Icelandic Language Technology Ten Years Later

Eiríkur Rögnvaldsson
Department of Icelandic, University of Iceland
Árnagarði við Suðurgötu, IS-101, Reykjavík, Iceland
E-mail: eirikur@hi.is

Abstract

We describe the establishment and development of Icelandic language technology since its very beginning ten years ago. The ground was laid with a report from a committee appointed by the Minister of Education, Science and Culture in 1998. In this report, which was delivered in the spring of 1999, the committee proposed several actions to establish Icelandic language technology. This paper reviews the concrete tasks that the committee listed as important and their current status. It is shown that even though we still have a long way to go to reach all the goals set in the report, good progress has been made in most of the tasks. Icelandic participation in Nordic cooperation on language technology has been vital in this respect. In the final part of the paper, we speculate on the cost of Icelandic language technology and the future prospects of a small language like Icelandic in the age of information technology.

1. Introduction

Ten years ago, Icelandic language technology (LT) virtually didn't exist. There was a relatively good spell checker, a not so good speech synthesizer, and that was all. There were no programs or even individual courses on language technology or computational linguistics at any Icelandic university or college, there was no ongoing research in these areas, and no Icelandic software companies were working on language technology.

All of this has now changed and Icelandic language technology has been firmly established. In the fall of 1998, the Minister of Education, Science and Culture, Mr. Björn Bjarnason, appointed a special committee to investigate the situation in language technology in Iceland. Furthermore, the committee was supposed to come up with proposals for strengthening the status of Icelandic language technology. The members of the committee were Rögnvaldur Ólafsson, Associate Professor of Physics, Eiríkur Rögnvaldsson, Professor of Icelandic Language, and Þorgeir Sigurðsson, electrical engineer and linguist.

The committee handed its report to the Minister in April 1999 (Ólafsson et al., 1999). It took a while to get things going, but in 2000, the Icelandic Government launched a special Language Technology Program (Arnalds, 2004; Ólafsson, 2004), with the aim of supporting institutions and companies to create basic resources for Icelandic language technology work. This initiative resulted in several projects which have had profound influence on the field. In this paper, we will give an overview of this work and other activities in the field during the past ten years, and then speculate on the prospects of language technology in Iceland and the future of the language in the age of information technology.

The purpose of the paper is to show how the authorities, industry, and academia can fruitfully cooperate to build language technology resources and tools from scratch in a relatively short time for a relatively small budget. We think our experience may be useful for other small language communities where language technology is in its infancy and needs to be established.

2. Proposals of the LT Committee

In the report of the Language Technology Committee (Ólafsson et al., 1999), four types of actions were proposed in order to establish Icelandic language technology:

- The development of common linguistic resources that can be used by companies as sources of raw material for their products.
- Investment in applied research in the field of language technology.
- Financial support for companies for the development of language technology products.
- Development and upgrading of education and training in language technology and linguistics.

This has all been done, to some extent at least (Arnalds, 2004; Ólafsson, 2004; Rögnvaldsson, 2005). An overview of the most important resources, research projects and language technology products is given in section 3 below.

In the fall of 2002, the University of Iceland launched a new Master's program in Language Technology. This is a two-year interdisciplinary program (120 ECTS credits), and the applicants can either have a B.A. degree in the humanities (languages and linguistics) or a B.Sc. degree in computer science (or electrical or software engineering). Due to lack of resources, both financial and human, students were only admitted to the program twice, in 2002 and 2003.

Last fall, the program was relaunched, now as a joint program between the Department of Icelandic at the University of Iceland and the School of Computer Science at Reykjavik University. We hope that this cooperation will enable the two universities, in cooperation with the Nordic Graduate School of Language Technology (NGSLT), to offer sound and solid education, and to recruit enthusiastic students who will engage in research and development on Icelandic language technology.

In addition to this, a few Icelandic students have studied language technology abroad in recent years, and the first Icelandic Ph.D. in the field received his degree last year from the University of Sheffield (Loftsson 2007).

3. Priority tasks and their implementation

The above-mentioned report on Icelandic language technology (Ólafsson et al., 1999) stated the following:

> For Icelanders, the main aim must be that it should be possible to use Icelandic, written with the proper characters, in as many contexts as possible in the sphere of computer and communication technology. Naturally, however, they will have to adjust their expectations to practical considerations. To make it possible to use Icelandic in all areas, under all circumstances, would be an immense task. Therefore, the main emphasis must be put on those areas that touch on the daily life and work of the general public, or are likely to do so in the near future.

Following this statement, the Language Technology Committee proposed a list of priority tasks for Icelandic language technology during the following five years. Those tasks are listed here in italics at the beginning of each subsection, and in the text that follows, we try to estimate to what extent each task has been fulfilled (cf. also Arnalds, 2004; Ólafsson, 2004; Rögnvaldsson, 2005).

3.1 Software translation

The main computer programs on the general market (Windows, Word, Excel, Netscape, Internet Explorer, Eudora,...) should be available in Icelandic.

In 2004, an Icelandic version of Windows XP (including Internet Explorer) and Microsoft Office 2003 came on the market. These versions do not seem to suffer from any technical bugs, as was the case with the first translation of Windows (Windows 98) into Icelandic a few years earlier. However, the translations have not met with great success, and most people, except perhaps the older generation, seem to prefer the English version. The reason is probably that people had grown used to having these programs in English and see no reason for adopting the Icelandic version. An Icelandic translation of Windows Vista and Microsoft Office 2007 has just been finished, and it will be interesting to see whether these versions gain more popularity than their predecessors.

In addition to this, special interest groups have been formed in order to translate open-source software for GNU/Linux. Thus, there exists an Icelandic version of the KDE (K Desktop Environment; http://www.is.kde.org/), and the new Hardy Heron version of the Ubuntu operating system (www.ubuntu.com) is currently being translated.

3.2 Icelandic characters

It should be possible to use the Icelandic non-ASCII characters (áéíóúýðþæöÁÉÍÓÚÝÐÞÆÖ) in all circumstances: in computers, mobile telephones, teletext and other applications used by the public.

When this was written, the ISO 8859-1 standard, which includes all the above-mentioned characters, had already been in existence for a number of years. However, many TV sets lacked special Icelandic characters in teletext pages, and mobile phones could not show any non-ASCII characters since they used a 7-bit character table. Nowadays, most TV sets and mobile phones can show all Icelandic characters although there seem to be some exceptions. Thus, the situation has improved considerably during the last decade.

3.3 Morphological and syntactic parsing

Work should proceed on the parsing of Icelandic, with the aim that it should be possible to use computer technology to analyze Icelandic texts grammatically and syntactically.

The Language Technology Project funded three major projects in this area. The Institute of Lexicography received a grant for building a full-form morphological database of Icelandic (Bjarnadóttir, 2005). This database is still growing and now contains around 259,000 lexemes and 5.6 million inflectional forms (iceland.spurl.net/tunga/VO/). In another project at the Institute of Lexicography, three data-driven taggers of different types (TnT, MXPOST and fnTBL) were trained and evaluated on a manually tagged Icelandic corpus of 500,000 words (Helgadóttir, 2005). A commercial company, Frisk Software (www.frisk.is), also received a grant for developing an HPSG-based parser with the future aim of building grammar and style checking software for Icelandic (Albertsdóttir and Stefánsson, 2004). Unfortunately, this latter project has not been finished.

The Language Technology Committee (Ólafsson et al., 1999) mentioned two prerequisites for further progress in this field, which are listed in 3.3.1 and 3.3.2.

3.3.1 A balanced corpus

A large computerized text corpus including Icelandic texts of a wide variety of types should be established.

In 2004, the Institute of Lexicography received a grant from the Language Technology Program for building a balanced morphologically tagged corpus of Modern Icelandic (Helgadóttir, 2004). This corpus will contain 25 million words of different genres, including transcribed spoken language, and shall be finished later this year.

3.3.2 A semantically annotated lexicon

A grammatically and semantically annotated lexicon should be established.

This lexicon was meant to be something similar to the PAROLE/SIMPLE lexicon (http://www.ub.es/gilcub/SIMPLE/simple.html). No such lexicon has been built yet. However, many types of raw material for building a lexicon of this type do exist, especially in various collections and databases at the Institute of Lexicography, such as the ISLEX database which is being built and will comprise 50,000 entries for Icelandic and their equivalents in Danish, Norwegian, and Swedish (www.lexis.hi.is/islex-ohvefur/islex-meira.html).

3.4 Spelling and grammar checkers

Good auxiliary programs should be developed for textual work in Icelandic, i.e. for hyphenation, spell-checking, grammar correction, etc.

When this was written nine years ago (Ólafsson et al., 1999), we had the spell-checking program *Púki* from Frisk Software, which has now been improved with support from the Language Technology Program (Skúlason, 2004). In 2002, the Dutch company Polderland (www.polderland.nl) developed a spell-checking program for the Microsoft Office package. Furthermore, there exists an open source spell checker for Icelandic based on Aspell (aspell.net/), which can be used with GNU/Linux applications. These programs (as most spell checkers) are word-based, and hence cannot cope with many common spelling errors.

No grammar checking or style checking programs exist, but current work on a context-sensitive spell checker mentioned in Section 5 below will presumably lay the ground for a basic grammar checker.

3.5 Text-to-speech system

A good Icelandic speech synthesizer should be developed. It should be capable of reading Icelandic texts with clear and comprehensible pronunciation and natural intonation that is understandable without special training.

A formant-based Icelandic speech synthesizer was originally made around 1990 (Carlson et al., 1990) and improved around 2000. Even though this synthesizer was very useful for blind and visually impaired people, its quality was far from being satisfactory for use in commercial applications for the general public.

The last project that the Language Technology Program supported was a new text-to-speech system, which was made in cooperation between the University of Iceland, Iceland Telecom, and Hex Software. The system was trained by Nuance and uses their technology. People seem to agree that the quality is very good. The system came on the market last year and appears to be a success, especially due to a recently launched online service which uses the system for reading web pages and text entered by users (http://www.hexia.net/upplestur).

3.6 Speech recognition

Work should be done on speech recognition for Icelandic, the aim being to develop programs that can understand normal Icelandic speech.

In 2003, the University of Iceland and four leading companies in the telecommunication and software industry joined efforts to build an isolated word speech recognizer for Icelandic, with support from the Language Technology Program and in cooperation with ScanSoft (now Nuance) (Rögnvaldsson, 2004). The performance of the system has turned out to be quite satisfying; the recognition rate appears to be at least 97% (Rögnvaldsson, 2004). However, no attempts have been made to develop a system for recognizing continuous speech.

3.7 Machine translation

Work should be done on the development of translation programs between Icelandic and other languages, one of the aims being to simplify searches in databases.

The development in this area has been limited, although some isolated experiments have been made. Just recently, Stefán Briem, an independent researcher, has launched a free web-based service, which offers translations between Icelandic and three other languages (English, Danish, and Esperanto; www.tungutorg.is). Furthermore, the Icelandic Technical Development Fund has given a grant to a private company that works on translation software for translating from Icelandic to English, but this software has not been marketed yet and the status of its development is unclear. Iceland has also taken part in a Nordic project which aims at enabling multilingual web search (Dalianis et al., 2007).

4. Nordic cooperation

Since 2000, Icelandic researchers and policy makers have taken active part in Nordic cooperation on language technology. This participation has been of major importance in establishing the field in Iceland. From 2001-2004, the Nordic Language Technology Research Programme (Holmboe, 2005) funded language technology Documentation Centers in the five Nordic countries (www.nordoknet.org). At the end of 2004, the Icelandic center merged with the website www.tungutaekni.is, which the Language Technology Program had been running since its start in 2001. This website is now run by the ICLT (see Section 5 below). Thanks to the documentation center, we now have a good and accessible overview of people, projects, products, materials, companies, organizations, etc. having to do with Icelandic language technology.

Through the documentation center, we have also made contacts with several people and institutions in the Nordic and Baltic countries (cf. Fersøe et al., 2005). As a result of those contacts, Icelandic researchers have participated in several applications to Nordic and European funding bodies during the past few years. Even though most of these applications have not been successful, we have gained invaluable experience from taking part in them and cooperating with Nordic colleagues.

Another important aspect of the Nordic cooperation in language technology is the Nordic Graduate School of Language Technology (NGSLT, www.ngslt.org), funded by NorFA – now NordForsk (Nordic Research Board, www.nordforsk.org). The activities of the school started in 2004 and will run for five years. Even though the school is primarily intended for doctoral students, master's level students from Iceland have been admitted to the courses. This is absolutely crucial for the Icelandic universities, since they do not have the capacity to give the students high-quality education in language technology at home.

Icelandic researchers also take part in other Nordic and Baltic activities in the field, such as the newly established Northern European Association for Language Technology (NEALT, omelia.uio.no/nealt), and the bi-annual Nordic conferences of computational linguistics (NODALIDA). In 2003, the 14th NODALIDA conference was held at the University of Iceland in Reykjavík.

5. The price and prospects of Icelandic LT

After the Language Technology Program ended by the end of 2004, researchers from three research institutes (University of Iceland, Reykjavik University, and the Árni Magnússon Institute for Icelandic Studies) decided to join forces in a consortium called Icelandic Centre for Language Technology (ICLT), in order to follow up on the tasks of the Program. During the past three years, these researchers, who had been involved in most of the projects supported by the Language Technology Program, have initiated several new projects, three of which should be especially mentioned: *IceTagger*, a linguistic rule-based tagger (Loftsson, 2006, 2007), *IceParser*, a shallow parser (Loftsson and Rögnvaldsson, 2007; Loftsson, 2007), and a context-sensitive spell checker which shall be finished later this year. These programs are seen as a contribution to the establishment of a BLARK (Basic Language Resource Kit; cf. Krauwer, 2003) for Icelandic, and the group has made plans for the next steps towards that goal.

These projects have been partly supported by the Icelandic Research Fund and the Icelandic Technical Development Fund. However, much more money is needed in order to create a BLARK for Icelandic. The Language Technology Committee estimated that it would cost around one billion Icelandic krónur, about ten million Euros, to make Icelandic language technology self-sustained (Ólafsson et al., 1999). After that, the free market should be able to take over, since it would have access to public resources that would have been created for money from the Language Technology Program, and that would be made available on an equal basis to everyone who were going to use these resources in their commercial products.

Even though the Language Technology Program was very successful and had a great impact on the development of Icelandic language technology, the fact remains that its total budget over the lifespan of the program (2000-2004) was only 133 million Icelandic krónur (Ólafsson, 2004), or around 1.35 million Euros – that is, 1/8 of the sum that the committee estimated would be needed. It should therefore come as no surprise that we still have a long way to go. There are only 300,000 people speaking Icelandic, and that is not enough to sustain costly development of new products. It costs just as much to build language resources for Icelandic as for languages with hundreds of millions of speakers. Therefore, we feel it is extremely important to continue public support for Icelandic language technology for some time, in order to make the most out of the money that has been spent up to now, and utilize the knowledge and experience that researchers and companies have gained.

One way to do this would be to make more use of free/open source licenses, both for software and linguistic resources. It has recently been argued convincingly by several authors (cf., for instance, Forcada, 2006; Streiter et al., 2007; Alegria et al., 2008) that it is essential for minor/non-central/less-resourced languages to adopt open source policy with respect to LT resources in order to survive the Information Age.

Unfortunately, many Icelandic resources such as dictionaries and corpora are privately owned, either by commercial companies or individual authors or researchers, and it can be difficult and expensive, or even impossible, to get permission to use them even for research, not to mention for commercial applications. All grants from the Language Technology program were given with the condition that the resources developed would be accessible for anyone wanting to use them in language technology products. However, these resources are not distributed under an open source license and most of them are not free. Even though the license to use them is usually not very expensive, the license fee acts as a barrier for the use of these resources in LT research and development. It would obviously be beneficial for the future of Icelandic LT to implement open source policy, and this has recently been strongly advocated (Trosterud, 2008; Gíslason, 2008).

6. Conclusion - LT and the future of Icelandic

In this paper, we have demonstrated how joined efforts of the government, research communities, and commercial companies, enhanced by Nordic cooperation, have succeeded in establishing the basis for Icelandic language technology in a relatively short time.

When we try to estimate the importance of Icelandic language technology we must realize that information technology has become an important and integrated feature of the daily life of almost every single Icelander. If Icelandic cannot be used within information technology, speakers will be faced with a completely new situation, without parallels earlier in the history of the language. We will have an important area of the daily life of ordinary people where they cannot use their native language. How is that going to affect the speakers and the language community? What will happen when the native language is no longer usable within new technologies and in other new and exciting areas; in fields of innovation and creativity; and in areas where new job opportunities are offered? We don't have to think long about this scenario to see the signs of imminent danger.

But the need for native language technology is not, and should not be, only driven by people's wish to protect and preserve their language. It is equally – or even more – important to look at this from the user's point of view. Ordinary people should not be forced to use foreign languages in their everyday lives. They have the right to be able to use their native language anytime and anywhere within their language community, in all possible contexts. Otherwise, they will be linguistically oppressed in their own language community.

7. Acknowledgements

Thanks to three anonymous reviewers whose comments have influenced (and hopefully improved) the paper considerably, especially by suggesting the addition of paragraphs on BLARK and open source license and translation.

8. References

Albertsdóttir, M., and Stefánsson, S.E. (2004). Beygingar- og málfræðigreinikerfi [A System for Morphological and Syntactic Parsing]. In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 16-19.

Alegria, I., Arregi, X., Artola, X., Diaz de Ilarraza, A., Labaka, G., Lersundi, M., Mayor, A., and Sarasola, K. (2008). Strategies for Sustainable MT for Basque: Incremental Design, Reusability, Standardization and Open-source. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyderabad, India, pp. 59-64.

Arnalds, A. (2004). Language Technology in Iceland. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 41-43.

Bjarnadóttir, K. (2005). Modern Icelandic Inflections. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2005*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 49-50.

Carlson, R., Granström, B., Helgason, P., Thráinsson, H., and Jensson, P. (1990). An Icelandic Text-to-Speech System for the Disabled. In *Proceedings of ECART (European Conference on the Advancement of Rehabilitation Technology)*. Maastricht, the Netherlands.

