Lesser Used Languages & Computer Linguistics (LULCL) II

"Combining efforts to foster computational support of minority languages"

European Academy Bozen/Bolzano, Italy
13th - 14th November 2008

The colloquium LULCL II aims at providing an overview of ongoing research activities on computational approaches for lesser used, lesser standardised and lesser resourced languages and strengthening the research community and its practices.

This year's LULCL colloquium puts a special focus on bringing together efforts from several related research communities, in order to join best practices, approaches and techniques and to add value to individual initiatives. In addition to lesser used languages, other types of language, including language varieties, sign languages, learner language and spoken language, pose similar issues for researchers, having to do with sparse resources, little standardisation, and challenges with automatic processing and building up of computational resources.

The colloquium will provide an opportunity to learn what tasks are analogous and shared among the different research communities, what practices and resources could be exchanged and generally how the groups could gain from working together and how they can bring forward lesser used languages.

Keynote speakers:

Karin Aijmer (University of Gothenburg, Sweden)
Dafydd Gibbon (Bielefeld University, Germany)

Scientific committee:

Andrea Abel (European Academy Bozen/Bolzano, Italy)
Stefanie Anstein (European Academy Bozen/Bolzano, Italy)
Christopher Culy (European Academy Bozen/Bolzano, Italy)
Dafydd Gibbon (Bielefeld University, Germany)
Christer Laurén (Vaasa University, Finland)
Marcello Soffritti (University of Bologna / European Academy Bozen/Bolzano, Italy)
Chiara Vettori (European Academy Bozen/Bolzano, Italy)
Paul Videsott (Free University of Bozen/Bolzano, Italy)

Organisation and contact:

Verena Lyding
Institute for Specialised Communication and Multilingualism, EURAC
Viale Druso 1, 39100 Bozen/Bolzano, ITALY
Tel: +39 0471 055127
Fax: +39 0471 055199
Email: communication.multilingualism@eurac.edu

P R O G R A M M E

Thursday 13th November 2008

08:00-09:00 Registration

Opening

09:00-09:30 Werner Stuflesser, President of EURAC

Barbara Perathoner, Cultura y Intendënza Ladina, Provinzia Autonoma de Balsan — Südtirol

Minority languages *Chair: Verena Lyding*

09:30-10:30 Keynote speech
Computing with or for minority languages? Aims, methods and responsibilities in computational linguistics.
Dafydd Gibbon (Universität Bielefeld, Deutschland)

10:30-11:10 **Research projects in the area of Ladin studies at the Free University of Bozen/Bolzano**
Paul Videsott (Freie Universität Bozen, Italien)

11:10-11:40 Coffee break

11:40-12:20 **Das Ladinische auf dem Weg eines zeitgemäßen Ausbaus**
Giovanni Mischì (Istitut Ladin Micurà de Rü, Talia)

12:20-13:45 Lunch break

Minority languages and language varieties *Chair: Chris Culy*

13:45-14:25 **The development and acceptance of electronic resources for the Welsh language**
Delyth Prys (Prifysgol Cymru, Bangor, UK)

14:25-15:05 **Corpus annotation and lexical analysis of African varieties of Portuguese**
Amália Mendes (Centro de Linguística da Universidade de Lisboa, Portugal)

15:05-16:30 Poster session and coffee break

Minority languages *Chair: Chris Culy*

16:30-17:10 **African Language Technology: the Data-Driven Perspective**
Guy De Pauw (Universiteit Antwerpen, België), Gilles-Maurice de Schryver (Universiteit Gent, België)

17:10-17:50 **Multimodal corpora and lesser-resourced/less standardized languages**
Jens Allwood (Göteborgs Universitet, Sverige)

Friday 14th November 2008

Learner language *Chair: Andrea Abel*

08:30-09:30 Keynote speech
Using spoken and written learner corpora for research
Karin Aijmer (Göteborgs Universitet, Sverige)

09:30-10:10 **VALICO - Varietà Apprendimento Lingua Italiana Corpus Online**
Elisa Corino (Università degli Studi di Torino, Italia)

10:10-10:45 Coffee break

Learner language and language varieties *Chair: Stefanie Anstein*

10:45-11:25 **Learner language as an under-resourced language**
Anke Lüdeling (Humboldt-Universität zu Berlin, Deutschland)

11:25-12:15 **Historical dictionaries as electronic research infrastructures: a presentation of their treatment and potential**
Julianne Nyhan (Universität Trier, Deutschland)
Andrea Rapp (Universität Trier, Deutschland)

12:15-14:00 Lunch break

Spoken language *Chair: Leonhard Voltmer*

14:00-14:40 **Creating and working with spoken language corpora in EXMARaLDA**
Thomas Schmidt (Universität Hamburg, Deutschland)

14:40-15:20 **Transcription bottleneck of speech corpus exploitation**
Caren Brinckmann (Institut für Deutsche Sprache (IDS), Mannheim, Deutschland)

15:20-15:45 Coffee break

Sign language *Chair: Leonhard Voltmer*

15:45-16:25 **Sign Language Corpora**
Onno Crasborn (Radboud Universiteit Nijmegen, Nederland)

16:25-17:05 **Deaf accessibility to educational content: development of a multilevel platform**
Eleni Efthimiou (ATHENA RC/ILSP –Sign Language Technologies Lab, Greece)

17:05-17:45 Closing session *Chair: Dafydd Gibbon*

Table of contents

Abstracts of oral presentations

Computing with or for minority languages?

Aims, methods and responsibilities in computational linguistics.

Dafydd Gibbon (Universität Bielefeld, Germany)

gibbon@uni-bielefeld.de

In order to examine the sense of applying computational linguistics to minority languages, I will start by discussing computational linguistics in a more general context of the humanities and the study of language and speech in the contemporary context, and then develop three theses as a basis for further discussion.

My starting point is the recognition that computing with language and speech is an activity which all linguists are involved in as soon as they apply their powers of reason to the systematic induction, abduction and deduction of generalisations about language and speech. Even the most die-hard hermeneutic scholar cannot escape this constraint on logical formality. Computing in the narrower sense, as in computational linguistics, adds a need to be very explicit about modelling conventions (ie. for which domain features which data structures and which algorithms are appropriate), and in both empirical and formal soundness and completeness (i.e. precision, consistency and the exact reproduceability of results), essentially in terms of the computational speed which permits quantitative expansion of results both in terms of the extensional coverage of a corpus of objects, and intensional coverage of their detailed properties.

These immanent features of computational linguistics are complemented by a technology and market driven push in both text and speech domains, leading to paradigm shifts in computational linguistics, with less focus on the 'small' units of language, and more on text and discourse. This has been happening since the computational revolution of the 1980s, with the move of practical computing from the large laboratory to the individual scientist, and then to the mass market, and from scientific calculation and modelling to office and games applications. This move was, in turn, the prerequisite for the development of the internet, its extensive social networking and commercial functionalities, and their repercussions for society, inducing massive changes in both personal and global communication habits.

I will present three theses about the relevance of these changes for computational linguistics and discuss their consequences in the context of minority - and in particular endangered - languages:

Thesis One: "Word processing, the Internet and Talking Information Systems are domains of Applied Computational Linguistics, and minority Languages of all kinds are - assuming a motivation for participating in the material benefits of the Global Information Society - no different from other languages in needing such applications."

