system, and that the maintenance and technical assistance service is provided. The user doesn't produce in an isolated manner, but in a collaborative work environment, which somewhat constrains his production, but, on the other hand, offers a conception of the whole and facilitates the integration of individual work.

The emergence of online environments, and within them, the possibility of producing and uploading digital documents, has changed the role of the user from passive user/reader to active user/author. In semi-autonomous mode, what is defined in advance concerns the common vocabulary, the visual aspect, and the structure of digital documents³.

The principal advantages of production in semi-autonomous mode are:

- the durability of the system and the possibility of its evolution (the contents can evolve more easily);
- the autonomy of handling (the use of 'fill-in' forms allows for the handling of the data by the users themselves);
- the minimal technical competence that is required (the systems remain intuitive and easy to handle); and,
- the common vocabulary.

It is in terms of these principles that we envisage the development of the digital document annotation model for Judeo-Spanish corpus edition.

3.3 Production in a Collaborative Mode

Production in a collaborative mode, already widely present in different forms as collective websites nourished by individual contributions (forums, blogs, wikis, etc.), seems particularly suited for the collaborative work of specialists of the documents in discussion. In this sense, annotation can play an important role in the evaluation, interpretation and production of a document, which thereby becomes itself dynamic and subject to evolution. The final (and collective) document obtained is the result of the contribution of individual fragments (but not necessarily the sum of the contributions), and it can also result from choices the user has made.

Annotation is something added to the document. It can be a remark, a comment, or, in our case, even a proposition of interpretation. Already in 1945, Vannevar Bush envisaged for the Memex (a device which was supposed to create links between related topics in different research papers) that the owner could add his own comments (Bush 1945). More recently, numerous developments in annotation management systems appeared with real promise in the direction of sharing and exchanging information. Several (wide public) office programs (some versions of MS-Word) or the W3C project Annotea in the Web domain (Annotae 2005) are just some of the examples of applications aimed at sharing annotations. A more exhaustive list can be found in Perry (2005).

3 We are not dealing here with the application of the XML (eXtensible Mark-up Language) which plays an important role in the data structuring and data exchange.

There are two types of annotations: semantic annotations (in the sense of standardised metadata Web annotations) and free annotations. The former are attached to the actual work on the Semantic Web, and are based on the development of metadata and/or ontologies used in the description of the document, with the purpose of facilitating their localization, identification and automatic recognition. Without neglecting this important issue, we will focus here on free annotations, because they are used - in philological and linguistic analysis - to interpret and comment upon documents, and we therefore consider that they can constitute an important factor in the development of collaborative digital production, improving, at the same time, communication among the specialists of the domain in question.

From our point of view, annotation can have two purposes. The first concerns the interpretation of the original document (how to translate or read it). The second adds comments to the one part of the document and/or to the interpretation already made. The annotation can be placed at several levels. The global level concerns the annotation made on the whole of the document put into discussion. This annotation can be based on free comments made on the entire document or can represent a reaction to the global annotation already made. We consider it useful, for the sake of analysis, that the zone of annotation is freely marked in the document. The smallest mark up unity is the character; therefore, a part of the word, a word, a line, several lines, a paragraph, or several paragraphs can also be the target of the mark up.

Collaborative work is situated in the context of semi-autonomous production. Every member of this collaborative community participates in a responsible way, benefits from the result of the work of the community, and receives feedback for his work. Two models of work where everyone's autonomy can be expressed are: *cooperative work* (wherein everyone accomplish a part of the work and shares it with others) and *collaborative work* (wherein several autonomous individuals work together in order to produce collectively). We'll see how the concept of a digital document and its production can help in our case.

4. Prototype Model for the Descriptions of Annotation

4.1 General Properties

Considering the problems related to the treatment of Judeo-Spanish texts and to the building of a corpus, and taking into account the theoretical approach to the digital document (particularly from the point of view of collaborative work), we propose a model that can respond to the needs we have identified. Our work focuses on the definition of needs without making choices that could constrain future program

implementation. In this sense, our contribution is situated on the analytic level prior to the concrete realisation of project.

One of the first preoccupations is the constitution of documentary corpus. In order to achieve this, the model must be conceived as a digital repository of documents that are described with specifications that are sufficiently fine-grained, but open to interoperability. The collaborative dimension must take into consideration the management of users. Our intention is to describe annotations and to build a typology of annotations that will appear as the system begins to function.

We will resume briefly here the requirements that the model should satisfy:

- the source document should be accessible, as a transliterated version, or, ideally, as a collection of image files;
- the transcribed version is given as a starting point of discussion/analysis;
- the metadata annotations (according to the widely accepted standard, Dublin Core) are provided with the transcribed version and image files;
- the authorised user can add free annotations on a global or any other text zone level, starting with the character; they may include new interpretation (may include corrections) and/or comments; and,
- the authorised user can export the result of his or another's work by making choices to use or not the annotations that he or another has made.

The annotation management system has still to be developed, and it will use, as a basis, the model here presented.

4.2 Data Description Model

We have seen that the particularity of Judeo-Spanish texts serves as the source of many methodological and technical problems. Among them, we'll concentrate on the annotation that represents the form of document production in collaborative mode. The annotation can help to handle various interpretations, as well as the comments made by reader-users. Here, we see the possibility of developing real support for documentary information and communication: the document in question remains the carrier of different contributions, and represents, at the same time, the archive of the exchanges made, as well as the basis for different reading possibilities.

We offer here a proposal of a data description model adapted to our needs in the study of Judeo-Spanish texts. We have chosen the representation model based on the Codd's relational model (Codd 1970), which presents the meaningful data in the form of relations, as grouped properties, and as a whole. The choice of representation

inspired by the relational model is guided by the desire to conserve the freedom in the definition of the necessary fields (type, size, etc.) in the future implementation.

The relational model for the annotation of Judeo-Spanish documents is constituted of four relations. The primary keys are in bold and underlined, the foreign keys are in bold and followed by the # sign.

- **ANNOTATION** (**annotation_num**, annotation_date_creation, annotation_date_lastmodified, annotation_comment_title, annotation_comment_text, annotation_language, annotation_position_begin, annotation_position_end, annotation_status, annotation_commented_num#, annotation_type_num#, document_num#, author_num#).

The relation ANNOTATION, whose identifier *annotation_num* (primary key) has to be created automatically, conserves the trace of the date of its creation (*annotation_date_creation*), as well as the date of the last modification (*annotation_date_lastmodified*). An annotation is described equally by the language of the author with the property *annotation_language*. The status carried by the property *annotation_status* allows the author to say whether the annotation is considered as active or not: value 1 is for active and public (default value), 0 signifies inactive or private (i.e., reserved to its author, who considers it of no utility for the public while in draft status, or for some other reason). The annotation can be deactivated, because it evolves over time and has no permanent character (notion of duration of annotation). The contribution added by the identified author (primary key *author_num*) over the given document (*document_num*) has a short title (*annotation_comment_title*), which will be used for the publication of the lists of comments, and a text field (*annotation_comment_text*) of variable size. The position of the annotation in the given text/document is determined by the starting point (*annotation_position_begin*) and the ending point (*annotation_position_end*). In the case where the annotation concerns the whole document and not only one of its fragments, position is indicated in the following manner: *annotation_position_begin* = *annotation_position_end* = 0.

The *annotation_commented_num* property allows for the formal identification of the annotation on which the comment is made. In the opposite case, where the annotation is not made over another annotation (without a link to another annotation), the value of *annotation_commented_num* is 0. The type of annotation can be specified (otherwise the value 0 is attributed) with the *annotation_type_num* property, which points to the common vocabulary shared by the members of the community.

- **ANNOTATION_TYPE** (**annotation_type_num**, annotation_type_vocabulary, annotation_type_description, annotation_type_mode_edit)

The property *annotation_type_mode_edit* indicates (with the value 0 or 1) whether the annotation aims (or not) at proposing that a text be substituted for the one that is initially put into discussion. This kind of annotation corresponds to the editing action in a document.

Some of the examples of *annotation_type_vocabulary* values would be: interpret, comment, refuse, confirm, accept, and so forth. This vocabulary can be established also *a posteriori* with the observation of users' practices, and with their help. The addition of elements in the future table can be made from proposals of the users.

The *annotation_type_description* allows for the inclusion of additional information, and for making the chosen vocabulary more precise.

- **AUTHOR** (author_num, name, email, login, password).

The AUTHOR relation describes the users that are authorised to interact with the document, to bring annotations, or to modify the existing ones (the author can only modify his own annotations). Whoever wants to propose a contribution must be identified.

- **DOCUMENT** (document_num, title, creator, keywords, description, publisher, contributor, date, resourcetype, format, identifier, source, language, relation, coverage, rightsmanagement).

The DOCUMENT properties follow the recommendations of the Dublin Core Metadata Initiative (Dublin Core 2005). The identifier *document_num* can serve for the denomination of different document resources (the original document can be presented as a collection of image files, but also as a transcription in ASCII format). The export formats envisaged here are HTML, XML or even PDF. Some difficulties are still to be overcome, since our annotation in theory allows for mark up overcrossing.

The documents are uploaded by the administrator on the proposal of one of the members of the community.

The program implementation must bring into consideration the different applications that are possible. The process of document annotation consists of two complementary phases. The first one comprises the contributive action of adding or modifying annotations (on the entire document or on one of its fragments). The second one concerns reading through the exploitation of existing annotations. The reading possibilities include the choice of exporting and saving files in different formats.

5. Conclusion and Prospects

The work on a documentary corpus as specific as Judeo-Spanish texts opens many questions concerning the design of the proposed electronic model.

In the context of this particular type of document, there is a necessity to share the results of study, remarks, comments and interpretation among the members of a relatively small and geographically dispersed scientific community. A possible solution is to develop a tool for the management and rationalisation of individual work.

In our approach to the problem, we have taken as a theoretical basis recently developed concepts related to digital documents, focusing chiefly on the programmable aspect of a digital document. Taking into account that the documents in question can be considered as digital documents (over which it is possible to act), we have worked on the modelling of contributions that can be added to these objects of study. This has led us to a model that describes the annotations made on documents collected in a digital repository.

The program implementation, which is still to be executed, must be situated in a full Web approach in order to satisfy the conditions of collaborative work and to remain easy to use with the help of the Web navigator.

Some questions that will certainly appear in the implementation phase are not accounted for by the proposed model, such as how to apply a modification starting from one sequence (the annotation that proposes that one sequence be replaced by another) to the whole document, or should all users have the same profile and same possibilities to act within the documents.

In this sense, the proposed model leaves many questions to be answered, but the direction in which we are pointing seems rather promising.

References

"Annotea Projet." Online at <http://www.w3.org/2001/Annotea>.

Briet, S. (1951). *Qu'est-ce que la documentation?* Paris: EDIT.

Buckland, M.K. (1998). "What is a 'Digital Document'?" *Document numérique* 2, 221-230.

Codd, E. F. (1970). "A Relational Model of Data for Large Shared Data Banks." *Communications of the ACM*. June 1970, 13(6), 377-387.

Dublin Core. Dublin Core Metadata Element Set, Version 1.1:Reference Description. <http://dublincore.org/documents/dces/>.

de Haan, P. (1984). "Problem-oriented Tagging of English Corpus Data." Aarts, J. & W. Meijs (eds) (1984). *Corpus Linguistics*. Amsterdam: Rodopi, 123-139.

Marshall, C.C. (1998). "Toward an Ecology of Hypertext Annotation." *Proceedings of 'Hypertext 98'*. New York: ACM Press.
<http://www.csdl.tamu.edu/~marshall/ht98-final.pdf>.

Minervini, L. (1992). *Testi giudeospagnoli medievali (Castiglia e Aragona)*. 2. Napoli: Liguori Editore.

Otlet, P. (1990). *International Organization and Dissemination of Knowledge: Selected essays*. (FID 684). Amsterdam:Elsevier.

Pascual Recuero, P. (1988). *Ortografía del ladino*. Granada: Universidad de Granada, Departamento de los Estudios Semíticos.

Perry, P. (2001). "Web Annotations."
<http://www.paulperry.net/notes/annotations.asp>.

Pédauque, R.T. (2003). "Document: Form, Sign and Medium, As Reformulated for Electronic Documents." *Version 3*, July 8, 2003.

http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/05/94/sic_00000594_02/sic_00000594.html.

Rouissi, S. (2005). "Production de document numérique en mode semi-autonome au service de la territorialité." *Colloque Les systèmes d'information élaborée*. Ile Rousse, juin 2005.

Rouissi, S. (2004). *Intelligence et normalisation dans la production des documents numériques. Cas de la communauté universitaire*. PhD Thesis, Bordeaux 3 University.

