Strategies for developing machine translation for minority languages

5th SALTMIL Workshop on Minority Languages
Tuesday May 23rd 2006 (morning), Genoa, Italy
Organised in conjunction with LREC 2006: Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, 24-26 May 2006

Workshop Programme

8:00 Registration and preparation of posters

8:15 Welcome: Briony Williams (University of Wales, Bangor, UK)

8:30 Stephen Krauwer (University of Utrecht, Netherlands)
"ENABLER, BLARK, what's next?"

9:00 Delyth Prys (University of Wales, Bangor, UK)
"The BLARK matrix and its relation to the language resources situation for the Celtic languages"

9:30 Christian Monson, Ariadna Font Llitjós, Roberto Aranovich, Lori Levin, Ralf Brown, Eric Peterson, Jaime Carbonell & Alon Lavie (Carnegie-Mellon University and University of Pittsburgh, USA)
"Building NLP Systems for Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua"

10:00 Mikel Forcada (University of Alacant, Spain)
"Open source machine translation: an opportunity for minor languages"

10:30 Anna Sågvall Hein & Per Weijnitz (University of Uppsala, Sweden)
"Approaching a new language in machine translation"

11:00 COFFEE

11:30 Daniel Yacob (Director, Ge'ez Frontier Foundation)
"Unicode Development for Under-Resourced Languages".

12:00 Maja Popović & Hermann Ney (RWTH Aachen University, Germany)
"Statistical Machine Translation with and without a bilingual training corpus"

12:30 POSTER SESSION

Tod Allman & Stephen Beale, "A Natural Language Generator for Minority Languages".

Saba Amsalu & Sisay Fissaha Adafre, "Machine Translation for Amharic: Where we are".

Carme Armentano-Oller and Mikel Forcada, “Open-source machine translation between small languages: Catalan and Aranese Occitan”.

Arantza Casillas, Arantza Díaz de Illarraza, Jon Igartua, R. Martínez & Kepa Sarasola, “Compilation and Structuring of a Spanish-Basque Parallel Corpus”.

Bostjan Dvorák, Petr Homola, Vladislav Kubon, “Exploiting Similarity in the MT into a Minority Language”.

Adrià de Gispert and José B. Mariño, “Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish”.

Jorge González, Antonio L. Lagarda, José R. Navarro, Laura Eliodoro, Adrià Giménez, Francisco Casacuberta, Joan M. de Val and Ferran Fabregat, “SisHiTra: A Spanish-to-Catalan hybrid machine translation system”.

Dafydd Jones and Andreas Eisele, “Phrase-based statistical machine translation between Welsh and English”.

Svetla Koeva, Svetlozara Lesseva and Maria Todorova, “Bulgarian sense tagged corpus”.

Mirjam Sepesy Mauèec and Zdravko Kaèiè, “Statistical machine translation using the IJS-ELAN Corpus”

Sara Morrissey and Andy Way, “Lost in Translation: The Problems of Using Mainstream MT Evaluation Metrics for Sign Language Translation”.

Alicia Pérez, Inés Torres, Francisco Casacuberta and Víctor Guijarrubia, “A Spanish-Basque weather forecast corpus for probabilistic speech translation”.

Kevin Scannell, “Machine translation for closely related language pairs”.

Oliver Streiter, Mathias Stuflesser and Qiu Lan Weng, “Models of Cooperation for the Development of NLP Resources: A Comparison of BLARK and XNLRDF”.

John Wogan, Brian Ó Raghallaigh, Áine Ní Bhriain, Eric Zoerner, Harald Berthelsen, Ailbhe Ní Chasaide and Christer Gobl, “Developing a Spoken Corpus and a Synthesiser for Irish”.

Pavel Zheltov, “An Attribute-Sample Database System for describing Chuvash affixes”.

13:30 <u>END OF WORKSHOP</u>

Strategies for developing machine translation for minority languages

5th SALTMIL Workshop on Minority Languages
Tuesday May 23rd 2006 (morning), Genoa, Italy

Workshop Organiser

Dr Briony Williams
University of Wales Bangor, UK

Programme Committee

Dr Briony Williams
University of Wales Bangor, UK

Dr Kepa Sarasola
University of the Basque Country, Spain

Dr Bojan Petek
University of Ljubljana, Slovenia

Ms Atelach Alemu Argaw
Stockholm University/KTH, Sweden

Dr. Julie Berndsen
University College Dublin, Ireland

Mikel Forcada
Universitat d'Alacant, Spain

Lori Levin
Carnegie Mellon University, USA

Anna Sågvall Hein
Uppsala University, Sweden

Strategies for developing machine translation for minority languages

5th SALTMIL Workshop on Minority Languages
Tuesday May 23rd 2006 (morning), Genoa, Italy

Table of Contents

A: Invited Papers

B: Contributed Papers

Open-source machine translation: an opportunity for minor languages

Mikel L. Forcada

Transducens group, Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant (Spain)
and
Prompsit Language Engineering,
Polígon Industrial de Canastell, Ctra. d'Agost, 77, office 3, E-03690 Sant Vicent del Raspeig (Spain)
mlf@ua.es

Abstract

I have explored the positive effects that the availability of machine translation may have on the status and development of minor languages, focusing in particular on those effects which are specific to *open-source* machine translation, examining the challenges that should be met, and illustrating it with a case study.

1. Introduction

In this paper I explore the opportunities offered by open-source machine translation to what will be referred to as *minor languages*, and examine the challenges ahead. To begin with, the concepts which make up the paper's title will be briefly reviewed in this introduction. The rest of the paper is organized as follows: section 2 discusses the effects of the availability of machine translation on minor languages; section 3 describes some of the limitations of commercial machine translation in this respect; section 4 focuses on the specific opportunities offered by open-source machine translation; section 5 examines the challenges faced, and section 6 illustrates some of the issues with a case study. Closing remarks are found in section 7.

1.1. Minor languages and minor language pairs

Most readers will be familiar with the concept of minor languages; I have, however, found it a bit harder than I thought to define exactly what I intended to refer to.

To start with, many different names are used more or less interchangeably with the one in the title. In addition to the denomination *minor languages*, I have observed also the following, related names (Google counts as of March 28, 2006 are given in parentheses): *minority languages* (as in SALTMIL, Speech and Language Technologies for Minority Languages, 885,000), *lesser-used languages* (as in EBLUL, the European Bureau for Lesser Used Languages, 93,100), *small languages* (57,100), *smaller languages* (22,900), *lesser languages* (932), *under-resourced languages* (212), *resource-poor languages* (116) *less-resourced languages* (17), etc. (the Google count for *minor languages* is 77,800). I will try to address the issue without considering all the different shades of meaning in each name (for example, a minority language in one country can be a large language in the world, such as Gujarati in the UK). I will use the term *minor language* to refer to a language exhibiting some, if not all of the following features:

- having a small number of speakers (or, as *translation* —of texts— is concerned, having a small number of *literate* speakers)
- being used far from normality (being more used at home or in family situations than in school, commerce or administration, being socially discriminated, politically neglected, under-funded, banned, or repressed, etc.)
- lacking a unique writing system, a stable spelling, or a widely-accepted standard variety of the language
- having a very limited presence on the Internet[1]
- lacking linguistic expertise
- lacking machine-readable resources such as linguistic data, corpora, etc. and being dependent on external technologies[2]

While the status of a single language may also be affected by the availability of machine translation for it, it should be made clear that machine translation deals with language *pairs* and that effects on the minor language will occur through other languages, for example, major languages having translation relationships with the minor or related minor languages (if two minor languages A and B are very related, it will be easier to build a machine translation system between them;[3] if there is already machine translation from a major language C to one of them, A, the other minor language B may benefit from the existence of (indirect) machine translation towards it[4]).

1.2. Open-source and free software

I will briefly review the concept of *open-source* software, or, to use its historical name still in use, *free software.* Free software (the reader will find a definition at http://www.gnu.org/philosophy/free-sw.html) is software that (a) may be freely executed for any purpose, (b) may be freely examined to see how it works and may be freely modified to adapt it to a new need or application (for that, source code must be available, hence the alternative name

[1] Or as Williams et al. (2001) would put it, being "nonvisible in information system mediated natural interactivity of the information age"

[2] Ostler (1998) rephrases Max Weinreich's famous "A shprakh iz a dyalekt mit an armey un flot" (yiddish for "a language is a dialect with an army and a navy") into "a language is a dialect with a dictionary, grammar, parser and a multi-million-word corpus of texts — and they'd better all be computer tractable."

[3] For instance, Irish Gaelic and Scottish Gaelic (Scannell 2006).

[4] See Dvorak et al. (2006), de Gispert et al. (2006)

open source), (c) may be freely redistributed to anyone, and (d) may be freely improved and released to the public so that the whole community of users benefits (source code must be available for this too). The Open Source Initiative establishes a definition (http://www.opensource.org/docs/definition.php) which is roughly equivalent for the purposes of this paper. I will use *open source* instead of *free* because of the ambiguity of the word *free* in English.

1.3. Machine translation

Machine translation (MT) software is special in the way it strongly depends on data. Rule-based machine translation (RBMT) depends on linguistic data such as morphological dictionaries, bilingual dictionaries, grammars and structural transfer rule files; corpus-based machine translation (such as statistical machine translation, for example) depends, directly or indirectly, on the availability of sentence-aligned parallel text. In both cases, one may distinguish three components: an *engine* (decoder, recombinator, etc.), *data* (either linguistic data or parallel corpora), and, optionally, *tools* to maintain these data and turn them into a format which is suitable for the engine to use.

This paper will focus on rule-based machine translation, not only because I am more familiar with rule-based approaches or because another invited paper (Ney 2006) specifically addresses corpus-based machine translation, but also because of another reason: in the case of minor languages it is quite hard to obtain and prepare the amounts of sentence-aligned parallel text (of the order of hundreds of thousands or millions of words) required to get reasonable results in "pure" corpus-based machine translation such as statistical machine translation (SMT); however, it may be much easier for speakers of the minor language to encode the language expertise needed to build a rule-based machine translation system.

1.3.1. Commercial machine translation

Most commercial machine translation systems are rule-based (although machine translation systems with a strong corpus-based component have started to appear[5]). Most RBMT systems have engines with proprietary technologies which are not completely disclosed (indeed, most companies view their proprietary technologies as their main competitive advantage). Linguistic data are not fully modifiable either; in most cases, one can only add new words or user glossaries to the system's dictionaries, and perhaps some simple rules, but it is not possible to build complete data for a new language pair and use it with the engine.

1.3.2. Open-source machine translation

On the one hand, for a rule-based machine translation system to be "open source", source code for the engine and tools should be distributed as well as the "source code" of linguistic data for the intended pairs. It is more likely for users of the open-source machine translation to change the linguistic data than to modify the machine translation engine; moreover, for the improved linguistic data to be used with the engine, tools to maintain them should also be distributed. On the other hand, for, say, a statistical machine translation system, source code both for the programs that learn the statistical translation models from parallel text as well as for the decoders that use these language models to generate the most likely translations of new sentences should be distributed *along with the necessary sentence-aligned parallel texts*.[6]

1.3.3. Machine translation that is neither commercial nor open source

So far I have mentioned commercial machine translation and open-source machine translation. The correct dichotomy would be open-source MT versus "closed-source" MT; indeed, there are a number of systems that do not clearly fit in the categories considered in the last two sections.

For example, there are MT systems on the web that may be freely used (with a varying range of restrictions); some are demonstration versions of commercial systems, whereas some other freely-available systems are not even commercial.[7]

Another possibility would be for the MT engine and tools not to be open-source (even using proprietary technologies) but just to be simply freely available and fully documented, with linguistic data being distributed openly (open-source linguistic data). This intermediate situation will be addressed later in this paper (see section 4.1.1).

2. Effects of the availability of MT on minor languages

The following sections address, without the aim of being exhaustive, a list of the effects of the availability of MT between surrounding major languages on the status of a minor language and its community, regardless of whether the MT system is open-source or not.

The objective is to "de-minorize" the minor language. Therefore, one should consider the effects on the indicators or features listed in section 1.1. Not all of the indicators are equally affected; the major effect would be on four of them, as follows.

2.1. Increasing "normality"

The availability of MT from one of the surrounding dominant languages may contribute to the increase of "normality" in the sense of extending the use of the minor language from familiar and home use to more formal social contexts such as schooling, media, administration, commercial relations, etc. Just to name a few examples:

[5] AutomaticTrans (www.automatictrans.es), Language Weaver (www.languageweaver.com).

[6] This last requirement may sound strange to some but is actually the SMT analog of distributing linguistic data for a RBMT system.

[7] This is the case, for example, of two non-commercial but freely available machine translation systems between Spanish and Catalan: interNOSTRUM (www.internostrum.com), which has thousands of daily users, and a less-known but powerful system called SisHiTra (González et al. 2006).

- Educational materials in one of the dominant languages could be translated into the minor language so that children may be schooled in this language.
- News releases from agencies in one of the dominant languages may be translated into the minor language to create written media for that language community.
- Laws, regulations, government informations, announcements, calls, etc. may be translated into the minor language.
- Companies would have it much easier to market new products in the minor language ("localization"), especially those in which the text component is important such as consumer electronics, mobile phones, etc.

Of course, it is assumed that it is feasible to post-edit the results of machine translation into adequate texts. Therefore, the positive effects mentioned will be more likely to occur when language divergences are small.

2.2. Increasing literacy

The increasing availability of text in the minor language, obtained through translation and subsequent elaboration of material originally written in a major language may motivate efforts to improve the levels of literacy of speakers of that language community.

2.3. Effects on standardization

The use of MT systems may contribute to the standardization of a language, for example, by promoting a particular writing system (a current debate in cases such as that of Tamazight[8]), a particular spelling system (Mason and Allen 2001), or a particular dialect (for example, the effort of the Catalan government in Spain to normalize the Aranese variety of Occitan[9] and to generate linguistic technology for it, as compared to the technological efforts addressing other varieties of this language, may increase the weight of this variety in a possible future standard for the whole language; see section 6).

2.4. Increasing "visibility"

The availability of MT from the minor language into one or more of the surrounding major languages may help the diffusion of material originally written in the minor language.

For instance, the content of websites could be authored and managed directly in the minor language and machine-translated for users of other major languages, either on-the-fly or after being revised by professionals.

3. Commercial MT systems and minor languages: limited opportunities

To start with, the main commercial MT systems are built by (usually multinational) companies whose business objectives concern major world languages, rather than minor languages. As a result, it is quite hard to find commercial MT for minor languages, and therefore the "generic" positive effects mentioned in section 2 will be hard to come by.

There are interesting exceptions: for instance, minor languages in Spain such as Catalan or Galician have commercial MT systems available; this may be due to the fact that laws grant linguistic rights to speakers of these languages, which are official in areas of Spain having a limited home-rule status, and are therefore becoming an interesting market for these companies. Most of these commercial initiatives have been partially funded by the corresponding local governments, as part of their local-language policies.

But, as has been mentioned, both the closed nature of the technologies used in their engines and the limitations to modify the linguistic data they use make it hard to adapt commercial machine translation systems to new language pairs.

4. Opportunities from open-source MT systems.

Open source machine translation systems have started to appear (an example is given in section 6); in fact, even a company in the commercial MT business has considered moving towards open-source distribution of their products.[10]

I will contend that open-source MT systems provide much better opportunities for minor languages than commercial, closed-source systems, as discussed in the following subsections. This is because, *in addition* to the "generic" positive effects mentioned in section 2, open-source MT may also have effects on the remaining indicators mentioned in 1.1.

4.1. Increasing "expertise" and language resources

A variety of different situations may occur when trying to build open-source machine translation for a new language pair involving a minor language. All of them involve to some extent a process of reflection about the minor language, leading to elicitation and subsequent fixation and encoding of monolingual and bilingual knowledge about it. The resulting linguistic expertise, in an open-source setting, would be made available to the whole language community. But the most important effect would be the generation of new, openly available language resources for the community of speakers of the minor language.

Let us consider a number of different situations.

4.1.1. Building data for an existing MT engine from scratch

The minimum set of resources needed to build a new language pair would be: (a) a freely available (even if not

[8] Also called Berber, a language spoken in North Africa (http://en.wikipedia.org/wiki/Berber_language).

[9] A language (http://en.wikipedia.org/wiki/Occitan) spoken in southern France, certain valleys of western Italy and a valley in nortwestern Spain, also known as Provençal or *Langue d'Oc*. It was one of the main literary languages in Middle-Age Europe but is now severely minorized ("patois") after centuries of neglect and active repression.

[10] LOGOS has recently released the sources to its OpenLogos MT system (www.logos-os.dfki.de).

open-source) engine for another language pair, (b) a freely available (even if not open-source) set of tools to manage linguistic data in connection with that engine, and (c) *complete documentation* on how to build, using the provided tools, linguistic data to use with that engine.

In this case, one could build from scratch a whole set of data for the new pair, but this is a very unfavorable setting, especially if one considers the initial lack of expertise, the need to study and understand a potentially complex documentation, and the difficulty of making initial decisions about the languages involved, such as defining the set of lexical categories, defining the set of indicators that will be used to represent their inflection, etc. Paralyzing symptoms of what one could call the "blank sheet syndrome" are likely to show.

If the initiative succeeded, the resulting data could be made open ("open source") and distributed to the community so that they could be improved, adapted for new applications or subject fields, or used to generate data for new language pairs, as discussed in the next section, multiplying the positive effect on the minor language.