Thesis Two: "The classic interests of linguists in documenting the languages of the world are in typological theory and in community payback, which are both impossible today without the "Workable, Efficient Language Documentation" (WELD) paradigm which computational language documentation and description provides, necessitating changes in the training of linguists."

Thesis Three: "Computational linguists have responsibilities to provide complete, efficient, state-of-the-art, affordable and fair results with and for the communities which use the languages we study. Responsibility in science is not limited to the avoidance of physical and biological catastrophes, but extends to non-exploitative use of resources of all kinds."

Research projects in the area of Ladin studies at the Free University of Bolzano/Bozen

Paul Videsott (Freie Universität Bozen, Italien)

Paul.Videsott@unibz.it

Among multilingual European universities, the Free University of Bolzano/Bozen takes an exceptional position since its multilingualism, at least in a part of its course offerings (primary school teacher education at the Faculty of Education), also includes a minority language: Ladin. As well, Ladin is a research emphasis. In our presentation we would like to present ongoing and planned projects in the area of Ladin studies in the Department of Ladin in the Faculty of Education, and put them up for discussion.

Das Ladinische auf dem Weg eines zeitgemäßen Ausbaus

Giovanni Mischì (Istitut Ladin Micurà de Rü, Talia)

giovanni@micura.it

Als die sprachliche Forschung im Bereich des Ladinischen vor etwa 150 Jahren einsetzte, richtete sich das Interesse in erster Linie auf die einzelnen ladinischen Idiome und deren phonologische und morphologische Eigenheiten sowie auf die Untersuchung älterer ladinischer Texte. Die Beschäftigung mit Sprachausbau, Wortschatzerweiterung bzw. ganz allgemein mit Wortbildung im heutigen Sinne schien hingegen auf wenig Interesse zu stoßen, im Gegenteil, man war teilweise sogar der Ansicht, Neuprägungen seien der Sprache alles andere als zuträglich.
Dem ist heute ganz anders: Das Hauptaugenmerk in der ladinischen Linguistik hat sich auf die Wortschöpfung (Wortschatzerweiterung) und auf die Sprachplanung (Standardisierung und Normierung) verlagert. Es entstehen Grammatiken und Wörterbücher, wobei der Zugriff auf diese wichtigen Instrumente längst auch über Computer und Internet möglich ist.
Durch den Einsatz moderner Informationstechnologien tun sich auch im Ladinischen neue Wege und Methoden auf. Anhand lexikalischer Datenbanken werden einige der wichtigsten Ansätzen vorgestellt.

English version:

The Ladin language on the path to modernisation

Giovanni Mischì (Istitut Ladin Micurà de Rü, Talia)

giovanni@micura.it

About 150 years ago, the first language studies on Ladin concentrated first and foremost on the single language variants and their phonological and morphological characteristics as well as on the analysis of old texts. At the time, issues of language development, vocabulary expansion or neologisms were considered of limited interest; on the contrary, neologisms were partly considered quite unwelcome.
The current situation is quite different: the main focus of Ladin linguistics has now moved towards neologisms (vocabulary expansion) and language planning (normalisation and standardisation). New important resources like grammars and dictionaries are being developed, both accessible also via the computer and Internet.
Modern information technologies also open up new paths and methods to Ladin . On the basis of lexical databases some of the most important approaches will be presented.

The development and acceptance of electronic resources for the Welsh language

Delyth Prys (Prifysgol Cymru, Banfor, UK)

eds017@bangor.ac.uk

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

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

Corpus annotation and lexical analysis of African varieties of Portuguese

Amália Mendes (Centro de Linguística da Universidade de Lisboa, Portugal)

amalia.mendes@clul.ul.pt

This presentation will focus on our recent experience of establishing fundamental linguistic resources for contrastive linguistic analyses of the African varieties of Portuguese (AVP), namely Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe. We will discuss the difficulties involved in the compilation of corpora for each variety, their annotation with POS information and lemmatization. Five contrastive lexicons have been corpus-extracted in order to establish for each variety a core and peripheral vocabulary and to study AVP-specific morphological processes. These are first steps towards an integrated description of the five varieties, together with contrastive analyses with European and Brazilian Portuguese.

African Language Technology: the Data-Driven Perspective

Guy De Pauw (CNTS – Language Technology Group, University of Antwerp, Belgium; School of Computing & Informatics, University of Nairobi, Kenya)

guy.depauw@aflat.org

Gilles-Maurice de Schryver (Department of African Languages and Cultures, Ghent University, Belgium; Xhosa Department, University of the Western Cape, Republic of South Africa; TshwaneDJe HLT, Pretoria, Republic of South Africa)

gillesmaurice.deschryver@aflat.org

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

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

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

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

- Language independence: we will illustrate how the same machine learning approach can be used to perform diacritic restoration for a variety of resource-scarce African languages (Ciluba, Gikuyu, Kikamba, Maa, Northern Sotho, Venda and Yoruba).
- Development Speed: we will show how a small annotated corpus can be used to develop a robust and accurate part-of-speech tagger for Northern Sotho.
- Adaptability: most data-driven techniques have a distinct Indo-European bias, but can easily be adapted to work for African languages as well, as exemplified by our work on a Swahili memory-based morphological analyzer.
- Empiricism: we show how the language technology components can be developed and evaluated using real-world data, offering a more realistic estimation of their usability in a practical application.

Multimodal corpora and lesser-resourced/less standardized languages

Jens Allwood (Göteborgs Universitet, Sverige)

Jens.Allwood@ling.gu.se

The talk describes some ideas for how multimodal corpora can be used in documenting and maintaining lesser resourced and less standardized languages. Examples from South Africa and Nepal, where I have done some work, will be discussed.

Using spoken and written learner corpora for research

Karin Aijmer (Göteborgs Universitet, Sverige)

karin.aijmer@eng.gu.se

Recently we have seen a lot of work on spoken and written learner corpora. The corpora have been used for research in a large number of areas including modality, particle verbs, information structure, and connectivity, metadiscourse, discourse markers (spoken learner corpora). The problems of compiling learner corpora will be discussed and the special methodologies they involve.

VALICO - Varietà Apprendimento Lingua Italiana Corpus Online

Elisa Corino (Università di Torino, Italia)

elisa.corino@unito.it

VALICO is an Italian international Learner Corpus freely available and searchable online.

The acronym VALICO stands for "Varietà di Apprendimento della Lingua Italiana: Corpus Online", i.e. 'Online Corpus of the Learning Varieties of the Italian Language'.

The name Valico (properly meaning '(mountain) pass') was chosen in 2003 for two main reasons:
- Piedmont, the region where our group operates, is a country of mountain passes.
- The mountain pass suggests the general metaphor of the learning process.

The main goals of this learner corpus are
- to offer a survey of text types that teachers of Italian around the world make their students write;
- to show how students of different age and mother tongue write in Italian;
- to work as a roundtable on methods to prevent the most frequent errors and avoidance strategies;
- to provide Italian Linguistics with new insights into variation and acquisition.