Dalianis, H., Rimka, M., and Kann, V. (2007). Using Uplug and SiteSeeker to Construct a Cross Language Search Engine for Scandinavian. Paper presented at the workshop *The Automatic Treatment of Multilinguality in Retrieval, Search and Lexicography*, Copenhagen, Denmark, April 26. (people.dsv.su.se/~hercules/papers/scanduplug.pdf)

Fersøe, H., Rögnvaldsson, E., and de Smedt, K. (2005). NorDokNet – Network of Nordic Documentation Centres – Contacts to Future Baltic Partners. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2005*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 13-23.

Forcada, M.L. (2006). Open Source Machine Translation: an Opportunity for Minor Languages. *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, May 23. (www.mt-archive.info/LREC-2006-Forcada.pdf)

Gíslason, H. (2008). Gögn og gaman: jarðvegur nýþróunar í tungutækni [The Ground for Innovation in Language Technology]. Paper presented at the workshop *Á íslenska sér framtíð innan upplýsingatækninnar?* [Does Icelandic Have a Future within Information Technology?], Reykjavík, Iceland, March 7.

Helgadóttir, S. (2004). Mörkuð íslensk málheild [A Tagged Icelandic Corpus]. In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 67-71.

Helgadóttir, S. (2005). Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2004*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 257-265.

Holmboe, H. (2005). *Nordisk sprogteknologisk forskningsprogram 2000-2004. Epilog*. NordForsk, Oslo, Norway.

Krauwer, S. (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*, Moscow, Russia, pp. 8-15.

Loftsson, H. (2006). Tagging a Morphologically Complex Language Using Heuristics. In Salakoski, T., Ginter, F., Pyysalo, S., and Pahikkala, T. (Eds.), *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Proceedings*. Turku, Finland, pp. 640-651.

Loftsson, H. (2007). Tagging and Parsing Icelandic Text. Doctoral dissertation, Department of Computer Science, University of Sheffield, UK.

Loftsson, H., and Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. In Nivre, J., Kaalep, H-J., Muischnek, K., and Koit, M. (Eds.), *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*. Tartu, Estonia, pp. 128-135.

Ólafsson, R. (2004). Tungutækniverkefni menntamálaráðuneytisins [The Language Technology Program of the Ministry of Education, Science and Culture]. In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 7-13.

Ólafsson, R., Rögnvaldsson, E., and Sigurðsson, Þ. (1999). *Tungutækni. Skýrsla starfshóps* [Language Technology. Report of a Committee]. Ministry of Education, Science and Culture, Reykjavík, Iceland.

Rögnvaldsson, E. (2004). The Icelandic Speech Recognition Project *Hjal*. In Holmboe, H. (Ed.), *Nordisk Sprogteknologi. Årbog 2003*. Museum Tusculanums Forlag, University of Copenhagen, Denmark, pp. 239-242.

Rögnvaldsson, E. (2005). Staða íslenskrar tungutækni við lok tungutækniátaks [The Status of Icelandic Language Technology at the End of the Language Technology Program]. *Tölvumál*, February 24.

Skúlason, F. (2004). Endurbætt tillögugerðar- og orðskiptiforrit Púka [Improved Suggestions and Hyphenations in the Púki Spell Checker]. In *Samspil tungu og tækni*. Ministry of Education, Science and Culture, Reykjavík, Iceland, pp. 29-31.

Streiter, O., Scannell, K.P., and Stuflesser, M. (2007). Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers. To appear in *Machine Translation*. (borel.slu.edu/pub/mt.pdf)

Trosterud, T. (2008). Grammar-based Language Technology as an Answer to the Challenges Facing Icelandic and other Circumpolar Languages. Paper presented at the workshop *Á íslenska sér framtíð innan upplýsingatækninnar?* [Does Icelandic Have a Future within Information Technology?], Reykjavík, Iceland, March 7.

Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources

Heather Simpson*, Christopher Cieri*, Kazuaki Maeda*, Kathryn Baker†, Boyan Onyshkevych†

*Linguistic Data Consortium
University of Pennsylvania
3600 Market St., Suite 810, Philadelphia PA, 19104, U.S.A.
{hsimpson, ccieri, maeda}@ldc.upenn.edu
†U.S. Department of Defense

Abstract

The REFLEX-LCTL (Research on English and Foreign Language Exploitation) program, sponsored by the United States government, was a medium-scale effort in simultaneous creation of basic language resources for several less commonly taught languages (LCTLs). To address some of the gaps in language technologies and resources, and to spur new research in this area, two REFLEX-LCTL sites constructed language packs for 19 LCTLs, and distributed them to research and development also funded by the program. This paper will focus on the work done at the Linguistic Data Consortium (LDC). LDC created language packs for 13 out of the 19 languages: Amazigh (Berber), Bengali, Hungarian, Kurdish, Pashto, Punjabi, Tamil, Tagalog, Thai, Tigrinya, Urdu, Uzbek, and Yoruba. Discussed are the goals and reasoning behind the language choice and language pack construction, and more in depth on the human resource and technology challenges in creating these language packs.

1. Introduction

The past decade has seen increased interest across multiple disciplines in resource creation for a growing number of languages. The new languages of focus have been grouped under several terms, including minority languages, less commonly taught languages, less resourced languages and endangered languages. Each term encodes differences in traditions, goals and approaches. A researcher working on an endangered language may seek to document that language and reinvigorate its use while a researcher working in less commonly taught languages (LCTLs) may seek to enable basic linguistic technologies or build language-aware applications.

The REFLEX-LCTL (Research on English and Foreign Language Exploitation) program, sponsored by the United States government, was a medium-scale effort in simultaneous creation of basic language resources for several LCTLs. To address some of the gaps in language technologies and resources, and to spur new research in this area, two REFLEX-LCTL sites constructed language packs for 19 LCTLs, and distributed them to research and development also funded by the program. The data sites are: the Linguistic Data Consortium (LDC), and the Computing Resource Laboratory (CRL) of the New Mexico State University (NMSU). This paper will focus on the work done at LDC.

The LCTL language packs address three goals. The first is to enable porting of existing technologies to new languages by providing training data and component technologies such as part-of-speech tagging and named entity extraction.. The second goal is to seed new research specifically on achieving better performance with fewer resources and on simplifying the process of porting of technologies to LCTLs when needed. Finally, the third goal is for the community to test and refine the choice, size and nature of the resources, contained in the language packs.

This third goal is directly related to the work of institutions ELSNET and ELRA (Evaluations and Language Resources Agency) in their definition of the BLARK (Basic Language Resource Kit) matrices. LCTL language packs contain 15 deliverable components including 6 of the 9 text resources and tools in 4 of the 15 text-based modules listed in the current BLARK matrix (ELDA, 2008).

2. Overview of Created Resources

2.1. Languages

LDC (http://projects.ldc.upenn.edu/LCTL) created resources for 13 of the 19 REFLEX-LCTL languages. These are: Amazigh (Berber), Bengali, Hungarian, Kurdish, Pashto, Punjabi, Tamil, Tagalog, Thai, Tigrinya, Urdu, Uzbek, and Yoruba.

CRL (http://crl.nmsu.edu/say) created resources for: Amharic, Burmese, Chechen, Guarani (spoken in Paraguay and Argentina), Maguindanao (Phillipines) and Uighur (Xinjiang, China).

The choice of REFLEX-LCTL targets addresses a number of criteria while still fitting within a fixed budget. All meet the basic criteria of being significant in terms of the number of native speakers but poorly represented in terms of available language resources.

Some of the languages (Thai, Urdu) were chosen to exercise a resource collection paradigm in which raw text is available digitally in sufficient quantity; others (Amazigh, Guarani, Maguindanao) were chosen to force the program to deal with cases in which it certainly is not. The cluster of Indic languages (Bengali, Punjabi, Urdu) was chosen to give researchers the opportunity to experiment with bootstrapping systems from material in related languages. Amazigh, Hungarian, Pashto, Tamil, and Yoruba were chosen to take advantage of existing collaborations in order to reduce costs.

Finally there was a general desire to select languages that are quite different from each other and from well-resourced

languages in order to maximize the generality of our methods. As a group, the LCTL languages are linguistically and geographically diverse; they include the national languages of fourteen different countries, representing eleven major language families, in Central, South and Southeast Asia, Austronesia, North, East and West Africa, the Middle East, Eastern Europe and South America.

2.2. Contents of Language Packs

The evolution of the planning of the LCTL language packs followed a path that has become somewhat familiar. The early phase was characterized by an appreciation of the difficulty of the endeavor and a strict balance in the distribution of resources across languages. As the work progressed, optimism inspired by some early successes and recognition of the differences in supply and demand of resources in the LCTLs led to modifications in the resource plan. The volume goals for some languages increased and specifications were refined to make the end result more useful across a broad range of HLTs, by converting found data from the original form into XML formats that were more easily integrated.

To control costs, we planned to take advantage of as much online data as possible. To this end we implemented a series of "Harvest Festivals"; intensive half day sessions where the entire LDC LCTL team, along with native speaker informants, convened to search the web for useful resources for each deliverable. By combining native speakers, linguists, programmers, information managers and projects managers in the same room, we were able to reduce communications latency nearly to zero, brainstorm jointly, and rapidly build upon each other's efforts.

This approach was generally quite successful, especially for the text corpora and lexica, and led us to some of our most useful data. Ideally the Harvest Festival would be the first step in language pack creation when the hope is to use raw online resources. Although it was not always possible to make it the preliminary step, we conducted a Harvest Festival at some point in the project for all but two of the 13 languages.

2.3. Text Corpora

Monolingual text serves as a basis for all of the other resources in the language pack and allows for small scale language modeling. For most of the LCTLs, this corpus was created by identifying and harvesting available resources from the internet, such as news and weblogs in the target language. Any source specific tags were removed from the harvested text, and it was converted into a standard digital representation for the LCTL, typically UTF8 encoded Unicode, and then tokenized.

Parallel Text supports the induction of translation lexicons and serves as both training and test material for machine translation technologies. Parallel text may be found and sentence aligned, or created from monolingual text by sentence segmenting and then having humans translate each sentence of source into one or more sentences in the target language. Our original concentration was on utilizing found Parallel Text, but we were not able to find a substantial amount for many of the LCTLs.

Additionally, although there are fewer steps involved in the found text processing, the alignment step can prove exceedingly difficult if there are deficiencies in either the segmentation in the original data, or in the sentence segmentation tool used to process the data.

In the end, most of our Parallel Text was created through outsourcing translation of our harvested Monolingual text to translation agencies. About 85,000 tokens of the Parallel Text for each language is English-to-LCTL translation. The English source text is shared across all 13 Language Packs, which will allow for comparison between these languages.

2.4. Lexica

Bilingual Lexicons support a variety of technologies including translation, tagging, information extraction and translingual information retrieval. The initial goal for this project was a lexicon, found or created, of at least 10,000 lemmas that included glosses and parts of speech. For most of the LCTLs, we were able to consult existing lexica, either digital or printed, to provide basic data for a subset of the lexical entries; however, in all cases we needed to process them substantially before they could be used efficiently. Processing steps included checking, normalizing and adding parts of speech and glosses, adding entire entries and removing irrelevant entries.

2.5. Tools for Conversion/Segmentation

The goal for this project was to include whatever encoding converters were needed to convert all of the raw text and lexical resources collected or created into the standard encoding selected for that LCTL.

Dividing text into individual sentences is a necessary first step for many processes including the human translation that dominated much of our effort. Simple in principle, LCTL sentence segmentation can prove tantalizingly complex. Our goal was to produce a sentence segmenter that accepts text in our standard encoding as input and outputs segmented sentences in the same encoding.

Word segmentation, or tokenization, is also relatively challenging for many LCTLs. Our goal for this project was to find or develop tokenizers that would produce word lists from texts in our standard format.

2.6. Annotated Corpora and Taggers

In order to support downstream processing, we also set out to produce three sets of internally coordinated resources: a part-of-speech tagger and tagged text, a morphological analyzer and tagged text and a named entity tagger and tagged text.

The project included the specific requirement that the morphological analyzer use the same tagset as the bilingual lexicon. Over time it became obvious that coordination among all of these resources was desirable and the work could be done most efficiently at the data sites. Unfortunately, we never found resources with this level of coordination. As a result we invested considerable time in creating or revising whatever resources we found for entity, part-of-speech, or morphology tagging. We found that at least 60,000 tokens of part-of-speech tagged text was the optimal amount for training our tagger, and we had to create this in-house

for almost every language. The named entity tagged text was also created in-house for all but the three outsourced languages.

2.7. Name Transliterators

The spelling of person names, particularly those foreign to the language under study, exhibit wide ranging variation in digital text and constitute a large percentage of the out-of-vocabulary terms in any HLT. To partially address this problem, we set out to create a personal name transliterator for each LCTL.

2.8. Grammatical Sketches

Finally, in order to identify for technology developers the challenges specific to the LCTLs, we undertook to create Grammatical Sketches for each. These are short outlines, approximately 50 pages, of the features of the written language and were based on existing grammars and experiences garnered in the work described above. The target audience included the other research groups participating in the REFLEX program, HLT developers who could be expected to have an understanding of basic concepts in linguistics.

2.9. Summary of LCTL Language Packs

We have completed a Language Pack for each of the 13 LCTL languages. 10 of them met our original requirements for project deliverables. Three of the Language Packs, Yoruba, Tigrinya, and Berber, fall short of our original requirements for some deliverables though they meet secondary requirements for others. Where these Language Packs do not meet original requirements, it was typically because the extreme dearth of resources existing for those languages made it impossible to do so given timeline and cost restraints. Table 1 and Table 2 summarize the contents of the Language Packs.[1]

Some of the Language Packs have already been distributed to REFLEX program members. Others are being held in reserve for possible use in technology evaluations. For example the Urdu Language Pack will be used in the NIST Open-MT evaluation campaign in 2008. Once a Language Pack has been exposed, it will be placed in the LDC publication queue for future release through the usual mechanisms.

3. Challenges and Solutions Toward Efficient Collaboration

3.1. Collaboration with Trained Researchers

As mentioned above, the extreme lack of available resources for Yoruba, Tigrinya, and Berber made it impossible for us to complete our requirements for some deliverables within the project's original time and budget.

For Yoruba and Berber, we found there simply was not enough harvestable digital text written in those languages to meet our Monolingual text requirement. We compensated for the lack of available Monolingual text by creating much of the data ourselves or under contract.

In the case of Yoruba, printed newspapers were physically collected and sent to us from Nigeria, which we then sent out to an outside agency to manually keyboard into digital text. The resulting corpus comprises 45% of our total Monolingual text for Yoruba.

In the case of Berber, we relied heavily upon our collaboration with the Institut Royal de la Culture Amazighe (IRCAM), in Morocco. IRCAM is working to develop and promote literacy and use of the Amazighe language. Two IRCAM researchers were able to come to LDC for a month, and shared their expertise and their resources with us. We were able to create tools to provide encoding conversion between IRCAM's standardized Latin-based transliteration of Berber, several other Latin-based transliterations, and Tifinagh, which we shared with IRCAM.

We also worked with Lori Levin at Carnegie Mellon University to help create our English-to-LCTL source text. She provided us with Elicitation Corpus, which she and her team specifically designed to elicit lexical distinctions in translations that do not occur in English (Alvarez et al., 2006).

Three of our LCTL Language Packs, Hungarian, Uzbek, and Kurdish, were entirely outsourced to the Media Research Centre at Budapest University of Technology and Economics (BUTE). This had the advantage that the team at BUTE was already working on or had access to many of the resources required for the language packs.

3.2. Working with Non-Specialist Native Speakers

We were dependent on finding native speaker assistance to create our annotated corpora and help identify harvestable online resources for most of the LCTL languages. Intensive recruiting efforts were conducted for native speakers of each non-outsourced LCTL language. Our recruiting strategy utilized such resources as online discussion boards and student associations for those language communities, and we were also able to capitalize on the diversity of the student/staff body of our host organization, the University of Pennsylvania, to recruit some native speakers internally.

We received a relatively high level of interest from most of our online advertising, from native speakers who seemed very excited that research attention was being paid to their languages. However, as might be expected, most of our respondents were not local to the Philadelphia area, and many were international. Though we did have support for remote work on some of our project tasks (as described in the Software Tools section below), we did not have the infrastructure to support complete outsourcing of annotation tasks to independent contractors. The creation of more comprehensive guidelines for non-specialist native speakers, and porting of more tasks into annotation tools such as the Annotation Collection Kit Interface (ACK), would perhaps make this a feasible option for a future effort of this kind.

We did find help from in-house native speakers all 10 non-outsourced languages. However, Berber and Yoruba were assisted by trained researchers who had limited time to spend on our particular needs, and our single Tigrinya native speaker informant also had time constraints. This resulted in a negative effect on completion of the Parallel Text, Part-of-Speech Tagger, and Named Entity Annotation

[1] The numbers represent the number of tokens.

	Large Languages		Small Languages				
	Urdu	Thai	Bengali	Tamil	Punjabi	Hungarian	Yoruba
Mono Text	14,804,000	39,700,000	2,640,000	1,112,000	13,739,000	1,414,000	363,000
Parallel Text (L ⇒ E)	1,300,000	694,000	237,000	308,000	221,000	70,000	
Parallel Text (Found)	947,000	1,496,000	243,000		230,000	2,338,000	78,600
Parallel Text (E ⇒ L)	65,000	65,000	65,000	65,000	65,000	65,000	65,000
Lexicon	26,000	232,000	482,000	10,000	108,000	182,400	128,200
Encoding Converter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sentence Segmenter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Word Segmenter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
POS Tagger	Yes	Yes	Yes	Yes	Yes	Yes	Yes
POS Tagged Text	5,000	5,000		59,000			
Morphological Analyzer	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Morph-Tagged Text	11,000			144,000			
NE Annotated Text	233,000	218,000	138,000	132,000	157,000	269,000	189,000
Named Entity Tagger	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name Transliterator	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Descriptive Grammar	Yes	Yes	Yes	Yes	Yes	Yes	

Table 1: LCTL Language Packs (Phase 1)

	Small Languages					
	Tagalog	Tigrinya	Pashto	Uzbek	Kurdish	Berber
Mono Text	774,000	617,000	5,958,000	790,000	2,463,000	181,000
Parallel Text (L ⇒ L)	203,000	139,000	180,000	206,000	163,000	26,000
Parallel Text (E ⇒ L)	65,000	65,000	65,000	65,000	65,000	65,000
Lexicon	18,000	0	10,000	25,400	6,500	Active
Encoding Converter	Yes	Yes	Yes	Yes	Yes	Yes
Sentence Segmenter	Yes	Yes	Yes	Yes	Yes	Yes
Word Segmenter	Yes	Yes	Yes	Yes	Yes	Yes
POS Tagger	Yes	Yes	Yes	Yes	Yes	
POS Tagged Text						
Morphological Analyzer	Yes	Active	Yes	Yes	Yes	Active
Morph-Tagged Text						
NE Annotated Text	136,000	123,000	165,000	93,000	62,000	60,000
Named Entity Tagger	Yes	Yes	Yes	Yes	Yes	Yes
Name Transliterator	Yes	Yes	Yes	Yes	Yes	Active
Descriptive Grammar	Yes	Yes	Yes	Yes	Yes	No

Table 2: LCTL Language Packs (Phase 2)

deliverable requirements for those three languages. Though we were able to find translation agencies who could deliver Parallel Text for Yoruba and Berber, turn-around and cost precluded us from meeting our goal quantities of text corpora.