Sampson, G. (1997). *Sistemas de escritura. Análisis lingüístico*. Barcelona: Gedisa. First published in 1985. Writing Systems. London: Hutchinson.

Stulic, A. (2002). "Recherches sur le judéo-espagnol des Balkans: l'exemple de la revue séfarade 'El amigo del pueblo'." (I, 1888, Belgrade). MS thesis, Bordeaux 3 University.

Schürmeyer, W. (1935). "Aufgaben und Methoden der Dokumentation." *Zentralblatt für Bibliothekswesen* 52, 533-543. Reprinted in FRA 78, 385-397.

Röscheisen, M., Mogensen, C. & Winograd, T. (1997). *Shared Web Annotations As A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples*. Technical Report CSDTR/DLTR. <http://dbpubs.stanford.edu:8091/diglib/pub/reports/commentor.html>

Bush, V. (1945). "As we may think." *The Atlantic Monthly*. July 1945. <http://www.theatlantic.com/doc/194507/bush>.

Wynne, M. (2003). *Writing a Corpus Cookbook*.
<http://ahds.ac.uk/litlangling/linguistics/IRCS.htm>.

“W3-CorporaProject” (1996-1998).

Online at http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/introduction.html.

Il ladino fra polinomia e standardizzazione: l'apporto della linguistica computazionale

Evelyn Bortolotti, Sabrina Rasom¹

Dolomite Ladin is a polynomic language: it is characterised by a rather large variety of local idioms, that have been undergoing a process of normalisation and standardisation. This process can be supported by the development of computer-based infrastructures and tools. The efforts of the major Ladin institutions and organisations have led to the creation of lexical and terminological databases, electronic dictionaries, concordancer tools for corpora analysis, and, eventually, to the development of spell-checkers and 'standard adapters/converters'.

1. Introduzione

Il ladino delle Dolomiti (Italia) è caratterizzato da una grande varietà interna, che ha reso necessario un intervento di normazione e standardizzazione, nel rispetto del carattere polinomico della lingua stessa.

Nelle cinque valli ladine dolomitiche si vanno formando lingue di scrittura, o standard di valle. Alcuni idiomi di valle sono piuttosto unitari ed è stato sufficiente codificarli, ma in Val Badia (con Marebbe) e in Val di Fassa la loro varietà ha portato alla proposta di una normazione che si sovrapponesse alle sottovarianti di paese. Ad esempio il *badiot unitar*, basato principalmente sull'idioma centrale (*San Martin*), ma aperto anche a elementi provenienti da idiomi di altri paesi, e similmente il *fascian standard*, orientato verso l'idioma *cazet*, la cui scelta come variante standard è giustificata anche dal fatto che questo idioma è molto più vicino nelle sue caratteristiche linguistiche agli altri idiomi dolomitici.

Infine si è sentito il bisogno di elaborare un livello ancora più alto di standardizzazione valido per l'intera Ladinia, sulle orme del *Rumantsch Grischun*, dando il via all'elaborazione del *Ladin Dolomitan*, o *Ladin Standard (LS)*.

Dal punto di vista della polinomia quindi, da una situazione linguistica molto differenziata, si è passati prima a un livello più alto di normazione che consente, prendendo la valle come unità di riferimento, di raccogliere più varietà in una norma unica. A seguire si è raggiunto un terzo livello che permette di avere a disposizione

1 I paragrafi "Introduzione" e "Risorse e infrastrutture linguistiche e lessicali" sono stati scritti da Evelyn Bortolotti; il paragrafo "Correttori ortografici con adattamento morfologico" è stato scritto da Sabrina Rasom.

un unico idioma di riferimento, una norma o lingua standard per tutte e cinque le vallate.

La standardizzazione ha riguardato in un primo tempo la forma grafica: si è cercato di adottare una grafia il più possibile comune alle diverse varianti ladine dolomitiche, per garantire, nella diversità, il riconoscimento dell'appartenenza alla stessa famiglia linguistica e un maggiore grado di coesione e uniformità del sistema.

L'utilizzo della lingua ladina scritta nelle scuole, nelle pubbliche amministrazioni, nella stampa ecc. comporta, a seconda del grado di standardizzazione del ladino utilizzato, un più o meno marcato sforzo di avvicinamento alla norma da parte dello scrivente e richiede una grande consapevolezza delle differenze fra la propria sottovarietà e lo standard utilizzato nello scrivere.

Di fondamentale importanza in questo processo di standardizzazione è stato e continua a essere l'apporto della linguistica computazionale. La diffusione della tecnologia informatica permette infatti la creazione e lo sviluppo di risorse linguistiche e di infrastrutture di supporto al trattamento automatico della lingua, soprattutto nell'ambito della lessicografia moderna e tradizionale e della terminologia settoriale basate su corpora e della standardizzazione linguistica. Inoltre favorisce la realizzazione di strumenti di aiuto alla scrittura che facilitino il passaggio verso la norma standard.

Nel caso del ladino delle valli dolomitiche, i vari progetti relativi all'informatizzazione delle risorse lessicali e allo sviluppo di strumenti per il trattamento automatico sono stati portati avanti attenendosi al principio di conservazione e valorizzazione della ricchezza e della varietà in una visione unitaria. Questo principio deriva dalla riflessione teorica del linguista corso Jean-Baptiste Marcellesi, che per primo ha introdotto il concetto di "lingue polinomiche" (*Langues Polynomiques*) [Chiorboli 1990].

Le principali istituzioni coinvolte nei progetti di modernizzazione e di trattamento automatico del ladino promossi o realizzati in collaborazione con l'Istitut Cultural Ladin "Majon di Fascegn" sono l'Union Generela di Ladins dles Dolomites e l'Istitut Ladin "Micurà de Rù".

I principali obiettivi perseguiti in campo linguistico computazionale sono:

- l'informatizzazione del patrimonio lessicale ladino con la creazione di una banca dati generale lessicale ladina (BLAD), di banche dati strutturate delle varietà locali e di una banca dati centrale dello standard;
- l'elaborazione di dizionari degli standard di valle (per il Fassano standard: DILF "Dizionario Italiano - Ladino Fassano / Dizionèr talian-ladin fascian", per il badiotto unitario: Giovanni Mischi, Wörterbuch Deutsch - Gadertaslich /

Vocabolar todësch - ladin (Val Badia); per il gardenese: Marco Forni, Wörterbuch Deutsch - Grödner-ladinisch / Vocabuler tudësch - ladin de Gherdëina) e del ladino standard (DLS "Dizionar dl Ladin Standard") anche in versione elettronica e alcuni consultabili online;

- la raccolta di glossari terminologici, parzialmente consultabili online (glossari di ambiente, botanica, materie giuridico-amministrative, medicina, architettura e costruzioni, pedagogia, musica e trasporto turistico);
- la creazione di corpora elettronici analizzabili tramite un'apposita interfaccia: il *web-concordancer*;
- la realizzazione di strumenti informatici per facilitare l'uso e l'apprendimento delle varianti standard: dizionario elettronico, e-learning, correttori ortografici e adattatori per il fassano standard e per il *Ladin Standard*.

2. Risorse e infrastrutture linguistiche e lessicali

2.1 BLAD: Banca dac Lessicala Ladina

La banca dati BLAD consente l'accesso:

- allo SPELL base, il database che raccoglie circa 15.000 schede con LS e idiomi di valle (lessico prevalentemente moderno), da cui è stato elaborato il DLS;
- alle banche dati locali di lessico tradizionale, per un totale di circa 90.000 schede, in cui sono confluiti i dati raccolti dai dizionari e dai database di lessico patrimoniale (per il fassano: Dell'Antonio 1972, Mazzel 1995, De Rossi 1999; per il badiotto: Pizzinini-Plangg 1966; per il gardenese: Lardschneider-Ciampac 1933 e 1992, Martini 1953; per il fodom: Masarei in stampa; per l'ampezzano: Comitato 1997) (descrittivi);
- alle banche dati dei dizionari moderni (normativi), per un totale di circa 250.000 schede (DILF, Mischì, Forni);
- alle banche dati terminologiche elaborate nell'ambito del progetto TERM-LeS, in cui sono raccolte circa 16.000 schede.

Fig. 1: L'interfaccia di ricerca della banca dati BLAD: la ricerca può essere effettuata in italiano, tedesco, LS e negli idiomi di valle.

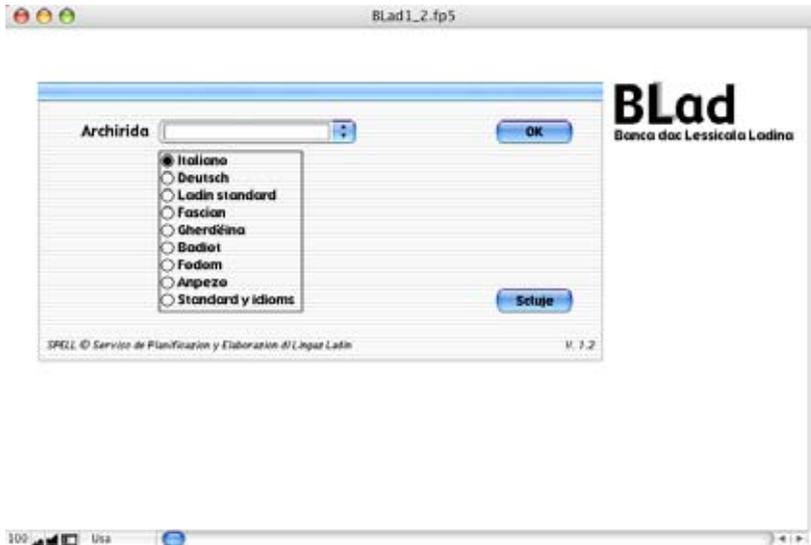


Fig. 2: Esempio di scheda: dal pannello "Idioms", in cui accanto al lemma in LS vengono riportate la traduzione italiana e tedesca e le forme corrispondenti negli idiomi di valle, si ha anche accesso alle singole banche dati locali.



2.2 I dizionari normativi: le versioni elettroniche online del DILF e del DLS

Il DILF e il DLS sono strumenti linguistici la cui accessibilità e semplicità d'uso consentono la facile consultazione di risorse lessicali di grande importanza per le

persone che si trovano a dover scrivere in fassano standard o in ladino standard.

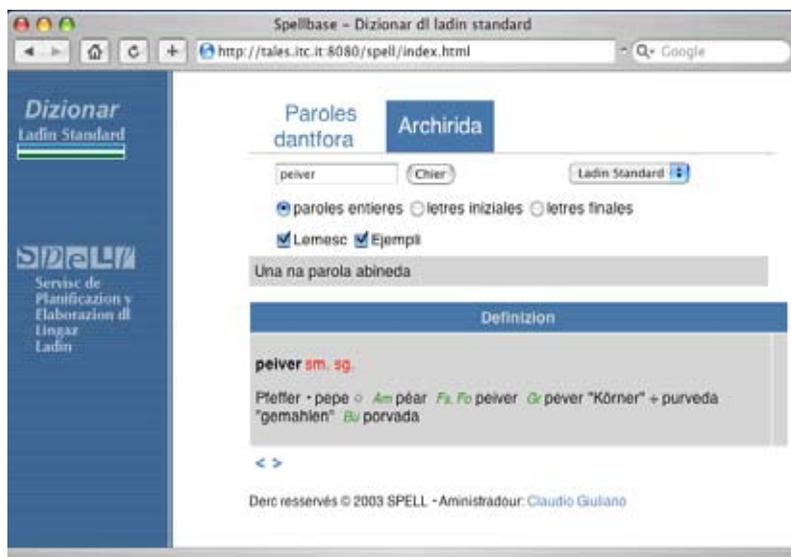
Nel DILF (Dizionario Italiano - Ladino Fassano / Dizionèr talian-ladin fascian) il repertorio lessicale tradizionale registrato nei dizionari descrittivi è stato integrato con un'ampia selezione di voci moderne il cui uso è ampiamente documentato nella produzione linguistica contemporanea. Questa versione elettronica, corrispondente alla seconda edizione cartacea (2001), è stata realizzata con la collaborazione dell'ITC-IRST di Trento, avviata nell'ambito di progetti relativi al trattamento automatico e allo sviluppo di infrastrutture informatiche per il ladino (progetto "TALES", iniziato nel 1999).

Fig. 3: DILF online: esempio di ricerca dal ladino fassano all'italiano, con visualizzazione dei risultati.