In our contribution we are going to discuss our experiences concerning methodological and technical challenges in learner corpus research, drawing a critical history of our corpus (choice and elicitation of texts – what and why, tagging and architecture, metadata…), comparing it with its English and German “relatives”, and giving an insight in current researches, results and materials derived from its analysis and exploitation.

Learner language as an under-resourced language

Anke Lüdeling (Humboldt-Universität zu Berlin, Deutschland)

Anke.Luedeling@rz.hu-berlin.de

Many varieties, such as learner language, dialects, spoken varieties etc. of otherwise well-resourced languages face the same problems often faced by under-resourced languages, namely lack of (orthographic or grammatical) standards, lack of (computer-readable) lexicons or grammars and, most importantly, lack of a theoretical description.
In this talk, I will use the German learner corpus Falko (Lüdeling et al. 2008) as an example to illustrate how design and annotation decisions influence the possibilities of evaluation. I will focus especially on the syntactic annotation. I will demonstrate how learner errors can be described and categorized and how this procedure can be applied to other varieties.

Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin & Walter, Maik (erscheint) Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache 2(2008)*, 67-73.

Historical dictionaries as electronic research infrastructures: a presentation of their treatment and potential

Julianne Nyhan (Universität Trier, Deutschland)
Andrea Rapp (Universität Trier, Deutschland)

The advantages and disadvantages of the application of XML to medieval inflected languages: a case study of electronic resources for medieval Irish

Julianne Nyhan (Universität Trier, Deutschland)
julianne.nyhan@ucc.ie

The Corpus of Electronic Texts (CELT, http://www.ucc.ie/celt) project at University College Cork is a multilingual on-line corpus of texts of Irish literature, politics and history. It is TEI conformant and encoded in SGML/XML. As of September 2008, the corpus has over 10 million words on-line. The doctoral work that I carried out at the project focused on the development of lexicographical electronic resources spanning the years c. AD 700-1700, and on the development of tools to integrate the corpus with these resources. The resulting XML framework allows advanced search and interrogation of Old, Middle and Early Modern Irish and a prototype of this work is freely available at (http://epu.ucc.ie/lexicon/entry). A further project that is ongoing, and that will also be integrated with the corpus and lexicographical resources for medieval Irish, is a digital edition of a modern Irish dictionary: Patrick S. Dinneen's *Foclóir Gaedhilge agus béarla*. Upon publication in late 2008, the Digital Dinneen will extend CELT's lexicographical coverage from the Old Irish period up to the modern period.
This paper will present a case study of these resources in order to discuss the main advantages and disadvantages of the application of XML to medieval inflected languages. While the findings will be made in the context of Old Irish, they will be shown to have applications to the wider domain of medieval inflected languages.

An introduction to the project 'Interactions between linguistic and bioinformatic operations, methods, and algorithms: modelings and mappings of variance in language and genome' (Federal Ministry of Education and Research, Germany)

Andrea Rapp (Universität Trier, Deutschland)
andrea.rapp@uni-trier.de

From the very beginning (historical) linguistics has tried to detect, document, record, and analyse variances in and varieties of language. Fixed points of reference, or benchmarks, that allow the mapping and organisation of variance are necessary not only in the recording of linguistic data, e.g. in dictionaries, but also for the analyses of corpora. While a concept based on semasiology limits the scope of reference, it also establishes an empirical and measurable base.

In our project we are using language data from more than 10 electronic dictionaries of german varieties (historical and regional) and genom databases (e.g. Single Nucleotide Polymorphism

Database with 34 million variances) to develop operations and algorithms that could describe, typecast and compare variances.

The variant keywords of the dictionaries will be mapped to a 'standard keyword' that derives from the modern german language corpora of the 'Institut für Deutsche Sprache' in Mannheim. This 'Meta-List' will form a network of lemma-variants that allows e.g. graphematic, phonologic, morphologic or lexical research on texts, dictionaries or corpora without equalising or altering the variance.

Creating and working with spoken language corpora in EXMARaLDA

Thomas Schmidt (Universität Hamburg, Deutschland)

thomas.schmidt@uni-hamburg.de

Spoken language corpora – as used in conversation analytic research, language acquisition studies and dialectology – pose a number of challenges that are rarely addressed by corpus linguistic methodology and technology. My talk will start by giving an overview of the most important methodological issues distinguishing spoken language corpus work from the work with written data. I will then show what technological challenges these methodological issues entail and demonstrate how they are dealt with in the architecture and tools of the EXMARaLDA system.

Transcription bottleneck of speech corpus exploitation

Caren Brinckmann (Institut für Deutsche Sprache (IDS), Mannheim, Germany)

brinckmann@ids-mannheim.de

While written corpora can be exploited without any linguistic annotations (e.g. Keibel and Belica, 2007), speech corpora need at least a basic transcription to be of any use for linguistic research or technological applications. The basic annotation of speech data usually consists of time-aligned orthographic transcriptions. To answer phonetic or phonological research questions phonetic transcriptions are needed as well. However, manual annotation is very time-consuming and requires considerable skill. The following approaches address the transcription bottleneck of speech corpus exploitation:

- *Crowdsourcing the orthographic transcription*: A web-based annotation tool (Draxler, 2005) can be used for collaborative transcriptions of dialectal speech data, which can be transcribed reliably only by native speakers of the dialect.
- *Automatic broad phonetic alignment*: Van Bael et al. (2007) showed that automatic canonical transcriptions are comparable to manually verified transcriptions. Automatically derived phonetic segments can in turn be subjected to automatic signal analyses.
- *Query-driven annotation*: Voormann and Gut (2008) suggest a cyclic corpus creation and annotation process that starts with the formulation of a query. Following the "agile corpus creation" approach only those parts of a corpus that are currently needed for a specific research question are annotated and analysed.

Current and future applications of all three methods within IDS speech corpus projects (e.g. Brinckmann et al., 2008; Raffelsiefen and Brinckmann, 2007) are presented and discussed.

Brinckmann, C., Kleiner, S., Knöbl, R., and Berend, N. (2008): German Today: an areally extensive corpus of spoken Standard German. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Draxler, C. (2005): WebTranscribe – an extensible web-based speech annotation framework. *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)*, Karlovy Vary, Czech Republic, 61-68.

Keibel, H. and Belica, C. (2007): CCDB: a corpus-linguistic research and development workbench. *Proceedings of Corpus Linguistics 2007*, Birmingham, United Kingdom.

Raffelsiefen, R. and Brinckmann, C. (2007): Evaluating phonological status: significance of paradigm uniformity vs. prosodic grouping effects. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI*), Saarbrücken, Germany, 1441-1444.

Van Bael, C., Boves, L., van den Heuvel, H. and Strik, H. (2007): Automatic phonetic transcription of large speech corpora. *Computer Speech and Language* 21 (4), 652-668.

Voormann, H. and Gut, U. (2008): Agile corpus creation. *Corpus Linguistics and Linguistic Theory* 4 (2), 235-251.