4.1.2. Building data for an existing MT engine from existing language-pair data

If open-source data are available for another language pair in which one of the languages is similar or related to the minor language in question, one can transform the existing data, which is much easier (as the "blank sheet syndrome" mentioned above is avoided). For example, one could use the same set of lexical categories and inflection indicators, and perhaps one could reuse many structural transfer rules which are not dependent on the particular lexicon; inflection paradigms in related languages tend to have similarities which could be exploited to build morphological dictionaries, etc. The resulting language-pair data could also be made open-source and distributed for use with the freely available engine and tools, as mentioned in the previous section. Note that neither the engine and nor tools have been required to be open-source, but just to be freely available, well documented, and built with a clear distinction between algorithms (engine, tools) and (linguistic) data.

4.1.3. Adapting an open-source engine or tools for a new language pair

If the source code is available for the engine and the tools, the community of experts of the minor language could enhance or adapt them, for example, to address features of the minor language that are not adequately dealt with by the current code.

For example, the code may not be prepared to deal with the particular character set of the language, or its transfer architecture may not be powerful enough to perform certain transformations which are needed to get adequate translations.

This poses additional problems to what would be simply building new linguistic data, but it may be the case that the rewriting of the engine and the tools could be tackled by programmers and computational linguists which do not have full command of the minor language (which would be needed to build lexicons, etc.) but who are aware of the linguistic issues in a more abstract way. Indeed, a good separation of (linguistic) data and (translation engine) algorithms becomes crucial for the success of this task.

The open distribution of the new engine, the new tools, and the linguistic data would contribute new linguistic technology resources as well as increase the expertise available to the minorized language community.

4.2. Increasing independence

An interesting side effect of the dissemination of open knowledge and open-source software for language pairs involving the minor language would make the users of this language community less dependent on a particular commercial, closed-source provider, not only for translation technologies, but perhaps for many other fields of linguistic technology.

5. Challenges

By now it must be rather clear that open-source machine translation generates opportunities for the growth of minor languages into normal, visible, and standard languages for communication in the Information Era. But to take advantages of these opportunities, the language communities involved have to face a number of challenges, some of which have already been mentioned. Let us review a few of them.

5.1. Standardization of the minor language

Section 2.3 discussed the benefits of having machine translation on the standardization of the minor language. But this potential may also have its downside: the lack of a commonly accepted writing system, spelling rules, or a reference dialect may actually pose a serious challenge to anyone trying to build a MT system for that minor language (one could call it "the pioneer syndrome").

5.2. Neutralizing technophobic attitudes

Even if a minor language has a very motivated and well-educated set of language activists, a connection must be made between this expertise and information-technology literacy. And this may be difficult; in the Catalan language community I have detected what I would call "technophobic attitudes": some highly educated, literate people distrust technologies because of their idealized view of language and human communication, and their low appreciation of non-formal or non-literary uses.[11] Any group of people endeavoring to build open-source machine translation systems for a minor language must be prepared to address this kind of, let us call them "socio-academic", adversities.

[11] Here is another possible explanation for some of these technophobic attitudes: many of these language professionals tend to focus more on usually highly improbable phenomena which are unique to the idiosyncrasy of a particular language (its "jewels"), which machine translation systems usually tend to treat incorrectly, rather than focusing on how these systems perform on common words and structures which make up 95% of everyday texts (its "building bricks").

5.3. Organizing community development[12]

One of the possible ways in which open-source machine translation technology could benefit a minor language by creating MT for a new language pair is through communities of volunteer developers. Many minor languages far from normality or officialness have activist groups, usually in the education arena, which include people whose linguistic and translation skills would allow them to collaborate in the creation of linguistic data (dictionaries and rules).

But language and translation skills and volunteered time, even if completely crucial in the case of minor languages, are not enough: volunteer work should be coordinated by a smaller group of people who master the details of the MT engine and tools used. Here are some ingredients of a possible way to organize such a project:

- Each language pair would have a coordinating team, that is, a small group of experts, which would lead the project (see below). This coordinating team could optionally have a *code captain* (dealing with installation, maintenance and possible modifications of the code of the engine or the linguistic maintenance tools) and a *linguistic captain* (responsible for the maintenance of linguistic data).
- A project server and website, which would serve both as the interface through which (registered) volunteers would contribute new linguistic data and as a way for users in the linguistic community involved to download or execute the latest build (version) of each module of the translator. The website would be administered by the coordinating team; ideally, the website should reside in a computer over which the coordinating team have complete control (installing software, adding users, etc.).
- A group of volunteers, ideally certified in some sense by the coordinating team to have the necessary linguistic and translation skills to make useful contributions to dictionaries.

A formula which can be worth exploring to start such a project may be to organize some kind of marathon or volunteer party in which a group of volunteers physically get together (for example, during a weekend) to build linguistic data (for example, generating entries for the first few thousand most frequent words in a corpus, or building bilingual dictionary data from the entries in a bilingual pocket dictionary, torn in similarly-sized portions which are given to each participant). The coordinating team would have to prepare a room with enough computers, install the necessary software for the effort, and arrange for meals and basic lodging. This scheme was used recently, for instance, to localize the open-source office suite OpenOffice.org 2.0 into Catalan.

5.4. Eliciting linguistic knowledge

This is one of the most important challenges, especially for very minor languages for which linguistic expertise is very hard to find. Speakers' knowledge of the language is usually rather intuitive, but to generate useful linguistic data this knowledge has to be made explicit, that is, elicited.

Admittedly, there are parts of the linguistic data that are more suitable for volunteer development than others. With a well-designed form interface capable of eliciting the linguistic knowledge of volunteers, it is possible to maintain the lexical data of the system. Volunteers could be asked to enter monolingual and bilingual dictionary entries through a form interface which would allow them to select inflection paradigms, make choices as to translation equivalents in either direction. etc.

However, one can argue that the design of certain portions of the linguistic data needed, such as structural transfer rules, does not lend itself so easily to volunteer work (elicitation of user knowledge in these cases is a research topic on itself; see, for instance Sheremetyeva and Nirenburg 2000, Font-Llitjós et al. 2005).

5.5. Simplicity of linguistic knowledge needed

Another issue to be considered in connection with the knowledge elicitation challenge is the following. To the extent that this is possible, the level of linguistic knowledge necessary to be able to build a new machine translation system should be kept to a minimum. This may not be possible for very advanced, deep transfer systems, but can easily be achieved for shallow transfer systems. The goal is to encode linguistic knowledge using levels of representation which can be easily learned on top of basic high-school grammar skills and concepts.

5.6. Standardization and documentation of linguistic data formats

As has been mentioned already in section 4.1.1, an adequate documentation of the format of linguistic data files is crucial. This implies carefully defining a systematic format for each source of linguistic data used by the system.

One of the best ways to define linguistic data formats is using the Extensible Markup Language XML:[13] in XML, (a) each data item is explicitly labeled with a descriptive, named tag which has a clear meaning attached; (b) each type of XML file (lexicon, rule file, etc.) has a structure which follows a certain document type definition (DTD) or schema against which it may be checked for validity; and (c) many technologies and applications exist that may be used to convert linguistic data of interest to and from XML formats (interoperability).

5.7. Modularity

For open-source machine translation engine and open linguistic data to be useful for different language pairs or different language technology applications, modularity is a must. A modular MT engine induces modularity in its linguistic data. For example, having an independent morphological analyser and an independent morphological dictionary for a certain language allows them to be used in another machine translation engine having the same source language and a different target language; but it could also be used to build "intelligent" search engines which would allow searching for the

[12] This section is largely based on part of Armentano-Oller et al. (2005).

[13] http://www.w3c.org/XML/ .

lemma of a word and would return all documents having any inflected form of the word.

6. Case study: Opentrad Apertium and Aranese

OpenTrad Apertium. or just Apertium (www.apertium.org) is an open-source shallow-transfer machine translation toolbox which makes it possible to build MT systems for "related" languages. Apertium currently comes with open linguistic data for Spanish—Catalan, Spanish—Galician and Spanish—Portuguese (Catalan and Galician may be considered "minor" languages in the sense given in this paper). At the time of writing this paper, Apertium is one of the few open-source MT systems that can be used for real-life purposes.

Recently, a linguist and I have started to use the architecture to generate a machine translation system between a small language (Catalan) and a *very* small language (the Aranese variety of Occitan, see section 2.3); a paper in these proceedings describes this in more detail (Armentano-Oller and Forcada 2006). In about two person-months, taking advantage of the remarkable similarities between Catalan and Occitan, using a few resources from the web and Catalan data from the Spanish—Catalan package for Apertium, we have been able to build an Aranese—Catalan MT system which already translates 88% of text and does so with error rates around 10%. This would be an example of what was described in section 4.1.2. Once a fully operational, bidirectional system is freely available and downloadable (having, for example, 98% text coverage, and a word error rate of, say, 7%), perhaps the amount of Aranese text on the web (visibility) will significantly increase; perhaps other Occitan speakers may adopt Aranese forms after using the translation on Catalan texts, increasing the contribution of the Aranese dialect to a future Occitan standard, as mentioned in section 4.1.2 (standardization); and, surely, open-source linguistic data for Aranese will be available to be used for other language technology applications.

7. Concluding remarks

I have explored the positive effects that the availability of machine translation may have on the status and development of minor languages (spreading the use of the language, increasing literacy, contributing to standardization, and increasing visibility), but, in particular, those effects which are specific to open-source machine translation (increasing the expertise of the language community, building reusable resources, and reducing technological dependency). For these effects to happen, however, there are a number of challenges that should be met (lack of a standard variety, technophobic attitudes, difficulties to encode linguistic knowledge, need for standard and interoperable formats for linguistic data, and the need for modularity), and I have tried to briefly outline them, adding a brief case study for illustration.

The reflections I went through and, above all, the discussions I had with my colleagues when writing this paper taught me a few interesting things, and made me even more convinced than when I started about the convenience of having open-source machine translation for minor languages. I hope the readers can say something similar after having read it.

Acknowledgements: This work has been partially funded by the Spanish government through grants TIC2003-08681-C02-01, FIT-340101-2004-3 and FIT-340001-2005-2. I thank Antonio M. Corbí-Bellot, Mireia Ginestí-Rosell, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, Carme Armentano-Oller, Míriam A. Scalco for their interesting comments and suggestions and the organizers of this Workshop for having invited me.

8. References

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., Sánchez-Martínez, F. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. In proceedings of *OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X*, September 12-16, 2005, Phuket, Thailand.

de Gispert, A., Mariño, J.B. (2006) Statistical machine translation without parallel corpus: bridging through Spanish. In *these proceedings*.

Dvorak, B., Homola, P., Kubon, V. (2006). Exploiting similarity in the MT into a minority language. In *these proceedings*.

Font-Llitjós, A., Carbonell, J.G., Lavie, A. (2005). A Framework for Interactive and Automatic Refinement of Transfer-based Machine Translation. In *Proceedings of EAMT 2005* (Budapest, 30-31 May 2005).

González, J., Lagarda, A.L., Navarro, J.R., Eliodoro, L., Giménez A., Casacuberta, F., de Val, J.M., Fabregat, F. (2006) SisHiTra: A Spanish-to-Catalan hybrid machine translation system. In *these proceedings*.

Mason, M. and Allen, J. (2001). Standardized Spelling as a Localization Issue. In *Multilingual Computing and Technology magazine*. Number 41, Vol. 12, Issue 5. July/August 2001, p. 37-40.

Ney, H. (2006) "Statistical Machine Translation with and without a bilingual training corpus". In *these proceedings*.

Ostler, N. (1998) "Review of the Workshop on Language Resources for European Minority Languages (Granada, Spain, May 27, 1988). Available at http://193.2.100.60/SALTMIL/history/review.htm

Scannell, K. (2006). Machine translation for closely related language pairs. In *these proceedings*.

Sheremetyeva, S.; Nirenburg, S. (2000). Towards a Universal Tool For NLP Resource Acquisition. In *Proceedings of The Second International Conference on Language Resources and Evaluation* (Greece, Athens, May 31-June 3, 2000).

Williams B., Sarasola K., Ó'Cróinin D., Petek B. (2001) Speech and Language Technology for Minority Languages. *In Proceedings of Eurospeech* 2001.

Approaching a New Language in Machine Translation

- Considerations in Choosing a Strategy

Anna Sågvall Hein, Per Weijnitz

Department of Linguistics and Philology, Uppsala University
Box 637, S-751 26 UPPSALA
anna@lingfil.uu.se, per.weijnitz@lingfil.uu.se

Abstract

As a contribution to the on-going discussions concerning what strategy to use when approaching a new language, we present our experience from working with Swedish in the rule-based and statistical paradigms. We outline the development of Convertus. a robust transfer-based system equipped with techniques for using partial analyses, external dictionaries, statistical models and fall-back strategies. We also present a number of experiments with statistical translation of Swedish involving several languages. We observe that the concrete language pair, translation direction and corpus characteristics have an impact on translation quality in terms of the BLEU score. In particular, we study the effects of the openness/closeness of the domain, and introduce the concept of corpus density to measure this aspect. Density is based on repetition and overlap of text segments, and it is demonstrated that density correlates with BLEU. We also compare a statistical versus a rule-based approach the translation of a Swedish corpus. The rule-based approach for which we use Convertus outperforms the statistical in a modest way. For both systems there is much room for improvement and it is likely that they both can be further developed to a BLEU score of 0.4 – 0.5 which seems good enough for post-editing to pay off. However, a major difference concerns the kinds of errors that are made and how they can be identified. The errors caused by Convertus can be easily traced and explained in linguistic terms and hence also avoided by extensions and modifications of the dictionaries and the grammars. The errors produced by the statistical system are, however, less predictable and difficult to pin-point and eliminate by further training. In particular, the many cases of omissions constitute a serious problem. Our conclusion will be that the investment made in developing a rule-based system, preferably backed up by a statistical system, will pay off in the long run. Thus it becomes an urgent issue to make rule-based systems available as open-source so that the development of new systems can be focused on creating the language resources.

1. Introduction

In the early days of machine translation, simplistic binary dictionaries were the only language resources that were used, and the results were poor. An increasing understanding of the importance of strategies for word sense disambiguation and for including morphological as well as syntactic knowledge emerged. A rule-based paradigm emerged, realized as direct translation with ad-hoc translation rules or transfer-based translation based on full syntactic sentence structures. Well-known shortcomings with the first strategy were due to difficulties in covering all contexts that were to be handled by translation rules (e.g. SYSTRAN), and with the transfer-based approach in covering the sentence structure in all its variation (see e.g. Hutchins, J. and Somers, H. 1992). A problem that was shared for all translation systems was developing full-covering dictionaries, and adapting them to specific domains hereby reducing the number of alternative interpretations. Initially, dictionaries were basically handcrafted, in particular as regards the definition of translation relations. A major step forward was taken with the development of strategies for aligning parallel text, bi-text, sentence-wise, word-wise, and phrase-wise. These strategies formed the basis for automatically extracting translation relations for dictionary-building purposes, and so-called statistical translation (Brown, P. F. et al. 1993), or example-based translation (see e.g. Way, A. and N. Gough. 2005). Translation was based on aligned parallel corpora and language models of the target language. Using the alignment strategies for building translation dictionaries for rule-based systems promoted the quality of these systems substantially. Another strategy that was introduced to overcome problems with insufficient coverage of source language analysis grammar in transfer-based systems was using partial analyses (see e.g. Weijnitz et al. forthcoming.). Two basic strategies, statistical mt, and rule-based MT making use of partial

analyses and the corpus-based translation dictionaries, emerged. Further, methods for automatically measuring similarity of the machine translated text with a reference translation, hereby estimating the quality of the machine translation text, were presented and heavily made use of in statistical as well as rule-based translation.

Each strategy has its shortcomings. A major problem with statistical MT concerns the identification of the bad translations, reasons behind them and ways of correcting errors of certain types, such as wrong or omitted lexical information, and syntactic errors. In contrast to rule-based systems, there are no individual linguistic rules to improve, rather global measures such as extending the language model or the translation model with more data, fine-tuning statistical parameters, and including external dictionaries. In other words, there are few linguistic ways of improving the translation. In rule-based translation, on the other hand, rules can be added and refined. Not surprisingly, current research aims at including linguistic knowledge into the statistical systems, and statistical models into rule-based systems. Ways of combining the two strategies into hybrid systems are explored (Hearne, M. 2005).

2. Outline

Here we will consider a situation where MT for a new language requires the building of a new system rather than extending an existing (commercial) system with a new language pair. The situation appears e.g. for minority languages and less used languages or languages that for other reasons are not considered to be commercially motivated. Based on the availability of language resources and tools, a decision has to be made between a rule-based, a statistical or an example-based approach. For an illustration of the kinds of issues that may have to be considered in making the choice, we will present our experience from working in the rule-based and the statistical paradigm with MT of Swedish and English. As regards the example-based paradigm it is, basically, outside of our experience and will not be further discussed in this paper. First we will outline the development of a rule-based system with fall-back strategies, Convertus[1], and then the achievements made with statistical MT from and to Swedish. In working with different domains and typologically different languages, we made some observations concerning translation quality in the different experimental settings. Typically, a setting is characterized by parameters such as language pair, translation direction, and domain. The domain is defined by a corpus, and we found it motivated to take into account, not only the size of the corpus, but also features concerning repetition and overlap of text segments. Based on these criteria we introduce the term corpus *density* for capturing the openness/closeness of the domain. We will get back to these issues and their implications for translation quality in terms of BLEU. We will then present an experiment of a rule-based and a statistical approach to the translation of Swedish into English, and discuss the pro's and cons' of the two approaches. Finally, we draw some conclusions that seem to have some general interest.