The screenshot shows the DILF online interface. On the left is a blue sidebar with the logo and text: "Dizionario italiano ladino fassano", "Dizionèr talian - ladin fascian", and "Istitut Cultural Ladin SPELL". The main content area has tabs for "Premessa", "Ricerca", "Note", and "Autori". The "Ricerca" tab is active, showing a search box with "piaz", a "Cerca" button, and a dropdown menu set to "Fassano-Italiano". Below the search box are radio buttons for "parole intere" (selected) and "lettere iniziali", and checkboxes for "lemmi" (checked), "esempi", and "accezioni". A status bar indicates "5 di 5 lemmi trovati".

Lemma	Abstract definizione
<u>spiaz</u>	sf. piaz, piac; plaza, -es...
spiazzo	sm. spiaz, spiac; piac, piac...
attraversare	v. t. traversèr, traversa, passèr via, passa; (in direzione del parlante) passèr cà, (andare attraverso qc.) passèr fora permez...
pattinaggio	
via1	sf. stràda, -es; via, vies...

Fig. 6: Esempio di scheda di lemma LS con traduzione italiana e tedesca e forme locali corrispondenti

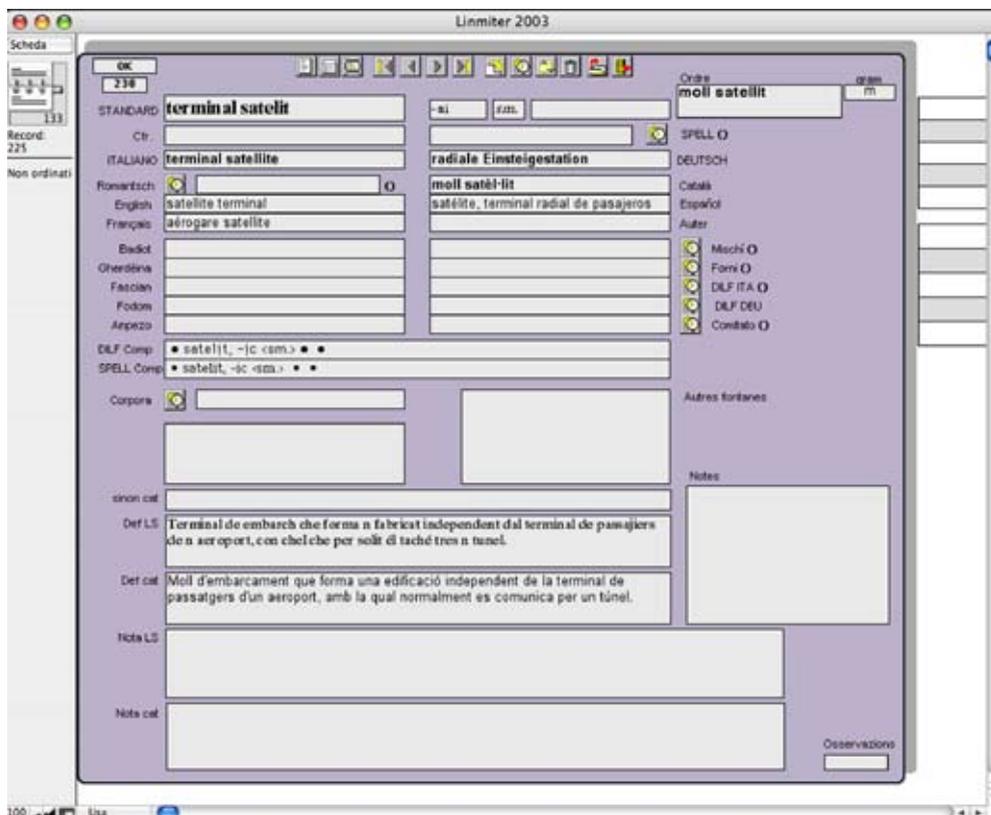


2.3 Il progetto TERM-LeS: Standardizzazione lessicale e terminologia per le lingue ladina e sarda²

Il progetto, condotto negli anni 2001-2003, ha previsto l'elaborazione di terminologia moderna e la creazione di banche dati terminologiche in ladino standard nei settori in cui l'uso della lingua ladina è obbligatorio (amministrazione) e in altri settori rilevanti per la realtà territoriale (architettura e costruzioni, ambiente, medicina, botanica, musica, pedagogia, trasporto turistico). Alcuni di questi glossari sono stati realizzati nel quadro del progetto Linmiter, promosso dalla Direzione Terminologia e Industrie della Lingua (DTIL) dell'Unione Latina, in coordinamento con altre minoranze linguistiche europee neolatine.

2 Il ladino e il sardo sono le lingue oggetto dello stesso progetto di standardizzazione terminologica e lessicale finanziato dalla Comunità Europea tra il 2001 e il 2003.

Fig. 7: Esempio di scheda terminologica: l'interfaccia di lavoro permette una visione sinottica sullo standard e sulle varianti. Da essa è inoltre possibile accedere direttamente alle banche dati degli idiomi di valle, ai dizionari moderni e ai corpora testuali.



Tanto nell'elaborazione lessicografica quanto in quella terminologica in ladino standard, la polinomia, la varietà e la diversità degli idiomi ladini, è la base di partenza per la standardizzazione; la lingua standard attinge quindi dalle varianti locali riassumendole in una norma comune, mirando nel contempo a essere non uno strumento per soffocare le differenze, ma al contrario un tetto, un ombrello di protezione contro gli influssi e le interferenze esterne, e un punto di collegamento fra i diversi idiomi per permetterne uno sviluppo parallelo e armonico. La struttura delle banche dati e l'interfaccia di lavoro tengono quindi conto dell'esigenza di avere facile e immediato accesso a tutte le risorse linguistiche utili e necessarie.

2.4 I corpora elettronici

Nell'ambito del progetto TALES sul trattamento automatico della lingua ladina sono state create delle raccolte organiche di testi ladini, sia nel ladino standard che nei singoli idiomi. I corpora raccolti (fassano, gardenese, badiotto e ampezzano) contengono complessivamente circa 6.500.000 parole. I testi selezionati coprono un periodo che va dal XIX secolo fino ai giorni nostri, con preponderanza di testi appartenenti alla seconda metà del XX secolo. Per garantire un certo equilibrio fra i vari generi, sono stati inseriti sia testi letterari (prosa, poesia, teatro, memorialistica, testi sul folclore e le tradizioni, libri di preghiere), sia testi non letterari (testi giuridici e amministrativi, modulistica, testi di informazione giornalistica e pragmatici, testi di divulgazione scientifica e culturale, testi scolastici). Attualmente il corpus fassano è quello nella fase più avanzata di elaborazione. La sua strutturazione, che fornisce per ogni testo informazioni rilevanti (data, luogo di provenienza, tipologia testuale, autore), permette di affinare la ricerca secondo una serie di criteri predeterminati.

I corpora sono consultabili tramite il *concordancer*, uno strumento elaborato *ad hoc* e rivolto anzitutto al linguista e allo studioso del ladino: esso permette l'analisi dei testi attraverso la ricerca di concordanze, collocazioni e frequenze secondo la modalità KWIC (*Keyword In Context*), ossia un sistema che permette di visualizzare la parola oggetto della ricerca con il suo contesto a corredo.

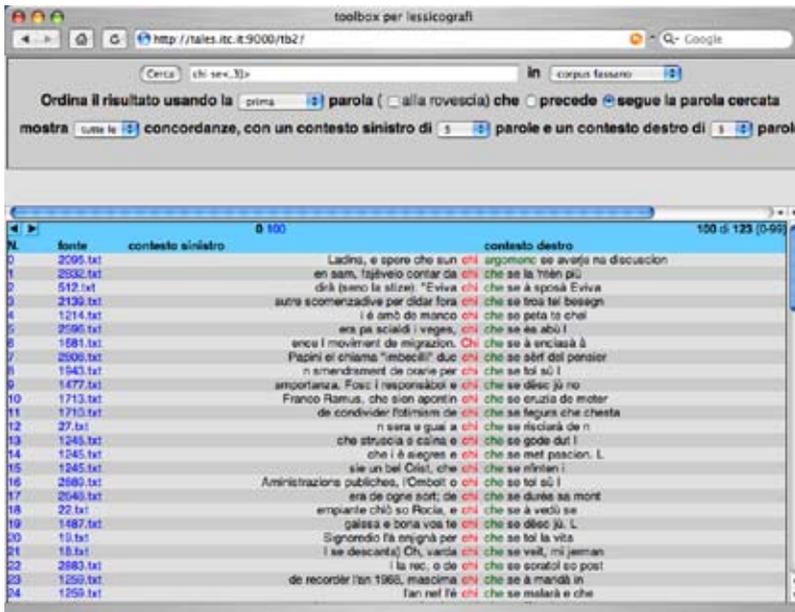
Una sezione del *concordancer* è dedicata ai corpora amministrativi bi- e trilingui allineati: questa raccolta è di particolare utilità nel lavoro di realizzazione di glossari settoriali.

Il lavoro preliminare per lo sviluppo dello strumento di analisi di corpora è consistito nella creazione di corpora testuali: i testi selezionati sono stati acquisiti elettronicamente oppure manualmente e sono stati elaborati rispettando precisi criteri di archiviazione. In seguito sono stati classificati in base alla loro appartenenza diatopica (individuazione della variante in cui sono scritti) e diacronica (dalle prime testimonianze scritte in ladino sino ai testi contemporanei) e alla tipologia testuale (testi letterari e non letterari con individuazione del genere specifico). Per ogni testo è stato creato un frontespizio elettronico che riassume tutte queste informazioni: periodo, autore, genere, nome del file, titolo originale, numero di parole, variante. Il frontespizio è stato linkato al testo corrispondente, cosicché le informazioni in esso contenute possano essere utilizzate per circoscrivere la ricerca.

I corpora consultabili attraverso il *concordancer* si rivelano una risorsa di fondamentale importanza per diversi campi di applicazione: per lo studio del lessico, della sintassi e della morfologia, per l'elaborazione di strumenti normativi e didattici,

per le operazioni di corpus planning, per i progetti relativi alla standardizzazione della lingua e per l'elaborazione di banche dati lessicografiche e di terminologia multilingue.

Fig. 8: Esempio di ricerca nel *concordancer*: la parola cercata viene visualizzata in un breve contesto e in rosso per essere facilmente riconosciuta. Anche la parola che la precede o segue può essere evidenziata in un colore diverso. L'interfaccia di ricerca permette all'utente di decidere quante parole devono apparire nel contesto.



3. Correttori ortografici con adattamento morfologico

Nell'ambito del progetto SPELL-TALES, nell'anno 2002, l'Istituto Culturale Ladino ha realizzato il correttore ortografico del ladino Fassano in collaborazione con l'ITC-IRST di Trento e col sostegno finanziario dell'Unione Europea, del Comprensorio Ladino di Fassa C11 e della Regione Trentino Alto-Adige. L'Istituto Culturale Ladino ha curato la parte linguistica del progetto riguardante la creazione delle regole morfologiche, mentre la parte informatica è stata seguita dall'ITC, nella persona del dott. Claudio Giuliano, che ha elaborato e applicato il programma di generazione delle forme. La realizzazione del software è poi stata affidata alla ditta Expert System di Modena. Nel corso del 2003 è stato messo a punto anche il correttore ortografico del ladino standard - SPELL-checker -, elaborato con le stesse modalità del correttore Fassano.

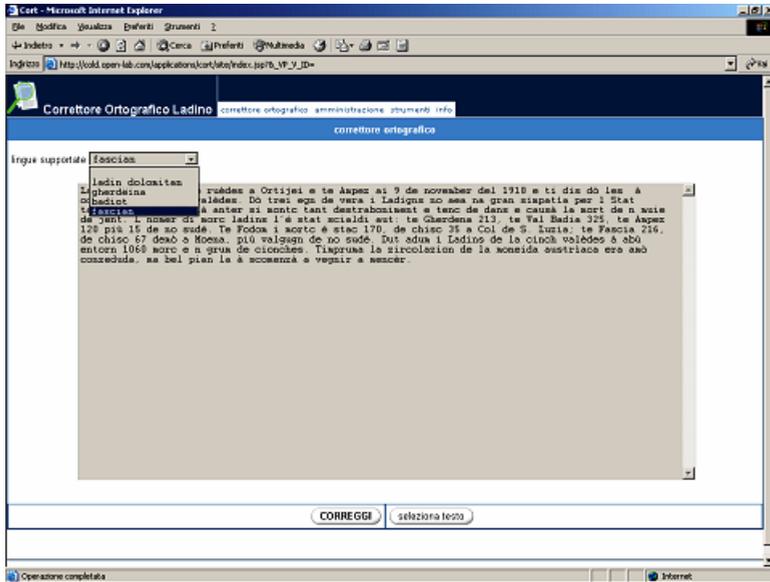
I due software di correzione sono realizzati in ambiente Windows e Macintosh per tutti gli applicativi Consumer della suite Microsoft Office e sono corredati di

installazione automatica e di guida e assistenza all'installazione. Le funzionalità previste da questi due strumenti, similmente ai correttori ortografici disponibili per le lingue maggioritarie, prevedono la correzione di errori di digitazione, di ortografia e di morfologia direttamente durante la redazione di un testo, oppure in un secondo momento, sottoponendo a verifica un testo già scritto. I correttori ortografici in questione si basano su forme ricavate dai dizionari di riferimento, rispettivamente il DILF per il fassano standard e il DLS per il ladino standard; il formario di base fassano è poi stato implementato con forme ottenute dallo spoglio di alcuni testi amministrativi e giornalistici (Usc di Ladins) esportati tramite il concordancer e con i dizionari personalizzati realizzati dagli utenti che hanno usato il correttore per circa un anno nell'ambito dell'amministrazione. Per quanto riguarda il formario del ladino standard l'implementazione è avvenuta attraverso il dizionario personalizzato creato dai redattori del sito Noeles.net e da export delle banche terminologiche.