Sign Language Corpora

Onno Crasborn (Radboud Universiteit Nijmegen, Nederland)

o.crasborn@let.ru.nl

The rise of YouTube marks the definitive arrival of video on the desktop. Only in recent years has video technology become flexible enough to allow for the easy handling of linguistic data for sign language research in the computer. For decades, film and video were patiently transcribed on paper without a direct link to the original data. This presentation describes the current status in the creation and validation of sign language corpora for linguistic research and the development of language technology. Covering both the challenges in working with video and the technical and linguistic complexities in creating annotations for signed languages will be discussed.

Deaf accessibility to educational content: development of a multilevel platform

Eleni Efthimiou (ATHENA RC / ILSP – Sign Language Technologies Lab, Greece)

Eleni_e@ilsp.gr

The presentation focuses on design requirements and implementation issues underlying a platform environment that allows development of various educational applications fully accessible by deaf users. Subject to Design for All primes, the environment to be discussed as showcase, is built on methodological principles, which adopt sign language as the basic means for communication of linguistically uttered educational content. It also makes extensive use of visual objects to support comprehension and navigation at all levels of human-computer interaction. Currently available instantiations of the environment have incorporated both video for content presentation and an avatar based dynamic sign synthesis mechanism. The educational applications to be referred to when discussing the design principles and user requirements taken into account, include one web-based and one off-line GSL teaching tool for the same school level (early primary school) as well as a vocational training tool for adult users.

Abstracts of poster presentations

English-Persian Parallel corpus as a Translation Aid

Tayebeh Mosavi Miangah (Payame Noor University of Yazd, Iran)

mosavit@pnu.ac.ir; *mosavit@hotmail.com*

1. Introduction

In recent years, corpus linguistics has provided new potentials for use in many applications in language sciences. The translator's workplace has changed dramatically over the last ten years or so, and today the computer is undoubtedly the single most important tool of the trade for a translator regardless of whether he or she is a literary translator working for a small publisher, a technical translator working for a translation agency or a legal translator.

Recently, large monolingual, comparable and parallel corpora have played a very crucial role in solving various problems of computational linguistics such as part of speech tagging (Brill, 1995), word sense disambiguation (Mosavi Miangah and Delavar khalafi, 2005), language teaching (Aston, 2000; Leech, 1997), phrase recognition (Cutting, et al., 1992), information retrieval (Mosavi Miangah 2008), statistical machine translation (Brown et al., 1990) and some other problems.

Before the present stage of ICT development, corpora and concordancing software were hardly available to translators in order to gain information about language, content, and translation practices. But now with continuous overload in translation work and massive production of translated texts corpus resources available to translators aroused an increased interest in their construction and use.

2. Constructing and Using Bilingual Parallel Corpus

In order to examine the superiority of corpus-based language analyses over traditional methods and compare the two kinds of translational resources, namely, bilingual parallel corpora and conventional bilingual dictionaries we had to construct our English-Persian parallel corpus.

Compiling bilingual corpora for high density languages such as English or French are very extensive and the results are very encouraging due to easy accessibility of the texts in these languages in digital forms including Websites. However, when a low or medium density language such as Persian comes to be one of the languages involved in a bilingual corpus, the case is much more difficult due to shortage of digitally stored materials as well as detectable parallel pages in World Wide Web.

For this experiment, we used our developmental English-Persian parallel corpus consisting of about three million words (more than 50,000 corresponding sentences in two languages). This is a kind of ongoing corpus, that is, an open corpus in which more material can be added as the need arises. Naturally, more rich the corpus in terms of the volume of the data and different kinds of annotations, more useful it will be for solving a variety of linguistic and translation problems. One of the main consequences of building such a corpus is to develop software for parallel concordancing in which a user can enter a search string in one language, and see all citations for that string in the search language as well as corresponding sentences in the target language.

3. Different Applications of Parallel Corpora

Although the range of applications of parallel aligned corpora in language sciences are wide, in this paper we only deal with some of the main applications of such corpora within the field of human translation.

One of the main applications of parallel corpora is to find different possible equivalents of certain words or collocations. That is, aligned translation units are simply displayed on the screen, offering the translator a range of similar contexts from a corpus of past translations. Usually finding appropriate and natural equivalent for different types of collocations is a difficult task especially in non-native language, and parallel corpora can be of great help in this respect.

In this connection, we prepared a set of one hundred English collocations and tried to find their appropriate Persian equivalents in the corpus. As far as bilingual dictionaries in most cases provide us with translational equivalents of single words and not collocations, bilingual corpora are considered as great help in this respect. That is, a parallel corpus is used to confirm the translation equivalent for a certain collocation where the majority of instances offer the same thing.
In some cases we may refer to a parallel corpus to verify, reject or supplement the equivalent(s) provided by bilingual dictionaries since it is believed that parallel corpora provide information that bilingual dictionaries do not usually contain. Using a bilingual dictionary for selecting a translation equivalent, the translator will decide about the appropriateness of different possible equivalents based of their definitions or a few examples given by the dictionary, while a parallel corpus offers the best possible translation equivalent based on natural evidences gained from past translations.

4. Conclusion

The method of using parallel corpora in finding translational equivalents for collocations not only has a great effect on improving the quality of translations produced by human translators, but also can be directly applied in machine translation systems. The other main potential of such corpora is to search for units above word level like collocations and phraseological units to extract correspondences between languages and make terminological databases. This further task can easily be realized with constructing specialized monolingual and bilingual corpora. It is hoped that our translators become more familiar with the valuable potentials of different types of corpora in their works.
The suggestion here is that modern technologies such as corpora and concordancing software should find their proper place in translator workbenches, and this ideal can be achieved provided that more corpus resources are accessible to the translators. Practical courses introduced by the translation trainers at the universities can be of great help in this respect.

References

Aston, G. (2000). "I corpora come risorse per la traduzione e l'apprendimento". In Silvia Bernardini and Federico Zanettin (eds.) *I corpora nella didattica della traduzione*. Bologna: CLUEB, 21-29.

Brill, E. (1995). Unsupervised learning of disambiguation rules for part of speech tagging. In *2nd Workshop on Large Corpora*, Boston, USA.

Brown, P., Cocke, S., Della Pietra, V., Della Pietra, S., Jelinek, F., Lafferty, J., Mercer, R. & Roosin, P. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics* 16:2, 79-85.

Cutting, D.; Kupiec, J.; Peterson, J. and Sibun. P. (1992). A practical part of speech tagger. In proceeding of 3rd Conference on Applied Computational Linguistics, Trento, Italy, PP. 133-140.

Leech, G. (1997). Teaching and language corpora: A convergence. In: A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (1-23). New York: Addison Wesley Longman

Mosavi Miangah, T. and Delavar Khalafi, A. (2005). Word sense disambiguation using target language corpus in a machine translation system. *Literary and Linguistic Computing*, 20(2), 237-249.

Mosavi Miangah, T. (2008). Automatic term extraction for cross-language information retrieval using a bilingual parallel corpus. Proceedings of the 6th International Conference on Informatics and Systems (INFOS2008), PP. 81-84, Egypt.