3.3. Software Tools

3.3.1. Overview

In creating the language resources included in the LCTL language packs, we developed a variety of software tools for helping humans provide data needed for the resource creation efforts. The following are some of the examples.

3.3.2. Annotation Collection Kit Interface (ACK)

Probably the most important of the annotation tools for the LCTL project was the Annotation Collection Kit Interface (ACK), developed by LDC (Maeda et al., 2008). ACK facilitates remote creation of multiple types of text-based annotation, by allowing individual "kits" to be uploaded onto a specific server URL which any remote user can access. Using this tool we were able to support native speaker annotators working on part-of-speech (POS) annotation from Thailand.

When annotators make judgments in ACK, they are stored in a relational database. The results can be downloaded in CSV (comma-separated value) or XML format, so anyone with secure access to the server can easily access the results.

Anyone with a relatively basic knowledge of a scripting language such as Perl or Python would be able to create the ACK annotation kits. They are essentially a set of data corresponding to a set of annotation decisions in the form of radio buttons, check boxes, pull-down menus, or comment

fields, so they are currently limited in scope, but creative use of this format can yield a great deal of helpful types of annotation.

For POS annotation, the annotators were given monolingual text from our corpus, word by word, in order, and asked to select the correct part of speech for that word in context. We also used ACK to add/QC glosses and parts of speech for lexicon entries and do morphological tagging, and many other tasks that require judgment from native speaker.

3.3.3. Named Entity Annotation Tool

LDC also developed an named entity (NE) annotation tool, called SimpleNET (Maeda et al., 2006). SimpleNET requires almost no training in tool usage, and annotations can be made with the keyboard or the mouse. The NE annotated text in the LCTL language packs was created with this tool.

3.3.4. POS and NE Taggers

The annotated text created with ACK and SimpleNET was used in the development of the part-of-speech (POS) taggers and named entity (NE) taggers included in the language packs. Most of these POS and NE taggers were created using a common development infrastructure, which was centered around the MALLET toolkit (McCallum, 2002). By using the common infrastructure, we minimized the duplicated effort in creating these tools.

3.3.5. Encoding Conversion Tools

We encountered difficulties relating to the lack of usage of standardized orthography for some of the LCTL languages, as mentioned earlier. Our Berber Encoding Converter supports conversion between 6 different romanizations/encodings, and there are still more out there that we did not have time or resources to include. There would have been more Berber Monolingual Text in our corpus if we had had the ability to decipher every idiosyncratic encoding and add to the converter.

4. Conclusion

Despite numerous challenges, we have successfully created large, and in some cases unique resources for each of the 13 LCTL languages that we hope will provide valuable support for research and technology development for these previously under-supported languages. At least some of the challenges we have undergone would surely be encountered during a similar effort with different LCTLs. We hope that others may be able to learn from our mistakes and from our solutions to make their project a more successful endeavor in HLT development for under-resourced languages.

5. References

Alison Alvarez, Lori S. Levin, Robert E. Frederking, Simon Fung, and Donna Gates. 2006. The MILE corpus for less commonly taught languages. In *Proceedings of HLT-NAACL 2006*.

ELDA. 2008. BLARK Resource/Modules Matrix. From Evaluations and Language Resources Distribution Agency (ELDA) web site http://www.elda.org/blark/matrice_res_mod.php , accessed on 2/23/2008.

Kazuaki Maeda, Haejoong Lee, Julie Medero, and Stephanie Strassel. 2006. A new phase in annotation tool development at the Linguistic Data Consortium: The evolution of the Annotation Graph Toolkit. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.

Kazuaki Maeda, Haejoong Lee, Shawn Medero, Julie Medero, Robert Parker, and Stephanie Strassel. 2008. Annotation tool development for large-scale corpus creation projects at the Linguistic Data Consortium. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Andrew Kachites McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Building resources for African languages

Karel Pala[1], Sonja Bosch[2], Christiane Fellbaum[3]

[1]Faculty of Informatics, Masaryk University, Botanicka 68a, 602 00 Brno, Czech Republic
pala@fi.muni.cz
[2]Department of African Languages, University of South Africa, PO Box 392, 0003 Pretoria, South Africa
boschse@unisa.ac.za
[3]Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544, USA
fellbaum@princeton.edu

Abstract

We report on work towards the creation of African Languages WordNet, comprised of interlinked semantic networks in several African languages that are known to have limited language resources. Adding these languages to the WordNet family will enable NLP applications for each of the languages in isolation. Moreover, linking the African Wordnets to one another and to the many global WordNets will make crosslinguistic information retrieval and question answering possible, and significantly aid machine translation. In this paper it is demonstrated how collaborative work between people, using existing tools, can contribute to the building of large text corpora and subsequently address the challenge of limited availability of language resources. The long term aim is the development of aligned WordNets for Bantu languages spoken in South Africa as multilingual knowledge resources which could be extended to include a wide variety of related languages from other parts of Africa.

1. Introduction

Many Natural Language Processing (NLP) applications that require word sense disambiguation rely on WordNet as an essential lexical resource. WordNets have been created for dozens of languages, primarily those spoken by large populations and in technologically advanced countries where funding for resource development is relatively easily available (Miller, 1995; Fellbaum, 1998; Vossen, 1998).

We report on work towards the creation of the African Languages WordNet, comprised of interlinked semantic networks in several African languages. Adding these languages to the WordNet family will enable NLP applications for each of the languages in isolation. Moreover, linking the African Languages Wordnets to one another and to the many global WordNets will make crosslinguistic information retrieval and question answering possible, and significantly aid machine translation. WordNets have also been shown to be very useful for language learning.

Besides these practical considerations, there are many purely linguistic motivations for building African languages WordNets. WordNets currently exist in some 50 languages, many of them typologically and historically unrelated. But no African language, and no language with the particular linguistic features of the languages of South Africa, has developed a WordNet. Doing so will force a new and broader perspective of the lexicon and will enrich our understanding of this component of human language.

1.1 WordNet

All present WordNets are modelled on the Princeton WordNet developed in the mid-1980s (Miller, 1995; Fellbaum, 1998). A WordNet is a large semantic network where words and groups of words are interlinked by means of lexical and conceptual relations represented by labelled arcs. Like a dictionary, WordNet's units are words, and its aim is to provide semantic information about words. This information is given in a form resembling a thesaurus, though the network of words is more rigorously structured than in a thesaurus.

WordNet's building blocks are unordered sets of synonymous words and phrases, dubbed "synsets". Synset members are denotationally equivalent and substitution of a synset member by another does not change the truth value of the context, though stylistic infelicity may result from such substitution. WordNet provides some information on how synset members are used; register tags are given ("colloquial," "slang" etc.), and example sentences accompany most synsets illustrating the synonyms' usage.

A synset is said to lexically express a concept. Examples of synsets are {mail, post}, {hit, strike} and {small, little}. All synsets further contain a brief definition. A domain label (sports, medicine, biology) marks many synsets.

Concepts expressed by nouns are densely interconnected by the hyponymy relation (or hyperononymy, or subsumption, or the ISA relation), which links specific concepts to more general ones. For example, the synset {gym shoe, sneaker, tennis shoe} is a hyponym, or subordinate of {shoe}, which in turn is a hyponym of {footwear, footgear}, etc. Hyponymy builds hierarchical "trees" up to fifteeen layers deep with increasingly specific "leaf" concepts growing from an abstract "root".

Crosslinguistic WordNets share the same structure and can be interlinked, allowing for the identification of

equivalent words and synsets and enabling translation. The technical instrument for interlinking and thus capturing multilinguality of WordNets is Interligual Index (ILI) developed in the EuroWordNet project (Vossen, 1998). Languages differ in their lexical make-up, and words (or entire areas of the lexicon) that are expressed in one language may be „missing“ in another. WordNet's systematic structure identifies both crosslinguistic matches as well as mismatches. WordNet construction is therefore a way to compare not only the lexicons of African languages with one another but also with those of dozens of other languages.

2. WordNet training workshop

In the light of the crucial contribution of global Wordnets to NLP, an infrastructure for WordNet development for African languages was created by means of a week long training workshop. The aim of the workshop was to develop a platform for WordNet development for African languages.

Seed research funding for the project was obtained from the Meraka Institute (2007) to enable facilitation by international experts namely Christiane Fellbaum (Wordnet, 2006) as one of the pioneers of WordNets, Piek Vossen (1998) as project coordinator of the EuroWordNet project, and Karel Pala as participant in the Czech WordNet (cf. Pala & Smrž, 2004) and developer of the lexicographer's editing tools DEBVisDic in particular (cf. DEBVisDic Manual, 2008). The project afforded linguists, translators and lexicographers representing the 9 official African languages in South Africa, as well as computer scientists, the opportunity of high level multi-disciplinary training.

The nine official African languages of South Africa are Zulu (isiZulu), Xhosa (isiXhosa), Swati (siSwati), Ndebelele (isiNdebele), Venda (Tšhivenda), Tsonga (Xitsonga), Southern Sotho (Sesotho), Northern Sotho (Sesotho sa Leboa) and Tswana (Setswana). These languages all belong to the Bantu language family and are grammatically closely related. The Nguni languages, i.e. Zulu, Xhosa, Swati and Ndebele form one group. The Sotho languages, viz. Southern Sotho, Northern Sotho and Tswana form another group with Venda and Tsonga being more or less on their own

2.1 Accomplishments

The facilitators each gave lectures on the area of their speciality that related to the methods, theory, and practical steps for WordNet construction. Christiane Fellbaum (Princeton) lectured on the design of WordNet and invited the participants to reflect on specific questions from the viewpoint of their native languages. A number of hands-on exercises were carried out, where the participants built "toy" WordNets for their languages. Piek Vossen (Amsterdam) lectured on his experiences with EuroWordNet, where he introduced some fundamental changes to the original Princeton WordNet.

Karel Pala (Brno) introduced his editing tool and the participants trained on it under his guidance.

The user manual for the editing tool DEBVisDic was updated after contributions made by workshop participants regarding the user friendliness of the software tool (cf. DEBVisDic Manual, 2008). Extensive reading matter was distributed among participants before, during and after the workshop. A CD ROM containing the reading matter as well the presentations of the facilitators was handed to participants after the workshop.

The African languages WordNets are still in a conceptualisation phase although experimental work on noun and verb synsets has begun (cf. le Roux et al., 2008).

2.2 Challenges specific to African languages

During the workshop various challenges for WordNets specific to the African languages were identified, the first and foremost being the morphological complexities of agglutinative languages centred around a noun class system and roots. WordNets for such languages pose novel challenges, especially with respect to the concept of "word," which must be defined to determine synset membership. The conjunctive and disjunctive orthographies of the various language groups contribute to this challenge. For example, the orthographic word *ngiyabathanda* (“I like them”) in Zulu corresponds to four orthographic words or separate orthographic entities in Northern Sotho, viz. *ke a ba rata* (“I like them”).

A feature particular to the Bantu languages is the POS known as “ideophone”, a term proposed by Doke (1935:118) for a word category which describes a predicate, qualificative or adverb in respect to manner, colour, sound, smell, action, state or intensity. In contrast to the linguistic word in the Bantu languages, which is characterised by a number of morphemes such as prefixes and suffixes, as well as a root or stem, the ideophone consists only of a root which simultaneously functions as a stem and a fully-fledged word. The following are some Zulu examples:

> *Bathula bathi **du*** (They kept **completely quiet**)
> *Ingilazi iwe yathi **phahla** phansi* (The glass fell **smashing** on the floor)
> ***Kubomvu klubhu!*** **(It is blood red)**

These can be accommodated in the WordNets in the following way. Often, they can map to the "canonical" parts of speech (nouns, verbs, adjectives, adverbs) in the existing WordNet. For example, the workshop participants cited over 200 verbs denoting manners of motion, many encoding ideophones. These can be entered as manner-specific subordinates ("troponyms") in the WordNets, and, wherever possible, mapped to the corresponding manner-of-walking verbs in other languages. Similarly, colour

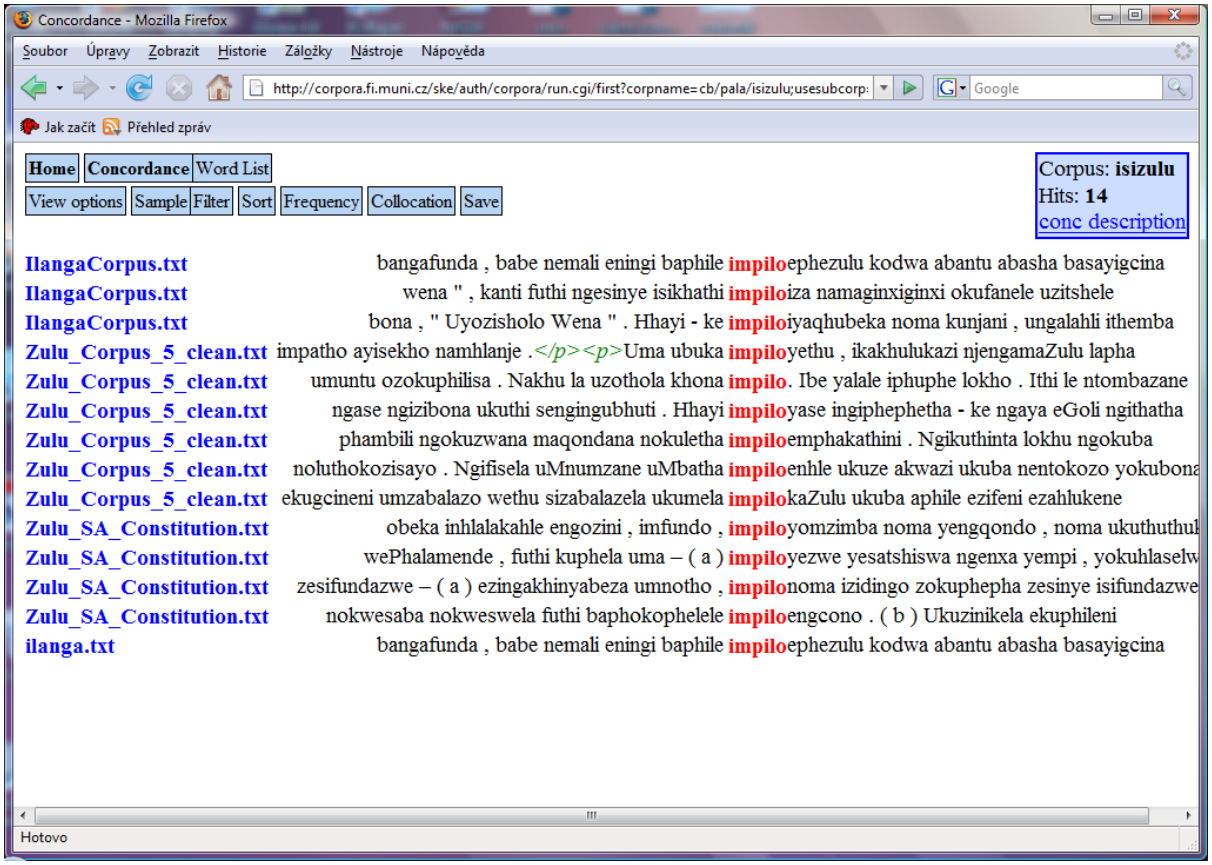

Figure 1: A concordance list from an experimental Zulu corpus

words involving ideophones can be linked to colour words (adjectives and nouns) in existing the WordNet. Wherever necessary, basic ideophones or words including an ideophone will be accommodated as a new lexical category.

In the course of the workshop, some noun and verb synsets were created in the various African languages. Secondly, limited availability of electronic language resources, such as large corpora, parallel corpora, electronic dictionaries and machine-readable lexicons was identified as a stumbling block, particularly in comparison to the generous availability of language recources for other WordNets in the world. Workshop participants relied on both monolingual and bilingual dictionaries available in their particular languages for semantic information. A need for corpus compilation was expressed, as corpora enable researchers to find and examine a particular word in context. Corpus data captures language use by many people in different contexts over time, and thus corpus data is more reliable than introspection by a few linguists or lexicographers. Corpus data is vital for determining the sense inventory of a language. General corpora are available for all nine mentioned African languages at the University of Pretoria (University of Pretoria, 2003), but with access restrictions which involve on site computer processing of the corpus and down-loading only the results of the analyses. The sizes of the various corpora currently range from 1 million tokens for Ndebele to 5.8 million tokens for Northern Sotho.

3. Working collaboratively to build text corpora with few existing language resources

In order to address the challenge of limited availability of electronic language resources, this section demonstrates how collaborative work between people, using existing tools, can contribute to the building of large text corpora.

3.1 Building corpora

Tools exist that allow us to almost automatically build text corpora for any language, and for African languages in particular. The only condition is the availability of a collection of texts in plain format. To demonstrate this we used the Corpus Builder tool developed in the NLP Centre at the Faculty of Informatics Masaryk University (Baroni et al., 2006) and created a small Zulu text corpus containing approximately 80 000 tokens in a very short time (approx. 30 minutes).

To visualise concordance lists we used a corpus manager tool, Manatee/Bonito2 (Rychlý, 2000). This tool is also integrated with the Corpus Builder, thus the newly built corpus can be immediately inspected.