3. Building a transfer-based system for translation of Swedish

As a result of more than ten years of research and development we arrived at a system for translating Swedish into English with a transfer-based core and complementary strategies for handling data outside the language description. It generates satisfactory results in several domains (automotive literature, agriculture, education). In the procedure leading to this system, the following main phases may be distinguished:

- Designing, implementing, and testing a modular, unification-based core engine, Multra (Beskow, B. 1993, Sågvall Hein 1994) with dictionaries and grammars of experimental size for translation of Swedish into English, German, and Russian; grammars and dictionaries were hand-crafted.
- Scaling up the system for one domain, e.g. automotive service literature, and one translation pair, Swedish to English; the scaling-up process was a major under-taking turning the prototype into a real system, Mats (Sågvall Hein, A. et al. 2003) capable of processing real-world documents. For the scaling-up effort a corpus of 16.1k sentence pairs (50,000 tokens) was established for training and testing, the so-called Mats corpus. Much effort was devoted to scaling up the dictionaries making use of word alignment techniques (Tiedemann, J. 2003), and to organizing the lexical material in a database with built-in morphology. The modularity of the core engine is reflected in the database, which includes a source dictionary, a target dictionary and a translation dictionary in terms of translations relations between lexical units. Options for evaluating the coverage of the language description (dictionaries, grammars, transfer rules) and tracing the processing at a detailed level were also built into the system (Sågvall Hein et al. 2003). It should be mentioned, however, that the goal of scaling up of the grammars to cover the training corpus was not achieved in the project.
- Compensating for gaps in the language description by adding techniques for using partial analyses, external dictionaries, statistical models and fall-back strategies (Weijnitz et al. forthc); building an evaluation center to support systematic evaluation of translation quality (Forsbom, E. 2003); this phase lead to a major reorganization of the architecture and the process control, and motivated a new name of the system, i.e. Convertus.
- Progressively training the system for several domains to a BLEU score on the training text of about 0.5 on an average; user feed-back in the various projects in which the training took place indicates that a BLEU score of 0.4-0.5 based on a single reference corpus, is good enough for post-editing.
- Work in process concerns the automatic extraction of grammar from corpora (Megyesi 2002, Nivre 2005) for experimenting with different parsers and different languages, hereby making the system easily adaptable to new language pairs.

In conclusion, a well functioning translation system for translating Swedish into English in several domains was developed with a substantial investment of man power during many years. The system as such is not limited to translation from Swedish to English, but so far there are no language resources of relevant size for other language pairs. Achievements of general interest include techniques for building monolingual and bilingual dictionaries from parallel text as well as a database technology for storing and maintaining lexical data with inbuilt morphology (Tiedemann 2002). Another language-independent achievement is a flexible architecture permitting the plugging-in of modules for analysis, transfer, and generation, for on-line consultation of external dictionaries, and for using fall-back strategies for recovering processing in different problematic situations, e.g. when grammars are insufficient. However, for high-quality translation, grammars are required, for analysis, as well as transfer and generation. This may be a bottleneck in applying the technology to new languages, in particular, less used languages. With the further development of machine-learning techniques for extracting grammar from text, this problem may be reduced. Several modules of Convertus might be presented as open-source software. However, before that, the modules need to be properly packaged and documented.

An alternative strategy when it comes to approaching new languages may be to use statistical machine translation. Thus, now let us turn to our experience of applying statistical machine translation to Swedish.

[1] http://stp.ling.uu.se/~gustav/convertus

4. Experiments with statistical machine translation of Swedish

To get a grasp of the perspectives of statistical machine translation from and to Swedish, we carried out a number of experiments. They indicate that language differences and translation direction have an impact on the translation quality measured by BLEU, in addition to corpus size and corpus density. The same system, to be described below, was used for all experiments.

4.1. The system

Phrase based systems work with both words and phrases, using at least two knowledge sources. Both the translation model and the language model are usually obtained automatically from parallel corpora. By using either the source-channel model, the more general direct maximum entropy translation model, or some other method, the translation model and language model are combined (Och, F. J. and Ney H. 2002).

Pharaoh is a beam search decoder implementing the best-performing methods for statistical machine translation as of year 2004 (Koehn 2004). The translation models were created using UPlug (Tiedemann, J. 2003), GIZA++ (Och, F. J. and Ney, H. 2000) and Thot (Ortiz-Martínez, D. et al. 2005). We used a basic set of models; a 3-gram target language model and a phrase translation model `P(target|source)`, and a length penalty parameter. The length and model parameters were automatically optimized on development corpora. For our experiments, we restricted the source phrase lengths to 4. The language models were created using SRILM (Stolcke, A. 2002).

4.2. Language differences and translation direction

The first experiments were run on the Mats corpus. It includes source documents in Swedish with translations into several languages, among them English and German. Swedish as well as English, belong to the Germanic language family. However, English is felt to be closer, and easier to learn for a Swede than German.

The system was trained on 15.8k sentence pairs per language, sv-en, and, sv-de, and for each language pair 300 sentence pairs were kept aside and used for testing. As expected, English to Swedish outperforms English to German in a significant way (table 1).

Language pair	BLEU
sv->en	0.627
en->sv	0.646
sv->de	0.491
de->sv	0.506

Table 1: BLEU scores for the Mats corpus

Reversing the translation direction, translating from Swedish to English and German, respectively, implies a slight decrease in the BLEU value (table 1), i.e. from 0.646 to 0.627 for English and from 0.506 to 0.491 for German (cf. Papineni, K. A. et al. 2002). In other words, reversing the translation direction does not seem to have any real importance. Still the data will be further examined, in particular from a linguistic point of view.

We also made an experiment with Swedish -> Turkish using a sub-domain (information about Sweden) of a Swedish-Turkish parallel corpus (Megyesi et al., LREC 06) for training. Swedish is the source language. All in all, the sub-domain comprises 1289 sentence pair and the test corpus 206. Evidently, the training corpus is too small for successful SMT, and, in addition, the domain is fairly open. Further, Turkish belonging to the Altaic language family, is typologically very different from Swedish, being a Germanic language in the Indo-European language family. All things taken together, we cannot expect much, and in table 2 we present the results in both directions. Here we observe that the results are slightly better for Swedish as the source language than as the target language, as opposed to the case with English and German. Still the difference is very small, and hardly statistically significant. We need more data to investigate this aspect.

Language pair	BLEU
sv->tr	0.170
Tr->sv	0.156

Table 2: BLEU scores for the Turkish corpus

The experimental findings presented above inspired us to proceed in investigating the implications of the concrete language pair with regard to the quality of SMT. Thus we made a number of experiments for Swedish and other languages using Europarl (Koehn, P. 2005)

BLEU from sv	BLEU to sv	lang. pair	size: sent.
0.2403	0.2090	sv-es	20893
0.2065	0.2074	sv-es	10630
0.1238	0.1401	sv-es	1601
0.2382	0.2192	sv-pt	20726
0.2218	0.2099	sv-pt	10663
0.1288	0.1249	sv-pt	1601
0.1750	0.2194	sv-nl	20690
0.1592	0.1969	sv-nl	10645
0.1020	0.1471	sv-nl	1602
0.1814	0.1910	sv-fi	20663
0.1670	0.1708	sv-fi	10632
0.1048	0.0874	sv-fi	1601

Table 3: BLEU scores for Europarl

We used three different corpus sizes for training, i.e. in terms of sentences pairs: approximately, 20k, 10k and 1,6k. (Cf. the Turkish experiment with a training corpus of 1,2k sentence pairs.) As expected, as the corpus size grows, so does the BLEU score. We may also observe, that the best results are achieved for Swedish – Spanish, closely followed by Swedish- Portuguese. For both these languages, BLEU scores higher for Swedish as a source language than as a target language. Spanish and

Portuguese are Romance languages, and as might be expected, behave in a similar way in relation to Swedish. As regards Dutch, a German language, and Finnish, a Finno-Ugric language, the results are not conclusive. With Swedish as the source language, BLEU scores slightly higher for Finnish, but with Swedish as the target, Dutch outperforms Finnish fairly well, 0.21 versus 0.19. We would have expected a higher score for Dutch being a Germanic language, and as such closer to Swedish; here the statistical, corpus-based data contradict the typological tradition. The data will be further investigated. It is fair to assume, that Swedish-English and Swedish-German should keep their positions as best and second best in the score ranking list, as evidenced by the experiments on the automotive corpus.[2] They were run on a corpus of comparable size, i.e. ~16k sentence pairs. However, a source of error may be due to the openness/closeness of the domain. Europarl is assumed to represent a more open domain than the Scania corpus, and in accordance with earlier research (see e.g. Weijnitz et al. 2004) we expect these aspects to have an impact on the translation results. We will get back to corpus size and density in 3.3. below.

4.2.1. Measuring language differences

In view of the impact of language differences on statistical translation, we would like to have access to a similarity measure for judging the potential of statistical machine translation when approaching a new language pair. In theoretical linguistics language differences are investigated in the sub-field of typology. Here focus is set on identifying distinguishing features such as word order, grammatical relations, case markings, animacy, etc. These features can hardly be translated into a formal, computable measure. What seems to be required is a corpus-based measure that can be calculated automatically.

An option that comes into mind is to base such a measure on word alignment scores. For Swedish – English and Swedish – German, F-values of word alignment scores, based on a gold standard, have been presented (Tiedemann 2003). The author reports an F-value of 83.0 for Swedish – English, and 79.3 for Swedish – German. The figures are derived from the Mats corpus, i.e. the same corpus as was used for training in the SMT experiment presented above. The difference in F-value is comparable to that of BLEU for the two languages, i.e. a BLEU score of 0.65/0.63 (depending on translation direction) for English versus an F-value of 83.0, and a BLEU score of 0.51/0.49 for German versus an F-value of 79.3. The idea of using word alignment scores as a basis for measuring language differences with a view on statistical MT may seem like a circular reasoning; the word aligment phase is crucial in SMT and constitutes a major part in implementing such a system. Still the idea seems worth exploring further. We may consider building a matrix of F-values of word alignment scores for a large variety of languages to be consulted when considering the potential of SMT for a specific translation task. It seems preferable to base this inventory of language similarity measures on properly balanced corpora. As SMT performance depends on training corpus alignment quality, it is important that all language pairs involved are equally well aligned. For a start Europarl (Koehn, P. 2005) or JRC-Acquis (Steinberger, R. 2006) could be useful. In addition, gold standards have to be provided, or other means for calculating the global success of the word alignment process.

4.3. Corpus size and density

Our hypothesis is that there are more training corpus features than size that influence SMT performance. By corpus density we mean to what extent the corpus is repetitive at the sentence and n-gram levels. Table 4 shows the characteristics of a set of corpora, and their BLEU scores obtained when used for SMT. Percents show the type/occurrence ratios of the sentences and n-grams. In this test, there is a negative correlation between the BLEU scores and the sentence ratios. There is also a negative correlation, albeit weaker, between BLEU and n-gram ratios, and larger n-grams mean stronger correlation. The measures relate to the source side of the parallel corpus, only. As for the alignment quality, a crucial issue in SMT, we have no separate data. Still we assume that density correlates not only with the BLEU score but also with alignment quality as such. This, however, remains to be demonstrated. We conclude that corpus density, in addition to size, is a useful criterion when judging the prospect of SMT based on a parallel corpus.

5. Rule-based and statistical translation

There is a huge difference in effort between creating the language resources for a rule-based translation system and for building a statistical translation system. On the other hand, an SMT critically depends on access to a large parallel corpus of high quality within a fairly closed domain. The required size of the corpus cannot be estimated in the general case. It depends on the language pair to be translated, the translation quality to be achieved, and the density of the corpus. Above we have given some clues that should be useful in this context. A high-quality parallel corpus is almost also a sine non qua in building the language resources for a rule-based system, and in developing and evaluating the system. Spending efforts in building such a corpus should pay off in the final end, regardless of what strategy is chosen.

5.1. An experiment with Convertus and Pharaoh

For a comparison between rule-based and statistical translation, we made an experiment using Convertus for the rule-based approach and Pharaoh, as above, for the statistical approach. The experiment was run on the Scania 98 corpus, and the translation direction was Swedish to English.

System	BLEU	Pair	Corpus
convertus	0.377	sv->en	Scania 98
pharaoh	0.324	sv->en	Scania 98

Table 5: BLEU for Convertus and Pharaoh

[2] Unfortunately, data on SMT for Swedish-English and Swedish-German with regard to Europarl were unavailable at the time of writing due to technical problems.

BLEU	BLEU				Sentence	1-gram		3-gram	4-gram
from sv	to sv	Corpus	Pair	Sentences	unique	unique	2-gram unique	unique	unique
0.6274	0.6460	mats	sv-en	16716	60.9%	9.1%	44.8%	68.2%	78.2%
0.5201	0.5267	plug	sv-en	22195	61.3%	10.0%	50.8%	79.2%	89.1%
0.4653	0.4710	agri	sv-en	40910	79.6%	6.6%	38.9%	68.6%	80.1%
0.3523	0.3270	plugfgord	sv-en	4997	99.6%	12.5%	52.5%	83.7%	95.2%
0.2951	0.2944	plugjoinall	sv-en	9204	99.8%	12.1%	53.0%	85.1%	96.1%
0.2297	0.2654	plugfbell	sv-en	4207	100.0%	17.2%	63.9%	91.8%	98.6%

Table 4: Corpus density and BLEU

The training of Pharaoh as well as Convertus was based on the Mats corpus, and there was a token overlap between training data and test data of 5.8%, and a type overlap of 5.3%. This is a small overlap to be compared to the token overlap of 31.7%, and type overlap of 29.9% in the experiment on the Mats korpus (Table 1). As expected, there is a substantial difference in the BLEU score, i.e. 0.324 (Scania 98) and 0.627 (Mats corpus). As for Convertus, the training of the language resources was, basically, limited to the vocabulary; thus there is a big potential for further training of the grammars for this text type (analysis, transfer, as well as generation). In spite of that, Convertus scores higher than Pharaoh, even though the difference is fairly modest.

5.1.1. Translation quality

As has been shown before (e.g. Weijnitz et al. 2004), there is a correlation between the BLEU score and human assessment of translation quality. Further, according to our previous experience, a BLEU score below 0.4 - 0.5 is not good enough for post-editing. An informal study of the two versions of the machine-translated text confirms this view.

Most of the errors produced by both systems are due to words unknown the systems. The standard action taken by them both is to leave the word un-translated. As a fall back strategy, Converts consults an external dictionary outside the domain. This sometimes leads to a wrong choice. However, Convertus optionally provides a list of unknown words including both words that were left un-translated and words for which an external translation was chosen. This list provides the basis for up-dating the dictionary. Unknown words seem to cause worse problems in the SMT system, since there are several cases where the unknown word is simply missing. There is no “place-holder” of the problematic word, which makes the error hard to trace and the translation sometimes incomprehensible. Typically, the statistical system has problems with the translation of Swedish compounds, due to the difficulties of recognizing multiword units in the word aligment phase. An example may be *lufttorkarens function* [air dryer operation, the function of the air dryer] translated as *air function*. Evidently, only the first part of the compound was recognized by the word aligner and the error came out as a case of missing word, and as such hard to identify. More often, however, Swedish compounds are left untranslated e.g. *regulatorfjädrarna* [governor springs]. As mentioned above, the training of Convertus with regard to the Mats corpus was, basically, limited to the vocabulary. Shortcomings concerning the transfer grammar appear in the translation of phrasal expressions, in particular phrasal verbs. Both systems expose structural errors and errors in the inflection of verbs and nouns. Convertus mainly encounters these problems when the subject is not located and when the noun is ambiguous in number. For the SMT system the distribution of these errors is less predictable.

The most striking difference between the two systems is the amount of omissions produced by the SMT. We didn’t calculate their number in this experiment, but in a previous study (Weijnitz et al. 2004) the SMT system exposes more than four times as many instances of omission as the rule-based system. There may be differences in the settings of the two experiments, but the general impression seems to hold. In general, the cause of the errors produced by Convertus can easily be traced and explained in linguistic terms. The SMT system on the other hand, often produces less predictable errors, such as *svåra personskador* [serious personal injury] translated as *svåra not followed.* There is no obvious way to trace the cause of these errors to parameters in the translation or language model.

The rule-based system achieved better results but there is much room for improvements in both systems and a lot of common problems to be solved. Convertus can be improved by adjusting the grammars and extending the dictionaries. The statistical tools require larger amounts of domain-specific training data for better coverage and higher translation quality.

6. Conclusions

We outlined the development of a robust, rule-based MT-system for Swedish, Convertus, equipped with techniques for using partial analyses, external dictionaries, statistical models and fall-back strategies. Most modules of the system are candidates for open-source software, but before that they have to be properly packaged and documented. The system has been progressively trained for several domains to a BLEU score on training data of ~0.5%. User feed-back indicate, that a BLEU score of 0.4-0.5% generated by Convertus represents a translation quality good enough for post-editing.

We also made several experiment withs SMT for Swedish to find out more about factors influencing the translation quality. The experiments showed that language differences are an important issue, with BLEU scores ranging from 0.175 to 0.240 on the same corpus (Europarl). Translation direction turned out to be of minor importance. Corpus size and density were also investigated, and a correlation was confirmed, not only between size and BLEU but also between density and BLEU. These factors should turn out to be useful in estimating the success of an SMT for a certain language pair and domain. However, our main conclusion will be that the investment made in developing a rule-based system, preferably backed up by a statistical system, will pay off in the long run. Thus it becomes an urgent issue to make rule-based systems available as open-source so that the development of new systems can be focused on creating the language resources. Regardless of strategy, the careful preparation of a parallel corpus appears as a crucial first step towards a high-quality MT system, where it is not readily available.