A Computational Approach to Language Documentation of Minority Languages in Bangladesh

Naira Khan (University of Dhaka, Bangladesh; Center for Research on Bangla Language Processing (CRBLP))

nairakhan@gmail.com

This paper is aimed at introducing a computational element in the documentation of minority languages of Bangladesh. Belonging to the Sino-Tibetan family, the minority languages of Bangladesh are a faction of a larger belt that has broken off from the main family and runs through Burma, India and Bangladesh where they exist in isolation as the minority. In Bangladesh the languages are some of the most neglected and under-documented and hence endangered languages of the world (Murshed 2003). Here we attempt to imbue ongoing documentation methods with a computational element in order to add a new dimension with the hopes of expediting documentation procedures as well as other benefits that result from the formalism used. For our current work we propose the use of the HPSG formalism as it is an easy-to-use simplistic yet rich framework that mirrors a high level programming language allowing us to describe languages in terms of hierarchical feature structures making it a powerful tool designed to simultaneously capture linguistic phenomena specific to languages along with cross-linguistic phenomena common to all languages (Sag and Wasow, 1999). Thus not only will we be able to document and describe the endangered languages, at the same time we will also be able to capture generalizations amongst these languages. A number of open-source, HPSG related computational resources are made available online by the Linguistic Grammars Online (LinGO) initiative at Stanford University, an ongoing collaboration, and includes grammars, lexicons, the Grammar Matrix- a framework to aid in the development of broad-coverage, precision, implemented grammars for diverse natural languages, and the Linguistic Knowledge Based (LKB) Grammar Engineering Platform (Sag and Wasow, 1999) - a powerful implementation platform comprising both a syntactic and a semantic level that allows the user to parse as well generate by using the formalism to code in linguistic rules through feature structures and feature unification (Copestake and Flickinger, 2000). The LKB system is a grammar and lexicon development environment for typed feature structure grammars. This is a powerful combination in that it allows a grammar developer to write grammars and lexicons that can be used to parse and generate natural languages (Copestake, 2002). Therefore the structure of the system itself requires that the grammatical forms and the lexicon i.e. vocabulary of the target language be documented which will then be used to produce successful parses, providing in its nature a system that can be readily used as a documentation tool. The benefit here will be bidirectional in that the framework will provide a tool for documentation while the documentation may actually serve to enrich the formalism as some of the languages, being extremely under-documented, may reveal unique linguistic structures or archaic forms that have no representation in the formalism and in turn will provide new data that might call for a new formula to be written. The other outcomes of using the LKB are the multiple uses of the parser and generator alongside documentation. The parser itself can be used for language education in order to reveal grammatical structure of sentences of the language being documented, as students can enter basic grammatical sentences and obtain parse trees that clearly exhibit the structure. On the other hand the generator allows for morphological variants to be generated from the root form. Such a tool can be used to learn various forms of the word classes i.e. verb-forms, nominal inflection, derivational forms etc. In terms of the documentation of phonology a proposition can be made to incorporate the non-procedural theory of phonology into HPSG, based on a model of Bird and Ellison (1992) which is a constraint-based phonological model congruent with HPSG and traverses the usual limitations by including a prosodic domain expressed through prosodic type hierarchy modeled on HPSG's lexical type hierarchy. Based on finite state phonology the framework encompasses non-linear

phonology and exemplifies interactions between phonology and morphology along with phonology and syntax where traditional prosodic domains are recast as types, and domain-bounded processes are regular expressions tied to the required prosodic type (Bird 1994). The encoding of the grammar itself calls upon the need for orthographic representation, hence if a language is devoid of a script a Romanized transliteration scheme will need to be immediately written leading to the automatic devising of a script as well. The most attractive feature of the Matrix/LKB framework is its high-end implementation for machine translation. As there are other languages that have been exhaustively described in this framework namely English, known as the English Resource Grammar (ERG) (Copestake, 2002), Japanese (JACY), Greek (Modern Greek Resource Grammar) etc. (http://lingo.stanford.edu/), hence the various grammars can be connected to create an automated machine translator as an immediate output of the documentation endeavor. In fact the Matrix and LKB are components of a suite of software known as Montage designed specifically to bring together grammar engineering and field linguistics through shared threads of commonality lying in descriptive methodology (Bender et al, 2004). The relevance of using the HPSG formalism as well as the Matrix/LKB platform is due to the fact that neither is new to Bangladesh as these are already being used to create a computational grammar of Bengali (Khan and Khan, 2008). This paper is therefore a description of an endeavor to introduce this computational dimension in the documentation of the endangered languages of Bangladesh with an aim of faster and better documentation of some of the most neglected languages of the world.

References:

Bender, Emily M., Dan Flickinger, Jeff Good and Ivan A. Sag. 2004. Montage: Leveraging Advances in Grammar Engineering, Linguistic Ontologies, and Mark-up for the Documentation of Underdescribed Languages. *Proceedings of the Workshop on First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, LREC 2004*, Lisbon, Portugal.

Bird, Steven. 1994. Finite-state phonology in HPSG. *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-92), 74-80*. Kyoto, Japan.

Bird, S. and Elison, T.M. 1992. One level phonology: autosegmental representations and rules as finite-state automata. *RP 51, University of Edinburgh, Centre for Cognitive Science*. Edinburgh, Scotland.

Copestake, A and Flickinger, D. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second conference on Language Resources and Evaluation (LREC-2000)*. Athens, Greece.

Copestake, A. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.

Khan N. and Khan M. 2008, *A Computational Grammar of Bangla using the HPSG formalism: Developing the First Phase*. The Dhaka University Journal of Linguistics. Dhaka, Bangladesh.

Murshed S. M., 2003. *Indigenous & Endangered language of Bangladesh*, ICAS -3, Singapore.

Sag, I and Wasow, T. 1999. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford.

VIS-À-VIS –
a System for the Comparison of Linguistic Varieties on the Basis of Corpora

Stefanie Anstein (Europäische Akademie Bozen, Italien)

Stefanie.anstein@eurac.edu

Main varieties of languages are usually far better investigated and described than their lesser-used counterparts. Even though varieties of a language in most cases do have similar linguistic characteristics, there are still differences that are to be extracted e.g. for variant lexicography (cf. Ammon et al. 2004), for standardising lesser-used varieties, or for aiding language teaching and learning. Differences on the lexical level usually consist in one-to-one equivalents such as "Kondominium" ("apartment building") in South Tyrolean German vs. "Mehrfamilienhaus" in Germany or in many-to-one equivalents such as "provisorische Ausfahrt" ("temporary exit") and "Behelfsausfahrt", respectively. More complex phenomena such as differing collocations or subtle semantic differences are more difficult to find, e.g. the additional meaning of "Mobilität" in South Tyrol, which is "unemployment" in addition to "mobility".

Corpora for the varieties of a language are a valuable basis for finding relevant differences. They are being compiled in projects such as the 'International Corpus of English' (ICE[1]) or are e.g. used for work on Portuguese varieties by Bacelar do Nascimento et al. (2006). Also for German, an initiative of research centres in Basel[2], Berlin[3], Bolzano[4], and Vienna[5] called 'C4' is developing variety corpora comparable with respect to contents and size. Related work has been done on the comparison of language over time, of originals and translations, of native and learner language, etc. Many of these studies were conducted manually for very specific phenomena. In addition, there are more statistical approaches to data extraction from parallel or even from unrelated monolingual corpora (e.g. Nazar 2008).