L'Istituto Culturale Ladino "Majon di Fascegn", in collaborazione con l'Istituto Ladino "Micurà de Rù" e con il supporto tecnico-informatico della Ditta Open Lab di Firenze, sta ora lavorando a una seconda generazione di correttori ortografici delle varietà dolomitiche (badiotto, fassano, gardenese) e del ladino standard, non più ancorata alla Suite Office di Microsoft. Si tratta di una scelta all'avanguardia che prevede la realizzazione di software e sistemi aperti (open source) disponibili in rete e non più dipendenti da programmi specifici.

Il motore alla base dei correttori delle diverse varianti sarà uno e il sistema totalmente internazionalizzato: l'interfaccia d'uso a lingua multipla permetterà di scegliere la lingua stessa di interfaccia e la lingua di correzione all'atto della configurazione. Le novità pratiche più importanti di questi strumenti stanno in un'accurata ricerca delle corrispondenze interne al formario, che non si presenterà più come una semplice lista di forme non ancorate fra loro, bensì avrà una sua coerenza interna, riconoscerà la categoria grammaticale a cui appartiene ogni forma, la rispettiva forma base di riferimento, la coniugazione o declinazione e la marca d'uso, per poi suggerire l'eventuale forma corretta. Inoltre, nel processo di sofisticazione delle opzioni di correzione che verranno fornite, le varietà ladine inserite nel correttore saranno corredate da uno specifico algoritmo fonetico - soundlike - che non sarà più quello Metaphone classico dell'inglese (usato fra l'altro dalla maggior parte dei correttori ortografici esistenti), ma verrà elaborato sui soundlike specifici delle varietà in questione, permettendo quindi opzioni di correzione più precise.

Fig. 9: Interfaccia del nuovo correttore ortografico open source accessibile direttamente da internet.



Nel progetto di elaborazione di questa nuova tipologia di strumenti di correzione l'Istituto Culturale Ladino "Majon di Fascegn" sta sperimentando un'ulteriore funzione nell'ambito del correttore ortografico *open source* per l'assistenza a chi scrive in ladino fassano e in ladino standard. Si tratta di una funzione di adattamento morfologico che permetterà di passare "automaticamente" dalla variante locale fassana (*cazet*, *brach*, *moenat*) alla variante fassana standard, oppure dalle varietà standard di valle (fassano standard, badiotto unificato e gardenese) al ladino standard durante la digitazione di un testo.

I nuovi strumenti di correzione si rendono quanto mai utili nel momento in cui una lingua polinomica viene riconosciuta come lingua ufficiale e si ritrova quindi a dover far fronte alle esigenze della comunicazione in ambito pubblico-amministrativo e nella scuola. Come è stato già osservato, l'apporto della linguistica computazionale nel processo di standardizzazione si è rivelato di primaria importanza per facilitare il passaggio dalla sottovarietà dello scrivente (impiegato, insegnante, studente o semplice appassionato) a una lingua standard ufficiale e unificata. I correttori ortografici sono quindi un passaggio fondamentale verso la realizzazione di strumenti ausiliari sempre più sofisticati per coloro che lavorano ogni giorno con la lingua ladina.

Bibliografia

Bortolotti, E. & Rasom, S. (2003). "Linguistic Resources and Infrastructures for the Automatic Treatment of Ladin Language." *Proceedings of TALN 2003. RECITAL 2003. Tome 2*. Batz-sur-Mer, 253-263.

Chiocchetti, N. & Iori, V. (2002). *Gramatica del Ladin Fascian*. Vigo di Fassa: Istitut Cultural Ladin "majon di fascegn".

Chiorboli, J. (ed.) (1991). *Corti 90: actes du Colloque international des langues polynomique*. PULA n° 3/4, Université de Corse.

Comitato del Vocabolario delle Regole d'Ampezzo (1997). *Vocabolario Italiano - Ampezzano*. Cortina d'Ampezzo: Regole d'Ampezzo e Cassa Rurale ed Artigiana di Cortina d'Ampezzo e delle Dolomiti.

Dell'Antonio, G. (1972). *Vocabolario ladino moenese - italiano*. Trento: Grop Ladin da Moena.

De Rossi, H. (1999). *Ladinisches Wörterbuch: vocabolario ladino (brach)-tedesco. A cura di Kindl, U. e Chiocchetti, F.* Vigo di Fassa: Istitut Cultural Ladin "majon di fascegn"/Universität Innsbruck.

Forni, M. (2002). *Wörterbuch Deutsch - Grödner-Ladinisch. Vocabuler Tudësch - Ladin de Gherdëina*. San Martin de Tor: Istitut Ladin "Micurà de Rü".

Giuliano, C. (2002). "A tool box for lexicographers." *Proceedings of EURALEX 2002*. Copenhagen: Center for Sprogteknologi (CST), 113-118.

Istitut Cultural Ladin "majon di fascegn"/SPELL (2001). *DILF: Dizionario italiano - ladino fassano con indice ladino - italiano = Dizioner talian-ladin fascian con indesc*

ladin-talian. *Dizionèr talian - ladin fascian*. II ed., 1. rist. Vigo di Fassa: Istitut Cultural Ladin "majon di fascegn"/SPELL.

Lardschneider-Ciampac, A. (1933). *Wörterbuch der Grödner Mundart*. (Schlern-Schriften ; 23). Innsbruck: Wagner.

Lardschneider-Ciampac, A. (1992). *Vocabulér dl ladin de Gherdëina: Gherdëina - Tudësch*. Übera. von Mussner, M. & Craffonara, L. San Martin de Tor: Istitut Ladin "Micurà de Rù".

Martini, G.S. (1953). *Vocabolario gardenese - Italiano*. Firenze: Francolini.

Masarei, S. (2005). *Dizionar Fodom - Talián - Todësch*. Colle Santa Lucia: Istitut Cultural Ladin "Cesa de Jan" - SPELL.

Mazzel, M. (1995). *Dizionario Ladino fassano(cazet) - Italiano: con indice italiano-ladino*. 5. ed. riv. e aggiornata (prima ed. 1976). Vigo di Fassa: Istitut Cultural Ladin "majon di fascegn".

Mischì, G. (2000). *Wörterbuch Deutsch - Gadertalisch. Vocabolar Todësch - Ladin (Val Badia)*. San Martin de Tor: Istitut Ladin "Micurà de Rù".

Pizzinini, A. & Plangg, G. (1966). *Parores Ladines. Vocabulare badiot - tudësk. ergänzt und überarbeitet von G. Plangg*. Innsbruck: L.F. Universität Innsbruck.

SPELL (2001). *Gramatica dl Ladin Standard*. Urtijëi: SPELL.

SPELL (2002). *DLS - Dizionar dl Ladin Standard*. Urtijëi: SPELL.

Schmid, H. (2000). *Criteria per la formazione di una lingua scritta comune della ladinia dolomitica*. San Martin de Tor/Vich: Istitut Ladin "Micurà de Rü"/Istitut Cultural Ladin "majon di fascegn".

Valentini, E. (2002). *Ladin Standard. N lingaz scrit unitar per i ladins dles Dolomites*. Urtijëi: SPELL.

Videsott, P. (1997). "Der Wortschatz des Ladin Dolomitan: Probleme der Standardisierung." Iliescu, Maria (Hrsg.) et al.: *Ladinia et Romania. Festschrift für Guntram Plangg zum 65. Geburtstag*. Vich/Vigo di Fassa: ICL 149-163. [Mondo Ladino, 21].

Il progetto “Zimbarbort” per il recupero del patrimonio linguistico cimbro

Luca Panieri

Some time ago, people living in the mountain territory between the rivers Adige and Brenta in northern Italy spoke a Germanic language usually known as ‘Cimbro.’ This language was brought into northern Italy by Bavarian colonists in the Middle Ages. Surrounded by Italian speakers, and isolated from the rest of the German-speaking world, Cimbro developed as an autonomous language, preserving many of its original old German features, but becoming strongly influenced by Italian lexis and syntax as well.

Since this language is nowadays commonly spoken only in Luserna (a village south of Trento), the local township has set up a project (presented in this paper) for the creation of a database of Cimbro lexis.

The main purpose of the project is to create a virtual memory of the Cimbrian language, where all known records of the Cimbrian language tradition can be stored. The first written records in Cimbro date back to around 1600, so the aim of the project is to give back to the Cimbrian language tradition its forgotten historical roots. We are sure that by looking into the deep historical layers of the language tradition, we will help the surviving Cimbrian community of Luserna to face the present.

Premessa

Con questo breve contributo si illustrano le linee guida di un progetto strategico finalizzato al recupero del patrimonio lessicale della tradizione linguistica cimbra. Tale progetto ha ottenuto l’approvazione del Comune di Luserna (l’isola linguistica cimbra più consistente), che ha erogato per l’anno in corso un primo finanziamento. Lo scrivente, membro del Comitato Scientifico dell’Istituto di Cultura Cimbra di Luserna è stato da esso designato Coordinatore del progetto.

L’Istituto di Cultura Cimbra, mediante la presentazione del progetto al Convegno Eurac “Lesser Used Languages and Computer” ha inteso soprattutto mettere a conoscenza gli esperti di linguistica computazionale dell’esistenza di tale iniziativa, illustrandone i contenuti, le finalità e la sua struttura operativa, allo scopo di sollecitare eventuali proposte sulle modalità tecniche della sua realizzazione. In tal senso, grazie all’occasione d’incontro con gli specialisti fornita dal Convegno di Bolzano, i promotori

del progetto sono effettivamente riusciti a suscitare vivo interesse e concrete proposte di collaborazione per la realizzazione della banca dati lessicale.

Si deve quindi premettere che il presente contributo non è che la trasposizione scritta della presentazione del progetto, inteso nei termini suddetti. Non si tratta quindi di un articolo specialistico di contenuto teorico o sperimentale, bensì della descrizione dell'iniziativa concreta che l'Istituto Cimbri intende promuovere per la salvaguardia del patrimonio lessicale della propria tradizione linguistica. Abbiamo demandato agli specialisti d'informatica il compito di indicarci le soluzioni tecnologiche più opportune alla sua realizzazione e gestione.

Quanto detto sul carattere di questo contributo spiega anche la mancanza quasi totale di riferimenti bibliografici, che sono tuttavia presenti in misura modesta nella sola introduzione, essendo essa finalizzata a portare a conoscenza del lettore la particolare realtà linguistica cimbra. Il resto della trattazione, invece, come già evidenziato, consiste nella semplice esposizione delle linee guida del progetto.

1. Introduzione

L'idea di questo progetto nasce dalla consapevolezza della situazione precaria in cui versano le tre isole linguistiche cimbre sopravvissute nei secoli fino ai giorni nostri: Giazza (VR), Roana-Mezzaselva (VI) e Luserna (TN). In particolare, la condizione relativamente rosea in cui fortunatamente ancora si trova la varietà cimbra di Luserna impone l'attuazione di ogni possibile strategia di difesa e consolidamento del patrimonio linguistico cimbro, essendo diventata Luserna l'ultima roccaforte di un gruppo etnico un tempo disseminato in tutto il territorio prealpino tra l'Adige e il Brenta.¹ Tale tradizione fu un tempo capace di trovare originale espressione letteraria e politico-amministrativa, in particolar modo sull'Altopiano d'Asiago, dove la Reggenza dei Sette Comuni riuscì a conservare la propria autonomia di governo locale per molti secoli, sopravvivendo all'avvicinarsi delle potenti signorie dell'Italia settentrionale e mantenendo una propria fisionomia linguistica e culturale anche nei confronti del vasto mondo di lingua tedesca, tanto geograficamente vicino.²

Ai nostri giorni, quando ormai l'area linguistica cimbra si è drasticamente ridotta, soppiantata quasi ovunque dal dialetto veneto o dalla lingua italiana, ed è rimasta vitale soltanto a Luserna, insorge la necessità di evitare che il patrimonio lessicale espresso dalla civiltà cimbra nel corso dei secoli cada per sempre nell'oblio. Non consideriamo

1 Tra i vari testi consultabili sulla questione dell'origine degli insediamenti "cimbrici" e sulla loro lingua rimane tuttora fondamentale lo studio del grande dialettologo bavarese Johann Andreas Schmeller (1985).