If appropriate collections of texts are available, for instance from Web pages that are freely accessible the corpus can be enlarged in next to no time.

When larger plain text corpora are built, the need arises to tag them. Thus, the next step is to build taggers and tagsets for African languages. Work on taggers for Northern Sotho is reported on in Prinsloo and Heid (2006) and de Schryver and de Pauw (2007). This is a relatively independent enterprise but it will contribute to enriching electronic African language resources considerably. For this purpose automatic morphological analysis and analysers are being developed (cf. Bosch et al., 2006). Without morphological analysers, building high quality mono- and multilingual lexical resources including WordNets and other lexical databases will not be possible. This applies especially to the Bantu languages, the rich morphology of which calls for these tools to be developed as soon as possible.

One further tool that needs to be mentioned is the BootCat (Baroni et al., 2006) which allows one to build rather small domain corpora directly from Web pages, if they are at one's disposal.

Finally, it should be remarked that the described way of building corpora can be applied to all African languages mentioned above since the techniques are language independent.

3.1.1. Corpora and WordNets

Experience with building WordNets in the Balkanet project (Pala & Smrž, 2004) has shown that the evidence obtained from corpora can be profitably exploited for making them empirically more reliable and descriptively adequate. This means that corpora are very helpful for compiling the representative list of synsets on the ground of frequency considerations obtained from corpora. It also means that the evidence obtained from corpora is useful for making decisions about the senses that have to be associated with the respective synsets. This applies fully to all the considered African languages – we have shown that corpora for these languages can be built cost effectively by using the tools that are easily accessible.

It should be noted that corpora exploited for developing WordNets have to be of a general nature. In other words, texts from which such corpora are created should come either from newspaper resources or they can be appropriately selected novel texts (in the Balkanet project it was the novel 1984 by G. Orwell which existed as a parallel corpus for all Balkanet languages). Specialized corpora containing specialized technical or terminologically oriented texts are not appropriate.

Obviously, the next step would be to try to build and then exploit parallel corpora for Bantu languages. This will be useful not only for developing African WordNets as such but also for promoting the cultural relations between them. More information about the mentioned corpus tools can be obtained from: http://www.textforge.cz/products, or http://www.fi.muni.cz/~thomas/corpora/cb_text_upload.htm

3.2 A tool for building WordNets – DEBVisDic

For editing and browsing WordNets one needs a tool that can serve this purpose. DEBVisDic (Horák et al, 2006) is a tool for building WordNets and it works with lexical data in XML format. It is a browser and editor exploiting client/server architecture so that more developers and/or lexicographers can work on their WordNets simultaneously. This is an important requirement for the people working on African languages WordNets, since the lack of resources calls for constant sharing of information. The tool is built on the DEB (Dictionary Editing and Browsing) platform and is equipped with all features for supporting the linguistic work on WordNets.

DEBVisDic uses a versatile interface that allows the user to arrange the work without any limitations. It displays the following main functions:

- multiple views of multiple WordNets (multilingual view)
- freely defined text views
- synset editing and introducing semantic relations between them
- building and visualising hypero-hyponymic trees
- query result lists
- plain XML view of a synset
- synchronization of the synsets
- inter-dictionary linking
- consistency checks and journaling
- user configuration ensuring exchange of data
- entry locking for concurrent editing
- links display preview caching (speeds up the processing)

Presently, DEBVisDic is being supplemented with several other features that are currently accessible only as separate tools or resources. This functionality includes:

- link to a morphological analyser (for languages, where it is available)
- connection to language corpora, including Word Sketches statistics (for languages with accessible Word Sketch Engine, (cf. Kilgarriff et al., 2004).
- access to any electronic dictionaries stored in XML format within the DEB server
- searching for literals within encyclopaedic web sites.

Existing WordNets as well as those under development are stored in the DB XML database which is well suited for processing complicated XML structures. Simple processing of the data (like export or import of the whole dictionary) is not a problem as the whole English WordNet export (over 100.000 entries) takes less than 1 minute. However, searching for values of specific subtags can take several seconds in such a large dictionary even when indexes are used. We are currently working on several solutions for this,

which include link caching, specific DB XML indexing and also experimenting with a completely different database backend.

The DEBVisDic client is continuously being developed in the NLP Centre, Faculty of Informatics Masaryk University according to specific needs of running projects. These needs are determined to a great extent by the people doing the lexicographic work in the projects. We can mention the work on the PolishWordNet (Pala, Vetulani et al., 2007) and on the Dutch Cornetto project (Vossen, 2007). The African languages are accommodated here as well, as mentioned previously. The DEBVisDic tool will be made available freely and is to be installed at the University of South Africa for the use of the WordNet builders in a next stage of the project.

More information about the DEBVisDic tools can be found in DEBVisDic Manual (2008) (see the References).

3. Future work

The long term aim of this project is the development of aligned WordNets for African languages spoken in South Africa (i.e. languages belonging to the Bantu language family) as multilingual knowledge resources which could be extended to include a wide variety of related languages from other parts of Africa.

The construction of WordNets in a number of African languages presents exciting prospects since these languages unfortunately have not yet been as widely studied as European and Asian languages. At the same time, their sophisticated properties are of great linguistic interest. Building WordNets will necessitate the careful investigation of linguistic phenomena like classifiers that have not yet been explored in the context of non-Asian languages and may force a re-examination and broadening of the WordNet structure. Undoubtedly, the crosslinguistic study of lexicalization patterns, an interest of the GWA (Global WordNet Association), will benefit greatly from the addition of the new perspectives afforded by African languages. Like the original Princeton WordNet as well as WordNets in many other languages, African WordNet will be made freely and publicly available. Information is available on the Global WordNet website (http://www.globalwordnet.org).

Needs that have been identified for the successful continuation of the project, are the following:

Language resources: corpora and (electronic) dictionaries

Hardware: laptops for lexicographers and coders

Human resources: database manager with appropriate technical training

Modules for future integration into comprehensive NLP systems: POS-taggers, morphological analysers and syntactic parsers.

Finally, such research and development would depend on the commitment of linguists, translators and lexicographers to continue the work begun with great enthusiasm, the co-operation of numerous language institutions, the availability of a variety of language resources as well as further financial support following the seed research funding.

4. Conclusion

Follow-up research funding by the National Human Language Technology Network of the Meraka Institute for the development of African languages WordNets has just been announced. WordNets consisting of either 10 000 synsets each for four of the above mentioned languages (two Nguni and two Sotho languages), or two Wordnets (for one Nguni and one Sotho language), containing 20 000 synsets each, will be developed. These decisions will be taken during the planning phase of the project and will depend on the availability of resources for the involved languages. Where deemed necessary and possible, international collaborators will also be involved. This will extend current networks of collaboration, and will also extend the knowledge-base of HLT in South Africa.

5. Acknowledgement

The African Languages Wordnet Project (2006-2007) was funded by the National Human Language Technology Network of the Meraka Institute (South Africa).

This work has also been partly supported by the Academy of Sciences of Czech Republic under the project 1ET200610406 and by the Ministry of Education of Czech Republic within the Center of Computational Linguistics Project LC536.

6. References

Baroni, M., Kilgarriff, A., Pomikálek, J., Rychlý, P. (2006). WebBootCat: a Web Tool for Instant Corpora. In Proceedings of the Twelfth EURALEX International Congress, Turin, 2006, pp. 123--132.

Bosch, S., Jones, J., Pretorius, L., Anderson, W. (2006). Resource Development for South African Bantu Languages: Computational Morphological Analysers and Machine-Readable Lexicons. In Proceedings on the Workshop on Networking the Development of Language Resources for African Languages 5th International Conference on Language Resources and Evaluation, 22 May 2006, Genoa, Italy, pp. 38--43.

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

DEBVisDic Manual. (2008). Available at: http://nlp.fi.muni.cz/trac/deb2/wiki/DebVisDicManual. Accessed on: 22 February 2008.

Doke, C.M. (1935). Textbook of Zulu Grammar. Johannesburg: University of the Witwatersrand Press.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Horák A., Pala, K., Rambousek, A., Povolný, M. 2006. First version of new client-server wordnet browsing and editing tool. In Proceedings of the Third International WordNet Conference – GWC 2006, pp. 325-328, Jeju, South Korea, Masaryk University, Brno.

Kilgarriff A., Rychlý P., Smrž P., Tugwell D. (2004). The Sketch Engine. In Proceedings of the Eleventh EURALEX International Congress, Lorient, 2004, pp. 105--116

le Roux, J., Moropa, K., Bosch, S., Fellbaum, C. (2008). Introducing the African Languages Wordnet. Proceedings for the Fourth Global WordNet Conference, Szeged, Hungary, January, 2008.

Meraka Institute. (2007). African Advanced Institute for Information and Communication. Available at: http://www.meraka.org.za/index.htm
Accessed on: 28 February 2008.

Miller, G..A. (1995). WordNet: a lexical database for English. In Communications of the ACM. 38(11), pp. 39--41.

Pala, K., Smrž, P. (2004). Building Czech WordNet. *Romanian Journal of Information Science and Technology*, Romanian Academy Bucharest, 7, 1-2, pp. 79--88.

Pala, K., Vetulani, Z., Horák, A., Rambousek, A., Konieczka, P., Marciniak, J., Obrębski, T., Rzepecki, P., Walkowska, J. (2007). DEB Platform tools for effective development of WordNets in application to PolNet. In Vetulani, Zygmunt (ed.) Human Language Technologies as a Challenge for Computer Science and Linguistics, Proceedings of 3rd Language and Technology Conference, Poznan, 2007, pp. 514--518.

Prinsloo, D.J., Heid, U. (2006). Creating Word Class tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping. In: Isabella Ties (Ed.)/ Proceedings of the Conference for Lesser Used Languages and Computer Linguistics, EURAC research, European Academy. Bolzano, Italy. 27th October - 28th October 2005, (Bolzano: EURAC), 2006, pp. 97-- 115.

Rychlý, P. (2000). Corpus managers and their effective implementation, Ph. D. Thesis, Masaryk University, Brno, Czech Republic.

University of Pretoria Department of African Languages. (2003). Available: http://www.up.ac.za/academic/humanities/eng/eng/afrlan/eng/initiative.htm
Accessed on 13 April 2006.

Vossen, P. (1998). (Ed.). EuroWordNet. Dordrecht: Holland: Kluwer.

Vossen, P. (2007). The Cornetto project. Available: http://www.let.vu.nl/onderzoek/projectsites/cornetto/start.htm
Accessed on 28 February 2008.

WordNet a lexical database for the English language. (2006). Available: http://wordnet.princeton.edu/
Accessed on 28 February 2008.

Extracting bilingual word pairs from Wikipedia

Francis M. Tyers[*], **Jacques A. Pienaar**[†]

[*]Grup Transducens,
Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant
03071 Alacant, Spain
ftyers@dlsi.ua.es

[*]Prompsit Language Engineering S.L.,
Pol. Ind. Canastell. Ctra. Agost, 77
03690 Sant Vicent del Raspeig, Spain

[†]Centre for Text Technology, North-West University
Potchefstroom 2531, South Africa
jacques.pienaar@nwu.ac.za

Abstract

A bilingual dictionary or word list is an important resource for many purposes, among them, machine translation. For many language pairs these are either non-existent, or very often unavailable owing to licensing restrictions. We describe a simple, fast and computationally inexpensive method for extracting bilingual dictionary entries from Wikipedia (using the `interwiki` link system) and assess the performance of this method with respect to four language pairs. Precision was found to be in the 69–92% region, but open to improvement.

1. Introduction

Bilingual dictionaries are an important resource for natural language processing, for example cross-language information retrieval, and especially machine translation. In machine translation they are central to rule-based systems and useful in statistical machine translation (Koehn and Knight, 2002).

While bilingual dictionaries exist for pairs of larger languages such as English – French, they are scarce resources for many smaller language pairs. This includes pairs where a smaller language is paired with a larger language, for example English – Afrikaans.

Wikipedia is an online, collaboratively edited encyclopaedia with articles available in 256 languages (Wikipedia, 2008b). Neither the addition nor maintenance of Wikipedia entries requires any specific expertise over being able to use a standard web browser and entering text using a simple markup language. These encyclopaedias are freely-editable, and freely-distributable, which makes them the ideal platform for developing encyclopaedias in all the world's languages. The content and structure of these encyclopaedias make them amenable to linguistic research, whilst the breadth of language coverage makes them appropriate and useful for creating linguistic resources for languages which lack them.

For each language a separate encyclopaedia exists with its own norms and the articles between the different encyclopaedias are not simply translations of one another.

Each Wikipedia article can provide links to other articles on the same subject in different languages, so for example on the English article for *socialism* there is a link to the same article, *socijalizam*, on the Serbo-Croatian Wikipedia. These links are between article titles, which may be a word or a phrase.

The links between the same article in different languages are called `interwiki` links and are periodically maintained by bots.[1] This maintenance can occur in either *supervised* or *unsupervised* mode and is intended to keep the consistency of the links between the various Wikipedias. The general functionality of both modes of execution are the same. For each article the bot first checks the existing `interwiki` links of the *source* article. If any are found it then retrieves the articles they point to. The bot then adds links from the *target* articles which were not included in the *source* article, to the *source* article. If more than one link is retrieved in supervised mode, for any given language pair, then the operator of the bot is asked to pick the correct one, whilst in automatic mode, ambiguous links are skipped. Harvesting these links provides useful translation equivalents for many different language pairs, and could provide a basis for further lexical acquisition techniques such as described by Koehn and Knight (2002).

It is expected that the method presented will be particularly useful for under-resourced languages which, in many cases, have an active and vibrant Wikipedia community.

2. Related work

The work presented in this paper is in the same vein as that by Koehn and Knight (2002) in that it focuses on attempting to create a translation lexicon from meagre resources. In their case these meagre resources were unrelated monolingual corpora. They cover a number of methods, some of which were based on linguistic knowledge, and others on statistics.

[1]As used on Wikipedia, a bot (short for robot) is a software program that makes automated changes to the Wikipedia.

Adafre and de Rijke (2006) describe an experiment in finding similar sentences between different language versions of Wikipedia and note that lexicons induced from Wikipedia titles are generally of high quality and there is "rarely conceptual mismatch" between pages linked by `interwiki` links. They propose two approaches, one using a machine translation system and the other using the hyperlinks between documents. Their second approach of working with the hyperlinks within a document is more general and involved than the method we propose here. They do not give any quantitative evaluation of the lexicon created, which includes both common nouns and proper nouns.

Wikipedia has also been used as a semantic resource in the vein of WordNet (Zesch et al., 2007a; Zesch et al., 2007b), and as a monolingual resource in developing systems for named entity and word sense disambiguation (Bunescu and Pașca, 2006; Cucerzan, 2007; Mihalcea, 2007).

3. Method

In this section we shall give a quick overview of the experiment and describe the algorithm used therein.

Our method, as described below, requires a monolingual word list of one of the languages in a translation pair.

Starting from a word list in the better sourced language of the pair is the logical and the recommended practise. In our experiment we had English in all of the language pairs and therefore used an English word list as seed for our method. This word list was extracted from the English–Catalan translation pair of Apertium,[2] an open-source, shallow transfer machine translation system (Armentano-Oller et al., 2005). The motivation behind using this specific wordlist was that the lexicons produced could be immediately useful in Apertium translation pairs. The word list[3] consisted of 11,393 lemmas,[4] all nouns, and was biased slightly towards technical and scientific terminology. The reason for choosing a list made up only of nouns was because Wikipedia titles are almost exclusively made up of nouns and proper nouns. The first ten words are shown below:

```
abandonment
abbey
abbot
abbreviation
abdomen
abduction
aberration
ability
abnormality
abolitionism
```

The total number of articles in the English Wikipedia which matched the entries in the word list was 10,024; this number represents the upper bound on the number of possible translation pairs.

The languages for which translations were attempted to be found were Macedonian (`mk`), Afrikaans (`af`), Iranian Persian (`fa`) and Swedish (`sv`). These choices were motivated by the availability of native speakers to evaluate the results, and the desire to cover a variety of language groups and Wikipedia sizes.

The bilingual word pair extraction algorithm, presented in pseudo-code in Figure 1, is very simple and computationally inexpensive.

```
EXTRACT-WORD-PAIRS()
1  for each w in Word-List
2  do
3     a ← RETRIEVE-PAGE( SourceWikipedia , w);
4     ℓ ← EXTRACT-LINKS(a);
5     for each t in Target-Languages
6     do if t in ℓ
7           then ADD-PAIR(w, ℓ[t]);
```

Figure 1: Description of the algorithm used.

In the algorithm (Figure 1) we iterate over both the word list and list of target languages, represent the extraction of the `interwiki` links with the function EXTRACT-LINKS and the target word/phrase of the `interwiki` link as the array ℓ.

Certain titles are ambiguous (can be associated with more than one topic) and are linked to a page that contain no content and only refers to other Wikipedia articles with which the user can resolve the conflict (Wikipedia, 2008a). In this case the title of these so-called disambiguation pages were taken as the translation. Information within parentheses of titles were uniformly removed from all page titles.

4. Results

The results for precision of this method are presented in Table 1. Also given is the *total* number of articles in the Wikipedia in question (on 9 February 2008), the total number of `interwiki` links retreived and the number of "correct" translations.

Table 1: Results for the language pairs

	Total	Links	Correct	Precision
`af`	9,183	444	354	79%
`mk`	14,887	779	631	81%
`fa`	32,194	1,605	1,487	92%
`sv`	273,291	4,913	3,428	69%
`en`	2,299,336	10,024	-	-

Precision was calculated by dividing the number of correct translations by the total number of possible translations retrieved. A correct translation was counted as an exact lemma-for-lemma translation and was judged by a native speaker. Note that the number of links retrieved, from the English word list of 10,024 entries, is rather low. The scientific and technical nature of the word list could be the cause hereof as more popular topics are added quicker and

[2] Available from `http://www.apertium.org/`

[3] The word list is under the same license as the linguistic package `apertium-en-ca`, and can be retrieved from `http://xixona.dlsi.ua.es/~fran/en-nouns.txt`.

[4] The lemma (or citation form, base form, head word, etc.) is the canonical form of a word, as is typically found in printed dictionaries.

revised more often. As one would expect, the number of pages in the target language's Wikipedia also greatly affects the number of links retrieved.