7. References

Beskow, B. (1993). Unification-Based Transfer in Machine Translation. In *RUUL 24.* Uppsala University. Department of Linguistics.

Brown, P. F. and Pietra, V. J. D. and Pietra, S. A. D. and Mercer R. L. (1993). The mathematics of statistical machine translation: parameter estimation. In *Computational Linguistics 19 2.* pp. 263-311. MIT Press, Cambridge, MA, USA

Forsbom, E. (2003). Training a Super Model Look-Alike: Featuring Edit Distance, N-Gram Occurrence, and One Reference Translation. In *Proceedings of the Workshop on Machine Translation Evaluation: Towards Systemizing MT Evaluation*, MT Summit IX, pp. 29-36. New Orleans, Louisiana, USA, 27 September.

Hearne, M. (2005). Data-Oriented Models of Parsing and Translation. PhD Thesis, Dublin City University, Dublin, Ireland.

Hutchins, W. J. and Somers, H. L (1992) *An Introduction to Machine Translation*, London: Academic Press. ISBN 012362830X

Koehn. P (2005). Europarl: a parallel corpus for statistical machine translation. In *Tenth Machine Translation Summit, AAMT.* Phuket, Thailand. November 2005.

Koehn. P (2004). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models, In *AMTA 2004*, pp. 115-124.

Megyesi, B. (2002). Data-Driven Syntactic Analysis - Methods and Applications for Swedish. Ph.D.Thesis. Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Nivre, J. (2005) Inductive Dependency Parsing of Natural Language Text. PhD Thesis, Växjö University.

Och, F. J. and Ney, H. (2000). Improved Statistical Alignment Models, In *ACL00* pp. 440—447. Hongkong, China, October 2000.

Och, F. J. and Ney H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *ACL 2002* pp. 295-302.

Ortiz-Martínez, D. and García-Varea, I. and Casacuberta , F. (2005). Thot: a Toolkit To Train Phrase-based Statistical Translation Models. In *Tenth Machine Translation Summit, AAMT.* Phuket, Thailand. November 2005.

Papineni, K.A. and Roukos, S. and Ward, T. and Zhu, W.J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL.*

Steinberger, R., and Pouliquen B. and Widiger, A. and Ignat, C. and E. Toma and Tufiş, D. and Varga, D. (2006). *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, vol. 2, pp. 901-904, Denver

Sågvall Hein, A., (1994) Preferences and Linguistic Choices in the Multra Machine Translation System. In: Eklund, R. (ed.), NODALIDA '93 *Proceedings of '9:e Nordiska Datalingvistikdagarna',* Stockholm 3-5 June 1993.

Sågvall Hein, A. and Forsbom, E. and Tiedemann, J. and Weijnitz, P. and Almqvist, I. and Olsson, L.-J. and Thaning, S. (2002). Scaling up an MT Prototype for Industrial Use - Databases and Data Flow. I Proceedings from the Third International Conference on Language Resources and Evaluation (LREC'02), pp. 1759-1766, Las Palmas de Gran Canaria, Spanien, 29-31 maj.

Sågvall Hein, A and Weijnitz, P, and Forsbom, E, and Tiedemann, J. and Gustavii, E. (2003). MATS - A Glass Box Machine Translation System. In *Proceedings of the Ninth Machine Translation Summit*, pp. 491 – 493. New Orleans, USA, September 23-27, 2003.

Tiedemann, J. (2003). Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing. Doctoral Thesis, Studia Linguistica Upsaliensia 1, ISSN 1652-1366, ISBN 91-554-5815-7

Tiedemann, J. (2002). MatsLex - a multilingual lexical database for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, volume VI. Las Palmas de Gran Canaria, Spain.

Way, A. and N. Gough. 2005. Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering* 11(3):295--309.

Weijnitz, P. and Sågvall Hein, A. and Forsbom, E. and Gustavii, E. and Pettersson, E. and Tiedemann. J. (forthcoming). The machine translation system MATS - past, present & future. In *Proceedings of RASMAT'04. Uppsala, Sweden, 22-23 April.*

Weijnitz, P. and Forsbom, E. and Gustavii, E. and Pettersson, E. and Tiedemann, J. (2004). MT goes farming: Comparing two machine translation approaches on a new domain. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04),* volym VI, s. 2043-2046. Lissabon, Portugal, May 26-28.

Stephen Krauwer

(University of Utrecht, The Netherlands)

"Update on BLARK developments"

Unfortunately, a written version of this paper is not available. The materials from the oral presentation will be available after the workshop on the SALTMIL website at:

http://isl.ntf.uni-lj.si/SALTMIL/index.htm

Building NLP Systems for Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua

Christian Monson[1], Ariadna Font Llitjós[1], Roberto Aranovich[2], Lori Levin[1], Ralf Brown[1], Eric Peterson[1], Jaime Carbonell[1], Alon Lavie[1]

1 Language Technologies Institute
School of Computer Science
Carnegie Mellon University

2 Department of Linguistics
University of Pittsburgh

Abstract

By adopting a "first-things-first" approach we overcome a number of challenges inherent in developing NLP Systems for resource-scarce languages. By first gathering the necessary corpora and lexicons we are then enabled to build, for Mapudungun, a spelling-corrector, morphological analyzer, and two Mapudungun-Spanish machine translation systems; and for Quechua, a morphological analyzer as well as a rule-based Quechua-Spanish machine translation system. We have also worked on languages (Hebrew, Hindi, Dutch, and Chinese) for which resources such as lexicons and morphological analyzers were available. The flexibility of the AVENUE project architecture allows us to take a different approach as needed in each case. This paper describes the Mapudungun and Quechua systems.

1. The AVENUE Project

The long-term goal of the AVENUE project at CMU is to facilitate machine translation for a larger percentage of the world's languages by reducing the cost and time of producing MT systems. There are a number of radically different ways to approach MT. Each of these methods of accomplishing machine translation has a different set of strengths and weaknesses and each requires different resources to build. The AVENUE approach combines these different types of MT in one "omnivorous" system that will "eat" whatever resources are available to produce the highest quality MT possible given the resources. If a parallel corpus is available in electronic form, we can use example-based machine translation (EBMT) (Brown et al., 2003; Brown, 2000), or Statistical machine translation (SMT). If native speakers are available with training in computational linguistics, a human-engineered set of rules can be developed. Finally, if neither a corpus nor a human computational linguist is available, AVENUE uses a newly developed machine learning technique (Probst, 2005) to learn translation rules from data that is elicited from native speakers. As detailed in the remainder of this paper, the particular resources that the AVENUE project produced facilitated developing an EBMT and a human-coded rule-based MT system for Mapudungun, and a hand-built rule-based MT system for Quechua. Automatic rule learning has been applied experimentally for several other language pairs: Hindi-to-English (Lavie et al. 2003) and Hebrew-to-English (Lavie et al. 2004).

The AVENUE project as a whole consists of six main modules (Figure 1), which are used in different combinations for different languages: 1) elicitation of a word aligned parallel corpus (Levin et al. in press); 2) automatic learning of translation rules (Probst, 2005) and morphological rules (Monson et al. 2004); 3) the run time MT system for performing source to target language translation based on transfer rules; 4) the EBMT system (Brown, 1997); 5) a statistical "decoder" for selecting the most likely translation from the available alternatives; and 6) a module that allows a user to interactively correct translations and automatically refines the translation rules (Font Llitjós et al. 2005a).

2. AVENUE and Indigenous Languages of the Western Hemisphere

Over the past six years the AVENUE project at the Language Technologies Institute at Carnegie Mellon University has worked with native informants and the government of Chile to produce a variety of natural language processing (NLP) tools for Mapudungun, an indigenous South American language spoken by less than 1 million people in Chile and Argentina. During the final year and a half of this time, the AVENUE team has also been developing tools for Quechua, spoken by approximately 10 million people in Peru, Bolivia, Ecuador, South of Colombia, and northern Argentina.

Electronic resources for both Quechua and Mapudungun are scarce. At the time the AVENUE team started working on Mapudungun, even simple natural language tools such as morphological analyzers or spelling correctors did not exist. In fact, there were few electronic resources from which such natural language tools might be built. There were no standard Mapudungun text or speech corpora, or lexicons, and no parsed treebanks. The text that does exist is in a variety of competing orthographic formats. More resources exist for Quechua but they are still far from what is needed to build a complete MT system.

In addition to these practical challenges facing construction of natural language systems for Mapudungun and Quechua, there are also theoretical and human factor challenges. Both Mapudungun and Quechua pose unique challenges from a linguistic theory perspective, since they have complex agglutinative morphological structures. In addition Mapudungun is polysynthetic, incorporating objects into the verb of a sentence. Agglutination and polysynthesis are both properties that the majority languages, for which most natural language resources have been built, do not possess. Human factors also pose a particular challenge for these two languages. Namely, there is a scarcity of people trained in computational linguistics who

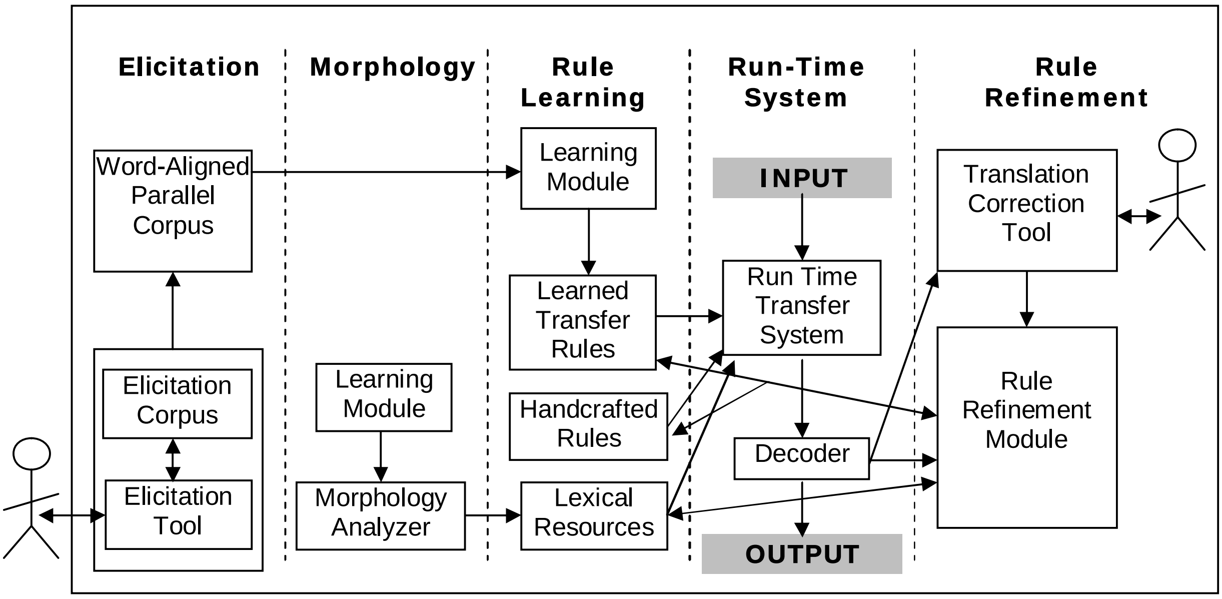


Figure 1: *Data Flow Diagram for the* AVENUE *Rule-Based MT System*

are native speakers or have a good knowledge of these indigenous languages. And finally, an often over looked challenge that confronts development of NLP tools for resource-scarce languages is the divergence of the culture of the native speaker population from the western culture of computational linguistics.

Despite the challenges facing the development of natural language processing systems for Mapudungun and Quechua, the AVENUE project has developed a suite of basic language resources for each of these languages, and has then leveraged these resources into more sophisticated natural language processing tools. The AVENUE project led a collaborative group of Mapudungun speakers in first collecting and then transcribing and translating into Spanish by far the largest spoken corpus of Mapudungun available. From this corpus we then built a spelling checker and a morphological analyzer for Mapudungun. For Quechua, we have created parallel and aligned data as well as a basic bilingual lexicon with morphological information. And with these resources we have built two prototype machine translation (MT) systems for Mapudungun and one prototype MT system for Quechua. This paper will detail the construction of these resources focusing on overcoming the specific challenges Mapudungun and Quechua each present as resource-scarce languages.

3. Mapudungun

Since May of 2000, in an effort to ultimately produce a machine translation system for Mapudungun and Spanish, computational linguists at CMU's Language Technologies Institute have collaborated with Mapudungun language experts at the Instituto de Estudios Indigenas (IEI - Institute for Indigenous Studies) at the Universidad de la Frontera (UFRO) in Chile and with the Bilingual and Multicultural Education Program of the Ministry of Education (Mineduc) in Chile (Levin et al., 2000).

From the very outset of our collaboration we battled the scarcity of electronic resources for Mapudungun. The first phase of the AVENUE collaboration was to collect and produce parallel Mapudungun-Spanish language data from which higher-level language processing tools and systems could be built.

One barrier we faced in the collection of Mapudungun language data is that there are currently several competing orthographic conventions for written Mapudungun. Early in the AVENUE Mapudungun collaboration the IEI-UFRO team established a set of orthographic conventions. All the data collected under the AVENUE project conforms to this set of orthographic conventions. If the Mapudungun data was originally in some other orthographic convention then we manually converted it into our own orthography. Recently, however, a different orthography, Azümchefi, has been chosen by the Chilean government for official documents. Portions of our data have been automatically converted into Azümchefi using automatic substitution rules.

3.1. Corpora and Lexicons

Recognizing the scarcity of Mapudungun language data, the AVENUE team began collecting Mapudungun data soon after our collaboration began in 2000. Directed from CMU and conducted by native speakers of Mapudungun at the Universidad de la Frontera in Temuco, Chile, our data collection efforts ultimately resulted in three separate corpora: 1) a small parallel Mapudungun-Spanish corpus of historical texts and newspaper text, 2) 1700 sentences in Spanish that were manually translated and aligned into Mapudungun (Elicitation Corpus) and 3) a relatively large parallel corpus consisting of 170 hours of transcribed and translated Mapudungun speech. These corpora are described by Monson et al. (2004).

3.1.1. Frequency Based Lexicons

As a first step toward higher level NLP resources for Mapudungun the AVENUE team converted the transcribed spoken corpus text into a lexicon for Mapudungun. All the unique words in the spoken corpus were extracted and then ordered by frequency. The first 117,003 most frequent of these fully-inflected word forms were hand-checked for spelling according to the adopted orthographic conventions for Mapudungun.

Amu -ke -yngün
go -habitual -3plIndic
They (usually) go

ngütrümtu -a -lu
call -fut -adverb
While calling (tomorrow),

Nentu -ñma -nge-ymi
extract-mal -pass -2sgIndic
you were extracted (on me)

ngütramka -me -a -fi -ñ
tell -loc -fut -3obj -1sgIndic
I will tell her (away)

Figure 2: *Examples of Mapudungun verbal morphology taken from the AVENUE corpus of spoken Mapudungun*

Because Mapudungun has a rich morphology, many NLP applications could benefit from knowing not just fully inflected word forms but also knowing lexical stems. To this end 15,120 of the most frequent fully inflected word forms were hand segmented into two parts: The first part consisting of the stem of the word and the second part consisting of one or more suffixes. This produced 5,234 stems and 1,303 suffix groups. These 5,234 stems were then translated into Spanish.

3.2. Basic Language Tools

With the basic Mapudungun corpora and lexicons in hand, the AVENUE team has developed two basic language tools for Mapudungun: a spelling checker (Monson et al., 2004) for use in a word processing application and a morphological analyzer that produces a syntactic description of individual Mapudungun words.

3.2.1. Mapudungun morphological analyzer

In contrast to the stand-alone spelling checker, the AVENUE team has also developed a morphological analyzer for Mapudungun designed to be integrated into machine translation systems. Mapudungun is an agglutinative and polysynthetic language. A typical complex verb form occurring in our corpus of spoken Mapudungun consists of five or six morphemes. Since each morpheme may alone contain several morpho-syntactic features, it is difficult for an MT system to translate Mapudungun words directly into another language. By identifying each morpheme in each Mapudungun word and assigning a meaning to each identified morpheme, a machine translation system can translate individually each piece of meaning. Figure 2 contains glosses of a few morphologically complex Mapudungun verbs that occur in the spoken corpus.

The morphological analyzer takes a Mapudungun word as input and as output it produces all possible segmentations of the word. Each segmentation specifies:

- A single stem in that word
- Each suffix in that word
- A syntactic analysis for the stem and each identified suffix.

To identify the morphemes (stem and suffixes) in a Mapudungun word, a lexicon of stems works together with a fairly complete lexicon of Mapudungun suffixes. The first version of the stem lexicon contains the 1,670 cleanest stems, and their Spanish translations, that were segmented and translated during the lexicon production for Mapudungun. Each entry in this lexicon lists the part of speech of the stem as well as other features associated with the stem such as lexical aspect in the case of verb stems. The suffix lexicon, built by hand by computational linguists on the AVENUE team, is fairly complete. Unlike the suffix groups used in the spelling checker, each suffix entry in the suffix lexicon for the morphological analyzer is an individual suffix. There are 105 Mapudungun suffixes in the suffix lexicon. Each suffix lists the part of speech that the suffix attaches to: verb, noun, adjective, etc. Each suffix also lists the linguistic features, such as person, number, or mood that it marks. The morphological analyzer performs a recursive and exhaustive search on all possible segmentations of a given Mapudungun word. The software starts from the beginning of the word and identifies each stem that is an initial string in that word. Next, the candidate stem from the word is removed. The software then examines the remaining string looking for a valid combination of suffixes that could complete the word. The software iteratively and exhaustively searches for sequences of suffixes that complete the word. Because the morphological analyzer also takes into account constraints on the allowable ordering of Mapudungun suffixes, most Mapudungun words for which the stem is in the stem lexicon receive a single analysis. A few truly ambiguous suffix combinations may cause a Mapudungun word to receive perhaps as many as five distinct analyses.