2 Per una sintesi efficace sulla storia istituzionale della comunità cimbra dei Sette Comuni dell'Altopiano d'Asiago, basata sulla documentazione, si veda anche Antonio Broglio (2000).

ciò solamente un'operazione dettata dal rispetto per la memoria storica di una civiltà, ma soprattutto un intervento preventivo di rilevante importanza strategica e finalizzato a salvaguardare la tradizione linguistica cimbra. Oggigiorno infatti la comunità di Luserna si trova in una situazione di bilinguismo nettamente sbilanciato, in cui la lingua italiana predomina come mezzo di comunicazione atto a esprimere il panorama concettuale astratto della cultura moderna, mentre il cimbro è soprattutto la lingua materna della sfera affettiva, quella che esprime con genuina spontaneità i moti dell'animo, il sentimento di appartenenza alla comunità e al suo territorio naturale. Per quanto questa ripartizione complementare dell'uso delle due lingue possa apparire accettabile, se non addirittura comoda, essa pone il cimbro in posizione debole nei confronti dell'italiano. I continui stimoli e cambiamenti socio-economici e culturali del mondo moderno e il loro influsso globalizzante scardinano la coesione tradizionale delle "piccole patrie" di un tempo e ne catapultano gli appartenenti in un contesto socio-culturale del tutto diverso e di più ampie dimensioni, il cui baricentro è al di fuori della stessa comunità che ne subisce l'influenza. Questo mondo si esprime soprattutto mediante le lingue nazionali della scolarizzazione di massa, come appunto l'italiano o il tedesco. La lingua cimbra rimane quindi legata e, purtroppo, confinata all'ambito delle relazioni socio-economiche e dei valori tradizionali della piccola comunità di un tempo. Ma con gli inevitabili e troppo repentini mutamenti di prospettiva dovuti alla modernizzazione, la lingua connaturata alla tradizione locale cede il passo a quella delle relazioni esterne, della cultura tecnologica, scientifica e amministrativa, sempre più preponderanti.

La sopravvivenza della tradizione linguistica cimbra dipende quindi dalla sua capacità di rinnovarsi ed espandere il proprio dominio espressivo agli ambiti concettuali tipici della cultura moderna.

2. Motivazioni strategiche e obiettivi

In considerazione di quanto sopra si è evidenziato, riteniamo necessario intervenire a tutela della lingua cimbra con un'operazione di consolidamento delle fondamenta storiche della stessa tradizione linguistica, mediante la realizzazione di una banca dati globale del patrimonio lessicale cimbro. In essa dovranno confluire i dati lessicali estrapolati da tutte le fonti scritte disponibili, a partire dalle prime attestazioni storiche di testi letterari quali il Catechismo cimbro del '600 fino ad arrivare alla lingua cimbra di oggi. L'idea di fondo è quella di creare una sorta di luogo virtuale della memoria linguistica collettiva della civiltà cimbra, che accolga il maggior numero di lemmi possibile, derivanti da tutte le varietà storiche del cimbro, oggi rappresentate dalle tre note isole linguistiche di Giazza, Roana-Mezzaselva e Luserna.

Oltre all'indubbio valore storico-documentario, tale operazione, sul piano strategico, consente di fornire alla lingua cimbra ancora in uso degli utili strumenti lessicologici per far fronte alla minaccia contingente di progressiva erosione del vocabolario originario. S'intende con ciò favorire il recupero delle risorse espressive della tradizione linguistica cimbra nel suo complesso, vedendo in essa il più valido punto di riferimento per consolidare la lingua di Luserna. Anche in relazione alla questione attuale della necessità di elaborare un lessico cimbro capace di esprimersi oltre l'ambito familiare e tradizionale, la sperimentazione di neologismi deve in prima istanza fare riferimento alla propria tradizione linguistica, sia pure intesa in senso lato, ancor prima che si faccia ricorso al modello italiano o tedesco. Entrambi sono da adottare solo se è accertata la mancanza di risorse linguistiche interne.

A tal riguardo si obietterà che attingere dal lessico storico cimbro per supplire alle deficienze semantiche della parlata attuale negli ambiti concettuali più astratti dell'espressione linguistica moderna potrebbe sembrare paradossale: come trovare nell'inventario lessicale del passato soluzioni adeguate a esprimere concetti che in molti casi non erano stati ancora immaginati da nessuno? Ad esempio, nell'ambito della tecnologia o in certi nuovi campi del sapere scientifico? Ovviamente non ci aspetteremo di "ritrovare" nel lessico storico cimbro la parola esatta per 'computer' o per 'ecologia', ma sicuramente non sarà difficile rendere tali concetti partendo dalle radici lessicali che per approssimazione semantica e/o per analogia strutturale meglio si prestano a descriverne il valore. Rimanendo negli esempi citati, considereremo il computer un 'calcolatore', perché tale è la sua funzione preminente, tale la sua prima denominazione italiana e tale il significato letterale del termine inglese preso in prestito. Si potrà proporre quindi di designarlo con un termine cimbro derivato dalla radice verbale tradizionale che indica il concetto di 'calcolare'. Per quanto riguarda il concetto di 'ecologia' occorrerà partire dalla sua possibile trasposizione in parole di uso comune che rendano chiaro il concetto, come 'scienza dell'ambiente', 'scienza della natura'. A questo punto avremo riportato il termine "moderno" negli ambiti concettuali già noti alla tradizione linguistica cimbra di 'sapere' e 'natura'. Ovviamente non si tratterà di imporre con l'autorità le soluzioni teoriche che si andranno proponendo, esse infatti si potranno realmente affermare nell'uso quotidiano solo se la comunità linguistica le avvertirà come utili alla comunicazione spontanea e in armonia con la percezione che ogni parlante nativo ha delle proprie radici linguistiche.

Certamente, rispetto alle più vaste comunità linguistiche nazionali, quella cimbra di Luserna, a fronte di tanti svantaggi, presenta almeno il vantaggio di una maggiore coesione tra le istituzioni e la cittadinanza. Di per sé ciò favorisce l'affermazione di

ogni iniziativa intrapresa dalle istituzioni locali, nelle quali il cittadino si rispecchia direttamente, in un clima di compartecipazione costruttiva. Ciò quindi gioca a favore anche degli interventi mirati di politica linguistica patrocinati dalle locali istituzioni.

3. Struttura operativa

La realizzazione della banca dati globale del lessico cimbro (progetto **Zimbarbort**) si articola essenzialmente nella fase di raccolta delle fonti primarie in lingua cimbra e nella fase di estrapolazione e inserimento dei singoli dati lessicali nel supporto informatico della banca dati stessa.

3.1 Raccolta delle fonti

In questa fase si procede al reperimento di ogni tipo di testimonianza linguistica del cimbro. Pur essendo questa fase logicamente preliminare rispetto a quella dell'extrapolazione e dell'inserimento dei dati nella banca virtuale, essa sarà destinata a protrarsi nel tempo fino all'esaurimento delle attestazioni storiche sulla lingua cimbra e continuerà seguendo a mano a mano gli eventuali sviluppi linguistici che si producono nel momento attuale. Poiché tale fase costituisce il momento di acquisizione alla "memoria virtuale collettiva" di ogni espressione lessicale integrata nella tradizione linguistica cimbra, essa sarà destinata ad arricchirsi progressivamente di ogni futuro neologismo che eventualmente si affermi nell'uso comune.

A prescindere dall'epoca a cui risalgono, le attestazioni della lingua cimbra si possono ripartire in due categorie, distinte dal diverso supporto in cui sono state registrate e tramandate ai giorni nostri:

- Fonti scritte

In quest'ambito rientra la moltitudine di attestazioni scritte in cimbro (interamente o parzialmente) nell'intero corso della storia, fino al tempo presente. Si tratta di testi scritti di tipologia e di epoca varia, che comprendono opere letterarie, quali poesie, racconti popolari o testi liturgici, scritti ad uso privato, quali le epistole, e opere finalizzate allo studio della lingua cimbra, quali grammatiche, glossari, studi toponomastici, ecc.

Ai fini del presente progetto si tratterà di individuare e raccogliere tutte le fonti scritte di cui si ha conoscenza per radunarle fisicamente in originale o almeno in copia fedele e inventariarle in modo ragionato, onde agevolarne la consultazione. Tra i criteri di catalogazione figureranno sicuramente il genere (poesia, grammatica, racconto popolare, epistola, ecc.) e il periodo storico.

- Fonti orali

Questo tipo di attestazioni comprendono tutte le registrazioni della voce viva dei parlanti nativi. Tali fonti sono della massima importanza per lo studio della fonologia e di tutti i fenomeni caratteristici del linguaggio parlato.

La disponibilità di questo genere di attestazioni si deve al progresso tecnologico avvenuto negli ultimi cento anni, in cui è andata progressivamente migliorando la qualità della riproduzione della voce viva, così come sono cambiati e si sono moltiplicati i supporti di registrazione (supporto radiofonico, magnetico, digitale, ecc.).

Anche in questo caso si tratterà di fare una ricognizione del materiale registrato esistente e di raccogliarlo in originale o in copia fedele. Esso sarà poi opportunamente inventariato con criteri che ne favoriscano la consultazione. In questo caso la tipologia delle attestazioni è però molto più omogenea, sia per epoca (durante l'ultimo secolo di storia) che per genere (per lo più interviste).

3.2 Estrapolazione e inserimento dei dati nella banca dati

Questa fase può avere inizio dal momento in cui un primo contingente di attestazioni, scritte e/o orali, sia stato raccolto e inventariato; a seguire la fase di raccolta e quella di estrapolazione e inserimento dei dati potranno proseguire anche in contemporanea.

Prima di dare avvio a questa fase è però indispensabile aver stabilito il formato in cui ogni dato sarà inserito nella banca dati virtuale. Con un termine tecnico chiameremo record ogni dato lessicale inserito con il suo corredo informativo (fonte di provenienza, significato in italiano, note grammaticali, fraseologiche, area semantica, riferimenti incrociati, ecc.).

- Scelta del formato e della struttura del record

Si dovrà porre particolare attenzione alla definizione preliminare dei parametri del corredo informativo che accompagnerà ogni dato inserito, poiché la scelta influenzerà la struttura globale della banca dati.

In linea di principio occorre tener presente il maggior numero possibile di informazioni attribuibile a un elemento lessicale. Dato che la rubricazione all'interno di ogni record assumerà, nel contesto informatico, la veste di 'campi', converrà attribuire a ogni categoria concettuale potenzialmente rilevante ai fini informativi un proprio campo. Il record esemplare sarà quello in cui tutti i campi informativi verranno compilati, ben sapendo che in numerosi casi non saranno disponibili tutti i dati. Se infatti, ad esempio, tra i parametri informativi accludiamo la trascrizione fonetica del dato lessicale, il campo destinato a questo parametro rimarrà certamente vuoto per tutte le voci del lessico cimbro risalenti a periodi storici molto antichi, non

essendo possibile stabilire con sufficiente sicurezza l'esatta pronuncia della lingua dell'epoca.

- Estrapolazione dei dati lessicali

L'operazione di acquisizione dei dati lessicali sarà più o meno complessa, a seconda della natura delle fonti esaminate. Ciò si rifletterà sul grado d'impegno lavorativo e sulle diverse competenze richieste allo svolgimento del compito.

Il caso più semplice è quello dello spoglio di un glossario, presentando già la fonte scritta di partenza i dati lessicali in forma di voci di entrata, con relativa traduzione e commento informativo. In questo caso l'inserimento dei dati lessicali nella banca dati può avvenire pressoché contemporaneamente alla loro estrapolazione dal testo in cui sono stati reperiti. Inoltre, sarà il testo stesso a fornirci importanti informazioni grammaticali e sul significato del lemma.