5. Analysis

The word lists were given to native speakers to check. A positive result is when the translation is judged as correct by a native speaker. That is when the word is in the right form, has the right sense and is in the appropriate register. If a word can have many possible translations, it is considered enough that it be among them, not necessarily being the most general or frequent. As all Wikipedia article titles are in uppercase, case distinctions were ignored. A rough typology for a negative results with examples can be found below:

1. Right sense, wrong surface form – *vandal* translated as *vandale* (vandals). This kind of error occurred when the lexical form of the word in the source language did not match the translation. For example a singular noun being translated as a plural noun.

2. Right sense, wrong register – *nephrolithiasis* translated as *njursten* (kidney stone). This is normally caused by a more scientific term or more specific term being a redirect to a more general article. The English Wikipedia guidelines recommend that the most common name, not the most correct name be used for the title of an article (Wikipedia, 2008c). This problem was also seen in the translation of acronyms, where the acronym typically redirects to the spelt-out form.

3. Wrong sense, right domain – *sociolinguist* translated as *sosiolinguistiek* (sociolinguistics). This type is also generally caused by redirects. Articles on professions, sub-fields, etc. are often redirected to a general article dealing with the whole field. This also occurs with derivations as shown above. It is worth noting that these are by no means regular, for example *bureaucrat* has its own article, while *bureaucratisation* redirects to *bureaucracy*. On the other hand, *colonist* redirects to *colony*, while *colonisation* has its own article.

4. Wrong sense, wrong domain – *solidarity* translated as *Solidarność*.[5] The fourth type of error occurred when an incorrect `interwiki` link was in place. These are caused either by badly configured bots, or human error. Examples of this kind of error are a proper name linked in the place of a common name.

Borderline situations also exist, for example, the translation of *amount* into Macedonian as *kvantitet* (literally 'quantity'). In this example, the translation found is not an exact translation, but refers to a similar and closely related word. These were marked as correct or incorrect translations at the discretion of the native speakers.
These errors were generally found to exist at approximately the same frequency, with none particularly more frequent than the other. No full quantitative analysis was done.

[5] A proper name referring to a trade union, later political party in Poland.

The increase over time in articles and `interwiki` links continue to gradually improve recall (the number of correct translations retrieved from a given word list). Therefore recall will be improved as the number of articles in each Wikipedia grows, along with the number of links between articles. Several techniques could improve precision:

- Double-check each pair – Ensuring that a retrieved link points back to the same source. The equivalent of cross-referencing in a paper dictionary. That is the `interwiki` links of the *target* page are checked for a link to the *source* page.

- Avoid following redirects – This would increase precision at the expense of reducing recall. Often differing orthographic conventions are linked through redirection, and if these links were not followed, the pages would not be retrieved.

- Analyse all links – A more complex strategy might involve retrieving the set of all the `interwiki` links from all the pages linked from the page in the *source* language, and choosing the most frequently linked translation in the *target language*. This is similar to what is done by Adafre and de Rijke (2006).

6. Discussion

We have presented a simple, computationally inexpensive and fast means of automatically obtaining bilingual word lists.
The accuracy of this method compares favourably with those of Koehn and Knight (2002), the lowest accuracy we achieved was 69% compared to the 39% accuracy they obtained in their experiment. But their method operates on unrelated, monolingual corpora and could potentially produce more word pairs.
Extracting word pairs from Wikipedia could prove useful for under-resourced languages, and for bootstrapping more complex induction techniques.
Further work would generally focus on improving the precision of results, although another avenue might be to work with trying to use additional information to provide sense disambiguation for the word pairs. Similar work has been some by Sammer and Soderland (2007), who use bilingual word lists and monolingual corpora to construct a sense disambiguated lexicon. Along with the `interwiki` links, Wikipedia articles are generally members of categories, which could be used for this task. Further disambiguation information comes from the page titles themselves, where there is more than one concept represented by a title, often they are disambiguated by means of a term in parentheses. These terms can be almost anything, indication of hierarchy (in the case of place names), of domain (in the case of nouns), or profession (in the case of people), etc.
Another possible use might be for automatically creating directories for named entities, containing places or people. Wikipedia has large numbers of articles on these topics and often, as they are quite formulaic, they are translated into quite a large number of languages. This strategy has been used in the expansion of dictionary entries for the Occitan – Catalan language pair in the Apertium machine translation

system to improve the coverage of place names. Indeed in further work it might be interesting to compare the accuracy of retrieval of translations of proper nouns to those of common nouns.

The wide range of the precision found, 69–92% would be another avenue for further investigation. Increasing the number of human evaluators of the output would likely provide a more accurate benchmark of translation quality.

A possible caveat with using Wikipedia in this manner is the licensing of the articles. The content of Wikipedia is uniformly released under the GNU Free Documentation Licence (GFDL),[6] which is incompatible with the GNU General Public License (GPL),[7] a licence under which much open-source software, including Apertium, is released. There has been an ongoing discussion of this problem in the Wikipedia mailing lists, however the most authoritative response comes from Mike Godwin, general counsel to the Wikimedia Foundation.[8] He argues that these, "... links and word pairs, standing alone, do not qualify as copyrightable, and thus fall outside the GFDL" (personal correspondence).

Acknowledgements

Many thanks to the people who evaluated the output of the process, Cenny Wenner, Slobodan Jakovski and Soroush Mesry.

7. References

Adafre, S. F. and de Rijke, M. 2006. Finding similar sentences across multiple languages in wikipedia. In *EACL 2006 Workshop on New Text–Wikis and Blogs and Other Dynamic Text Sources*, March.

Armentano-Oller, C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Bonev, B., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., and Sánchez-Martínez, F. 2005. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. In *OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X*, pages 23–30, September.

Bunescu, R. and Pașca, M. 2006. Using encyclopedic knowledge for named entity disambiguation. *Proceedings of EACL*, pages 9–16.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on wikipedia data. *The EMNLP-CoNLL Joint Conference. Prague*.

Koehn, P. and Knight, K. 2002. Learning a translation lexicon from monolingual corpora. *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, 34.

Mihalcea, R. 2007. Using wikipedia for automatic word sense disambiguation. *Proceedings of NAACL HLT 2007*, pages 196–203.

Sammer, M. and Soderland, S. 2007. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. *Proceedings of Machine Translation Summit XI*.

Wikipedia. 2008a. Disambiguation — Wikipedia, the free encyclopedia. [Online; accessed 9 February 2008].

Wikipedia. 2008b. List of wikipedias — Wikipedia, the free encyclopedia. [Online; accessed 9 February 2008].

Wikipedia. 2008c. Naming conventions (common names) — Wikipedia, the free encyclopedia. [Online; accessed 9 February 2008].

Zesch, T., Gurevych, I., and Mühlhäuser, M. 2007a. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Data Structures for Linguistic Resources and Applications*, pages 197–205. Gunter Narr, Tübingen, Tuebingen, Germany.

Zesch, T., Gurevych, I., and Mühlhäuser, M. 2007b. Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 205–208.

[6] Available at `http://www.gnu.org/licenses/fdl.html`.

[7] Available at `http://www.gnu.org/licenses/gpl.html`.

[8] The Wikimedia Foundation operates Wikipedia and related sites.

Building a Basque/Spanish bilingual database for speaker verification

Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola, Juan J. Igarza and Inma Hernaez
AhoLab Signal Processing Group, Department of Electronics and Telecommunications,
University of the Basque Country (UPV/EHU)
Alda. Urquijo s/n, 48013 Bilbao
E-mail: {ikerl, eva, inaki, ibon, ion, igor, jigarza, inma}@ aholab.ehu.es

Abstract

Research groups aiming to record new speech databases for minority languages have to face a series of difficulties, such as the lack of previous resources, the scarcity of fluent speakers and the shortage of funding for the project. Sometimes it is possible to take advantage of recording campaigns for other projects, and extend them in order to make some recordings in that language for every contributor that is able to speak it. In this way, new databases can be recorded with little extra effort, as the campaign is already prepared and funded. Using this technique, a new Basque/Spanish bilingual database has been created. Thanks to this database, some research is being undertaken on Basque/Spanish bilingual speaker verification systems. We present a description of the resulting database and the difficulties encountered during its acquisition.

1. Introduction

Currently many systems need some kind of user authentication procedure, in order to verify the users' identity. Most of them use password-type authentication, but passwords may be forgotten or stolen. Nowadays biometric authentication is the best alternative because it can provide extremely accurate and secure access to information (Jain et al., 2006). Furthermore, biometric characteristics cannot be lost nor forgotten and are very difficult to imitate. This kind of authentication can already be seen in different applications such as laptops with fingerprint controlled access and hand geometry based access to certain buildings. Speech is a biometric feature that is non intrusive, has a high degree of acceptability among users and is suitable for long distance verification over data and voice networks. For the development of these speech-based authentication systems, speech databases containing recordings from different speakers are needed.

As a biometric authentication method, a speaker verification system has to decide whether or not a person is who she or he claims to be, using one or more spoken utterances produced by this person (Campbell, 1997). In a generic speaker verification system two modules can be distinguished: the enrolment or training module (which produces a model of every user of the system) and the verification or test module (which decides whether a test utterance has been spoken by an specific speaker)(Naik, 1990)(Bimbot et al., 2004). Usually the language of the training and testing utterances is forced to be the same. But in some environments where bilingual speakers are common, it is desirable that the users of the speaker verification application may be able to utilise any of their known languages to address the system. Therefore, in the last few years various researchers have focused their attention on speaker recognition systems in multilingual environments, where speaker models may be trained with recordings in one language but testing is performed with utterances spoken in another language (Nordstrom et al., 1998)(Faundez-Zanuy & Satue-Villar, 2006). This multilingual environment involves some additional difficulties for the verification system. On the one hand language mismatch between enrolment and verification data produces a degradation in the results of the speaker verification system (Ma & Meng, 2004). On the other hand language mismatches between the target speaker and the world model in a GMM speaker verification system make its performance worse (Auckenthaler et al., 2001).

This bilingual environment is found in the Basque Country. The Basque Country extends north and south of the western side of the Spanish-French border. In the Basque Autonomous Community, situated in the south of the Basque Country, both Basque and Spanish are spoken. Basque is a minority language and therefore there is a scarcity of linguistic resources in this language (Diaz de Illarraza et al., 2003). Specifically, there is no public speech database in Basque available for the development of speaker verification systems.

This paper presents the work and difficulties of recording a new bilingual speech database in the Basque Country for the development of bilingual speaker verification systems in both Spanish and Basque. Section 2 analyses the problems associated with the recording of new speech databases for minority languages. Section 3 describes the recorded bilingual database and its contents, while in section 4 the difficulties that arose during its acquisition are described. Finally, some conclusions are extracted in section 5.

2. Dealing with minority languages in the acquisition of speech databases

Although they are of great social interest, minority and endangered languages are usually not economically interesting, i.e. as there are very few speakers using those languages, it is not worth investing big amounts of money

for research and development of new resources, let them be new corpora compilations or speech databases. This means that finding funding resources for these projects is usually difficult and very often this funding is short in the case of getting any. Moreover, typically there are not many research groups working in these languages, and frequently only one or two groups have interest in a specific language. So collaborative work among research groups in order to distribute work load and expenses is difficult too.

In the case of speaker verification databases, some of the requirements make the process even harder. On the one hand, the recordings in these databases should span along a time period long enough to collect the intra-speaker variation of the voice (Kenny & Dumouchel, 2004). This means that each speaker should be recorded more than once, in different time-spaced sessions, which makes the recording process longer and more expensive. On the other hand, it is desirable that the age and gender distributions of the speakers in the database approximately match the distribution of the expected users. This restriction, together with the fact that being a minority language there may be few speakers available, makes the recruitment of contributors more difficult.

In order to make the acquisition of new databases for minority languages feasible, it can be interesting to take advantage of recording campaigns organized for other projects and use the opportunity to make additional recordings in the minority language too, even though that is not the main objective of the campaign. In this way, new databases can be obtained with little extra effort, as the whole campaign is already prepared.

In the AhoLab Signal Processing Laboratory of the University of the Basque Country research on speech technologies for Basque is being carried out, mainly related to text to speech (TTS), automatic speech recognition (ASR) and speaker recognition. For the development of this research, speech databases in Basque are needed. Sometimes, when there is a database recording campaign for other projects (mainly Spanish speech recordings), contributors are asked to make some extra recordings in Basque, in order to complete a parallel Basque database.

3. Description of the database

The new Basque/Spanish bilingual speaker database was recorded together with a multimodal biometric database acquired in five different Universities all along Spain, including the University of the Basque Country (Galbally et al., 2007). In this database different biometric features were acquired, like fingerprints, signature, handwriting, images of the iris and speech (in Spanish). Seizing the opportunity, the contributors to the database in the University of the Basque Country that were fluent in Basque were also recorded in this language. In this way a small bilingual speaker verification database could be built with little extra effort.

3.1 Design of the database

The recording protocol included four sessions distributed along the time to ensure the capture of the intra-speaker variations that arise as time goes by. There is a difference of two weeks between the recording of the first and second sessions, four weeks between the second and third sessions and six weeks between the third and fourth sessions.

The content of the recordings in each session is:

- Session 1: 4 isolated sentences and 4+3 numeric sequences
- Session 2: 2 isolated sentences and 4+3 numeric sequences
- Session 3: 2 isolated sentences and 4+3 numeric sequences
- Session 4: 2 isolated sentences and 4+3 numeric sequences

The numeric sequences are formed by 8 digits that the speaker reads as she or he prefers. Every speaker has a unique numeric sequence that she or he repeats four times in each session. In addition, she or he records the numeric sequence assigned to three other speakers, which are different in each session, in order to be used as impostor trials. All numeric sequences recorded in each session are common for Spanish and Basque.

The isolated sentences are phonetically rich and balanced, i. e., the distribution of phones in the sentences is similar to the one found in the language. These sentences are the same for all the speakers and are changed from session to session. Obviously, the sentences are different for Spanish and Basque. Therefore the recorded corpus includes 10 different phonetically rich sentences for Spanish and 10 for Basque. In order to select the most appropriate sentences a big corpus used as a reference of the language must be compiled. Once this initial corpus for each language was collected and analysed, the sentences were selected using a software tool called CorpusCrt, made by the TALP research group from the UPC [1], which produces a reduced set of sentences keeping the original frequency of the phonemes as far as it is possible.

The recordings were made in the half-silent environment of a research laboratory, using a Plantronics DSP-400 headset microphone. They were sampled at 44.1KHz and quantified using 16 bits per sample.

3.2 Additional data

The recordings in the database were further processed in order to extract some additional information, namely voice activity and pitch curves.

Voice activity estimation is necessary in order to reject those frames in which there is no vocal information. In this way, noise level during speech silences will not corrupt the features calculated for the speaker verification system. A voice activity detector (VAD) was implemented, based on the computation of the long term spectral deviation (LTSD) between vocal and noisy frames. The implemented system is based in the one presented in (Ramirez et al., 2004), in which an adaptive decision threshold is used in order to get

[1] Universidad Politécnica de Catalunya. http://www.talp.upc.es

the best performance for each signal to noise ratio.
For the calculation of pitch curves a tool developed at AhoLab Signal Processing Group has been used (Luengo et al., 2007). This tool uses dynamic programming with cesptral coefficients in order to estimate the pitch curve.

4. Difficulties encountered

4.1 Scarcity of bilingual speakers

Collecting a database in Basque is not easy, as many people in the Basque Country do not speak Basque or they do not speak it fluently. Table 1 presents the number of Basque speakers in the Basque Autonomous Community in 2001, distributed according to age[2]. In this table active as well as passive bilingual speakers have been considered, i.e. it includes data about those speakers whose primary language at home is Basque and about those whose primary language at home is not Basque. The language competence among the latter is not good in all cases, as they include people who speak Basque with difficulty or do not speak it at all, although they understand or read it well.

Age range	Total	Percentage
16-24	170 453	23.1%
25-34	171 608	23.3%
35-49	175 522	23.8%
50-64	104 055	14.1%
>=65	115 442	15.7%
TOTAL	737 080	100.0%

Table 1: Distribution of active and passive bilingual speakers according to age in the Basque Autonomous Community in 2001.

The knowledge and use of Basque by the inhabitants of the Basque Autonomous Community vary according to the age range. Table 2 shows the percentage of monolingual and bilingual speakers for each age range. The proportion of Basque speakers is higher among young people.
Furthermore, the fact that our bilingual speaker verification database was recorded as an extension of another biometric database that was part of another project has its own drawbacks. The main specifications of the biometric database, such as the number of volunteers, their age distribution and delivery dates had to be respected. As the Basque recordings were not part of the main project, the specifications set for the biometric database did not take into account the special requirements needed to match the goals of the bilingual speaker verification database. It was a priority to fulfil the specifications of the biometric database, even if this meant to deviate from the optimal specifications for the bilingual database. For example, it was not possible to reject a volunteer just because she or he did not speak Basque, as this would had made the recruiting more difficult and extended the delivery dates of the whole database. This is the reason why, although all 55 volunteers recruited were recorded in Spanish , only 30 of them were recorded in Basque, as the remaining ones were not bilingual or fluent in this language.

Age range	Monolingual	Bilingual
16-24	31.4%	68.6%
25-34	50.5%	49.5%
35-49	63.3%	36.7%
50-64	72.5%	27.5%
>=65	67.4%	32.6%

Table 2: Percentage of monolingual and bilingual speakers according to their age range in the Basque Autonomous Community in 2001.

4.2 Deviation from age distribution

In a speaker verification database, the population should be well represented. It is important that the database includes examples representative of all the potential users of the system. This is the reason why this kind of databases are usually balanced in age ranges and gender. To achieve this balance in the recording of the database, a target distribution of speakers is proposed following the expected distribution of potential users, and the selection of contributors to the database is made according to it. Table 3 shows the goal distribution of recordings by age range, and the real distributions achieved among the recorded Spanish and Basque speakers.

Age range	Goal	Spanish	Basque
18 to 25 years	30%	32.7%	33.3%
25 to 35 years	20%	40.0%	53.3%
35 to 45 years	20%	12.7%	10.0%
45 to 55 years	20%	7.3%	3.3%
More than 55 years	10%	7.3%	0.0%

Table 3: Distribution of speaker's age range in the bilingual database.