Once the morphological analyzer has found all possible and correct segmentations of a word, it combines the feature information from the stem and the suffixes encountered in the analyzed word to create a syntactic analysis that is returned. For an example, see Figure 3.

3.3. Machine Translation Systems

3.3.1. Example-Based Machine Translation system

Example-Based Machine Translation (EBMT) relies on previous translations performed by humans to create new translations without the need for human translators. The previous translations are called the training corpus. For the best translation quality, the training corpus should be as large as possible, and as similar to the text to be translated as possible. When the exact sentence to be

pekelan pe-ke-la-n
Lexeme = pe (see)
subject person = 1
subject number = singular
mode = indicative
negation = +
aspect = habitual

Figure 3. *Example showing the output of the morphological analyzer for Mapudungun.*

translated occurs in the training material, the translation quality is human-level, because the previous translation is re-used. As the sentence to be translated differs more and more from the training material, quality decreases because smaller and smaller fragments must be combined to produce the translation, increasing the chances of an incorrect translation.

As the amount of training material decreases, so does the translation quality; in this case, there are fewer long matches between the training texts and the input to be translated. Conversely, more training data can be added at any time, improving the system's performance by allowing more and longer matches. EBMT usually finds only partial matches, which generate lower-quality translations. Further, EBMT searches for phrases of two or more words, and thus there can be portions of the input which do not produce phrasal translations. For unmatched portions of the input, EBMT falls back on a probabilistic lexicon trained from the corpus to produce word-for-word translations. This fall-back approach provides some translation (even though of lower quality) for any word form encountered in the training data. While the translation quality of an EBMT system can be human-level when long phrases or entire sentence are matched, any mistakes in the human translations used for training—spelling errors, omissions, mistranslations—will become visible in the EBMT system's output. Thus, it is important that the training data be as accurate as possible. The training corpus we use for EBMT is the spoken language corpus described earlier. As discussed in section 1.1.1, the corpus of spoken Mapudungun contains some errors and awkward translations.

Highly agglutinative languages pose a challenge for Example Based MT. Because there are so many inflected versions of each stem, most inflected words are rare. If the rare words do not occur in the corpus at all, they will not be translatable by EBMT. If they occur only a few times, it will also be hard for EBMT to have accurate statistics about how they are used. Additionally, word-level alignment between the two languages, which is used to determine the appropriate translation of a partial match, becomes more difficult when individual words in one language correspond to many words in the other language. We address both issues by using the morphological analyzer for Mapudungun to split words into stems and suffixes. Each individual stem and suffix is more common than the original combination of stem and suffixes, and the individual parts are more likely to map to single words in Spanish.

We currently have a prototype EBMT system trained on approximately 204,000 sentence pairs from the corpus of spoken Mapudungun, containing a total of 1.25 million Mapudungun tokens after splitting the original words into stems and one or more suffixes. This separation increased BLEU scores (Papineni et al., 2001) on a held out portion of the speech corpus by 5.48%, from 0.1530 to 0.1614. We expect further increases from improved use of morphological analysis, the inclusion of common phrases in the corpus, and fixing translation errors and awkward translations in the corpus.

3.3.2. Rule-Based MT system

Simultaneous to the development of the example based machine translation system for Mapudungun we have been working on a prototype rule-based MT system. Rule-based machine translation, which requires a detailed comparative analysis of the grammar of source and target languages, can produce high quality translation but takes a longer amount of time to implement. Hand-built rule-based MT also has lower coverage than EBMT because there is no probabilistic mechanism for filling in the parts of sentences that are not covered by rules.

The rule-based machine translation system is composed of a series of components and databases. The input to the system is a Mapudungun sentence, phrase or word, which is processed in different stages until a Spanish string is output. The MT system consists of three main components: the Mapudungun morphological analyzer discussed in section 3.2.1, the transfer system, and the Spanish morphological analyzer. Each of these programs makes use of different data bases (lexicons or grammars). The transfer system makes use of a transfer grammar and a transfer lexicon, which contain syntactic and lexical rules in order to map Mapudungun expressions into Spanish expressions. The output of the transfer system is a Spanish expression composed of uninflected words plus grammatical features, which constitutes the input for the Spanish morphological generator. The morphological generator makes use of a Spanish lexicon of inflected words (developed by the Universitat Politècnica de Catalunya). Each of these programs and databases, as well as their interactions, will be described in more detail in the following sections of this paper.

3.3.2.1. Run-time Transfer System

At run time, the transfer module translates a source language sentence into a target language sentence. The output of the run-time system is a lattice of translation alternatives. The alternatives arise from syntactic ambiguity, lexical ambiguity, multiple synonymous choices for lexical items in the dictionary, and multiple competing hypotheses from the transfer rules (see next section).

The run-time translation system incorporates the three main processes involved in transfer-based MT: parsing of the source language input, transfer of the parsed constituents of the source language to their corresponding structured constituents on the target language side, and generation of the target language output. All three of these processes are performed based on the transfer grammar – the comprehensive set of transfer rules that are loaded into the run-time system. In the first stage, parsing is performed based solely on the SL side, also called x-side, of the transfer rules. The implemented parsing algorithm is for the most part a standard bottom-up Chart Parser, such as described in Allen (1995). A chart is populated with all constituent structures that were created in the course of parsing the SL input with the source-side portion of the transfer grammar. Transfer and generation are performed in an integrated second stage. A dual TL chart is constructed by applying transfer and generation operations on each and every constituent entry in the SL parse chart. The transfer rules associated with each entry in the SL chart are used in order to determine the corresponding constituent structure on the TL side. At the word level, lexical transfer rules are accessed in order to seed the individual lexical choices for the TL word-level entries in the TL chart. Finally, the set of generated TL output strings that corresponds to the collection of all TL chart entries is collected into a TL lattice, which is then passed on for decoding (choosing the correct path through the lattice of translation possibilities.) A more detailed description of the runtime transfer-based translation subsystem can be found in Peterson (2002).

```
{NBar,1}                          (identifier)
Nbar::Nbar: [PART N] -> [N]       (x-side/y-side
                                  constituent structures)
((X2::Y1)                         (alignment)
((X1 number) =c pl)               (x-side constraint)
((X0 number) = (X1 number))       (passing feature up)
((Y0 number) = (X0 number))       (transfer equation)
((Y1 number) = (Y0 number))       (passing feature down)
((Y0 gender) = (Y1 gender)))

                                  (passing feature up)
```

Figure 4. Plural noun marked by particle pu. Example: pu ruka::casas ('houses')

3.3.2.2. Transfer Rules

The function of the transfer rules is to decompose the grammatical information contained in a Mapudungun expression into a set of grammatical properties, such as number, person, tense, subject, object, lexical meaning, etc. Then, each particular rule builds an equivalent Spanish expression, copying, modifying, or rearranging grammatical values according to the requirements of Spanish grammar and lexicon.

In the AVENUE system, translation rules have six components[1]: a. rule identifier, which consists of a constituent type (Sentence, Nominal Phrase, Verbal Phrase, etc.) and a unique ID number; b. constituent structure for both the source language (SL), in this case Mapudungun, and the target language (TL), in this case Spanish; c. alignments between the SL constituents and the TL constituents; d. x-side constraints, which provide information about features and their values in the SL sentence; e. y-side constraints, which provide information about features and their values in the TL sentence, and f. transfer equations, which provide information about which feature values transfer from the source into the target language.

In Mapudungun, plurality in nouns is marked, in some cases, by the pronominal particle pu. The NBar rule below (Figure 4) illustrates a simple example of a Mapudungun to Spanish transfer rule for plural Mapudungun nouns (following traditional use, in this Transfer Grammar, NBar is the constituent that dominates the noun and its modifiers, but not its determiners).

According to this rule, the Mapudungun sequence PART N will be transfered into a noun in Spanish. That is why there is only one alignment. The x-side constraint is checked in order to ensure the application of the rule in the right context. In this case, the constraint is that the particle should be specified for (number = pl); if the noun is preceded by any other particle, the rule would not apply. The number feature is passed up from the particle to the Mapudungun NBar, then transferred to the Spanish NBar and passed down to the Spanish noun. The gender feature, present only in Spanish, is passed up from the Spanish noun to the Spanish NBar. This process is represented graphically by the tree structure showed in Figure 5.

Some of the problems that the Transfer Grammar has to solve, among others, are the agglutination of Mapudungun suffixes, that have been previously segmented by the morphological analyzer; the fact that tense is mostly unmarked in Mapudungun, but has to be specified in Spanish; and the existence of a series of grammatical structures that have a morphological nature in Mapudungun (by means of inflection or derivation) and a syntactic nature in Spanish (by means of auxiliaries or other free morphemes).

3.3.2.3. Suffix Agglutination

The transfer grammar manages suffix agglutination by constructing constituents called Verbal Suffix Groups (VSuffG). These rules can operate recursively. The first VSuffG rule turns a Verbal Suffix (VSuff) into a VSuffG, copying the set of features of the suffix into the new constituent. Notice that at this level there are no transfer of features to the target language and no alignment. See Figure 6.

The second VSuffG rule combines a VSuffG with another VSuff, passing up the feature structure of both suffixes to the parent node. For instance, in a word like pe-fi-ñ (pe-: to see; -fi: 3rd. person object; -ñ: 1st. person singular, indicative mood; 'I saw he/she/them/it'), the rule {VSuffG,1} is applied to -fi, and the rule {VSuffG,2} is applied to the sequence -fi-ñ. The result is a Verb Suffix Group that has all the grammatical features of its components. This process could continue recursively if there are more suffixes to add.

3.3.2.4. Tense

Tense in Mapudungun is mostly morphologically unmarked. The temporal interpretation of a verb is determined compositionally by the lexical meaning of the verb (the relevant feature is if the verb is stative or not) and the grammatical features of the suffix complex. Figure 7 lists the basic rules for tense in Mapudungun. Since tense should be determined taking into account information from both the verb and the VSuffG, it is managed by the rules that combine these constituents (called VBar rules in this grammar). For instance, Figure 8 displays a simplified version of the rule that assigns the past tense feature when necessary (transfer of features from Mapudungun to Spanish are not represented in the rule for brevity). Analogous rules deal with the other temporal specifications.

3.3.2.5. Typological divergence

As an agglutinative language, Mapudungun has many grammatical constructions that are expressed by morphological, rather than syntactic, means. For instance, passive

[1] This is a simplified description, for a full description see Peterson (2002) and Probst et al. (2003).

```
{VSuffG,1}                           {VSuffG,2}
VSuffG::VSuffG : [VSuff] -> [""]     VSuffG::VSuffG : [VSuffG VSuff] -> [""]
((X0 = X1))                          ((X0 = X1)
                                     (X0 = X2))
```

Figure 6. *Verbal Suffix Group Rules.*

Lexical/grammatical features	Temporal interpretation
a. Unmarked tense + unmarked lexical aspect + unmarked grammatical aspect	past (kellu-n::ayudé::(I)helped)
b. Unmarked tense + stative lexical aspect	present (niye-n::poseo::(I)own)
c. Unmarked tense + unmarked lexical aspect + habitual grammatical aspect	present (kellu-ke-n::ayudo::(I)help)
d. Marked tense (for instance, future)	future (pe-a-n::veré::(I)will see)

Figure 7. *Tense in Mapudungun.*

```
{VBar,1}
VBar::VBar :  [V VSuffG] -> [V]
((X1::Y1)                              (alignment)
((X2 tense) = *UNDEFINED*)             (x-side constraint on morphological tense)
((X1 lexicalaspect) = *UNDEFINED*)     (x-side constraint on verb's aspectual class)
((X2 aspect) = (*NOT* habitual))       (x-side constraint on grammatical aspect)
((X0 tense) = past) …)                 (tense feature assignment)
```

Figure 8. *Past tense rule (transfer of features omitted)*

voice in Mapudungun is marked by the suffix -nge. On the other hand, passive voice in Spanish, as well as in English, requires an auxiliary verb, which carries tense and agreement features, and a passive participle.

For instance, pe-nge-n (pe-: to see; -nge: passive voice; -n: 1rst. person singular, indicative mood; 'I was seen') has to be translated as fui visto o fue vista. The rule for passive (a VBar level rule in this grammar) has to insert the auxiliary, assign it the right grammatical features, and inflect the verb as a passive participle. Figure 9 shows a simplified version of the rule that produces this result (transfer of features from Mapudungun to Spanish are not represented in the rule for brevity).

3.3.2.6. Spanish Morphology generation

Even though Spanish is not as highly inflected as Mapudungun or Quechua, there is still a great deal to be gained from listing just the stems in the translation lexicon, and having a Spanish morphology generator take care of inflecting all the words according to the relevant features. In order to generate Spanish morphology, we obtained a morphologically inflected dictionary from the Universitat Politècnica de Catalunya (UPC) in Barcelona under a research license. Each citation form (infinitive for verbs and masculine, singular for nouns, adjectives, determiners, etc.) has all the inflected words listed with a PAROLE tag (http://www.lsi.upc.es/~nlp/freeling/parole-es.html) that contains the values for the relevant feature attributes.

In order to be able to use this Spanish dictionary, we mapped the PAROLE tags for each POS into feature attribute and value pairs in the format that our MT system is expecting. This way, the AVENUE transfer engine can easily pass all the citation forms to the Spanish Morphology

```
{VBar,6}
VBar::VBar :  [V VSuffG] -> [V V]      (insertion of aux in Spanish side)
((X1::Y2)                              (Mapudungun verb aligned to Spanish verb)
((X2 voice) =c passive)                (x-side voice constraint )
((Y1 person) = (Y0 person))            (passing person features to aux)
((Y1 number) = (Y0 number))            (passing number features to aux)
((Y1 mood) = (Y0 mood))                (passing mood features to aux)
((Y2 number) =c (Y1 number))           (y-side agreement constraint)
((Y1 tense) = past)                    (assigning tense feature to aux)
((Y1 form) =c ser)                     (auxiliary selection)
((Y2 mood) = part)                     (y-side verb form constraint)
  …)
```

Figure 9. *Passive voice rule (transfer of features omitted).*

Generator, once the translation has been completed, and have it generate the appropriate surface, inflected forms.

When the Spanish morphological generation is integrated with the run-time transfer system the final rule-based Quechua-Spanish MT system produces output such as the following:

```
sl: kümelen (I'm fine)
tl: ESTOY BIEN
   tree: <((S,5 (VPBAR,2 (VP,1 (VBAR,9 (V,10
            'ESTOY') (V,15 'BIEN') ) ) ) ) )>

sl: ayudangelay (he/she were not helped)
tl: NO FUE AYUDADA
   tree: <((S,5 (VPBAR,2 (VP,2 (NEGP,1 (LITERAL
            'NO') (VBAR,3 (V,11 'FUE')
            (V,8 'AYUDADA') ) ) ) ) ) )>
tl: NO FUE AYUDADO
   tree: <((S,5 (VPBAR,2 (VP,2 (NEGP,1 (LITERAL
            'NO') (VBAR,3 (V,11 'FUE')
            (V,8 'AYUDADO') ) ) ) ) ) )>

sl: Kuan ñi ruka (John's house)
tl: LA CASA DE JUAN
   tree: <((S,12 (NP,7 (DET,10 'LA') (NBAR,2 (N,8
            'CASA') ) (LITERAL 'DE')
            (NP,4 (NBAR,2 (N,1 'JUAN') ) ) ) ) )>
```

4. Quechua

Data collection for Quechua started in 2004, when the AVENUE team established a collaboration with bilingual speakers in Cusco (Peru). In 2005, one of the authors (Ariadna Font Llitjós) spent the summer in Cusco to set up basic infrastructure and to develop a first Quechua-Spanish MT prototype system, with the main goal to have an initial system for testing the Translation Correction Tool (Font Llitjós & Carbonell, 2004) and the Rule Refinement module (Font Llitjós et al., 2005a). Translation and morphology lexicons were automatically created from data annotated by a native speaker using Perl scripts. A small translation grammar was written. Additionally, a preliminary user study of the correction of Quechua to Spanish translations was also conducted using the Translation Correction Tool (TCTool), an online user-friendly interface.

4.1. Text Corpora

As part of the data collected for Quechua, the AVENUE Elicitation Corpora (EC) were translated and manually aligned by a both a native Quechua speaker and a linguist with good knowledge of Quechua. The EC is used when there is no natural corpus large enough to use for development of MT. The EC resembles a fieldwork questionnaire containing simple sentences that elicit specific meanings and structures. It has two parts. The first part, the Functional Elicitation Corpus, contains sentences designed to elicit functional/communicative features such as number, person, tense, and gender. The version that was used in Peru had 1,700 sentences. The second part, the Structural Elicitation Corpus, is a smaller corpus designed to cover the major structures present in the Penn Treebank (Marcus et al., 1992). Out of 122,176 sentences from the Brown Corpus section of the Penn Treebank, 222 different basic structures and substructures were extracted; namely, 25 AdvPs, 47 AdjPs, 64 NPs, 13 PPs, 23 SBARs, and 50 Ss. For more information about how this corpus was created and what its properties are, see Probst and Lavie (2004). The final Structural Elicitation Corpus which was translated into Quechua had 146 Spanish sentences.