Ben più complessa sarà invece l'estrapolazione di dati lessicali derivanti da fonti orali registrate. Qui l'operazione sarà particolarmente difficile nel caso di registrazioni di qualità scadente e/o di provenienza dialettale diversa da Luserna. In questo caso il gruppo di lavoro dovrà cimentarsi nella comprensione di varianti del cimbro ormai vicine all'estinzione ed essere in grado di individuare, dal contesto di un discorso parlato, i singoli costituenti lessicali riconoscendone la loro reciproca relazione grammaticale. Gli operatori dovranno poi trasporre la propria interpretazione dei dati lessicali in forma scritta, operando una scelta ragionata sulla loro rappresentazione grafica, e da qui procedere al loro inserimento nella banca dati.

- Inserimento dei dati lessicali nella banca dati

L'operazione d'inserimento dei dati, come abbiamo già evidenziato, presuppone la creazione di un formato uniforme per tutti i record della banca dati. Per ogni dato lessicale (lemma) estrapolato sarà creato un record specifico all'interno del quale il dato sarà corredato di varie annotazioni informative ripartite nei rispettivi campi. L'operatore, inserito il lemma nel suo record, dovrà riempire i campi con le informazioni di cui dispone al momento, lasciando vuoti gli altri campi. Ad esempio, si potrebbe presentare il caso in cui l'operatore inserisca un lemma risalente a una fonte antica dal cui contesto non sia possibile risalire al genere grammaticale. In tale circostanza lascerà vuoto il campo relativo all'informazione grammaticale sul genere dei sostantivi.

Questa procedura lascia aperta la possibilità di successive revisioni dei record, finalizzate a integrare il corredo informativo dei lemmi ogniqualvolta emergano nuove informazioni sui medesimi. Per rimanere nell'esempio citato, può darsi il

caso che, successivamente, lo spoglio di altre fonti porti alla conoscenza del genere grammaticale di quello stesso lemma.

Naturalmente la continua acquisizione di fonti da sottoporre ad analisi porta spesso a estrapolare dati lessicali già noti da altre attestazioni precedentemente esaminate. La ricorrenza multipla di uno stesso lemma porta automaticamente alla revisione del record in cui è stato inizialmente inserito, aggiungendovi via via le nuove informazioni desunte dal contesto della fonte.

Oltre a questa revisione “automatica” in corso d’opera, è tuttavia raccomandabile affiancare all’operatore che al momento svolge il lavoro d’inserimento dei dati lessicali un revisore che controlli nell’immediato la compilazione dei record, poiché in molti casi il grado di completamento delle note informative sui lemmi dipende, oltre che dal contesto in cui sono stati reperiti, anche dalla competenza specialistica di chi svolge il compito.

Bibliografia

Broglia, A. (2000). *La proprietà collettiva nei Sette Comuni. Aspetti storico-normativi*. Roana: Istituto di Cultura Cimbra.

Schmeller, J.A. (1985). Über die sogenannten Cimbern der VII und XIII Communen auf den Venedischen Alpen und ihre Sprache, 1811, 1838, 1852, 1855†, Curatorium Cimbricum Bavarense, Landshut.

Stealth Learning with an Online Dog (Web-based Word Games for Welsh)

Gruffudd Prys and Ambrose Choy

This paper describes issues surrounding developing web-based word games in a minority language setting, and is based on experience gained from the development of a project designed to improve the language skills of fluent Welsh speakers undertaken at Canolfan Bedwyr at the University of Wales, Bangor.

This project was conceived by the BBC as an entertaining way of improving the language skills of fluent Welsh-speakers, especially those in the 18-40 age range. Funded by ELWa, the body responsible for post-16 education and training outside higher education in Wales, it was to form part of BBC Wales' "Learn Welsh" website.

The BBC's Welsh language web pages are immensely popular, attracting a high proportion of younger Welsh-speakers. A survey conducted by the BBC in April and May 2003 revealed that 43% of the BBC Welsh language online news service "Cymru'r Byd" belonged to the 15-34 age group, with a high level of workplace usage, peaking at lunchtimes. The project was to provide this audience with word games, a self marking set of language improvement exercises, and an online answering service dealing with grammatical and other language problems. In order to appeal to the target audience, it was important that they be entertaining and attractive in addition to being educational. It was also intended that the project should emphasise progressive youth culture rather than old-fashioned Celtic themes, and this would be incorporated into the design and feel of the games.

This paper will concentrate specifically on the development of the interactive online games and puzzles, showing how digital language resources originally created for previous digital language projects were adapted and recycled, allowing the e-Welsh team at the University of Wales, Bangor, to produce a working website within a few short months. It will also detail some of the new innovations created as part of the project, with a view of building a modularized set of components that will provide a versatile resource bank for future projects.

1. The leithgi Name

Welsh has a peculiar word for people intensely interested in language. It is *ieithgi*, the literal translation of which would be 'language dog.' Perhaps 'language terrier' would be a meaningful image for English speakers, as it denotes someone, who, having got hold of a particularly tasty bone to gnaw, is unwilling to let it go. It may be a question of some obscure Welsh grammar rule, or the origin of some Welsh place-name, but the *ieithgi* will not let the subject drop without knowing the answer.

By coincidence, a project aimed at Welsh learners was using an animated dog, called Cumberland, and his owner, Colin, to introduce Welsh to new audiences. In the *Colin and Cumberland* storyline, Colin has no Welsh, whereas his dog Cumberland, is a fluent, knowledgeable and slightly pompous Welsh speaker. As *Colin and Cumberland* was aimed at the same demographic age group as the *leithgi* project, and possessed a design that was modern, contemporary and attractive, it was therefore a short step for Cumberland, the know-all dog in the animated cartoons, to become the namesake and mascot of the *leithgi* project, on hand to answer questions on Welsh grammar as well as guide users through the games and exercises.

2. Macromedia Flash and XML

The brief received from the BBC specified that the games were to be created using Macromedia Flash. Flash is a multimedia authoring program that creates files that can be played on any computer, Mac or PC, where Flash Player is installed (Macromedia claim a coverage of 98% of all desktops worldwide).

Flash can combine vector and raster graphics, and uses a native scripting language called actionscript which is similar to Java. It can communicate with external XML files and databases, and, when used intelligently, produces small files which are quick to download. Flash also allows easy collaboration between a software engineer and a designer.

Figure 1: Colin and Cumberland - The BBC Cartoon for Learners



3. Technical Challenges

The main technical challenge posed by the games was the need to adapt game formulas already existing in English to work with the characteristics of the Welsh language. This meant that new code specific to the needs of Welsh had to be created. The lack of ready-made Welsh language components available to form the building blocks needed to create the word games was a significant disadvantage when compared with developers creating similar games in a major language. These building blocks for Welsh had to be created as part of the project.

In order to keep down costs, the project hoped to reuse resources developed originally for previous digital language projects undertaken by Canolfan Bedwyr. This is one way that a minority language such as Welsh can keep costs down and make frugal use of existing components in an attempt to keep pace with greater resourced languages.

4. Resource Audit

Over the years, as part of its mission to address the needs of Welsh language in a digital environment, Canolfan Bedwyr has built up a library of language resources, including digital dictionaries, spelling and grammar checkers as well as the assorted components such as lexicons and lemmatizers that combine to create such tools. Many of these resources are either useful or essential when attempting to create games

such as leithgi; although seemingly quite different, digital dictionaries share many prerequisites with word games.

As the leithgi project was a low budget, tight deadline project, it was imperative that we make as much use as possible of our existing resources, as opposed to reinventing the wheel. However, we also recognised that new tools for manipulating the Welsh language would also have to be forged in order for some aspects of Welsh to function properly in a digital online setting.

Below is a list of the relevant resources available to Canolfan Bedwyr and the games in which they would be used:

- *Lexicon*: To be used in Cybolfa (conundrum) and Dic Penderyn (hangman)
- *Place-name databases (AMR and Enwau Cymru)*: To be used in Rhoi Cymru yn ei lle (locate and identify place-names);
- *Proverb database*: To be used in Diarhebol (guess the proverb);
- *Alphabet order sorter*: To be used in Cybolfa, Diarhebol, Dic Penderyn, Pos croeseiriau (crossword) and Ystyrlon (identify the correct meaning).

5. The Games

Six games were to be produced for the leithgi project. Of these six, three were to be open-ended games. These games draw randomly from a large list of words or phrases each time the game is played, giving the user a fresh challenge every time they start a new game, and ensuring that the games have enormous replay value. Each instance of a closed game, on the other hand, must be created manually by a games designer, and this means in practice that there are fewer unique instances of closed games than of open games. However, conversely, the content of closed games can be more complex, as they do not need to be designed to conform to such tight technical constraints.

Below is a list of the games divided by category:

- Open Ended
 - Dic Penderyn (hangman)
 - Cybolfa (conundrums)
 - Diarhebol (guess the proverb)
-

-
- Closed
 - Pos croeseiriau (crosswords)
 - Rhoi Cymru yn ei Lle (locate and identify place-names)
 - Ystyrlon (identify the correct meaning)

6. Open Ended Games

From a technical point of view, the open-ended games posed the greatest challenge. Cybolfa, Dic Penderyn and Diarhebol all make use of XML word lists that are used to supply the games with random words or phrases that test the player's language skills.

6.1 Dic Penderyn

Dic Penderyn, named after a Welsh folk hero, is our version of the popular Hangman game. Drawing a word at random from an XML file, Dic Penderyn gives the person playing the game ten attempts to guess the word before a set of gallows are built and a caricature of Colin, Cumberland's owner, is hung, signalling 'Game Over.'

From an educational point of view, Dic Penderyn nurtures spelling ability by having the player think in terms of the letter patterns present in the language in order to correctly identify the game word. The game also increases the player's vocabulary by sometimes suggesting unfamiliar words (as the word list contains words of varying degrees of familiarity).

The XML wordlist was drawn from the lexicon compiled for the BBC's Learn Welsh dictionary, which had been created previously for the BBC by Canolfan Bedwyr. This had the bonus of making it possible to link each of the words in the wordlist to a definition on the dictionary's Webpage. The link would appear each time the player failed to identify the word, increasing the educational value of the game by providing definitions of words that had proved unfamiliar.

The lexicon itself included words taken from Corpws Electroneg o Gymraeg (CEG), the tagged 1 million word Welsh language corpus developed at the University of Wales Bangor in the early nineties.

Having a part-of-speech tagged lexicon proved extremely useful as it enables a game designer to tweak the content of the word list that would be created from it. After some initial playtesting, it was decided that conjugated verbs would be excluded from the wordlist. These are sometimes included in English versions of Hangman, as English has limited conjugation possibilities. In Welsh, however, as in Romance

languages, most verbs follow a regular pattern of conjugation, with separate but regular conjugations for the different persons as well as tenses.

Table 1: Conjugation of rhedeg (to run)

Present	Imperfect	Past	Pluperfect	Subjunctive	Imperfect Subjunctive	Imperative
rhedaf	rhedwn	rhedais	rhedaswn	rhedwyf	rhedwn	rhed, rheda
rhedi	rhedit	rhedaist	rhedasit	rhedych	rhedit	rheded
rhed, rheda	rhedai	rhedodd	rhedasai	rhedo	rhedai	rhedwn
rhedwn	rhedem	rhedasom	rhedasem	rhedom	rhedem	rhedwch
rhedwch	rhedech	rhedasoch	rhedasech	rhedoch	rhedech	rhedent
rhedant	rhedent	rhedasant	rhedasent	rhedont	rhedent	rheder
rhedir	rhedid	rhedwyd	rhedasid	rheder	rhedid	

As is apparent from Table 1, many of the verb forms above are far too similar for a player to differentiate between them in a game of hangman. Coupled with the fact that conjugated verbs seem unfamiliar outside the context of a sentence, this meant that their inclusion would have made the game too difficult and unrewarding from a playability perspective.

6.1.1 Mutation

The lemmatizer also allowed us to prevent the initial consonant mutation that is a feature of Welsh words within sentences from making its way from sentences in the corpus to words in the word list.

A word such as *ci* (dog) can have the following mutations:

ci **nghi** **gi** **chi**

For example:

Fy **nghi**

Dy **gi**

Ei **chi**

Eu **ci**

My **dog**

Your **dog**

Her **dog**

Their **dog**

Mutations never occur in words when they appear in isolation (as they do in Dic Penderyn). It was therefore inappropriate to have them included in the XML word list.