For a systematic and comprehensive comparison of corpora on different levels of linguistic description, semi-automatic tools are needed. Manual work has to be supported and reduced by the automatic filtering of statistically produced lists containing suggested 'candidates' for differences or particularities. Trivial characteristics of a variety or knowledge already investigated and confirmed (e.g. collections of proper names or regionalisms such as the above mentioned "Kondominium") can be automatically removed from such candidate lists. Experts can then concentrate on the evaluation and interpretation of the remaining, new phenomena, which will always have to be done manually.

The toolkit VIS-À-VIS is being developed in a doctoral thesis in the framework of the project 'Korpus Südtirol'. It will be evaluated mainly with German varieties, but the resulting system will be language independent. The tools aim at providing support to linguists for the systematic comparison of varieties on the basis of corpora. This support consists in methods to filter huge amounts of data and present to the expert only probably relevant material. With this approach, less manual work is necessary and quantitative methods can be combined with qualitative ones. In addition, the data to be evaluated manually is presented in a user-friendly and intuitive way to facilitate the interpretation and further processing.

As input to VIS-À-VIS, users give the corpora to be compared as well as, if available, lists with previous knowledge as described above. The corpora are then annotated with standard tools, which is where difficult cases for the tools or errors produced by them can identify the first set of candidates for special variety characteristics, since the tools are usually created for the main varieties. In the following modules, the corpora are analysed and compared with a combination of existing as well as new or adapted tools for e.g. concordancing or frequency statistics. The lexical level is the first and most promising linguistic area to explore; further studies will elaborate on collocations and phrases up to more subtle semantic or pragmatic

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

A description of the first approaches to this toolkit can be found in Abel and Anstein (2008). In this poster, the overall VIS-À-VIS architecture and workflow is to be demonstrated. Since ongoing work is being described, it includes discussion on possible alternative detail solutions and future work.

Abel, Andrea; Anstein, Stefanie (2008): Approaches to Computational Lexicography for German Varieties. In: Proceedings of the XIIIth Euralex International Congress, Barcelona; pp. 251-260.

Ammon, Ulrich et al. (2004): Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Walter de Gruyter, Berlin.

Bacelar do Nascimento, Maria F. et al. (2006): The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon. In: Calzolari, Nicoletta et al. (eds.), Proceedings of the Vth International LREC, Genoa; pp. 1791-1794.

Nazar, Rogelio (2008): Bilingual terminology acquisition from unrelated corpora. In: Proceedings of the XIIIth International Euralex Congress. European Association for Lexicography, Universidad Pompeu Fabra.

[1] http://www.ucl.ac.uk/english-usage/ice

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

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

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

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

Ranked Terms List for Computer Assisted Translation

Atelach Alemu Argaw (Stockholm University / KTH Stockholm, Sweden)

atelach@dsv.su.se

Abstract

We present an approach for a Computer Assisted Translation tool which ranks possible translations in the order of appropriateness given the current context. We use an Amharic to English translation as an example. The resources required are a machine readable dictionary or some form of lexicon, and a corpus in the target language. A word space is generated using the random indexing methodology, and similarity of context vectors in the word space is used to rank the translations.

1. Introduction

Computer Assisted Translation (CAT) tools are designed to increase the speed and accuracy of translation. Unlike Machine Translation (MT) systems, all decisions are made by the user, and the computer's task is to assist the user by providing appropriate information. A CAT tool can range from spell checkers and simple access to machine readable dictionaries (MRD), glossaries, terminology databases, etc to concordances, and translation memories. Due to the lack of resources such as lexicon, morphological analyzers, Part of Speech (POS) taggers, parsers, parallel corpus, etc, we have a long way to go before we can produce a full fledged MT system or an advanced CAT system for minority languages. While keeping the effort towards enabling MT for minority languages, we could also make use of the small resources that are available to provide a simplistic CAT system using minimal resources. Such availability would in turn facilitate the creation of more and more translated text in the language that could then be used as a training material for statistical MT systems, or translation memories.

Computational linguistic resources are not well developed for Amharic[1]. There are no MT or CAT systems developed for the language to date, but there has been a continuing effort to develop computational tools are resources for the language. In relation to MT, some work has been done in the area of automatic lexicon acquisition. Argaw et al (2004) report a bilingual lexicon construction using co-location measures, Amsalu (2006) reports lexicon acquisition through statistical language modeling, to give some examples. See (Amsalu & Adafre, 2006) for a description of the current state of Amharic MT and recommendations.

We present an approach for producing an Amharic CAT that requires a MRD and running text in the target language. The system provides fast access to information in the MRD that would have taken longer to find in paper dictionaries, as well as through the provision of a ranked list of possible translations from the MRD, users would get suggestions for word choices appropriate for the current context. Given one contextual word in English, the system provides a ranked list of possible English translations for an Amharic word based on vector similarity in a word space model generated by the Random Indexing methodology (Sahlgren, 2006), from about 160,000 English news articles.

2. Experiments and Results

For a given word in a citation form in an Amharic sentence and one translated content bearing word in the same sentence, the system is designed to suggest a ranked list of possible translations for the Amharic word. Amharic sentences written using the Amharic script *fidel* were transliterated to SERA[2], a convention for the transcription of *fidel* into the ASCII format, using a file conversion utility called g2[3]. This is done for compatibility reasons since the MRD and the sample text we used are written using different fonts (there is no standard font for

[1] Amharic is a semitic language spoken by an estimated 20-30 million people in Ethiopia.

[2] http://www.abyssiniacybergateway.net/fidel/sera-94.html

[3] g2 was made available to us through Daniel Yacob of the Ge'ez Frontier Foundation (http://www.ethiopic.org/)

Amharic yet and texts are written using many fonts that are not compatible with one another). We selected ambiguous Amharic terms from the sample sentences and tried to automatically induce a ranked list of the possible translations from those given in an Amharic-English MRD (Aklilu, 1981) containing 15,000 entries. The ranking is done by using context vector similarity measures in a word space. A large document collection (160,000 news articles from the CLEF[4] English document collection) is used to create a word space where semantically related words appear in close proximity. We used GSDM[5] (Guile Sparse Distributed Memory) and a morphologically normalized (using the Porter[6] stemming algorithm) CLEF collection to generate the word space using the random indexing methodology and train context vectors for syntagmatic word spaces. See (Sahlgren, 2006) for a detailed description of this process. We calculate the cosine similarity between the context vectors of a pair of words using the functions provided in GSDM. When the translation equivalent is a multi word description, the similarity is calculated with the content bearing words in the definition. Some examples are given in Table 1 below. If we take the Amharic word "mewTat", it has the translations "go out, leave, climb" in the MRD utilized. When it appears in a sentence that also has "terara" which is translated as "mountain" in the dictionary (or alternatively as given by the user), we calculate the similarity between the context vectors of the word "mountain" and each of the possible translations and rank them based on their closeness to "mountain" as given by the vector similarity values in the word space generated. The ranked list is then presented to the translator/user to assist in the translation process.