Besides the Elicitation Corpora, there was no other Quechua text readily available on electronic format, and thus three books which had parallel text in Spanish and Quechua were scanned: Cuento Cusqueños, Cuentos de Urubamba, and Gregorio Condori Mamani. Quechua speakers examined the scanned Quechua text (360 pages), and corrected the optical character recognition (OCR) errors, with the original image of the text as a reference.

4.2. A Rule-Based MT Prototype

Similar to the Mapudungun-Spanish system, the Quechua-Spanish system also contains a Quechua morphological analyzer which pre-processes the input sentences to split words into roots and suffixes. The lexicon and the rules are applied by the transfer engine, and finally, the Spanish morphology generation module is called to inflect the corresponding Spanish stems with the relevant features (Section 3.3.2.6).

4.2.1. Morphology and Translation Lexicons

In order to build a translation and morphology lexicon, the word types from the three Quechua books were extracted and ordered by frequency. The total number of types was 31,986 (Cuento Cusqueños 9,988; Cuentos de Urubamba 12,223; Gregorio Condori Mamani 12,979), with less than 10% overlap between books. Only 3,002 word types were in more than one book.[2] Since 16,722 word types were only seen once in the books (singletons), we decided to segment and translate only the 10,000 most frequent words in the list, hoping to reduce the number of OCR errors and misspellings. Additionally, all the unique word types from one of the versions of the Elicitation Corpora were also extracted (1,666 types) to ensure basic coverage.

10,000 words were segmented and translated by a native Quechua speaker. The (Excel) file used for this task contained the following fields: Word Segmentation, Root translation, Root POS, Word Translation, Word POS and Translation of the final root if there has been a POS change. The reason for the last field is that if the POS fields for the root and the word differ, the translation of the final root might have changed and thus the translation in the lexical entry actually needs to be different from the translation of the root. In Quechua, this is important for words such as “machuyani” (I age/get older), where the root “machu” is an adjective meaning “old” and the word is a verb, whose root really means “to get old” (“machuyay”)[3]. Instead of having a lexical entry like V-machuy-viejo (old), we are interested in having a lexical entry V-machu(ya)y-envejecer (to get old).

[2] This was done before the OCR correction was completed and thus this list contained OCR errors.

[3] -ya- is a verbalizer in Quechua.

```
{S,2}
S::S : [NP VP] -> [NP VP]
(  (X1::Y1)   (X2::Y2)

  ((x0 type) = (x2 type))
  ((y1 number) = (x1 number))
  ((y1 person) = (x1 person))
  ((y1 case) = nom)

; subj-v agreement
  ((y2 number) = (y1 number))
  ((y2 person) = (y1 person))

; subj-embedded Adj agreement
  ((y2 PredAdj number) = (y1 number))
  ((y2 PredAdj gender) = (y1 gender)))
```

```
{SBar,1}
SBar::SBar : [S] -> ["Dice que" S]
( (X1::Y2)
  ((x1 type) =c reportative) )

{VBar,4}
VBar::VBar : [V VSuff VSuff] -> [V]
( (X1::Y1)
   ((x0 person) = (x3 person))
   ((x0 number) = (x3 number))
   ((x2 mood) = (*NOT* ger))
   ((x3 inflected) =c +)
   ((x0 inflected) = +)
   ((x0 tense) = (x2 tense))
   ((y1 tense) = (x2 tense))
   ((y1 person) = (x3 person))
   ((y1 number) = (x3 number))
   ((y1 mood) = (x3 mood)))
```

Figure 13. *Manually written grammar rules for Quechua-Spanish translation..*

```
Interj |: [alli] -> ["a pesar"]      ((X1::Y1))
((X1::Y1))
```

Figure 11. *Automatically generated lexical entries from segmented and translated word*

```
; "dicen que" on the Spanish side
Suff::Suff |: [s] -> [""]
((X1::Y1)
((x0 type) = reportative))

; when following a consonant
Suff::Suff |: [si] -> [""]
((X1::Y1)
((x0 type) = reportative))

Suff::Suff |: [qa] -> [""]
((X1::Y1)
 ((x0 type) = emph))

Suff::Suff |: [chu] -> [""]
((X1::Y1)
((x0 type) = interr))
```

```
VSuff::VSuff |: [nki] -> [""]
((X1::Y1)
((x0 person) = 2)
((x0 number) = sg)
((x0 mood) = ind)
((x0 tense) = pres)
((x0 inflected) = +))

NSuff::NSuff |: [kuna] -> [""]
((X1::Y1)
((x0 number) = pl))

NSuff::Prep |: [manta] -> [de]
((X1::Y1)
((x0 form) = manta))
```

Figure 12. *Manually written suffix lexical entries.*

From the list of segmented and translated words, a stem lexicon was automatically generated and manually corrected. For example, from the word type "chayqa" and the specifications given for all the other fields as shown in Figure 10, six different lexical entries were automatically created, one for each POS and each alternative translation (Pron-ese, Pron-esa, Pron-eso, Adj-ese, Adj-esa, Adj-eso). In some cases, when the word has a different POS, it actually is translated differently in Spanish. For these cases, the native speaker was asked to use || instead of |, and the post-processing scripts were designed to check for the consistency of || in both the translation and the POS fields. The scripts allow for fast post-processing of thousands of words, however manual checking is still required to make sure that no spurious lexical entries have been created. Some examples of automatically generated lexical entries are presented in Figure 11. Suffix lexical entries, however, were hand-crafted, see Figure 12. For the current working MT prototype the Suffix Lexicon has 36 entries. Cusihuaman's grammar (2001) lists a total of 150 suffixes.

4.2.2. Translation Rules

The translation grammar, written with comprehensive rules following the same formalism described in subsection 3.3.2.2 above, currently contains 25 rules and it covers subject-verb agreement, agreement within the NP (Det-N and N-Adj), intransitive VPs, copula verbs, verbal suffixes, nominal suffixes and enclitics. Figure 13 shows a couple of examples of rules in the translation grammar.

Below are a few correct translations as output by the Quechua-Spanish MT system. For these, the input of the system was already segmented (and so they weren't run by the Quechua Morphology Analyzer), and the MT output is the result of inflecting the Spanish citation forms using the Morphological Generator discussed in section 3.3.2.6:

sl: taki sha ra ni (I was singing)

tl: ESTUVE CANTANDO
tree: <((S,1 (VP,0 (VBAR,5 (V,0:0 "ESTUVE") (V,2:1 "CANTANDO")))))>

sl: taki ra n si (it is said that s/he sang)
tl: DICE QUE CANTÓ
tree: <((SBAR,1 (LITERAL "DICE QUE") (S,1 (VP,0 (VBAR,1 (VBAR,4 (V,2:1 "CANTÓ")))))))>

sl: noqa qa barcelona manta ka ni (I am from Barcelona)
tl: YO SOY DE BARCELONA
tree: <((S,2 (NP,6 (NP,1 (PRONBAR,1 (PRON,0:1 "YO")))) (VP,3 (VBAR,2 (V,3:5 "SOY")) (NP,5 (NSUFF,1:4 "DE") (NP,2 (NBAR,1 (N,2:3 "BARCELONA")))))))>

4.3. Preliminary User Studies

A preliminary user study of the correction of Quechua to Spanish translations was conducted where three Quechua speakers with good knowledge of Spanish evaluated and corrected nine machine translations, when necessary, through a user-friendly interface called the Translation Correction Tool (TCTool). This small user study allowed us to see how Quechua speakers used the TCTool and whether they had any problems with the interface. It showed that the Quechua representation of stem and suffixes as separate words does not seem to pose a problem and that it was relatively easy to use for non-technical users.

5. Conclusions and Future Work

The "first-things-first" approach the AVENUE project has taken to building NLP systems for scarce-resource languages has proven effective. By first focusing effort on producing basic NLP resources, the resultant resources are of sufficient quality to be put to any number of uses: from building all manner of NLP tools to potentially aiding linguists in better understanding an indigenous culture and language. For both Mapudungun and Quechua, separate work on morphology analysis and on a transfer grammar modularized the problem in a way that allowed rapid development. Besides a spelling-checker for Mapudungun, the AVENUE team has developed computational lexicons, morphology analyzers and one or more Machine Translation systems for Mapudungun and Quechua into Spanish.

The AVENUE team has recently put many of the resources we have developed for Mapudungun online at http://www.lenguasamerindias.org/mapudungun. The AVENUE interactive website, which is still in an experimental phase, contains a basic Mapudungun-Spanish lexicon, the Mapudungun morphological analyzer, and the example-based MT system from Mapudungun to Spanish.

The AVENUE team continues to develop the resources for both Mapudungun and Quechua. We are actively working on improving the Mapudungun-Spanish rule-based MT system by both increasing the size of the lexicon as well as improving the rules themselves. The Mapudungun-Spanish example-based MT system can be improved by cleaning and increasing the size of the training text. For the next version of the MT website, we plan to plug in the Translation Correction Tool to allow bilingual users interested in translating sentences to give us feedback about the correctness of the automatic translation produced by our systems in a simple and user-friendly way.

6. Acknowledgements

Many people beyond the authors of this paper contributed to the work described. The Chilean AVENUE team includes Carolina Huenchullan Arrúe, Claudio Millacura Salas, Eliseo Cañulef, Rosendo Huisca, Hugo Carrasco, Hector Painequeo, Flor Caniupil, Luis Caniupil Huaiquiñir, Marcela Collio Calfunao, Cristian Carrillan Anton, and Salvador Cañulef. Rodolfo Vega was the liaison between the CMU team and the government agency and indigenous community in Chile. People that participated in the development of Quechua data are Marilyn Feke, Salomé Gutierrez, Irene Gómez and Yenny Ccolque.

This research was funded in part by the DARPA TIDES program and by NSF grant number IIS-0121-631.

7. References

Allen, James. (1995). Natural Language Understanding. Second Edition ed. Benjamin Cummings.

Brown, Ralf D., Rebecca Hutchinson, Paul N.Bennett, Jaime G. Carbonell, and Peter Jansen. (2003). "Reducing Boundary Friction Using Translation-Fragment Overlap", in Proceedings of the Ninth Machine Translation Summit, New Orleans, USA. pp. 24-31.

Brown, Ralf D. (2000). "Example-Based Machine Translation at Carnegie Mellon University". In The ELRA Newsletter, European Language Resources Association, vol 5:1, January-March 2000.

Cusihuaman, Antonio. (2001). Gramatica Quechua. Cuzco Callao. 2a edición. Centro Bartolomé de las Casas.

Font Llitjós, Ariadna, Jaime Carbonell, Alon Lavie. (2005a). A Framework for Interactive and Automatic Refinement of Transfer-based Machine Translation. European Association of Machine Translation (EAMT) 10th Annual Conference. Budapest, Hungary.

Font Llitjós, Ariadna, Roberto Aranovich, and Lori Levin (2005b). Building Machine translation systems for indigenous languages. Second Conference on the Indigenous Languages of Latin America (CILLA II). Texas, USA.

Font Llitjós, Ariadna and Jaime Carbonell. (2004). The Translation Correction Tool: English-Spanish user studies. International Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal.

Frederking, Robert and Sergei Nirenburg. (1994). Three Heads are Better than One. Proceedings of the fourth Conference on Applied Natural Language Processing (ANLP-94), pp. 95-100, Stuttgart, Germany.

Lavie, A., S. Wintner, Y. Eytani, E. Peterson and K. Probst. (2004) "Rapid Prototyping of a Transfer-based Hebrew-to-English Machine Translation System". In Proceedings of the 10th International Conference on

Theoretical and Methodological Issues in Machine Translation (TMI-2004), Baltimore, MD, October 2004. Pages 1-10.

Lavie, Alon, Stephan Vogel, Lori Levin, Erik Peterson, Katharina Probst, Ariadna Font Llitjós, Rachel Reynolds, Jaime Carbonell, and Richard Cohen. (2003). Experiments with a Hindi-to-English Transfer-based MT System under a Miserly Data Scenario". ACM Transactions on Asian Language Information Processing (TALIP), 2(2).

Levin, Lori, Alison Alvarez, Jeff Good and Robert Frederking. (In Press). Automatic Learning of Grammatical Encoding. To appear in Jane Grimshaw, Joan Maling, Chris Manning, Joan Simpson and Annie Zaenen (eds) Architectures, Rules and Preferences: A Festschrift for Joan Bresnan , CSLI Publications.

Levin, Lori, Rodolfo Vega, Jaime Carbonell, Ralf Brown, Alon Lavie, Eliseo Cañulef, and Carolina Huenchullan. (2000). Data Collection and Language Technologies for Mapudungun. International Conference on Language Resources and Evaluation (LREC).

Mitchell, Marcus, A. Taylor, R. MacIntyre, A. Bies, C. Cooper, M. Ferguson, and A. Littmann (1992). The Penn Treebank Project. http://www.cis.upenn.edu/ treebank/home.html.

Monson, Christian, Lori Levin, Rodolfo Vega, Ralf Brown, Ariadna Font Llitjós, Alon Lavie, Jaime Carbonell, Eliseo Cañulef, and Rosendo Huesca. (2004). Data Collection and Analysis of Mapudungun Morphology for Spelling Correction. International Conference on Language Resources and Evaluation (LREC).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022).

Peterson, Erik. (2002). Adapting a transfer engine for rapid machine translation development. M.S. thesis, Georgetown University.

Probst, Katharina. (2005). Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario. PhD Thesis. Carnegie Mellon.

Probst, Katharina and Alon Lavie. (2004). A structurally diverse minimal corpus for eliciting structural mappings between languages. Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-04).

Probst, Katharina, Ralf Brown, Jaime Carbonell, Alon Lavie, Lori Levin, and Erik Peterson. (2001). Design and Implementation of Controlled Elicitation for Machine Translation of Low-density Languages. Proceedings of the MT2010 workshop at MT Summit

Statistical Machine Translation with a Small Amount of Bilingual Training Data

Maja Popović, Hermann Ney

Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University
Ahornstrasse 55, 52056 Aachen, Germany
{popovic,ney}@informatik.rwth-aachen.de

Abstract

The performance of a statistical machine translation system depends on the size of the available task-specific bilingual training corpus. On the other hand, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires a lot of time and effort, and, for some language pairs, is not even possible. Besides, small corpora have certain advantages like low memory and time requirements for the training of a translation system, the possibility of manual corrections and even manual creation. Therefore, investigation of statistical machine translation with small amounts of bilingual training data is receiving more and more attention. This paper gives an overview of the state of the art and presents the most recent results of translation systems trained on sparse bilingual data for two language pairs: Spanish-English, already widely explored with a number of (large) bilingual training corpora available, and Serbian-English - a rarely investigated language pair with restricted bilingual resources.

1. Introduction

The goal of this paper is to give an overview of the state of the art in statistical machine translation using a small amount of bilingual training data and to illustrate it with the most recent results obtained on the Spanish-English and Serbian-English language pairs.

2. Statistical Machine Translation with Sparse Bilingual Training Data

The goal of statistical machine translation is to translate a source language sequence into a target language sequence by maximising the posterior probability of the target sequence given the source sequence. In state-of-the-art translation systems, this posterior probability usually is modelled as a combination of several different models, such as: phrase-based models for both translation directions, lexicon models for both translation directions, target language model, phrase and word penalties, etc. Probabilities that describe correspondences between the words in the source language and the words in the target language are learned from a bilingual parallel text corpus and language model probabilities are learned from a monolingual text in the target language. Usually, the larger the available training corpus, the better the performance of a translation system. Whereas the task of finding appropriate monolingual text for the language model is not considered as difficult, acquisition of a large high-quality bilingual parallel text for the desired domain and language pair requires a lot of time and effort, and for some language pairs is not even possible. In addition, small corpora have certain advantages: the possibility of manual creation of the corpus, possible manual corrections of automatically collected corpus, low memory and time requirements for the training of a translation system, etc. Therefore, the strategies for exploiting limited amounts of bilingual data are receiving more and more attention. In the last five years various publications have dealt with the issue of sparse bilingual corpora.

(Al-Onaizan et al., 2000) report an experiment of Tetun-English translation with a small parallel corpus, although this work was not focused on the statistical approach. The translation experiment has been done by different groups including one using statistical machine translation. They found that the human mind is very capable of deriving dependencies such as morphology, cognates, proper names, etc. and that this capability is the crucial reason for the better results produced by humans compared to corpus based machine translation. If a program sees a particular word or phrase one thousand times during the training, it is more likely to learn a correct translation than if it sees it ten times, or never. Because of this, statistical translation techniques are less likely to work well when only a small amount of data is given.

(Callison-Burch and Osborne, 2003) propose a co-training method for statistical machine translation using the multilingual European Parliament corpus. Multiple translation models trained on different language pairs are used for producing new sentence pairs. They are then added to the original corpus and all translation models are retrained. The best improvements have been achieved after two or three training rounds.

In (Nießen and Ney, 2004) the impact of the training corpus size for stastistical machine translation from German into English is investigated, and the use of a conventional dictionary and morpho-syntactic information for improving the performance is proposed. They use several types of word reorderings as well as a hierarchical lexicon based on the POS tags and base forms of the German language. They report results on the full corpus of about sixty thousand sentences, on the very small part of the corpus containing five thousand sentences and on the conventional dictionary only. Morpho-syntactic information yields significant improvements in all cases and an acceptable translation quality is also obtained with the very small corpus.