6.1.2 Dic Penderyn XML Word List Example

Sample taken from list of 3,000+ six letter words

```
<Rhestr_dicpenderyn>  
<Gair>gormes</Gair>  
</Rhestr_dicpenderyn>  
<Rhestr_dicpenderyn>  
<Gair>gormod</Gair>  
</Rhestr_dicpenderyn>  
<Rhestr_dicpenderyn>  
<Gair>goroer</Gair>  
</Rhestr_dicpenderyn>
```

6.1.3 Digraphs

The Welsh alphabet contains digraphs:

ch dd ng ll ph rh

These digraphs count as single letters rather than a combination of two separate letters. This means that Welsh, unlike most other languages that use the Roman alphabet, has two-character letters in addition to single-character letters.

Take for example the word llefrith (milk), which has eight characters:

L, L, E, F, R, I, T, H

But six letters:

LL, E, F, R, I, TH

6.1.4 Digraph Problems

The existence of digraphs in Welsh creates a number of problems:

- Simple character count functions can't be used to count letters.

Due to the existence of digraphs, a function that simply counts the number of characters in a word will not be able to accurately count the number of letters in a Welsh word. Using a simple character count function to create the XML six letter word word list would not have included words such as llefrith that have six letters but have over six characters, and would erroneously include words with less than six letters

but that possessed six characters. In order to be able correctly count the letters, a digraph filter was created.

Every word in the lexicon had to be passed through the filter. The filter identifies the characters that form Welsh digraphs (ch, dd, ng, ll, ph, rh) and treats them as single characters. The amount of letters in a word can then be counted correctly, so that only six letter words are added to our XML list of six letter words, whether they contain digraphs or not.

Here is an example of the code used:

```
public static int welshCharSplit(string word, ArrayList charArray)
{
    int letterCount=0,x=0;

    charArray.Clear();
    word = word.ToUpper();
    string digraff = String.Empty;

    for (x=0; x<word.Length; x++)
    {
        digraff = String.Empty;
        if (x!=0)
        {
            // Check for Rh,Th,Ph,Ch
            if (word[x] == 'H')
            {
                if ((word[x-1] == 'R') || (word[x-1] == 'T') ||
                    (word[x-1] == 'P') || (word[x-1] == 'C'))
                {
                    digraff = word[x-1].ToString() + word[x].ToString();
                }
            }
            //Check for Ng
            if (word[x] == 'G')
```

```

    {
        if (word[x-1] == 'N')
        {
            digraff = word[x-1].ToString() + word[x].ToString();
        }
    }
    //Check for Dd, Ff, Ll
    if ((word[x] == 'D') || (word[x] == 'F') || (word[x] == 'L'))
    {
        if (word[x-1] == word[x])
        {
            digraff = word[x-1].ToString() + word[x].ToString();
        }
    }
}

string buff = String.Empty;
if (digraff != String.Empty)
{
    charArray.RemoveAt((charArray.Count)-1);
    buff = digraff;
}
else
{
    buff = word[x].ToString();
}

charArray.Add(buff);
}
letterCount = charArray.Count;
return letterCount;
}

```

This creates a word list containing six-letter words, including digraphs.

- Some combinations of characters can be a digraph or two separate letters.

Bangor = (compound word of *ban* + *côr*) pronounced 'n-g'

angor (meaning *anchor*) pronounced 'ng'

Fortunately, digraph look-up lists had previously been developed at Canolfan Bedwyr in order to correctly sort dictionaries according to the Welsh alphabet (ng follows g in the Welsh alphabet, so ng and n are sorted quite differently). These look-up lists could then be used to prevent confusion between digraphs and similar character combinations.

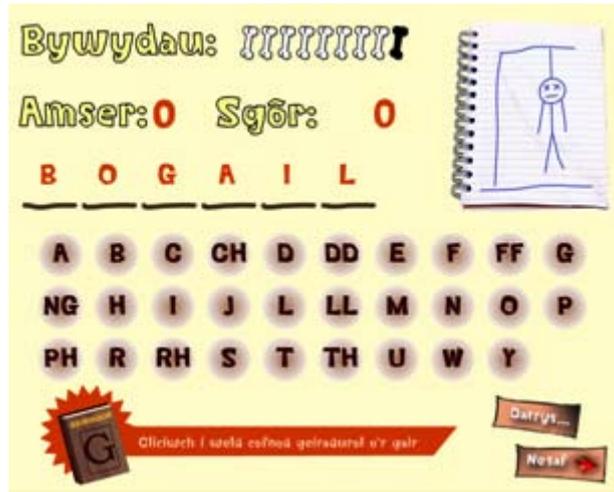
- Inputting letters using the keyboard becomes more complicated.

Designing an online interface that can differentiate elegantly between an inputted d and an inputted dd is a challenge. There is no support for digraphs in Welsh's UTF-8 character set, and no specific Welsh keyboard that has digraph keys (UK English QWERTY keyboards are generally used). In practice this makes a keyboard-based approach to inputting awkward, especially when playing against the clock as in many of the leithgi games.

It was decided that a visual interface would be devised in order to allow the user to input these characters quickly and efficiently. This came in the form of an on-screen keyboard featuring all the letters of the Welsh alphabet. Although the keyboard takes up some of the game's screen space, it gives the player valuable feedback such as which letters have been chosen and which letters remain, as well as serving as a visual reminder to users more familiar with the English alphabet that Welsh considers digraphs to be single letters.

The on-screen keyboard is used in Cybolfa, Diarhebol and Pos Croeseiriau in addition to Dic Penderyn, shown below:

Figure 2: Screenshot showing Dic Penderyn after an Unsuccessful Attempt.



6.2 Cybolfa

The digraphs pose another problem when generating the Welsh words for Cybolfa, a game where the player must attempt to create words from a jumbled set of letters. Cybolfa uses the same six-letter XML word list as Dic Penderyn to supply the main six-letter game word. However, Cybolfa must then scramble the word so that it is difficult for the player to recognise. In English, this could be done fairly quickly by scrambling each individual character in a word. The same method can not be applied to the Welsh words because of the existence of digraphs, therefore the word must be passed through a filter that identifies any Welsh alphabetical letters in the word before scrambling it. To return to the earlier example of *llefrith*, an actionscript digraph filter within the Flash file identifies the digraphs as distinct letters. so that all six letters can be identified (LL, E, F, R, I, TH).

Below shows the function called `WelshFilter` written in actionscript 2.0 in Flash MX, it receives the word in the form of an array and checks for the existence of any digraphs and returns the length of the word. If a digraph is found, it merges the letters to become a single element within the array.

```
_global.welshFilter = function(WordArray) {  
    for (x=1; x<=WordArray.length; x++)  
    { // Check for Rh,Th,Ph,Ch  
        if(WordArray[x] == "H")
```

```

    { if ((WordArray[x-1] == "R") || (WordArray[x-1] == "T") ||
      (WordArray[x-1] == "P") || (WordArray[x-1] == "C"))
      { WordArray[x-1] = WordArray[x-1]+ WordArray[x];
        WordArray.splice(x,1);
      }
    }
  //Check for Ng
  if (WordArray[x] == "G")
    { if (WordArray[x-1] == "N")
      { WordArray[x-1] = WordArray[x-1]+ WordArray[x];
        WordArray.splice(x,1);
      }
    }
  //Check for Dd, Ff, Ll
  if ((WordArray[x] == "D") || (WordArray[x] == "F") || (WordArray[x] == "L"))
    { if (WordArray[x-1] == WordArray[x])
      { WordArray[x-1] = WordArray[x-1]+ WordArray[x];
        WordArray.splice(x,1);
      }
    }
  }
}
return(WordArray.length);
}

```

Once both digraphs and single-character letters have been identified as single elements, the word can be scrambled and displayed to the player in an unfamiliar letter order whilst still retaining the digraph integrity (TH, F, R, LL, I, E).

6.3 Anagram Maker

As described previously, the word list for the Cybolfa games is derived from the Dic Penderyn word list. Each time Cybolfa is played, a random six-letter word is drawn from the list and an anagram maker within the actionsript code generates a list of all possible anagrams for that word. This is achieved by cross-referencing Canolfan Bedwyr's Welsh spellchecker list with the original word's possible letter combinations.

In programming terms, this is done by a one-to-one mapping of letter values to prime numbers, allowing words to be represented as composite numbers by multiplying together the primes that map each letter in the word. Words formed from the same letters, regardless of order, will then map to the same composite number. Therefore, if a word's number divides exactly into another word's, the first word's letters must all appear in the second word.

For example, take the word *gwelwi*.

```
<Dicpenderyn>
  <Geiriau>gwelwi</Geiriau>
</Dicpenderyn>
```

By looking up the spellchecker list and using the anagram checker function, the following list is generated in XML:

```
<root>
  <Gair>
    <Anagram>GWELWI</Anagram>
    <Anagram>GLIW</Anagram>
    <Anagram>ELI</Anagram>
    <Anagram>ELW</Anagram>
    <Anagram>EWIG</Anagram>
    <Anagram>GWELW</Anagram>
    <Anagram>GLEW</Anagram>
    <Anagram>GWIWI</Anagram>
    <Anagram>IGLW</Anagram>
  </Gair>
</root>
```

This list then is passed through and used as one of the games in Cybolfa.

Below is a screenshot of a completed game where *polisi* was the six-letter game-word.

Figure 3: Screenshot of Cybolfa Showing all Possible Anagrams.



6.4 Diarhebol

Diarhebol is in essence very similar to Dic Penderyn, the main difference being that rather than guessing a random six-letter word, the player must attempt to guess a Welsh proverb. Once again, players have a limited number of chances to achieve their objective before the game ends. If needed, a clue is provided in the form of an English translation of the proverb, and, whilst guessing a whole proverb may at first seem daunting, the higher probability of a sentence as opposed to a word containing a specific letter ensures that the game is of a similar level of difficulty.

An XML proverb list replaces the XML word list used by both Dic Penderyn and Cybolfa, and an example is shown below.

```
<Diarhebion>
<Dihareb>Yr afal mwyaf yw'r pydraff ei galon</Dihareb>
<Esboniad>The biggest apple has the rottenest heart</Esboniad>
</Diarhebion>
<Diarhebion>
<Dihareb>Yr euog a ffy heb neb yn ei erlid</Dihareb>
<Esboniad>The guilty flees when no-one chases him</Esboniad>
</Diarhebion>
<Diarhebion>
<Dihareb>Yr hen a wyr, yr ieuanc a dybia</Dihareb>
```

<Esboniad>The old know, the young suppose</Esboniad>
 </Diarhebion>
 <Diarhebion>
 <Dihareb>A fo'n ddigwilydd a fo'n ddigolled</Dihareb>
 <Esboniad>The shameless will be without loss</Esboniad>
 </Diarhebion>

Figure 4: Screenshot of Successful Attempt at Diarhebol



7. Closed Games

Unlike the open-ended games, which draw their content from a list, the content for closed games must be manually created in advance due to the more involved nature of their content.

7.1 Pos Croeseiriau

Pos Croeseiriau is an online Welsh crossword puzzle. Crossword puzzles have been popular for some time in Welsh language publications such as local papers, where the custom of representing digraphs as a single letter within a single square has long been established. Due to the complexity of creating crosswords, both the clues and the answers have been hardcoded into the code. However, *Cysgeir*, Canolfan Bedwyr's electronic dictionary can be used to aid in the creation of crosswords, as it can suggest words that contain specific letters in specific positions within the word. Take for instance a situation where the crossword designer has decided on the two

words *cyfarth* and *cynffon* to form the answers for 1 and 2 across, and needs a word that will fit in the space for 1 down:

Figure 5: TITLE

		1				
¹ C	Y	F	A	R	TH	
² C	Y	N	FF	O	N	

The designer has to only type in ??F??N??? into *Cysgeir* (where '?' represents an empty square) to be provided with a list of compatible words. In this case *Cysgeir* provides 32 different words that fulfil our requirements, of which we choose *elfennol* 'elementary.'

Figure 6: title

		¹ E				
		L				
¹ C	Y	F	A	R	TH	
		E				
		N				
² C	Y	N	FF	O	N	
		O				
		L				

Pos Croeseiriau used a slightly modified version of the on-screen keyboard found in the open-ended games to give players the ability to delete unwanted letters, and the resulting interface is simple and easy to use despite the complications caused by the Welsh digraphs.

Figure 7: Pos Croeseiriau Screenshot Showing a Completed Game.



7.2 Rhoi Cymru yn ei Lle

Rhoi Cymru yn ei Lle was designed as a game that would educate people as to the geographical location of Welsh place-names. Players must attempt to drag a place-name to its correct position on a map, with themed clues relevant to each place providing some assistance. There are various themes, including sport, religion, culture, and history, so that the player learns a little about different aspects of their country as they play, and gain satisfaction from being able to locate an unfamiliar place on a map.