The recruitment of the speakers was mainly done among the students and staff of the Faculty of Engineering of the University of the Basque Country. The average age in this group is quite low, as reflected in the deviation from the goal distribution that is observed in the 25 to 35 and 45 and up year ranges both for Spanish and Basque. Furthermore, it is greatly difficult to recruit people older than 35 years for a Basque/ Spanish bilingual database, because most of them are monolingual, as shown in Table 2. That is why there are so few Basque speakers in the database in higher age ranges.
The deviation of Basque speakers' distribution from the target values is higher than that achieved for Spanish speakers. Once more, the main reason for this is that during recruitment it was a priority to keep the age distribution for the main biometrical database, in which

[2] Source EAS (Language Indicator System of the Basque Country)
http://www1.euskadi.net/euskara_adierazleak/zerrenda.apl?hizk=i&gaia=25&sel=64

Spanish recordings were included. But then, when non-bilingual people were dropped, the new age distribution for Basque speakers did not match the objective.

The balance in gender was easier to achieve. In table 4 the goal distribution sought in gender and the real distributions both for the Spanish and Basque parts are shown. These real distributions do not differ significantly between Spanish and Basque.

Gender	Goal	Spanish	Basque
Male	50%	47.3%	43.3%
Female	50%	52.7%	56.7%

Table 4: Distribution of speaker's gender in the bilingual database.

5. Conclusions

Taking into account that Basque is a minority language the development of new spoken resources for this language is difficult and the funding for them is usually scarce. Under these circumstances, the acquisition process of a database in a majority language represents an opportunity that can be seized to build another database in the minority language. Using this strategy, a new database for bilingual speaker verification in Spanish and Basque has been created. Its features are not ideal, because the acquisition process has not been designed explicitly for it, but nonetheless it is a new and useful spoken resource. Currently it is being used in the building of a bilingual speaker verification system with success.

6. Acknowledgements

This work has been partially founded by Basque Government under grant IE06-185 (ANHITZ project, http://www.anhitz.com/) and by the University of the Basque Country and EJIE S.A. under grant EJIE07/02 (MULTILOK project).

The authors would also like to thank all the contributors that took part in the acquisition of the biometric database.

7. References

Auckenthaler, R., Carey, M.J., Mason, J.S.D. (2001). Language dependency in text-independent speaker verification. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* vol. 1, Salt Lake City, UT, USA, pp. 441--444.

Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz,D., Reynolds, D.A. (2004) A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing* vol. 4, pp. 430--451.

Campbell, J.P. (1997) Speaker Recognition: A tutorial. *In Proceedings of the IEEE*, 85, pp. 1437--1462.

Díaz de Ilarraza A., Sarasola K., Gurrutxaga, A., Hernaez, I., Lopez de Gereñu, N (2003). HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities. *In Proceedings of the Workshop on NLP of Minority Languages and Small Languages*, Nantes, France.

Faundez-Zanuy, M. Satue-Villar, A. (2006) Speaker Recognition Experiments on a Bilingual Database. *In Proceedings of the 14th European Conference on Signal Processing (EUSIPCO)*, Florence, Italy.

Galbally, J., Fierrez, J., Ortega-Garcia, J. et. al. (2007), BiosecurID: a Multimodal Biometric Database. *In Proceedings of the User-Centric Technologies and Applications Workshop*, Salamanca, Spain, pp. 68-76.

Jain, A.K.; Ross, A.; Pankanti, S. (2006) Biometrics: a tool for information security. , *IEEE Transactions on Information Forensics and Security*, Vol. 1 (2), pp. 125 --143.

Kenny, P.; Dumouchel, P. (2004) Disentangling speaker and channel effects in speaker verification. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* vol. 1, Montreal, Quebec, Canada, pp. 37--40.

Luengo, I., Saratxaga, I., Navas, E., Hernáez, I., Sanchez, J., Sainz, I. (2007) Evaluation Of Pitch Detection Algorithms Under Real Conditions. *In Proceeding of 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Honolulu, HI, USA, pp. 1057--1060.

Ma, B., Meng, H. (2004) English-Chinese bilingual text-independent speaker verification. *In Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)* vol. 5, Montreal, Quebec, Canada, pp. 293--296.

Naik, J.M. (1990) Speaker verification: a tutorial. *IEEE Communications Magazine*, vol. 28(1), pp. 42--48.

Nordstrom, T., Melin, H., Lindberg, J. A Comparative Study of Speaker Verification Systems using the Polycost Database. *In Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia, pp. 1359--1362.

Ramirez, J., Segura, J., Benitez, C., de la Torre, A., Rubio, A. (2004) Efficient Voice Activity Detection Algorithms Using Long Term Speech Information, *Speech Communication* 42, pp. 271--287.

Language resources for Uralic minority languages

Attila Novák
MorphoLogic Ltd.
1126 Budapest Orbánhegyi út 5., Hungary
novak@morphologic.hu

Abstract

Most members of the Uralic language family are small minority languages spoken on the territory of the Russian Federation, which all are endangered. In past and ongoing projects, computational morphologies and annotated corpora have been and are being created for several of these Uralic minority languages: Udmurt, Komi-Zyrian, Eastern Mari, Northern Mansi, and the Kazym and Synya dialects of Khanty, Tundra Nenets and Nganasan. This article presents the morphological analyzers and other annotation tools and the resources developed and used during the projects.

1. Introduction

Besides the national languages spoken by several million speakers: Hungarian, Finnish and Estonian, the Uralic language family includes several minority languages with significantly smaller speaker communities, the majority of which is spoken on the territory of the Russian Federation. In a series of projects[1], computational morphologies and annotated corpora have been and are being created for several of these languages.

2. The projects

One aim of these projects is to make linguistic data concerning these languages available for research to a broader community of linguists, not only the Uralist specialists, and to make corpus-based investigation of these languages possible. Many of these languages exhibit phenomena that would be exciting to explore for a variety of linguists, such as theoreticians specializing in any module of grammar or those interested in language typology. Annotated corpora make it possible to carry out research on various aspects of the language without a long preliminary study of the language itself.

One of the most important lessons that we learned form the first project during which morphologies of six Uralic minority languages (Udmurt, Komi-Zyrian, Eastern Mari, Northern Mansi, Tundra Nenets and Nganasan) were created was that since many details of the description which often remain vague in written grammars must unavoidably made explicit in a computationally implemented grammar, the process of creating the implementations as well as the resulting programs themselves shed light on inconsistencies and gaps in the available descriptions of the phonology and morphology of the language, and often help correcting them.

Moreover, while examining linguistic models with regard to exactness and completeness by hand is an impossible task, the computational implementation makes an exhaustive testing of the adequacy of our grammatical models possible against a great amount of real linguistic data. Systematic comparison of word forms generated against model paradigms has pinpointed errors not only in the computational implementation (which were then eliminated) but also in the model paradigms or the grammars the computational implementation was based on.

Another fact makes a more thorough documentation of these languages urgent is that due to the nature of Russian minority policy, the school system, the great degree of dispersion, the low esteem of the ethnic language and culture and the general lack of an urban culture of their own, all these languages are endangered. On the other hand, there are significant differences among these languages concerning the number of speakers and the exact sociolinguistic situation they are in.

3. Moribund languages

Some of the languages can be categorized as moribund, with virtually no chance of the language still being spoken in another 50 years, not only due to the low number of speakers (some of these languages have existed and developed as the communication medium of small nomadic communities of about a thousand people for thousands of years without an immediate risk of disappearance), but because one generation of speakers has already failed to pass on the language to the next and thus hardly any children speak it. In the case of these languages, an example of which is the Nganasan language of the Northern Samoyedic branch of the Uralic family with about 400 middle-aged and elderly speakers, the most we can do is trying to document as much of the language as possible. Documenting these languages is not a trivial task though, not only because of the extreme complexity of some of them (e.g. in terms of their morpho-phonology), but also because the speaker communities are disintegrating into a small assembly of individuals with more and more uncertain language skills and a heavy influence from their parallel knowledge of the majority language, Russian, that seems to impact not only the syntactic structures they

[1] ('Complex Uralic Linguistic Database', NKFP 5/135/2001), 'Development of Komi and Udmurt morphological analyzers' (OTKA-T 048309) and 'Development of a Ngananasan morphological analyzer' (OTKA / K 60807)

use[2] but even the morpho-phonology[3]. (According to the 2002 census data, there are only 9 monolingual Nganasan speakers, who are all elderly people over 70 living in practically inaccessible spots). The language becomes thus just a collection of idiolects which presumably all differ significantly from both the Nganasan that was spoken by monolingual speakers 60 years ago and from each other. Whose idiolect are we to document? The complexity of the language (e.g. that of the morpho-phonology of Nganasan, or that of the intricate system of verbal moods and evidentiality) might partially account for the fact that no outsiders, including the linguists doing research on the language have managed to master Nganasan. But these languages are not only very difficult to learn for anybody but babies, but they are not very useful to know, either. They have lost much of their function when these nomadic peoples were forced to settle as a minority in settlements inhabited by people speaking another language and to give up their traditional way of life, their rituals and practices. Their tame reindeer herds were collectivized (which subsequently fell victim to epidemics) and they were practically prohibited from reindeer hunting. But the fatal blow on these languages was the schooling of minority children in boarding schools hundreds of kilometres away from their home where the language of education was exclusively Russian. The children had no contact at all with their parents and their home community during the school year, and both their knowledge and their esteem of their mother tongue deteriorated significantly. This was the generation that growing up failed to pass on the language to their children.

There is another factor that makes the documentation of some of these languages difficult. During the Soviet era, making field trips to areas where many of these small minority languages are spoken was only possible for linguists from within then Soviet Union. In the nineties, during the Yeltsin era, an unprecedented freedom of movement made it possible also for foreign linguists to travel freely to the areas previously inaccessible to them and do research there. Fortunately, this is still true for many areas (such as the region of the River Ob, where the Mansi and Khanty live). Certain areas of the northern Arctic regions where some of these minority languages are spoken, however, (the Taymyr Peninsula in particular, where the Nganasans live) have unfortunately been declared divisions of restricted access. Foreign linguists intending to do field work in the region must apply for an entrance permit at the local security authorities which they may fail to issue. This might make it necessary to find alternatives to field trips such as carrying native speakers to places accessible for the researchers as well.

4. Minority languages having a chance of survival

Another group of the languages mentioned do not seem to be threatened by an immediate language death, but even within this group there are significant differences. Although Udmurt and Mari have a similar number of speakers according to the census data, Mari seems to have a different sociolinguistic status than Udmurt due to the native speakers' different attitude toward their mother tongue. While the Mari are proud of their language and their cultural heritage, Udmurts have a rather low esteem of their mother tongue, which they consider inferior to Russian. On the other hand, Maris tend to have more conflicts with the Russian majority than Udmurts for the same reason.

In the case of these languages, the computational tools we create can also be adapted for practical purposes, such as providing the speaker communities with spell checkers and electronic dictionaries in their native language in the hope that the existence of such applications can help to raise the prestige of these languages. In order to be able to create applications of good quality we will need to collaborate with native speakers. Cooperation with publishing houses is vital so that we can obtain corpora that can be used in the process of the development and testing of the tools as well as for linguistic annotation, since on-line resources in these language are rather scarce. On the other hand, there is a stable output of books and newspapers from local publishing houses in all of the languages belonging to this group. The fact that we managed to obtain the manuscript of a 31000 word Komi–Russian dictionary in an electronic form from the company that published it shows that publishers are willing to cooperate. It is important that we make it clear that our goal is to give rather than to take something away from them.

5. Computational morphologies

In our first project, computational morphologies for six languages (Udmurt, Komi, Eastern Mari, Northern Mansi, Tundra Nenets and Nganasan) were created and tested on small corpora. These morphologies were based on Latin script based phonological transliterations generally used by linguists dealing with Finno-Ugric and in general with Uralic languages instead of the standard Cyrillic orthographies of the languages, since the tools were intended for linguistic annotation. This also made our lives easier avoiding an inherently non-phonological characteristic of Russian Cyrillic orthography, where palatalized consonants, the *j* phoneme and most vowels are represented by the orthography in a context sensitive manner. Inherently the same system is applied to the palatal consonants of all

[2] A contrast for example between Nenets and Nganasan focus constructions (preverbal vs. postverbal focus) can probably attributed to an adaptation of Russian post verbal focus by the Nganasan.

[3] Among the entries of the Nganasan–Russian dictionary which formed the basis of the stem lexicon of our Ngansan morphological analyzer, we have found about a dozen infinitives that according to our model of Nganasan morpho-phonology cannot be well-formed Nganasan infinitives. All of these 'ungrammatical' forms end in *s'a,* an allomorph of the Nganasan infinitive marker that happens to coincide with the infinitive ending of Russian reflexive verbs instead of some other allomorph that should appear there for the words to be well-formed infinitives.

non-Slavic languages of Russia in their respective Cyrillic orthographies with additional letters or diacritics to represent phonemes that do not exist in Russian.

However, especially in the case of the languages where orthographic texts (newspaper articles, books, etc.) are available, it is desirable that we can directly annotate these, so in a follow-up project, the goal of which was the enhancement of the Komi (Zyrian) analyzer, we created a version of the analyzer that can directly analyze orthographic text. In addition, the stem database of the analyzer was significantly enhanced by incorporating the entries form a 31000 word Komi–Russian dictionary (Beznosikova, 2000). Using standard orthography is of course also a prerequisite if we want to create spell checkers for these languages.

In another follow-up project that has just started this year, we are to create morphologies and annotated and glossed corpora for various dialects of the two Ob-Ugric languages: Khanty[4] and (Northern) Mansi. These analyzers will be based on the Latin script based phonological transliterations generally used in the lingustic works dealing with these languages.

Uralic languages are of the agglutinating type with a high frequency of words containing long suffix sequences and several thousands of possible word forms for each stem in the open word classes. We used two morphological development and analysis toolsets both of which are capable to handle this type of morphologies.

Of the six computational morphologies in our first project, the ones describing Finno-Ugric languages, Komi, Udmurt, Mari and Mansi were created using the formalism of the Humor ('High speed Unification MORphology') morphological analyzer engine of MorphoLogic (Prószéky and Kis, 1999), while the tools for two Samoyed languages, Nganasan and Tundra Nenets were developed using xfst ('Xerox Finite State Tool') of Xerox (Beesley and Karttunen, 2003). We plan to implement the additional Ob-Ugric analyzers using the Humor formalism.

The following table summarizes properties of the morphologies created in our first project and the follow-up Komi analyzer project. The size of the affix lexicons is indicated as a number of morphemes and lexicalized morpheme sequences in the source lexicon.

Language	stem lexicon (lemmas)	affix lexicon (morphemes)
Komi1	2100	156
Komi2	31000+2800 names	156
Udmurt	14100	238
Mari	2200	189
Mansi	1800	270
Nganasan	4150 non-derived	334
Tundra Nenets	19 500	254

[4] We are to create resources for two Northern Khanty dialects: Kazym and Synya Khanty, each named after that tributary of the Ob River along which the dialect is spoken.

5.1 The Humor analyzer

The Humor analyzer performs an 'item-and-arrangement' (IA) style analysis segmenting the input word into a sequence of morphs. The analyzer contains a regular word grammar and it produces flat morph lists as possible analyses. The program performs a search on the input word form for possible analyses looking up morphs in its lexicon that both match the beginning of the yet unanalyzed part of the input and satisfy all morph adjacency constraints of the previous morph. In addition, the candidate morph must form, together with the already analyzed part, the beginning of a possible word construction in the given language. Possible word structures are represented by an extended finite-state automaton in the analyzer.[5]

The morphological database that the Humor engine uses is not directly created and maintained manually, since for the analyzer to work efficiently, the data structures it uses must contain redundant data, which are both hard to read and hard to maintain for humans. The linguistic resources used by the Humor engine explicitly contain allomorphs instead of descriptions of morphemes, along with data structures such as binary vectors and continuation matrices that describe morph adjacency constraints. These resources are created using a morphological description development environment from a feature-based high level human readable description that contains no redundant information and is thus easy to maintain. The system transforms it to the redundant representations that the analyzer uses in two steps.

First, a lexical representation is created that already explicitly contains all the allomorphs of each morpheme along with all their properties and adjacency constraints (using a feature-based formalism) in a human-readable form, which can thus be checked easily by a linguist. This transformation is based on implicational relations, formulated as rules, which either define how redundant properties and requirements of allomorphs can be inferred from their already known (lexically given or previously inferred) properties (including their shape), or define default properties. These rules also describe how allomorphs should be created for each morpheme and what properties and constraints the individual allomorphs have (in addition to morpheme level properties and constraints).

The human readable redundant representation is then transformed to the format used by the analyzer using an encoding definition description, which defines how each of the features used in the description should be encoded for the analyzer.

In addition to the analyzer, the toolset contains a lemmatizer and a word form generator.