Amharic Word	Context	Translations with Ranks
lsat	bEt: house	fire (1), intelligent (2), brilliant (3)
xlmat	selam: peace	present (2), prize (1)
mesfafat	Eds: AIDS	spread (1), develop (4), expand (2), distend (3)
texkerkari	gebeya: market	vehicle (1); s/t which is rolled along (2) [roll (2)]
mewTat	terara: mountain	go out (3) [go (3)], leave (2), climb (1)

Table 1: Translations with corresponding ranks based on context

3. Concluding Remarks

We have presented preliminary efforts towards creation of resources for CAT of Amharic. The approach requires a MRD and a corpus in the target language, and can be applied to any language pair. Such resources are more readily available for minority languages than most resources that are required for advanced CAT and MT systems. Although we need to make large scale evaluations, preliminary results show that it is possible to accurately rank the possible translations that are found in a MRD, given the context of a word. Such information can play a crucial role in minimizing the time taken by a translator or a user.

Acknowledgments

The GSDM tool and the Guile functions were provided by Anders Holst and Magnus Sahlgren at the Swedish Institute of Computer Science (SICS).

References

Aklilu, A. (1981). *Amharic English Dictionary*, volume 1. Mega Publishing Enterprise, Addis Ababa, Ethiopia.

Amsalu, S. (2006). Data-driven Amharic-English bilingual lexicon acquisition. *In Proceedings of LREC-2006: Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy.

[4] http://www.clef-campaign.org/

[5] GSDM is an open source C-library for Guile, designed specifically for the Random Indexing methodology, written by Anders Holst at the Swedish Institute for Computer Science.

[6] http://tartarus.org/~martin/PorterStemmer/index.html

Amsalu, S & Adafre, S. F. (2006). Machine Translation for Amharic: Where we are. *In Proceedings of LREC-2006: Fifth International Conference on Language Resources and Evaluation*. 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages", Genoa, Italy.

Argaw , A. A., Asker, L., & Eriksson, G. (2004). Building an Amharic Lexicon from Parallel Texts (Poster presentation), *In Proceedings of First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*, a Workshop at LREC 2004, Lisbon, Portugal.

Sahlgren, M. (2006). *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Doctoral thesis, Stockholm University.

Towards an online semi automated POS tagger for Assamese

Pallav Kr. Dutta, Shakuntala Mahanta, Gautam Barua

(Indian Institute of Technology Guwahati, India)

pkdutta@iitg.ernet.in smahanta@iitg.ernet.in gb@iitg.ernet.in

Developing annotated tagged corpora for a language with limited electronic resources can be very demanding. A number of other components like a dictionary, a stemmer, and a morphological analyzer have to be prepared, prior to the development of a semi automated tagged corpora. Since these tools require a lot of human input, they are time consuming and are also prone to error. This kind of approach has been undertaken in the development of tagged annotated corpora for some of the major languages spoken in India, for example, Hindi [1]. Although Assamese is a language spoken by about 15 million people in the Indian state of Assam as a first language, the development of electronic resources for the language has been lagging behind other Indian languages. In order to fill this gap, we have designed a POS Tagger of Assamese. Our approach is to manually handle stemming and morphological analysis of words in a base corpus. Words in this corpus will be analysed, broken down, and tagged by a team of experts. Words so tagged are stored in a database which contains a dictionary, a list of affixes, and other such information. This database is then used as a resource in the second stage where a large body of native speakers assist the system in tagging a larger corpus. Our premise is that if we manually tag a core set of words in a language, we can ascertain the tagging required of a new word by consulting the database of information already tagged words and with the help of untrained, native speakers. We therefore do away with the complex tasks of building a morphological analyser and stemmer for a language. We rely on pattern matching of new instances with what is already available with native speakers providing verification.

We first take an untagged corpus and each unique word is stored in a database. In subsequent analyses, if a word already exists, it will not be stored in the database. Assamese is morphologically rich, which means the morphological analyzer will have to take into account a plethora of linguistic information concerning noun and verb paradigms and prefixes, suffixes and classifiers. Therefore, each word is analyzed in detail, so that all the lexical and grammatical information are successfully identified – breaking it into root and affixes (prefix or suffix), quantifiers, classifiers etc and these details are stored in the database (Fig-2). During this step, expert users provide inputs to the system. Separate tables of affixes, quantifiers etc. are also created as they are encountered. The dictionary contains fully tagged information of words. One word has to be tagged only once unless ambiguity arises. During the analyzing process of each word, the occurrence of the word in the sentences of the corpus is displayed. If a particular word seems to be ambiguous, then that word has to be analyzed multiple times and the user can enable this by setting an ambiguous flag to “yes” (Fig-1).

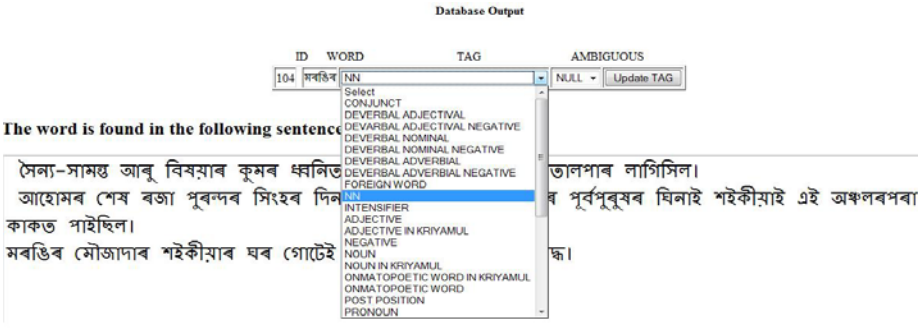

Fig- 1

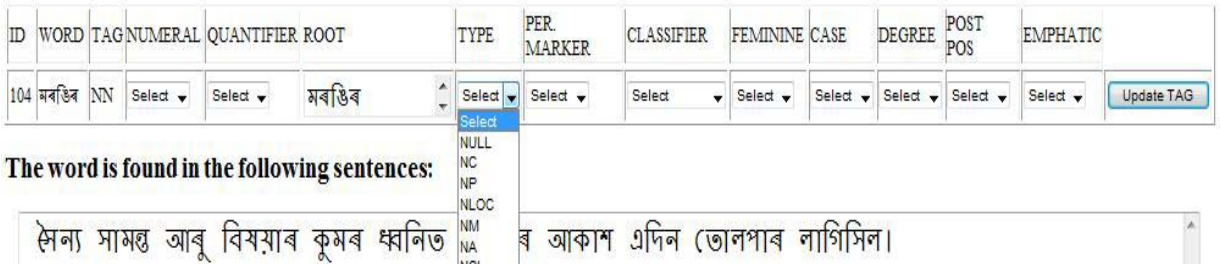

Fig- 2

In this word analysis process we also gather the lexical and grammatical rules which will be used for the automatic tagging process. For example, noun root (NN) can be followed by some inflectional suffix (Classifier- Singular/Plural, Feminine, Case, Degree) or an emphatic suffix, or a post position.

Later, when an untagged corpus is submitted for tagging, each word will be looked up in the dictionary and if it exits in the dictionary, confirmation of the base tags and meaning will be obtained from the user (to detect multiple uses of words). If the word is not in the dictionary, it will try to find word(s) with similar affixes in the database and if found, will break up the word as per the matching word and ask the user for confirmation of this break-up (and so the tagging). The user may or may not be able to vet the analysis depending on his individual knowledge. Later on the user verified words are vetted by an expert. The expert has to only read the tagged word and not the whole text and this will reduce the amount of work that will be required. If a similar type of word is absent in the database, the word will be added in the database for future tagging/ analyzing by an expert. During this stage, the user does not need to be an expert in linguistics. An untrained native speaker is all that is required.