When creating the content, *Cronfa Archif Melville Richards* and *Enwau Cymru* (developed by Canolfan Bedwyr) were invaluable in aiding in the identification and placement of place-names and their associated clues. *Cronfa Archif Melville Richards* is a fully searchable online database of historic Welsh place-name forms that contains location information and grid references, whilst *Enwau Cymru* is an online database of modern Welsh place-names dealing in particular with bilingual place-names and again giving location information. As with *Pos Croeseiriau*, due to its complexity, the game content is coded into the game itself.

Figure 8: Ystyrlon Screenshot Showing a Game in Progress



7.3 Ystyrlon

Ystyrlon is similar to the popular game *Call my Bluff* in that the player is given an uncommon word (that is hopefully unfamiliar to him), and is then asked to guess the correct definition from a choice of three. From a technical viewpoint, this is a very simple game, the hard work being the creation of original content, choosing the unfamiliar words, and creating humorous and misleading definitions that will entertain those who play the game and keep them on their toes.

As the content, once created, is quite simple, it is stored as an XML file that is then referenced by the Flash game file. This aids the production of new games, as it enables the creation of new content without having to use or understand the Flash programming application.

Figure 7: Screenshot of Ystyrion Following an Incorrect Guess



8. Results

Usually, academic establishments do not undertake commercial projects such as leithgi, concentrating on research that can then be exploited and taken forward by the private sector. However, in a minority language situation, the technical expertise and experience needed to create such language-specific products may not exist in the private sector, or the financial returns may not be high enough to justify the investment of time and money. In such a situation, centres such as Canolfan Bedwyr that see their goal as catering to the needs of a modern, living minority language, must sometimes fulfil both roles if the language is ever to see such products.

The successful realisation of such a product has been one positive result of this venture.

The leithgi project has also led to the creation of new digital resources, including a Welsh anagram maker and digraph filter, as well as a process for integrating resources through XML into Flash; these add to and enhance the resources available to Canolfan Bedwyr for future projects.

The need to repackaging existing digital resources to facilitate further reuse as part of future projects has also been identified, leading to a new programme of modularization of lexical components for future projects.

A sure sign of a successful product is one that results in further commissions, and the success of leithgi has resulted in a further commission to develop a similar set

of stealth educational Welsh language online games targeted at adults with below average literacy.

It is hoped by Canolfan Bedwyr that the leithgi project will serve as an example of how to make a little go a long way, and that building up language resources and corpora can benefit a minority language in more ways than by producing dictionaries and spellcheckers, allowing existing resources to stretch further.

References

"Archif Melville Richards Historical Place-name Database." Online at <http://www.bangor.ac.uk/amr>.

"Canolfan Bedwyr Website." Online at <http://www.bangor.ac.uk/ar/cb/>.

"Colin and Cumberland Website." Online at <http://www.bbc.co.uk/colinandcumberland/>.

Corpws Electronig o Gymraeg (CEG). "A 1 Million Word Lexical Database and Frequency Count for Welsh."

http://www.bangor.ac.uk/ar/cb/ceg/ceg_eng.html.

"Cysgeir Electronic Dictionary Information Website." Online at http://www.bangor.ac.uk/ar/cb/meddalwedd_cysgair.php.

Davies, G. (2005). "Beginnings, New Media and the Welsh Language." *North American Journal of Welsh Studies*, 5(1).

"Enwau Cymru Modern Place-name Database." Online at <http://www.e-gymraeg.org/enwaucymru>.

Hicks, W.J. (2004). "Welsh Proofing Tools: Making a Little NLP go a Long Way." *Proceedings of the 1st Workshop on International Proofing Tools and Language Technologies*. Greece: University of Patras.

"Learn Welsh-The BBC's Website for Welsh Learners." Online at <http://www.bbc.co.uk/wales/learnwelsh/>.

Prys, D. & Morgan, M. (2000). "E-Celtic Language Tools." *The Information Age, Celtic*

Languages and the New Millenium. Ireland: University of Limerick.

“The leithgi Website.” Online at <http://www.bbc.co.uk/cymru/leithgi/>.

Alphabetical list of authors & titles with keywords

Victoria Arranz (& Elisabet Comelles, David Farwell)

Speech-to-Speech Translation for Catalan

Keywords: Catalan, Multilingualism, Speech-to-Speech Translation, Interlingua.

Ermenegildo Bidese (& Cecilia Poletto, Alessandra Tomaselli)

The relevance of lesser used languages for theoretical linguistics: the case of Cimbrian and the support of the TITUS corpus

Keywords: Cimbrian, clitics, Wackernagelposition, Agreement, TITUS.

Evelyn Bortolotti (& Sabrina Rasom)

Il ladino fra polinomia e standardizzazione: l'apporto della linguistica computazionale

Keywords: lessicografia, terminologia, corpus testuale, correttore ortografico, strumenti per la standardizzazione.

Sonja E. Bosch (& Elsabé Taljard)

A Comparison of Approaches towards Word Class Tagging: Disjunctively vs Conjunctively Written Bantu Languages

Keywords: word class tagging, Bantu languages, disjunctive writing system, conjunctive writing system, morphological analyser, disambiguation rules, tagsets.

Ambrose Choy (& Gruffudd Prys)

Stealth Learning with an on-line dog Keywords: Web-based Word Games for Welsh

Keywords: Stealth learning, Welsh, on-line games.

Elisabet Comelles (& Victoria Arranz, David Farwell)

Speech-to-Speech Translation for Catalan

Keywords: Catalan, Multilingualism, Speech-to-Speech Translation, Interlingua.

David Farwell (& Elisabet Comelles, Victoria Arranz)

Speech-to-Speech Translation for Catalan

Keywords: Catalan, Multilingualism, Speech-to-Speech Translation, Interlingua.

Olya Gurevich

Computing Non-Concatenative Morphology: the Case of Georgian

Keywords: computational linguistics, morphology, Georgian, non-concatenative, construction grammar.

Ulrich Heid (& Danie Prinsloo)

Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping

Keywords: POS-tagger, Bantu-languages, Taggerlexicon, Tagging reference cp.

Dewi Jones (& Delyth Prys)

The Welsh National On-line Database

Keywords: terminology standardization, Welsh, termbases, terminology markup framework.

Cecilia Poletto (& Ermenegildo Bidese, Alessandra Tomaselli)

The relevance of lesser used languages for theoretical linguistics: the case of Cimbrian and the support of the TITUS corpus

Keywords: Cimbrian, clitics, Wackernagelposition, Agreement, TITUS.

Danie Prinsloo (& Ulrich Heid)

Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping

Keywords: POS-tagger, Bantu-languages, taggerlexicon, tagging reference cp.

Luca Panieri

Il progetto “Zimbarbort” per il recupero del patrimonio linguistico cimbro

Keywords: cimbro, lessico, patrimonio linguistico.

Delyth Prys (& Dewi Jones)

The Welsh National On-line Database

Keywords: terminology standardization, Welsh, termbases, terminology markup framework.

Gruffudd Prys (& Ambrose Choy)

Stealth Learning with an on-line dog Keywords: Web-based Word Games for Welsh

Keywords: Stealth learning, Welsh, on-line games.

Nicoletta Puddu

Un corpus per il sardo: problemi e prospettive

Keywords: corpus planning, corpus design, sardinian, non-standardized languages, XML.

Sabrina Rasom (& Evelyn Bortolotti)

Il ladino fra polinomia e standardizzazione: l’apporto della linguistica computazionale

Keywords: Lessicografia, terminologia, corpus testuale, correttore ortografico, strumenti per la standardizzazione.

Soufiane Rouissi (& Ana Stulic)

Annotation of Documents for Eletronic Edition of Judeo-Spanish Texts: Problems and Solutions

Keywords: electronic corpus, Judeo-Spanish, collaborative production, digital document.

Clau Solèr

Spracherneuerung im Rätoromanischen: Linguistische, soziale und politische Aspekte

Oliver Streiter

Implementing NLP-Projects for Small Languages: Instructions for Funding Bodies, Strategies for Developers

Oliver Streiter (& Mathias Stuflesser)

XNLRDF, A Framework for the Description of Natural Language Resources. A proposal and first implementation

Keywords: XNLRDF, metadata, writing system, Unicode, encoding.

Mathias Stuflesser (& Oliver Streiter)

XNLRDF, A Framework for the Description of Natural Language Resources. A proposal and first implementation

Keywords: XNLRDF, metadata, writing system, Unicode, encoding.

Ana Stulic (& oufiane Rouissi)

Annotation of Documents for Eletronic Edition of Judeo-Spanish Texts: Problems and Solutions

Keywords: electronic corpus, Judeo-Spanish, collaborative production, digital document.

Elsabé Taljard (& Sonja E. Bosch)

A Comparison of Approaches towards Word Class Tagging: Disjunctively vs Conjunctively Written Bantu Languages

Keywords: word class tagging, Bantu languages, disjunctive writing system, conjunctive writing system, morphological analyser, disambiguation rules, tagsets.

Alessandra Tomaselli (& Ermenegildo Bidese, Cecilia Poletto)

The relevance of lesser used languages for theoretical linguistics: the case of Cimbrian and the support of the TITUS corpus

Keywords: Cimbrian, clitics, Wackernagelposition, Agreement, TITUS.

Trond Trosterud

Grammar-based language technology for the Sámi languages

Keywords: Sámi, transducers, disambiguation, language technology, minority languages.

Chinedu Uchechukwu

The Igbo Language and Computer Linguistics: Problems and Prospects

Keywords: language technology, lexicography, computer linguistics, linguistic tools.

Ivan Uemlianin

SpeechCluster: a speech database builder's multitool

Keywords: annotation, speech data, Welsh, Irish, open-source.

Alphabetical list of contributors & contact addresses

Victoria Arranz

ELDA-Evaluation and Language
Resources Distribution Agency
arranz@elda.org

Evelyn Bortolotti

Istitut Cultural Ladin “majon di
fascegn”
rep.ling@istladin.net

Ambrose Choy

Canolfan Bedwyr
Univeristy of Wales
a.choy@bangor.ac.uk

David Farwell

Institució Catalana de Reserca i Estudis
Avançats TALP-Centre de Tecnologies i
Aplicacions del Llenguatge i la Parla
Universitat Politècnica de Catalunya
farwell@lsi.upc.edu

Ulrich Heid

IMS-CL, Institut für maschinelle
Sprachverarbeitung
Univerität Stuttgart
uli@ims.uni-stuttgart.de

Luca Panieri

Istituto Cimbri di Luserna
luca.panieri@fastwebnet.it

Danie Prinsloo

Department of African Languages
University of Pretoria
danie.prinsloo@up.ac.za

Ermenegildo Bidese

Università di Verona/ Philosophisch-
Theologische Hochschule Brixen
ebidese@lingua.univr.it

Sonja E. Bosch

University of South Africa
boschse@unisa.ac.za

Elisabet Comelles

TALP-Centre de Tecnologies i
Aplicacions del Llenguatge i la Parla
Universitat Politècnica de Catalunya
comelles@lsi.upc.edu

Olya Gurevich

UC Berkeley
olya@berkeley.edu

Dewi Jones

Language Technologies
Canolfan Bedwyr
University of Wales, Bangor
d.b.jones@bangor.ac.uk

Cecilia Poletto

Padova-CNR
cecilia.poletto@unipd.it

Delyth Prys

Canolfan Bedwyr
Univeristy of Wales
d.prys@bangor.ac.uk

Gruffudd Prys
Language Technologies
Canolfan Bedwyr
University of Wales, Bangor
g.prys@bangor.ac.uk

Sabrina Rasom
Istitut Cultural Ladin “majon di
fascegn” (ICL)
lengaz@istladin.net

Clau Soler
Universität Genf
clau.soler@bluewin.ch

Ana Stulic
University of Bordeaux 3 AMERIBER
etchevers@tele2.fr

Elsabé Taljard
University of Pretoria
elsabe.taljard@up.ac.za

Trond Trosterud
Universitetet i Tromsø
trond.trosterud@hum.uit.no

Ivan Uemlianin
Language Technologies
Canolfan Bedwyr
University of Wales, Bangor
i.uemliani@bangor.ac.uk

Nicoletta Puddu
University of Pavia
attel76@hotmail.com

Soufiane Rouissi
University of Bordeaux 3 CEMIC-GRESIC
Soufiane.Rouissi@u-bordeaux3.fr

Oliver Streiter
National University of Kaohsiung
ostreiter@nuk.edu.tw

Mathias Stuflesser
European Academy of Bolzano
mstuflesser@eurac.edu

Alessandra Tomaselli
Università di Verona
alessandra.tomaselli@univr.it

Chinedu Uchechukwu
Universität Bramberg, Germany
neduchi@netscape.net