The lemmatizer, built around the analyzer core, outputputs simplified analyses of word forms consisting of a lemma and morphosyntactic category tags that, in contrast

[5] One can use feature variables in the automaton in to check long distance dependencies a fashion rather similar to flag diacritics in the Xerox tools.

to the more verbose analyses produced by the core analyzer, do not reveal the internal structure of words: compound members and derivational suffixes do not appear as independent items in the output of the lemmatizer.
The output of the lemmatizer and the analyzer is compared in the example below (analyses of the derived Komi word form *kylanly*):

```
analyzer>kylanly
 kyv[S_V]=kyl+an[D=A_PImpPs]+ly[I_DAT]
 kyv[S_V]=kyl+an[D=N_Tool]+ly[I_DAT]

lemmatizer>kylanly
 kylan[N][DAT]
 kylan[A][DAT]
```

The analyses produced by the lemmatizer are well suited for such tasks as corpus tagging, indexing and parsing.
The generator produces all word forms that could be realizations of a given morpheme sequence. The input for the generator is a lemma followed by a sequence of category labels that express the morphosyntactic features the word form should expose. The word form generator is not a simple inverse of the corresponding analyzer: it can generate the inflected and derived forms of any multiply derived and/or compound stem without explicitly referring to compound boundaries and derivational suffixes in the input even if the whole complex stem is not listed in the source stem lexicon (like in the case of the Komi derived nominal stem *kylan*):

```
generator>kylan[N][DAT]
 kylanly
generator>kyv[V][_Tool][DAT]
 kylanly
```

5.2 The Xerox Tools

The two level morphological toolset of Xerox contains various formalisms to create morpheme lexicons and phonological and morpho-phonological rule systems. Morpheme inventories can be created using the *lexc* formalism by defining sublexicons. A sequential phonological rule-system can be defined using the formalism of *xfst* resembling the form used in classical generative phonology as a set of context dependent re-write rules. Using *xfst*, one can compose the rules and the lexicon and during composition the program automatically eliminates intermediate levels of representation created by individual rules. The emerging single two-level finite-state transducer, called a lexical transducer, is a full morpho-phonological description of the language, which can be efficiently used both for analysis and generation. While *xfst* is a compiler for lexical transducers, actual morphological analysis and generation is performed by another program called *lookup*. Lookup may be invoked with either a single transducer, or a script containing an ordered sequence of transducer chains. The chains are applied to the input in order until one produces analyses, so each chain represents a fallback strategy to be applied if all previous strategies have failed. The default strategy is usually simple lookup with the lexical transducer of the language, others may include a chain of a case normalization transducer and the lexical transducer etc. The last fallback strategy can be a guesser, a lexical transducer featuring an extremely underspecified stem lexicon of open word classes besides the normal phonology and suffix grammar of the language. The fact that *lookup* is able to handle chains of transducers as individual strategies instead of just single transducers is important because normally the composition of e.g. a case normalization transducer and a lexical transducer would yield an enormous single transducer.
The two Samoyed languages: Tundra Nenets and Nganasan have a particularly complex phonology with a great abundance of very productive and quite complex phonological and surface phonetical processes. In both of these languages, the combination of phonological and morpho-phonological alternation processes can quite easily result in a single mono- or disyllabic suffix having as many as 20 different allomorphs and stems also tend to have several allomorphs. In the case of these languages, the exact form of a morpheme required by the morpho-phonology of the language cannot in general be determined by considering only local constraints between morphs, because the very intricate well-formedness constraints on syllable structure may involve phonological segments in non-adjacent morphemes. Formalizing these non-local phonological constraints would have been difficult in a formalism based on morph adjacency constraints. Since the descriptions we based our computational morphologies on used a sequential rewrite rule system formalism that was much easier to convert to an *xfst* grammar than to a Humor rule system, we decided to use the Xerox tools for the implementation of these two morphologies.
The feature-based Humor formalism proved to be an efficient means of describing morphological constraints. We also extensively used the corresponding flag diacritics feature of the Xerox tools to describe selectional restrictions between morphemes, such as morphological root selection in Nganasan, suffixes attaching to perfective or imperfective verbal roots; suffixes of verbs requiring an Agent; suffixes attaching to transitive verbs etc. Many of these constraints are local. The flags corresponding to the local constraints can be eliminated from the networks without a size penalty. They are just a convenient way to describe the constraints. The flags constraining long distance dependencies, on the other hand, help to keep the network sizes manageable.

5.3 The two morphological tool sets

Both the Humor analyzer and the Xerox tools are proprietary commercial software. Since Humor was developed by MorphoLogic, it was a natural choice for us to use in these projects. The Xerox tools were published on a CD accompanying Beesley and Karttunen (2003) published in June 2003, accompanied by license that made the version published with the book freely available for non-commercial purposes.
The Xerox tools have an advantage in terms of analysis speed over Humor of a factor of 1.5–4 at an expense of a

significant compile time and runtime memory requirement overhead. Depending on the complexity of the language and the structure of the word grammar, the runtime memory requirement of the Xerox lookup tool may be 10 times as much as that of the Humor analyzer for the same language (even when using Flag diacritics and transducer chains to reduce the memory requirements of the Xerox analyzer). The ratio of compile time memory requirement seems to be at least another order of magnitude higher (i.e. xfst may require more than a hundred times as much memory as the Humor lexicon compiler). 17 years ago, when the Humor analyzer was conceived, the compile time and even the runtime memory requirements of the finite-state tools would have been unfeasibly high. With today's RAM sizes, even a 30 MB analyzer lexicon does not seem to be a serious problem anymore. The Humor analyzer, however, seems to be more applicable in environments with limited memory resources.

The compile time memory requirement of xfst depends significantly on the compilation scenario used. The standard procedure suggested in (Beesley and Karttunen, 2003) of compiling the rule component separately by compiling and composing all the rules using xfst and then composing it with the lexicon compiled by lexc completely failed in a 512 MB machine for lack of memory when first trying to compile our Nganasan morphology. Finally, we managed to tackle this problem by changing the procedure of creating the final transducer: we composed the rules one by one with the lexicon. The lexicon constrained the space of possible underlying representations from the very beginning and thus the size of the network remained manageable throughout the whole compilation process.

6. A web based corpus annotation tool

Although morphological analyzers can be used to rapidly analyse huge amounts of text, they cannot be used alone to create morphosyntactically annotated corpora, because there is always a great degree of morphological ambiguity in the texts. In addition, corpora always contain a number of out of vocabulary word forms that the morphological analyzer is not able to recognize. Usually, some kind of morphological guessing may be used to solve this latter problem, but that usually leads to a disambiguation problem again: that of the possible guessed analyses. The morphological annotation needs to be disambiguated. Although there are standard (statistical) techniques of automatic disambiguated morphosyntactic (part of speech) tagging, these tagging tools must always be trained on manually disambiguated texts. And in fact for the automatic tagging to be of an acceptable accuracy, a huge amount of manually tagged training data is needed (and even then there will be tagging errors). Another problem with standard part of speech taggers is that they do not identify the lemma of words (only the part of speech tag), which is only half of the annotation that we would like to have. Moreover, the word form and the part of speech tag does not always identify the lemma unambiguously, because the paradigms of different lemmas quite often partially overlap at the same paradigm slots[6]. In those cases the lemma cannot be identified fully automatically from the part of speech tagged text. Thus manual disambiguation is inevitable (for at least a subset of the corpus). So a tool is needed that makes the manual disambiguation task as efficient as possible.

We have created a tool that can be used for the morphosyntactic annotation and manual disambiguation of corpora. In order to make the use of this tool efficient, we implemented it as a web application so that it can be concurrently used by linguists/native speakers remotely. It can of course also be installed on and used locally form a local web server.

After tokenizing and morphologically analyzing the text uploaded to the web server, the tool presents individual sentences to the user along with their context clearly indicating ambiguous and unanalyzed words, with the possibility of manually adding analyses of unknown words, removing bogus nonsense analyses (regular expressions can be used to override whole classes of unwanted analyses). The program uses statistical methods to initially rank analyses so that the automatically top ranked analysis of ambiguous words rarely need to be manually overridden. The program learns the decisions of the user. Initial ranking of the analysis candidates can be based on the output of a tagger, the accuracy of which can be incrementally enhanced by adding more and more texts to its training set. In addition to annotating words with their lemmas and morphosyntactic tags, the tool can be configured to add glosses in various languages. When, after making the needed adjustments, the top ranked analysis and glossing candidates are all deemed correct, the user can accept the sentence as correctly analyzed. Manually overridden ranking is always recorded as such. For each disambiguated sentence, the user id of the annotator is logged. Manual correction of typos in the original text is also possible. The user can also mark sentences as problematic. If an update of the database of the morphological analyzer is needed, the corpus can be reanalyzed using the recompiled analyzer without the already disambiguated and accepted sentences being affected.

7. Conclusions

In this paper, we have presented the results of completed projects as well as work in progress the goal of which is to create electronic linguistic resources for several minority languages spoken in Russia belonging to the Uralic language family, also comparing strengths and weaknesses of the two morphological toolsets used in the projects. We have also described a web based corpus annotation workbench that we developed.

A lesson that we learned from the projects is that the need of strict formalization when creating computational grammars may play an important role in creating more adequate grammatical descriptions. We have also found that classical linguistic fieldwork might not be the only

[6] E.g. most forms of the Hungarian verbs '(felül)múl' and 'múlik' coincide. There are many similar lemma pairs.

way to acquire linguistic data in endangered languages. Moreover, we think that further projects with the goal of providing tools such as spell checkers and electronic dictionaries to speaker communities of minority languages (and publishers of books and newspapers) could be a reasonable sequel to these projects.

8. Acknowledgements

The individual analyzers were created by Attila Novák in co-operation with László Fejes (Komi, Udmurt, Mari, Mansi, the grammar being mostly László Fejes's work), with Beáta Wagner-Nagy and Zsuzsa Várnai (Nganasan) and with Nóra Wenszky (Tundra Nenets). The Tundra Nenets analyzer is based on Tapani Salminen's work (his dissertation, Salminen (1997) and his morphological dictionary, Salminen (1998), which he kindly made available to us in a machine readable form) and was created in close on-line co-operation with him. The projects were funded by the National Research and Development Programmes of Hungary ('Complex Uralic Linguistic Database', NKFP 5/135/2001) and by OTKA ('Development of Komi and Udmurt morphological analyzers' (OTKA-T 048309) and 'Development of a Ngananasan morphological analyzer' (OTKA / K 60807)).

9. References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology.* CSLI Publications, Ventura Hall.

Ljucija Beznosikova, ed. 2000. *Komi-Roča Kyvčukör.* Syktyvkar.

N. T. Kost'erkina, A. Č. Momd'e, and T. Ju. Ždanova. 2001. *Slovar' nganasansko-russkij i russko-nganasanskij.* Prosvesčen'ije. Sankt-Pet'erburg.

Prószéky, Gábor and Balázs Kis. 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* pp. 261–268. College Park, Maryland, USA.

Tapani Salminen. 1997. *Tundra Nenets inflection.* Mémoires de la Société Finno-Ougrienne 227, Helsinki.

Tapani Salminen. 1998. *A morphological dictionary of Tundra Nenets.* Lexica Societatis Fenno-Ugricae 26, Helsinki.

Eslema. Towards a Corpus for Asturian

Xulio Viejo[‡], Roser Saurí[*], Ángel Neira[†]

[‡]Departamento de Filología Española
[†]Computer Science Department
Universidad de Oviedo
{jviejo, neira}@uniovi.es
[*]Computer Science Department
Brandeis University
roser@cs.brandeis.edu

Abstract

We present *Eslema*, the first project devoted to building a corpus for Asturian, which is carried out at Oviedo University. *Eslema* receives minor funding from the Spanish government, which is fundamental for basic issues such as equipment acquisition. However, it is insufficient for hiring researchers for a reasonable period of time. The scarcity of funding prompted us to look for much needed resources in entities with no institutional relation to the project, such as publishing companies and radio stations. In addition, we have started collaborations with external research groups. We are for example initiating a project devoted to developing a wiki-based platform, to be used by the community of Asturian speakers, for loading and annotating texts in Eslema. That will benefit both our project, allowing to enlarge the corpus at a minimum cost, and the Asturian community, causing a stronger presence of Asturian in information technologies and, as a consequence, boosting the confidence of speakers in their language, which will hopefully contribute to slow down the serious process of substitution it is currently undergoing.

1. Introduction

We present *Eslema*, the first project devoted to building a corpus for Asturian, which is carried out by the Research Group on Asturian Philology (*Seminariu de Filoloxía Asturiana*) within the Spanish Philology Department at Oviedo University (http://www.uniovi.es/eslema). Its goal is compiling a corpus representative of this Romance language, hence containing documents of a varied typology, including both written and oral discourse, and texts from the different historical periods of the Asturian language, starting from the medieval ages.

The name of the project, *Eslema*, is an Asturian word not documented in any dictionary but in a collection of oral Asturian folk tales (Suárez López, 2000). *Eslema* refers to an impression drawn in a stone indicating the existence of a hidden treasure. Aside from its obvious poetic force, this word also expresses the great value that a complete collection of documents, adequately structured and organized, has for research in linguistic-related disciplines.

The paper is organized as follows: The next section offers a small introduction to Asturian and its current situation, and section 3 presents the project in some detail. Then, section 4 comments on the resources (both material an human) that make Eslema possible. And finally, section 5 touches upon the social contribution of the project.

2. The Asturian language

Asturian (or, equivalently, Asturian-Leonese, Asturleonese, or Bable) is the autochthonous language of most of the territory in the Principality of Asturias, the provinces of Leon and Zamora, in Spain, and the district of Miranda do Douro (Portugal). Nowadays it is most actively used in the Asturian territory proper and in the north of Portugal, where the local version, Mirandese, achieved co-oficial status in 1998. The most reliable estimates of the status and vitality of Asturian nowadays calculate the community of speakers corresponds to approximately a third of the population. That is, there are about 300,000 speakers of Asturian in a community of 1,000,000 people, including about 50,000 speakers of Galego-Asturian concentrated in the westernmost part of Asturias. These figures bode ill for the future of the language since Asturian competence is notably reduced among young people, in such a way that the generational transmission of the language is seriously threatened. In fact, there has been a 20% loss of native speakers during the last decade (Llera Ramo, 2002).

The legal status of Asturian is very different from that of other languages in Spain, or other European minority languages such as Romansh, Irish, and Frisian. The Asturias Autonomy Statute (1981) recognizes the Asturian linguistic specificity in its fourth article, and establishes that "Bable will enjoy protection. Its use, diffusion in the mass media, and teaching will be promoted, respecting in every case the local variants and the voluntary nature of its learning." However, it does not bestow oficial status upon Asturian, or in other words, it does not recognize the civil rights of its speakers as such. Increasing social demand for linguistic normalization triggered an intense political debate about the inclusion of the official status of the language in the reformation of the Autonomy Statute in 1998 (blocked by the mainstream parties). It also led local politicians to enact the so-called "Law for the protection of Bable/Asturian", which explicitly includes some linguistic rights (e.g., street signs in Asturian, publication of administrative documentation and official advertising brochures in Asturian, etc.) and was unanimously approved by all political parties. So far, however, this law's normative intentions have not been significantly developed.

One of the main linguistic demands since the 1970s has been the possibility of teaching Asturian at schools. So socially deep-rooted is that demand that, as a matter of fact, it is specifically contained in the fourth article of the Autonomy Statute, as mentioned above. The first school-teaching initiatives took place in 1984. Nowadays, the teaching of Asturian has attained a remarkable expansion in primary schools (86% of the public schools, with 15,227 students in 2002), and a much more modest one in secondary school (20% of the institutions, with 2,171 students in that same year). Nonetheless, Asturian is conceived as a marginal subject and not as part of the core school curriculum, and it is often taught at inconvenient or unpopular times. In fact, its optional status implies that students frequently have to choose between studying Asturian or other subjects essential for their background, such as a foreign second language or computer programming. Viejo Fernández (2004) offers a detailed view of Asturian socio-linguistic situation.

Currently, Asturias is undergoing a new reformation of its statute, during which the issue of the Asturian legal status will probably be debated again. In spite of some slight advancements concerning the normalization of Asturian in the past few years, its officilization still appears to be a remote possibility given the opposing stance of the two main political parties. This situation is indeed of absolute relevance to Eslema. Not having a satisfactory political and legal framework seriously restricts the possibilities of an adequate linguistic normalization and, consequently, of any potential research on Asturian aiming at effective and practical outcomes. On the other hand, this precarious situation makes it even more necessary developing new strategies for ensuring the use of Asturian in everyday situations, as well as for boosting its prestige as a a modern language, capable of coping with the technology-based communication needs of present-day society.

3. Eslema. A General Corpus for Asturian

3.1. Project goal

Eslema was conceived as a framework to develop several subcorpora –hence the title General Corpus for the Asturian Language (*Corpus xeneral de la llingua asturiana*), which is part of the project's official name. Our initial goal is the construction of three of these subcorpora: the Corpus of Medieval Asturian-Leonese (including documents from the 13th-15th centuries), the Corpus of Classical Literature (between 1639 and 1950), and the Corpus of Present-day Asturian.

Being the corpus of a minorized language, the main objective of Eslema is helping set the foundation for fully normalizing Asturian as the language of use in any possible social context. At first impression, medieval and classical Asturian texts do not appear to be able to contribute to this goal, even less so considering the remarkable differences that exist between the historical and the current language (Viejo Fernández, 2003). We nevertheless decided to compile language samples from those periods. Aside from the fact that they are inherently interesting for linguistic and philological research in general, this decision follows two strategic reasons and thus aspires to have some degree of normalizing effect as well.

Asturian is generally seen through a Castilian perspective and considered a mere dialect, an approach that undervalues, or simply ignores, the long and solid written tradition that exists in that language. Sometimes its historical texts have been presented (with scarce scientific rigor) as either samples of "Old Spanish" or "dialectal literature" epiphenomena. As a matter of fact, most historical corpora for Spanish include medieval Asturian-Leonese texts, adhering to no linguistic criteria other than a unitarizing discourse of clear ideological and nationalistic objective, often times fostered by the Spanish philological tradition itself. In that sense, documenting in a coherent way the linguistic tradition of Asturian and its corresponding historic evolution is a priority issue.

There is yet an additional reason for compiling a historical corpus. Current Asturian, mainly relegated to oral linguistic exchanges, is affected by the strong pressure of Spanish, with which it tends to blend in an ongoing process of dissolution. Amidst this adverse situation, Asturian is currently undergoing the codification process that healthier languages (e.g., French, Spanish, English) experienced at an earlier time, mostly during the Renaissance. And for the success of that process, the existence of historical text has proven to be crucial. As it happens, oftentimes establishing the autochthonous vocabulary is approached from the mere intuition of intellectuals and writers, leading to unnatural or unpopular solutions which end up being rejected by the speakers. Therefore, having historical text dating previous to the Castilianization suffered in the past half century will substantially help codify and stabilize the language. What is more, medieval and classical Asturian texts provide consolidated models of specific domains and their corresponding terminology, such as administrative and juridic registers, which are fundamental when normalizing a language used mainly in oral contexts.

3.2. Subcorpora

3.2.1. Corpus of Medieval Asturian-Leonese

It contains several hundreds of texts in medieval Asturian and Leonese (two dialectal variations of the same language), all of them of legal typology. These documents are not available online yet, given that they are still pending annotation. Currently, we are working on the tag set design.

Aside from specific issues related to medieval legal language (e.g., what is an adequate textual typology, etc.), we need to address two additional problems. First, determining what kind of texts we consider as representative of the medieval Asturian-Leonese language; that is, setting a border between Latin and Romance, an issue that affects a considerable number of documents that date back to a time earlier than 1260. And second, addressing the process of hybridation of Asturian with respect to Spanish, which took place from the end of the 14th century and throughout the whole 15th century.