Statistical machine translation of spontaneous speech with a training corpus containing about three thousand sentences has been dealt with in (Matusov et al., 2004). They propose acquiring additional training data using a n-gram coverage measure, lexicon smoothing and hierarchical lexicon structure for improving word alignments as well as several types of word reorderings based on POS tags.

Statistical machine translation of the Spanish-English and Catalan-English language pair with sparse bilingual resources in the tourism and travelling domain is investigated in (Popović and Ney, 2005). The use of a phrasal lexicon as an additional language resource is proposed as well as introducing expansions of the Spanish and Catalan verbs. With the help of the phrasal lexicon and morphological information, a reasonable translation quality is achieved with only one thousand sentence pairs from the domain.

The Serbian-English language pair is investigated in (Popović et al., 2005). A small bilingual corpus containing less than three thousand sentences was created and statistical machine translation systems were trained on different sizes of the corpus. The obtained translation results are comparable with results for other language pairs, especially if the small size of the corpus and rich inflectional morphology of the Serbian language are taken into account. Morpho-syntactic information is shown to be very helpful for this language pair.

Statistical machine translation of the Czech-English language pair and the impact of the morphological information are investigated in (Goldwater and McClosky, 2005). As with Serbian-English, morphological transformations have an important role for the translation quality.

The problem of creating word alignments for languages with scarce resources i.e. Romanian-English, Inuktikut-English and Hindi-English has been adressed in (Lopez and Resnik, 2005; Martin et al., 2005).

3. Recent Translation Results

The translation system used in our most recent experiments with sparse training data is based on a log-linear combination of seven different models, the most important ones being phrase models (Vilar et al., 2005; Zens et al., 2005). For each language pair, several set-ups with different amount of bilingual data and several types of morpho-syntactic transformations were defined. The morpho-syntactic transformations have been implemented as a preprocessing step, therefore modifications of the training or search procedure were not necessary. For all experiments, the language model has been trained on the largest target language corpus because acquisition of monolingual data is not a particularly difficult issue. The evaluation metrics used for assessment of the systems are WER (Word Error Rate), PER (Position-independent word Error Rate) and BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002).

3.1. Spanish-English

The translation systems for this language pair are tested on the European Parliament Plenary Sessions (EPPS) corpus which is also used in the TC-Star project evaluation. A description of the corpus can be found in (Vilar et al., 2005). In order to investigate sparse training data scenarios, two sets of a small corpora have been constructed by random selection of sentences from the original corpus. The small corpus referred to as 13k contains about 1% of the original large corpus. The corpus referred to as 1k contains only 1000 sentences - such a corpus basically can be produced manually in relatively short time.

Training		Spanish	English
1.3M	Sentences	1281427	
	Running Words+PM	36578514	34918192
	Vocabulary	153124	106496
	Singletons [%]	35.2	36.2
13k	Sentences	13360	
	Running Words+PM	385198	366055
	Vocabulary	22425	16326
	Singletons [%]	47.6	43.7
1k	Sentences	1113	
	Running Words+PM	31022	29497
	Vocabulary	5809	4749
	Singletons [%]	60.8	55.3
dict.	Entries	52566	
	Running Words+PM	60964	62011
	Vocabulary	31126	30761
	Singletons [%]	67.7	67.4
Test	Sentences	840	1094
	Running Words+PM	22774	26917
	Distinct Words	4081	3958
	OOVs (1.3M) [%]	0.14	0.25
	OOVs (13k) [%]	2.8	2.6
	OOVs (1k) [%]	10.6	9.4
	OOVs (dict.) [%]	19.1	16.2

Table 1: Corpus statistics for the Spanish-English EPPS task (PM = punctuation marks)

In addition, a conventional Spanish-English dictionary collected from the web which is not related to any particular task is used. The dictionary contains about fifty thousand entries and thirty thousand distinct words for each language.

The statistics for all corpora can be seen in Table 1. For the large corpus, the number of OOVs in the test is very small, much less than 1%. This number grows up to 10% by reducing the bilingual corpus, and for the dictionary alone it reaches 19% for Spanish and 16% for English.

Morpho-syntactic transformations: Adjectives in the Spanish language are usually placed after the corresponding noun, whereas for English it is the other way round. Therefore, for this language pair we applied local reorderings of nouns and adjective groups in the source language as described in (Popović and Ney, 2006). If the source language is Spanish, each noun is moved behind the corresponding adjective group. If the source language is English, each adjective group is moved behind the corresponding noun. An adverb followed by adjective (e.g. ”more important”) or two adjectives with a coordinate conjuntion in between (e.g. ”economic and political”) are treated as an adjective group. In addition, Spanish adjectives, in contrast to English, have four possible inflectional forms depending on gender and number. This might introduce additional data sparseness problems, especially if only a small amount of training data is available. Thus we replace all Spanish adjectives with their base forms.

Translation results: The following set-ups are defined for the Spanish-English language pair:

Spanish→English		WER	PER	BLEU
dict	baseline	60.4	49.3	19.4
	+reorder adjective	59.4	47.4	20.1
	+adjective base	56.4	46.8	23.8
1k	baseline	52.4	40.7	30.0
	+dictionary	48.0	36.5	36.0
	+reorder adjective	45.0	35.3	39.8
	+adjective base	44.5	34.8	40.9
13k	baseline	41.8	30.7	43.2
	+dictionary	40.6	29.6	46.3
	+reorder adjective	38.5	29.2	48.9
	+adjective base	38.3	29.0	49.6
1.3M	baseline	34.5	25.5	54.7
	+reorder adjective	33.5	25.2	56.4

Table 2: Translation results [%] for Spanish→English

English→Spanish		WER	PER	BLEU
dict	baseline	67.6	55.9	14.1
	+reorder adjective	66.3	55.2	15.7
	+align adjective base	65.7	54.5	16.5
1k	baseline	60.1	47.4	23.9
	+dictionary	56.0	43.2	28.3
	+reorder adjective	54.0	42.0	30.5
	+align adjective base	53.9	42.0	30.6
13k	baseline	49.6	37.4	36.2
	+dictionary	48.6	36.3	37.2
	+reorder adjective	47.4	36.0	38.6
	+align adjective base	47.3	35.7	39.1
1.3M	baseline	39.7	30.6	47.8
	+reorder adjective	39.6	30.5	48.3

Table 3: Translation results [%] for English→Spanish

- training only on a conventional dictionary (dict);
- training on a very small task-specific bilingual corpus (1k);
- training on a small task-specific bilingual corpus (13k);
- training on a large task-specific bilingual corpus (1.3M).

The language model for all set-ups is trained on the large corpus.

Table 2 presents the results for the translation from Spanish to English. It can be seen that the error rates of the system trained only on the dictionary are high and that morpho-syntactic transformations improve the performance. Although the final error rates are still high, they might be acceptable for tasks where only the gist of the translated text is needed, like for example document classification or multilingual information retrieval. Additional morpho-syntactic transformations such as treatment of Spanish verbs could further improve the performance.

When only a very small amount of task-specific bilingual parallel text is used (1k), all error rates are decreased and the BLEU score is increased in comparison to a system trained on the dictionary alone, although they are still rather high. Further, it can be seen that the dictionary is very helpful as an additional training corpus and the morpho-syntactic transformations have a significant impact so that the final error rates are reduced by about 15% relative in comparison to the baseline system. By increasing the size of the task-specific training corpus (13k) all error rates are further decreasing and can be further reduced with help of the dictionary and morpho-syntactic transformations.

The best results obtained with the large corpus are about 12% (relative) better than the best results with the small corpus (13k) and about 25% better in comparison with the very small corpus (1k). These differences seem to be very large, but we have to keep in mind how large the differences between the corpus size are, especially in terms of the time and effort necessary for collection and handling of large corpora.

It should be noted that the impact of a dictionary has not been tested for the full corpus since the corpus itself is sufficiently large. The improvements by replacing Spanish adjectives with their base forms are rather insignificant on this corpus and therefore are not reported.

The translation results for the other direction can be seen in Table 3. All error rates are higher due to the inflectional morphology of the Spanish language, and the effects of the training corpus size, dictionary and morpho-syntactic transformations are very similar. The improvements from the morpho-syntactic transformations are slightly smaller than for the translation into English due to the following reason: noun-adjective reordering is less important for the translation into Spanish because the adjective group is not always situated behind the noun. Therefore some reorderings in English are not really needed. As for the Spanish adjective inflections, for this translation direction alignment has been trained using adjective base forms, whereas the translation models have been trained on the original corpus. This enables better learning from the corpus to some extent, but finding a correct inflection of a Spanish adjective still remains relatively difficult.

3.2. Serbian-English

The Serbian-English parallel corpus used in our experiments is the electronic form of the Assimil language course described in (Popović et al., 2005). The full corpus is already rather small, containing about three thousand sentences and twenty five thousand running words. In order to investigate extremely sparse training material, a reduced corpus containing 200 sentences reffered to as 0.2k has been randomly extracted from the original corpus. For this corpus, a set of short phrases has been investigated as additional bilingual knowledge.

Table 4 presents the corpora statistics. It can be seen that even for the full corpus the number of OOVs is high, about 5% for English and almost 12% for Serbian (due to the rich inflectional morphology of this language). For the extremely small training corpus, the number of OOVs is about 3 to 4 times higher.

Morpho-syntactic transformations: The inflectional morphology of the Serbian language is very rich for all

Training		Serbian	English
2.6k	Sentences	2632	
	Running Words+PM	22227	24808
	Vocabulary	4546	2645
	Singletons [%]	60.0	45.8
0.2k	Sentences	200	
	Running Words+PM	1666	1878
	Vocabulary	778	603
	Singletons [%]	79.4	65.5
phrases	Entries	351	
	Running Words+PM	617	730
	Vocabulary	335	315
	Singletons [%]	71.3	66.3
Test	Sentences	260	
	Running Words+PM	2100	2336
	Distinct Words	891	674
	OOVs (2.6k) [%]	11.7	4.9
	OOVs (0.2k) [%]	35.2	21.8

Table 4: Corpus statistics for the Serbian-English Assimil task (PM = punctuation marks)

Serbian→English		WER	PER	BLEU
0.2k	baseline	65.5	60.8	8.3
	+phrases	65.0	59.8	10.3
	+base forms	59.2	54.8	13.9
	+verb POS+neg	60.0	52.6	14.8
2.6k	baseline	44.5	37.9	32.1
	+base forms	42.9	37.4	35.4
	+verb POS+neg	41.9	34.7	34.6

Table 5: Translation results [%] for Serbian→English

English→Serbian		WER	PER	BLEU
0.2k	baseline	73.4	68.4	6.8
	+phrases	71.9	67.5	9.3
	+remove article	66.7	62.2	9.4
2.6k	baseline	51.8	45.8	23.1
	+remove article	50.4	44.6	24.6

Table 6: Translation results [%] for English→Serbian

open word classes, but information contained in the inflection usually is not relevant for translation into English. Therefore, converting all Serbian words into their base forms is proposed. Nevertheless, inflections of Serbian verbs might contain relevant information about the person, which is especially important if the pronoun is omitted. Apart from this, there are three Serbian verbs which are negated by appending the negative particle to the verb as a prefix. Thus the following treatment of the Serbian verbs is applied: each verb is converted into a sequence of its base form and the part of the POS tag referring to a person. If the negative form is built by appending a prefix, the prefix i. e. the negative particle is separated.

For the other translation direction, since the articles are one of the most frequent word classes in English, but on the other hand there are no articles at all in Serbian, the articles are removed from the English corpus.

Translation results: For this language pair the following set-ups are defined:

- training on an extremely small task-specific bilingual corpus (0.2k);
- training on a small task-specific bilingual corpus (2.6k).

Since the largest available corpus is already small and the external phrase book is even smaller, we have not investigated translation using only the phrase book, but we used it as additional training material for the extremely sparse training corpus. The language model for all set-ups was trained on the full (2.6k) corpus.

Error rates for the translation from Serbian into English are shown in Table 5. As expected, the error rates of the system trained on an extremely small amount of parallel corpus are high. Performance of such a system is comparable with a system trained only on a conventional dictionary. Adding short phrases is helpful to some extent, and replacing words with base forms has the most significant impact. Further improvements of PER and BLEU score are obtained by the verb treatment although WER is slightly deteriorated. Increasing the size of the bilingual training corpus to about three thousand sentences and applying morpho-syntactic transformations leads to an improvement of about 30% relative. Using a conventional dictionary and additional morpho-syntactic transformations could further improve the performance.

Table 6 shows results for the translation from English into Serbian. As expected, all error rates are significantly higher than for the other translation direction since the translation into the morphologically richer language always has poorer quality.

The importance of the phrases seems to be larger for this translation direction. Removing English articles improves the translation quality for both set-ups. As for the other translation direction, increasing the size of the training corpus results in up to 30% relative improvement.

4. Conclusion

Strategies for statistical machine translation with limited amount of bilingual training data are receiving more and more attention. Past and recent experiences have shown that an acceptable translation quality can be achieved with a very small amount of task-specific parallel text, especially if conventional dictionaries, phrasal books, as well as morpho-syntactic knowledge are available. Translation systems built only on a conventional dictionary or on extremely small task-specific corpora might be usefull for applications such as document classification or multilingual information retrieval.

5. Acknowledgement

The reported work is based on the projects that have been suported by the German Science Foundation (DFG) and by the European Project TC-Star (FP6-506738).

6. References

Yaser Al-Onaizan, Ulrich Germann, Ulf Hermjakob, Kevin Knight, P. Koehn, Daniel Marcu, and Kenji Yamada. 2000. Translating with scarce resources. In *The Seventeenth National Conf. on Artificial Intelligence*, pages 672–678, Austin, TX, July.

Chris Callison-Burch and Miles Osborne. 2003. Co-training for statistical machine translation. In *Proc. of the 6th Annual CLUK Research Colloquium*, Edinburgh, UK, January.

Sharon Goldwater and David McClosky. 2005. Improving stastistical machine translation through morphological analysis. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Vancouver, Canada, October.

Adam Lopez and Philip Resnik. 2005. Improved hmm alignment for languages with scarce resources. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 83–86, Ann Arbor, MI, June.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 65–74, Ann Arbor, MI, June.

Evgeny Matusov, Maja Popović, Richard Zens, and Hermann Ney. 2004. Statistical machine translation of spontaneous speech with scarce resources. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 139–146, Kyoto, Japan, September.

Sonja Nießen and Hermann Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wie-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

Maja Popović and Hermann Ney. 2005. Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, pages 212–218, Budapest, Hungary, May.

Maja Popović and Hermann Ney. 2006. POS-based word reorderings for statistical machine translation. To appear in *Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC)*, Genova, Italy, May.

Maja Popović, David Vilar, Hermann Ney, Slobodan Jovičić, and Zoran Šarić. 2005. Augmenting a small parallel text with morpho-syntactic language resources for Serbian–English statistical machine translation. In *43rd Annual Meeting of the Assoc. for Computational Linguistics: Proc. Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 41–48, Ann Arbor, MI, June.

David Vilar, Evgeny Matusov, Saša Hasan, Richard Zens, and Hermann Ney. 2005. Statistical machine translation of european parliamentary speeches. In *Proc. MT Summit X*, pages 259–266, Phuket, Thailand, September.

Richard Zens, Oliver Bender, Saša Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang, and Hermann Ney. 2005. The RWTH phrase-based statistical machine translation system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.

The BLARK Matrix and its relation to the language resources situation for the Celtic languages

Delyth Prys
Language Technologies Unit, Canolfan Bedwyr
University of Wales, Bangor, UK
E-mail: d.prys@bangor.ac.uk

Abstract

BLARK (Basic Language Resource Kit) was originally developed as a concept to specify the minimum corpora, tools, and skills needed to engage in pre-competitive research for a language. It was then meant to help identify existing resources, and co-ordinate action to fill any gaps. The BLARK matrix has so far been used for Dutch and Arabic, both fairly well-endowed languages in terms of resources, government support and commercial potential. This paper seeks to explore the usefulness of BLARK for languages in a much weaker position, having very few resources, and little, if any, official support. It examines in particular the six Celtic languages spoken today and asks whether BLARK can be used as a tool to assess their resource needs. It suggests that the lack of basic raw materials, such as the absence of daily newspapers in these languages, is a serious drawback for the development of basic resources, and is not adequately covered in the present BLARK matrix. However, identifying the raw materials needed, and finding alternatives where these do not exist, is in itself a valid exercise. This paper proposes therefore the creation of a preliminary matrix, or PRELARK, to aid the development of resources for these languages.

1. BACKGROUND TO BLARK

The Basic Language Resource Kit or BLARK was originally conceived of by Steven Krauwer and proposed as a cooperative initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association). The hope was that “with such an action, every European Language, inside or outside the European Union, could have its own BLARK” (Krauwer 1998). This led to the adoption of the BLARK concept for the Dutch language (Cucchiarini, Daelemans & Strik 2001). The BLARK matrices were then developed and published on the web by ELDA (http://www.elda.org/blark/) with an exercise to gather information on Arabic resources from the NEMLAR project (Maegaard 2004) completed. The BLARK website is an interactive one, with an open invitation to all languages to engage with the exercise and fill in the matrices, or send in proposals for a new matrix and other relevant comments.