While this use of humans to assist in the tagging may seem that it is not going to be of much use, we have observed that providing simple aids to humans improves their productivity tremendously. Our system should therefore speed up the process of tagging a great deal. We also intend to make the tool available online, so that we can tap the large community of native speakers to contribute towards the development of a tagged corpus of Assamese. As further work, we will seek to generalize this process so that it becomes applicable to any Indian language, and we aim to create such a system for at least one more language.

References:

1. Shrivastava, Agrawal, Mohapatra, Singh and Battacharya, "Morphology based Natural Language Processing tool for Indian languages", paper presented in the 4th Annual Inter Research Institute Student Seminar in Computer Science (IRISS05), April 2005, IIT Kanpur. (*www.cse.iitk.ac.in/users/iriss05/m_shrivastava.pdf)*

2. Baskaran, S Et al., "Designing a common POS-Tagset Framework for Indian Languages", Proceedings of the 6th Workshop on Asian Language Resources (ALR 6), January 2008, Hyderabad., pp. 89

3. Monisha Das, S. Borgohain and Juli Gogoi and S B Nair, "Design And Implementation of a Spell Checker for Assamese", Proceedings of the Language Engineering Conference (LEC'02), 2002, pp. 156.

4. Sirajul Islam Choudhury, L. Sarbajit Singh, S. Borgohain and P. K. Das, "Morphological Analyzer for Manipuri: Design and Implementation", Applied Computing, 2004, pp. 123-129.

5. T. N Vikram and Shalini R Urs, "Development of prototype Morphological Analyzer for the South Indian Language of Kannada", Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, 2007, pp. 109-116.

The Latgalian Component in the Latvian National Corpus

Aleksey Andronov (St. Petersburg State University, Russia)
Everita Andronova (University of Latvia, Latvia)
baltistica@gmail.com

The year 2008 is marked by the initiative of the National Library of Latvia to create the Latvian National Corpus. The National Corpus will be a complex system of separate subcorpora, both synchronic and diachronic, monolingual and multilingual, developed by different partners such as the University of Latvia, the Institute of Mathematics and Computer Science, etc. The Corpus is expected to represent the Latvian language in full. According to the State Language Law, there are two standardized varieties of Latvian: the Latvian literary language and the Latgalian written language [VVL 1999]. Thus, the Latvian National Corpus cannot be considered complete without the Latgalian component, which is the topic of the present report.

What is Latgalian?

The Latgalian written language is a standardized variety of a language used by a part of Latvians living mostly in eastern Latvia (Latgale). Due to the history of the region the native population of Latgale differs a lot from the other Latvians — not only in language, but also in ethnography, cultural life and religion (Latgalians are mostly Roman Catholics, while other Latvians are mostly Protestants). There is no common agreement on the linguistic status of the language spoken in Latgale: it is considered either one of the three main dialects of the Latvian language or a separate Baltic language on equal terms with Latvian and Lithuanian [Brejdak 2006: 195]. Further in the text it will be referred to as just Latgalian, and its standardized variety as Standard Latgalian.

The linguistic distinction between standard Latgalian and standard Latvian is big enough to complicate mutual understanding. The differences are mainly found in the phonological system, as well as in the vocabulary, but certain important deviations exist also in morphology and syntax. Latgalian has a well established written tradition dating back to 1753 (assumingly more than 750 books have been published). There are several linguistic descriptions of Latgalian (practical grammars and dictionaries) reflecting deliberate work on developing a literary norm. Precise statistic evaluation is difficult, but according to the Research institute of Latgale some 150–200 thousands people are using Latgalian as an everyday means of communication (http://dau.lv/ld/latgale(english).html).

Why is the Latgalian corpus necessary?

In spite of considerable usage of Latgalian in fiction and mass media (including radio broadcasting and internet) it is paid no special attention to by the government and lacks linguistic research, which makes the language endangered in Latvia today. There are several courses on the Latgalian written language and its history in universities, but practical language is not taught at schools. Several linguistic resources and tools should be developed for Latgalian in order to raise its prestige and to ensure its development. A standard dictionary and grammar, school books and readers, a spell-checker and a morphological analyzer, together with a linguistic corpus are necessary. In light of the social linguistic situation a modern language corpus is perhaps the first task. However, the written tradition stretching more than 250 years back provides a good foundation for developing a diachronic corpus in the future. The Modern language period began together with the National awakening and the reestablishment of the Republic of Latvia in 1991, which gave a new impulse to the rebirth of Latgalian after its being almost neglected during the years of the Soviet rule.

Corpus creation problems

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

This restricted usage and lack of some functional styles or genres affect the size, representativeness and balance of the CMSLg. To start with, some 2–5 mill. running words can be processed in the corpus, although estimating the size is problematic before one has compiled a complete list of sources and studied their availability (issues of authorship, etc). Thus, composing a comprehensive bibliography of Latgalian publications is a prerequisite which can be a topic for a separate project. The main part of a corpus of modern language usually consists of texts from periodicals, but CMSLg is quite different in this respect. Seemingly, fiction (mainly original) will be the main source of data. Latgalian lacks or has a very little amount of medical, juridical, business and technical texts.

Data acquisition and processing in the CMSLg can be solved on the same grounds as in the Latvian part of the National corpus [Koncepcija 2005], while text selection and sampling procedures might differ. One should pay special attention to the input data quality. Many Latgalian texts are created just by mere phonetic transponations from Latvian according to sound correspondence rules, which gives an inadequate impression of the authentic lexicon, morphology and syntax. These texts cannot serve a source of the CMSLg.

Several practical grammars and a few dictionaries of Standard Latgalian published in the 20th century together with special commissions elaborating the literary norm are still not able to fight the lack of generally accepted orthography and a considerable variation in morphology and lexicon (to say nothing about pronunciation, which is not yet even touched by the literary standard). The problem of mixing odd elements coming from tradition and those promoted by linguistic authorities should be solved to ensure the automatic processing of the corpus. An intelligent search engine is necessary to identify the spelling variants (cf. recent orthography rules — LPN 2008).

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

Bibliogrpahy

Koncepcija 2005 — Latviešu valodas korpusa koncepcija / Latvijas Universitātes Matemātikas un informātikas institūts. — Rīga, 2005 (unpublished, available on request at: agentura@valoda.lv)

LPN 2008 — Latgaliešu pareizrakstības noteikumi / Tieslietu ministrijas Valsts valodas centrs. — Rīga — Rēzekne, 2008.

VVL 2000 — Valsts valodas likums // Latvijas Vēstnesis 428/433 (1888/1893), 21/12/1999. (The English translation is available at: http://isec.gov.lv/normdok/oflanglaw.htm)

Brejdak 2006 — A. B. Brejdak. Latgal'skij jazyk // Jazyki mira: Baltijskie jazyki / RAN. Institut jazykoznanija. — Moskva: Academia, 2006. — p. 193–213.