3.2.2. Corpus of Classical Literature

It is a wide collection of literary texts (aside from few exceptions) belonging to the genre of poetry. These are selected according to criteria concerning their linguistic qual-

ity given that, together with pieces of work of absolute linguistic quality, since the 19th century there exists a profusion of folk texts of minor quality, characterized by the use of very poor Asturian and often times written from Castilian hence, just imitative of the Asturian voice. We considered that this type of texts is not representative of the actual Asturian used at that period. However, determining the samples that are truly representative based on criteria that go beyond our mere subjective intuitions as speakers and philologists, is a challenging task.
The most representative part of this corpus is already in digital format, although it is not yet annotated and cannot be accessed online.

3.2.3. Corpus of Present-Day Asturian

This corpus collects documents in Asturian that have been produced during the last three decades. Currently, it is the most visible part of the project, although some great amount of it is still under development.
The Corpus of Present-Day Asturian consists of two subcorpora, a written and an oral one, each containing documents of a varied nature. The documents in the written subcorpus are available in text format. As for those in the oral subcorpus, we plan to have them available in audio format and accompanied with their corresponding transcription.
Documents in both subcorpora are of different types. There are mainly literary texts (from all genres), although we also have a significant sample of journalistic language (news reports, op-eds, etc.) and, to a lesser extent, of expert domain language (i.e., scientific, technical, administrative, etc.). Similarly, we aim at gathering a balanced sample of documents covering different geographical and social dialects although most of the text collected so far is in normative central Asturian.

Written documents. At the linguistic level, we are currently working on the annotation at the different relevant layers –mainly, morphology, part-of-speech, and basic syntax. This work is becoming particularly slow, given that we are unable to hire professional computer scientists (or computational linguists) who can work full time in the project. The only advancement in this area has been done thanks to the collaboration of computer science students. However, we foresee that we will be able to report initial results along this direction within the few coming months.

Oral documents. At present, Eslema has a wide collection of audio files, which will be available online momentarily. Some of them are representative of a more formal (or controlled) use of Asturian –basically, radio programs, broadcasting of sport events, news reports, etc. Some others, on the other hand, represent the colloquial (and spontaneous) use of Asturian. The progress made for this part of the corpus is slow mainly due to two reasons. First, the need for digitalizing the original recordings, which are in cassette support (they started being recorded in the 80s). And second, the fact that our goal is not only to make this material available online in audio format, but to provide their corresponding transcription as well.
In addition to these materials, which were generated previously to (and independently from) the beginning of Eslema, we are currently working on compiling further new oral documents, mainly dialogue, by means of surveys designed to that purpose. Throughout this past year, we have managed to record several hours of spoken language of high quality, from different dialectal areas in Asturias.

3.3. Document annotation

3.3.1. Metadata

Each document is described at the metadata level with XML tags codifying a subset of the 15 basic elements in the Dublin Core scheme, as customized by the Open Language Archives Community (OLAC). OLAC Metadata Set extends the Dublin Core in order to express community-specific needs, namely, those concerning archiving resources that document or describe languages –e.g., dictionaries and grammars (Bird & Simons, 2001).
We decided to adopt OLAC Metadata Set in order to adhere to an international standard that can potentially facilitate the future sharing of our data with other groups. Eslema does not consist of resources describing a language, but resources that exemplify and actualize it instead. In that respect, its contents do not correspond to the type of documents targeted by OLAC. However, because Eslema's focus is on representing language as well, the set of basic elements and attributes considered by OLAC are fully adequate for describing its documents.
The following list provides the name of the OLAC elements used in Eslema to encode document metadata, with a brief description if deemed necessary:

- **Contributor.** The entity that provided the document.
- **Creator.** The person (or institution) responsible for the content of the document (e.g., author, translator).
- **Date.** The document creation time.
- **Description.** Free description of the document content.
- **Identifier.** Unambiguous reference to the resource (e.g., ISBN)
- **Language.** Asturian.
- **Publisher.**
- **Relation.** A reference to a related resource, if applicable.
- **Subject.** Keywords describing the topic of the content.
- **Title.**
- **Type.** Sound (for recordings of oral discourse) and text (for written documents).

In addition, documents are characterized according to a textual typology we have created, which classifies them according to their genre and the geographical and social dialect they represent. Through this classification, we can control the representativity and balance of the corpus, facilitate the process of information search, and contribute to future research projects concerning the Asturian language. The following distinctions are made:

- **Channel.** The original physical medium of the document contents: oral or written.
- **Textual tipology.** With possible values: *colloquial, literary, journalistic, political, advertising, scientific-professional, legal-administrative,* and *ritual*.
- **Diasistemic properties.** Identifying the dialectal variety of the text.

3.3.2. Linguistic information

The encoding of linguistic information is currently one of the main bottlenecks of our project, due to the scarcity of funding and the lack of (volunteering or poorly paid) personal with an adequate formation for developing and customizing linguistic processing tools.

We are currently working on the POS and morphological annotation. So far, we have already delimited the POS tagset to be used, and are exploring what is the best solution for text tagging given existing tools for similar languages. The tagset is mostly based in the one for Spanish adopted in the FreeLing platform of linguistic analyzers (Carreras et al., 2004),[1] inherited from the EAGLES initiative for annotating morphological and syntactic information in lexicon and corpora, in all European languages.

Furthermore, the specificity of certain grammatical elements in Asturian made it necessary to minimally extend the Spanish POS tagset. For example, the gender attribute in nouns, adjectives, and determinants needs to account for the possibility of the neutral value as well.

3.4. Eslema search tool

Eslema is provided with a basic search tool, which has access to a collection of around 8,000,000 words from the written part of the Corpus of Present-Day Asturian. It allows for two different types of search, simple and complex, which so far can only be formulated in terms of the attributes encoding information at the document metadata level. This tool, however, is about to be substituted by a more efficient index-based tool, more in accord with standard approaches to search engines. In addition to the already available search features of the current tool, the new one will allow for context- and linguistic-based searches, as it is expected from any corpus query tool. Figure 1 illustrates the interface of the of the current version (for simple queries).

Figure 1: Current Eslema's search tool

[1] FreeLing is presented at http://garraf.epsevg.upc.es/freeling/. The specifics of the Spanish tagset can be found in: http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html.

4. The few resources of a less-resourced language

4.1. Funding sources

Eslema was born in the Research Group on Asturian Philology (*Seminariu de Filoloxía Asturiana*), within the Department of Spanish Philology at Oviedo University, and was initially supported by specific grants from the Culture Ministry of the Asturian Government, through the Office for Linguistic Planning. Since 2006, it is benefiting from funding from the National Plan for Research and Development of the Spanish Ministry of Culture and Education (period 2006-2008). Such funding is fundamental for basic issues such as equipment acquisition, or for covering for specific activities related to the project. However, it is fairly limited in the sense that it is insufficient for hiring researchers or project assistants for a reasonable period of time.

This situation has very negative consequences on the pace of the project, and has prevented us from making the progress we had initially planned. It is for that reason that, currently, more than a corpus appropriately annotated and provided with an adequate set of tools for accessing and managing its information, Eslema is simply a repository of documents with a basic search tool associated to it.

At present, we plan to apply for the yearly research grants offered by the Asturian Government for work on the autochthonous language, which would allow us to pay students and hire collaborators who can then work on Eslema full time, even if it is at a yearly basis.

4.2. Eslema's human face

In spite of the economic situation of the project, we consider that the amount of work accomplished up to now is remarkable. This has been possible thanks to the collaboration of professors and researchers from different departments at Oviedo University (namely, Spanish Philology and Computer Science), the assistance from researchers at Brandeis University and Georgetown University, in the Unites States of America, the specific contribution of graduated philologists as well as computer science students, and most remarkably, the unconditional dedication of very faithful supporters of the project, who cannot always be paid for their work but whose contribution is truly essential.

Furthermore, we benefit from the interest shown by some students in the undergraduate degree of Philology. In exchange for working on enlarging the corpus (mainly, collecting oral discourse and generating its corresponding transcription), we let them use Eslema as the framework within which to develop research projects funded by specific student scholarships.

Integrating all these collaborators in a project with such limited amount of resources is a challenging and arduous task. Even more so given the fact that, in the Asturian scientific scene, there was an absolute lack of experience in corpus linguistics, not to mention computational linguistics –cf. Saurí Colomer (2004). Furthermore, most of the people showing some interest in the project were coming from fairly unrelated backgrounds. They were either computer scientists with no knowledge whatsoever on linguistics, or

philologists with no computational background who, in addition, were specialized in very different areas of the field –e.g., language history, diachronic linguistics, literature, grammar.

Because of this difficult balance among so varied backgrounds, we are presently looking for additional funding for supporting initiatives to further educate our community of researchers, students, and collaborators on aspects that are of central relevance within the scope of the project, mainly by organizing short term courses that can address these topics.

4.3. External data contributors

The scarcity of funding prompted us to look for much needed resources to entities with no institutional relation to the project, but which nevertheless got involved and contributed in a very remarkable way. In particular, we are referring to the publishing companies Trabe and Ámbitu, which provided us with an important collection of digitalized text archives, and to the broadcasting station Radio Sele, which contributed a set of audio archives containing discourse of a genre not very common for the Asturian language –broadcastings of sport events, news reports, interviews, etc. Aside from their obvious value in terms of material resources, these contributions have been an important source of encouragement to us, and have helped project our work outwards at the social scene.

4.4. Collaborating with other research groups

Research funding in general is fairly neglected by our local administration. Because of that, we have decided to move to a more practical phase, focused on developing tools and products that can be of interest to private companies but also, and very specifically, to the public administration. For that purpose, we are currently in contact with the Culture Ministry of the Asturian Government concerning the creation of a Spanish-Asturian/Asturian-Spanish translator, a tool that would be inspired on the Opentrad translator between Spanish and the co-official languages under the Spanish administration (i.e., Catalan, Euskera, and Galician) (Alegría Loinaz et al., 2006), and which would effectively contribute to the much needed normalization of Asturian.

Given the limited material and economic conditions of our project, this constitutes a clear challenge. For that reason, we initiated collaboration contacts with groups with similar research interests. In particular, Xavier Guinovart, the head of the Research Group on Computational Linguistics (*Seminario de Lingüística Informática*) at Vigo University, has shown great interest in our translation project, to which his group could contribute the expertise it acquired during the development of the Galician part of Opentrad.

In parallel with that, we are also in collaboration with two research groups at Brandeis University (the Lab for Linguistics and Computation, and the Group for Research On Usage and Pragmatics), as explained in the next section.

4.5. Opening Eslema to the Asturian community. Wiki technology

Beyond the inherent interest Eslema offers to researchers on Asturian, the corpus was conceived with a true vocation of helping in both codifying the language and normalizing its use, or in other words, aiming at assisting Asturian speakers to be granted their linguistic rights.

As it happens, however, not only Eslema can benefit the Asturian community, but benefit from it as well. As is the case with many minorized languages in Western societies, a good number of Asturian speakers feels a fair degree of commitment to Asturian and its survival, an attitude that certainly plays in our favor.

Added to that, there is the fact that the role of information technology in Asturias has increased notably in the past few years. According to the latest report by Telefónica on the use of this type of technology within the borders of the Spanish administration, internet access in Asturian homes increased from 21.38% in 2003 to 41.4% in 2006, placing this community slightly above the Spanish average (39%).[2]

The two necessary ingredients for involving Asturian society to the project are therefore at place: a motivated and committed attitude of speakers towards their language, and the basic technology for them to collaborate. Furthermore, that there is an interest on the population side towards making Asturian present in modern technologies can be attested by, for example, its incidence in the popular Wikipedia. Out of the 255 languages covered there, Asturian is ranked as the 74th in terms of number of articles (11,130), the 75th in terms of collaborators (833), and the 62th in terms of editing activity (242,204).[3]

Encouraged by these favorable conditions, we are currently collaborating with the Lab for Linguistics and Computation and the Group for Research On Usage and Pragmatics, both at Brandeis University, in developing a wiki-based platform for loading and annotating texts in Eslema. The relevance of wiki technology for facilitating online collaboration among the members of community of practice is already well-known. Wikis are easy to use, flexible in nature, and provide with a set of core components which can promote interaction among collaborators in various ways.

Previous research developed at Brandeis has focused on the use of wikis for different educational and research tasks (Larusson & Alterman, 2007). A product resulting from that research is a wiki environment called WikiDesignPlatform (WDP) (Larusson & Alterman, 2008), which is conceived as an online meeting space for members in the same community of practice and, at the same time, a central repository and workspace. In addition, the WDP has been designed so that it can be tailored for other kind of collaboration activities.

So far, the emphasis of that research has been put on using wiki technology to support class work. Our collaboration will therefore be beneficial to both sides. Eslema will become the testbed for expanding WDP to different domain

[2] Refer to *La sociedad de la información en España 2007*, available at: http://sie07.telefonica.es/aplicacion_sie.html.

[3] http://meta.wikimedia.org/wiki/List_of_Wikipedias#10.2B_articles. Data obtained on March 30th, 2008.

needs and, at the same time, the resulting WDP will benefit our project both at the material level, allowing to enlarge the corpus at a minimum cost, and socially, making Asturian more active within information technologies and, consequently, contributing to slow down the serious process of substitution it is currently undergoing.

5. Contribution of the project

5.1. Increasing linguistic resources in Asturian

The availability of Eslema as a digitalized and structured collection of documents in Asturian is of great significance for linguistic research on this language. Eslema has opened venues for advancing in the knowledge of Asturian lexicon and grammar at the theoretical level. Within the Eslema framework, for example, a programmatic work on Asturian syntax has just been published (Viejo Fernández, 2008). This kind of work does not represent any novelty in more stable languages, but it is certainly a new contribution to the Asturian linguistic community. Along similar lines, there is a doctoral dissertation currently under development devoted to the analysis of prepositions, oblique complements, and verbal periphrasis in Asturian (Hinojal Díaz, frth). And finally, it is about to be published a wide anthology of historical texts (from Medieval Ages until nowadays) written in the Occidental Asturian dialect (Cueto & Viejo, frth).
Furthermore, Eslema also plays a central role in the task of setting the autochthonous vocabulary and terminology, thus contributing to its codification. The adverse socio-political conditions of Asturian during the past centuries and its resulting diglossic situation caused all lexicographic enterprise to fail (Arias Cabal, 1996), at the time when other languages underwent the process of codification that guaranteed their nowadays healthy status; e.g., English, French, or Spanish. The sample of medieval and classical Asturian texts in Eslema, which date previous to the strong Castilianization suffered in the past half century, is of great use for recovering Asturian lexicon. As a matter of fact, the language from those periods provides consolidated models of specific domains and their corresponding terminology, such as administrative and juridic registers.

5.2. Social projection: boosting the confidence in our language

An incipient social projection of Eslema has manifested, for instance, through references to our work in *Radio Sele*, or through an article published in the Asturian weekly magazine *Les Noticies*, which caused an increase in the number of online visits to our corpus. In terms of the small social scale that defines the Asturian community, such public projection has generated some degree of expectation at two different levels. On the one hand, at the administration level, situation that can be of help to us when looking for funding in order to pursue new lines of research. On the other, among the general public, which tends to appreciate Asturian from a diglossic bias and now has the chance to start looking at it as an adequate and modern language because it is capable of benefiting from information technology. This boosting of confidence in Asturian will hopefully contribute to slow down, or even reverse, the serious process of substitution it is currently undergoing.

References

Alegría Loinaz, I., Arantzabal, I., Forcada, M. L., Gómez Guinovart, X., Padró, L., Pichel Campos, J. R., & Walino, J. (2006). OpenTrad: Traducción automática de código abierto para las lenguas del estado español. *Procesamiento del Lenguaje Natural*, *37*, 357–358.

Arias Cabal, A. (1996). La lesicografía asturiana. Cronoloxía de doscientos años d'intentos. *Lletres Asturianes*, *6*, 41–63.

Bird, S. & Simons, G. (2001). The OLAC Metadata Set and Controlled Vocabularies. In *Proceedings of the ACL Workshop on Sharing Tools and Resources for Research and Education*, (pp. 7–18)., Toulose. ACL.

Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). FreeLing: An open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisboa.

Cueto, M. & Viejo, X. (frth). *Antoloxía de testos n'asturianu occidental*. Asociación de Escritores en Asturiano Occidental. Forthcoming.

Hinojal Díaz, R. (frth). *Complejos verbales y regimen preposicional de los verbos asturianos*. PhD thesis, Universidad de Oviedo. Forthcoming.

Larusson, J. A. & Alterman, R. (2007). Tracking online collaborative work as representational practice: Analysis and tool. In Steinfield, C., Pentland, B., Ackerman, M., & Contractor, N. (Eds.), *Communities and Technologies 2007: Proceedings of the Third Communities and Technologies Conference*, (pp. 245–264)., Michigan.

Larusson, J. A. & Alterman, R. (2008). Wiki technology for collaborative learning. In *Supporting and Tracking Collective Cognition in Wikis Symposium at ICLS 2008*, Utrecht. Forthcoming.

Llera Ramo, F. (2002). II Estudiu siciollingüísticu d'Asturies. Avance de datos. *Lletres Asturianes*, *89*, 181–197.

Saurí Colomer, R. (2004). Un corpus pal asturiano. les tecnoloxíes llingüístiques na consolodación de les llingües minorizaes. *Revista de Filoloxía Asturiana*, *3/4*, 135–174. Spanish version: http://www.cs.brandeis.edu/ roser/pubs/rfa_corpus_sp.pdf.

Suárez López, J. (2000). *Tesoros, ayalgas y chalgueiros. La fiebre del oro en Asturies*. Muséu del Pueblu d'Asturies.

Viejo Fernández, X. (2003). *La formación histórica de la llingua asturiana*. Uviéu: Trabe.

Viejo Fernández, X. (2004). Asturian: Resurgence and impeding demise of a minority language in the Iberian Peninsula. *International Journal of the Sociology of Language*, *170*, 169–190.

Viejo Fernández, X. (2008). *Pensar asturiano. Ensayos programáticos de sintaxis asturiana*. Uviéu: Trabe.