2. WELSH LANGUAGE RESOURCES

Interest in using the BARK matrices to help the Celtic languages was first expressed at a workshop at the Language Technologies Unit[1], based at Canolfan Bedwyr in the University of Wales, Bangor, in December 2005. This Unit had, for many years been involved in the creation of language resources for Welsh. These include terminology and lexical databases, spelling and grammar checkers, and more recently, speech processing resources. The Unit was, and remains, entirely self-funded, and had therefore undertaken projects on a needs-led basis, responding to invitations to tender where it had the necessary skills to offer, and developing grant proposals where suitable funding opportunities could be identified. However, it was acutely aware of the ad hoc nature of these developments and felt the need for a 'roadmap' or audit of resources available and needed in order to plan future projects in a more comprehensive and orderly manner. The BLARK concept provided it with an off the peg solution to its needs.

BLARK was seen as providing greater detail and more systematic coverage, compared to other attempts to provide a coherent framework for Welsh language technologies. Specifically, it compared favourably to the Welsh Language Board IT Strategy Document, published in March 2006. This document "aspires to the normalisation of the Welsh language in the world of Information Technology". It lists 64 targets and policy statements in order to fulfill its aims, and are relevant to the creation of a digital infrastructure. However, they are conceived of as discrete targets, with the focus on encouraging, discussing, facilitating, engaging in dialogue, examining possibilities etc, towards developing specific applications, such as machine translation, rather than the establishment of a cohesive, coherent infrastructure to support the proposed activities. There is therefore no consideration of the importance of basic resources, such as written and speech corpora, which may be reused and recycled in many applications, thus saving time and money.

3. OTHER CELTIC LANGUAGES

Some of the projects being undertaken by the Language Technologies Unit were of a multilingual nature, involving other Celtic languages. These included two projects funded under the EU Interreg IIIA Wales/Ireland Programme. One was WISPR (Welsh and Irish Speech

[1] Originally founded as the Centre for the Standardization of Welsh Terminology, it subsequently became part of the e-Welsh Unit which has recently been renamed to reflect more accurately its multilingual role. The Canolfan Bedwyr website may be found at http://www.bangor.ac.uk/ar/cb/index.php

Processing Resources), and the other was Lexicelt, an on-line interactive Welsh/Irish Phrasebook and Dictionary. The question was therefore asked whether, if the BLARK matrices could prove useful for Welsh, it could also be extended to the other five Celtic languages. All six Celtic languages share common socio-political situations, as well as a common linguistic structure and heritage (Ó Néill, 2005). With the exception of Irish, which has recently received full status as an official language of the EU, none of the Celtic languages has an official status in the European Union, and all of them, including Irish, have relatively few native-standard speakers[2].

More important than the official status or number of speakers however was the lack of some basic materials from which to create language resources. For example, language corpora of various kinds are core elements for a number of language resources. Written corpora depend on large amounts of text, such as that provided by daily newspapers. However, none of the Celtic languages currently has a daily newspaper, and therefore, the task of collecting material for a basic, balanced corpus is problematic. Material collected from the internet is increasingly being used for the creation of large scale corpora, and software such as Kevin Scannells' *An Crúbadán* is able to crawl the web and compile corpora automatically from web pages in minority languages. It has been argued that using the web as a huge on-line corpus does away with the need for balanced, representitively chosen samples, as everything will be included. While this may work for larger languages with comprehensive coverage of material on the web, it is less satisfactory for very small languages, for example Cornish, where the material may be mainly the work of a very small core of enthusiasts, with, for example, very little 'official' texts. Neither does it distinguish between original and translated material. In the case of Welsh, for example, public bodies have to maintain bilingual websites, but this material almost always originates in English, and the Welsh versions are translations. Translated material may be of poor linguistic standard, and unsuitable for use as a basis for applications such as semantic analysis. One positive aspect of the extensive use of translation in such circumstnaces is that bilingual text corpora at least should be relatively easy to build, paving the way for applications such as machine translation and bilingual lexicons.

Modules such as transcription of broadcast news are also difficult to gather if there is no, or very little, radio and telvision news being broadcast, as is the case with Irish and Manx, and to a lesser degree, Breton and Scots Gaelic. Under such circumstnces new stratergies may need to be devised, such as the recording of live plays or the commissioning of specially prepared scripts to be read aloud and recorded.

The lack of teaching and research into some of these Celtic languages is also detrimental to the creation of basic resources. Cornish and Manx do not have institutes of higher education on their territories, and there are no undergraduate courses fully dedicated to teaching these languages. This leads to lack of basic materials such as contemporary phonetic descriptions, up-to-date grammars, and accurate place-name information.

4. THE NEED FOR A PRELIMINARY BLARK

While BLARK was designed to be language independent, and to be geared towards a "smaller language of Europe" like Dutch, (Krauwer, 2003), it is still geared towards languages which are in a different league from minority languages such as the Celtic ones which are unlikely to receive much private or public funding for language technologies. While Krauwer did envisage the scenario of some languages having to "start from scratch" to create BLARK components, the reality of being faced with a matrix where the score is repeatedly zero as to availability, and high in terms of prioritized need, can be daunting.

What would be helpful therefore, would be a cut-down version of BLARK, or a PRELARK (Preliminary Language Resources Kit) which would identify the very first components needed to provide entry level applications into language technologies. This would focus not only on the intended basic applications envisaged, but also on the limitations placed the lack of available raw materials. While this would prove useful to many smaller minority languages in Europe, the Celtic languages would be well placed to take immediate advantage of it, to build on their existing networks of cooperation, and use it to share and develop common resources in the wider European context.

5. BIBLIOGRAPHICAL REFERENCES

Krauwer, Steven, (2003). The Basic Language Resource Kit as the First Milestone for the Language Resources Roadmap. http://www.elsnet.org/dox/krauwer-specom2003.pdf

LEXICELT Welsh/Irish on-line Phrasebook and Dictionary http://www.lexicelt.org/

Ó Néill, Diarmuid (ed.) (2005). Rebuilding the Celtic Languages. Y Lolfa Press.

Scannell, Kevin P. (2004). Corpus Building for Minority Languages. http://borel.slu.edu/crubadan/

Welsh Language Board (2006). Information Technology and the Welsh Language: A Strategy Document. Cardiff

WISPR (Welsh and Irish Speech Processing Resources) website http://www.bangor.ac.uk/ar/cb/wispr.php

[2] Note however, that the Isle of Man is not a part of the European Union, but a dependency of the British Crown.

Unicode for Under-Resourced Languages

Daniel Yacob

The Ge’ez Frontier Foundation
7802 Solomon Seal Dr., Springfield, VA 22152, USA
yacob@geez.org

Abstract

Best known as a standard for character encoding, Unicode should now be understood as a collection of resources for textual computing that can aid solution of NLP problems. When an under-resourced language (U-RL) makes use of symbols or an entire system of writing that is not supported by electronic standards, Unicode is by design the best option for the temporary encoding the written symbols before a standard can be formalised. Leveraging Unicode for such U-RLs involves more than just symbol encoding and the NLP researcher must also anticipate developing the most basic of computer resources such as fonts, keyboard and transliteration systems. The paper describes in an overview the work entailed and goes further to cover raising the work products of a personal NLP project to the level of a national and international ICT standard where the native speaker community at large will also benefit directly.

1. What is Unicode?

Very often software engineers not familiar with text representation techniques will perceive “Unicode” as something used to support non-English fonts. Many early Unicode fonts did in fact append the term “Unicode” to their base name[1] and fonts are likely the most frequently encountered utilization of Unicode that a person will be made consciously aware of. However, Unicode is now as pervasive in applications and computer systems as text itself –aware of it or not, you are already using it and have been for some time.

“Unicode” is a very broadly used term, having many contexts, and will mean different things to different people. Historically, “Unicode” refers to the Unicode Standard which defines character encodings for the writing systems of the world. The character encoding standard is synchronised with a parallel standard known as “ISO/IEC 10646” that also addresses character encoding. “Unicode” is also used interchangeably with “ISO/IEC 10646” by all but experts.

“Unicode” most commonly refers to only the Basic Multilingual Plane (BMP) of the original 16 bit standard for encoding modern writing systems. In the right context “Unicode” may also include the 16 additional supplementary planes now a part of the Unicode Standard where historical scripts and scripts used by smaller communities are encoded.

“Unicode” may refer to the Unicode Consortium[2] the governing body that defines Unicode Standard. The consortium membership is comprised of members of the computer industry, academic institutions, and private citizens with an interest in the activities of the consortium. The International Standards Organization (ISO) and the International Electrotechnical Commission (IEC) jointly maintain the ISO/IEC 10646 standard and participants are almost entirely appointed by a government recognised national standards body. The Unicode consortium works so closely with the ISO/IEC that it has become the primary driving force behind revisions of the standard. To further distinguish between these standards bodies and their unified standards is not important unless one intends to work very deeply on defining the standards themselves.

“Unicode” may also refer to the Unicode Character Database (UCD) which defines the names and properties of characters. The database is critical for text processing tools such as regular expressions libraries and layout engines. Similarly, “Unicode” may refer to a collection of annexes, technical reports and technical standards that further define textual properties and behaviours such as text boundaries, collation and equivalence classes.

“Unicode” may refer to the projects under the Unicode Consortium such as the Common Locale Data Repository (CLDR) that defines basic vocabulary and cultural conventions used in operating systems. Likewise “Unicode” can refer to the International Components for Unicode (ICU), a project involving many of the same people, which implements most of the specifications of the Unicode Consortium and data such as CLDR.

Finally, “Unicode” may refer to the Unicode home page portal, mail lists, community, technical committee (UTC), twice annual International Unicode Conference (IUC), or the process by which these aspects work together to further develop the Unicode Standard and its family of specifications.

For the purposes of this paper “Unicode” will refer then to the family of standards and technologies associated with the Unicode Consortium that can be utilised for working with a written language in a computer environment. For the NLP researcher, “Unicode” need be no more than a means to some other end. How Unicode can be applied to help solve problems in NLP, and in particular for Under-Resourced Languages (U-RLs) that may not yet be supported in Unicode, will be the focus of this paper.

2. Working *With* Unicode

More than anything else, what Unicode offers NLP is a resource for representing written languages. To a lesser degree, Unicode can also be an aid for tokenizing spoken language with phonetic symbology and for text corpora processing from comprehensive properties provided for all written symbols.

[1] e.g. in Microsoft Windows the “Lucida Sans Unicode” font is the companion to “Lucida Sans” and “Arial Unicode MS” the companion to “Arial”.

[2] Unicode Consortium Homepage: http://unicode.org/

A review of software under the Natural Language Software Registry[3] reveals that many applications have already migrated to Unicode which makes them accessible to U-RLs already. Typically, Unicode based tools are designed to be language neutral and will be able to process language specific information through modules (classes, xml data, etc) containing settings and rules applicable for that language. So adapting a tool for a specific language is often a matter of developing the needed module.

NLP tools and applications are most often engineered to work on a very specific problem set so it will not be the purpose here to recommend one NLP resource over another. Rather the objective is to make an overview of the most likely tasks entailed to customise an existing application to support a given language via Unicode as well to leverage Unicode in applications that you author yourself.

2.1. Operating Systems

Applying Unicode, the first thing that you will need is a recent computer operating system (OS) with native support for the encoding standard. Fortunately, this side of the 21st century it is harder and harder to find an operating system that does *not* support Unicode. The newer the operating system the better but you will want to consider upgrading if your computer runs with Apple MacOS earlier than version 9.2, Microsoft Windows other than CE, NT, XP, Vista or 2000, Solaris earlier than 2.8, or GNU/Linux with glibc earlier than 2.2.2.

Linux systems are provided by many different distributors with as many different configurations and so deserve a little more attention. To determine if the version of glibc is of version 2.2.2 or later enter the following at the command line:

```
% ls -l /lib/libc.so.* /lib/libgtk*
```

which should return a response containing libraries similar to:

```
  lrwxr-xr-x  1 root root  13 Nov 29
2004 /lib/libc.so.6 -> libc-2.3.2.so*
```

libc.so.6 is a symbolic link to the current version of glibc on the system which in the above example is version 2.3.2. If the operating system is not at the version level indicated here this does not mean that you will be unable to work with Unicode. The consequence will be that you are less assured that you will be able to view or enter Unicode text in terminals or use it with some applications. You may in fact still apply the encoding for your needs but there are fewer guarantees and you will likely have to try several editor applications before finding one that can display Unicode text as expected (consider Yudit discussed later).

The window desktop environment in a GNU/Linux system is entirely separate from the OS. If your system is recent enough to have glibc 2.2.2 or later, then it very likely also has GNOME 2.0 or KDE 2.0 or later which are Unicode safe window environments.

2.2. The IPA

When working with phonology, Unicode offers the symbols of the International Phonetic Alphabet. (IPA) The IPA is maintained by the International Phonetic Association and defines a unique symbol for every phoneme used in spoken languages. The IPA is itself a standard for the linguistics community to apply for phonetic transcription that will assure mutual comprehension within the field in the present and future. The IPA should be applied in the NLP community wherever spoken language must be tokenised and avoids creating your own system where data can only be understood by your own applications. Using the IPA under Unicode will allow you to take advantage of Unicode based software and reduces the new resources that you would otherwise have to invest time to develop.

The SIL Doulos font presents all the IPA symbols with excelling quality using typeface resembling the Times Roman[4].

2.3. The PUA

When developing a new orthography, extending an existing one, or have found that Unicode does not have the symbol that you need –you can still use Unicode. Unicode has a built in way to support additions to its own character repertoire via the Private Use Area (PUA). In this region of the character encoding standard you may define your own symbols. This is helpful when working with an experimental orthography or an as yet unsupported writing system, two frequent occurrences with U-RLs.

You may define additional letters in this region of the standard and use them safely with your own applications and others. If you want to visualise the symbols that you have encoded in the PUA you will need to modify a font to contain these symbols. Since you have made your own "private encoding" within Unicode you risk losing some portability. Sending a document to a colleague that contains your additions, you will also need to send along your modified font. If your PUA additions are processed in a multilingual system where another researcher has also made PUA extensions, there is the possibility of encoding collisions. This is an exceedingly rare occurrence but becomes possible when you go from the private use case to one that is public.

The PUA may also be used to encode other useful tokens, for example tags and other markers that you would like to have in a text stream that you would processes yourself. You may not need to view tokens of this type but if you chose to do so it is again a matter of modifying a font.

2.4. Fonts

Fonts provide us with an instance of Unicode encoded characters. Fonts can have many styles (typefaces) and the Unicode standard does not tell us how letters should appear (glyphs) or how they should be typed (hardware dependent). The standard simply assigns numeric addresses (in computer memory these are byte sequences) to the abstraction notion of a

[3] NL Software Registry: http://registry.dfki.de/

[4] SIL Doulos Font: http://scripts.sil.org/DoulosSILfont

"letter". A "Unicode font" is one that applies these assignments (encoding) to the letters that it contains. Not every "Unicode font" will contain all letters of Unicode, in fact few do. The reasons here are primarily the labour required to produce a complete coverage of Unicode's 51,980 graphical characters as well as the undesirably file sizes that result (fonts are files) which can be on the order of 20 Mb. Few people really need all characters of all alphabets and so most Unicode fonts will only support a few writing systems needed by a target community. This approach works fine so long as you are aware of the supported character range within the font, when letters are not available an application may instead substitute a blank space, a dot, a question mark or commonly a rectangle (▯).

When working with a U-RL there is a greater likelihood that the orthography of the language is not yet supported in Unicode. In this case you will need to create your own fonts in order to view text in the native script. This need not be an obstacle because there are freely available tools to create and edit fonts, but it will require an investment of your time that will vary depending on number of letters that need be created.

Fonts come in two types they are either bitmaps – a matrix of on or off dots like a tile design, or they are outlines of the shapes of the letters. The outlined versions ("TrueType" is the most common) are more portable and scale to different sizes with better quality than do the bitmap fonts, but they can take more time to produce. If your intent is not to market the fonts commercially than you do not need to be an artist to be able create letters you need only be able to use a mouse.

A number of commercial and free tools are available for creating and manipulating fonts. Working with bitmap fonts "gbdfed"[5] is a very easy to use tool that runs natively under Linux but can also be used on other operating systems where the GTK library has been ported to. "FontForge"[6] is an open source outline font editor that can run on most every major operating system (some additions may be required) that also supports some bitmap formats. It will always be a less intensive effort to start with an open source font and to modify it for your needs than to start completely from scratch. Fonts that do not also contain the Latin letters can later be problematic to work with, some systems will refuse to use them, so including the ASCII range of letters is always recommend. Seemingly countless free and open source fonts can be found readily with an internet search.

2.5. International Components for Unicode

By far the most extensive, complete and Unicode specific resource is the International Components for Unicode (ICU)[7]. ICU comes in the form of both a C/C++ library and a Java JAR. The resource was initially a project of IBM before going Open Source, it has the participation of many of the key Unicode personnel and offers reference implementations of the Unicode family of standards.

The ICU homepage described the resource as "…a mature, widely used set of C/C++ and Java libraries for Unicode support, software internationalization and globalization (i18n/g11n). It grew out of the JDK 1.1 internationalization APIs, which the ICU team contributed, and the project continues to be developed for the most advanced Unicode/i18n support. ICU is widely portable and gives applications the same results on all platforms and between C/C++ and Java software." (ICU)

ICU should be considered as a resource for Unicode text processing, some of its services will be touched on in following sections.

2.6. Transliteration

Transliteration is the systematic conversion of one system of writing onto another. It is most often used in NLP for converting text to and from some encoding system into a Romanised form that legacy resources can then understand. As more resource become Unicode enabled there is proportionally less reliance on the conversion technique. Transliteration will however always be useful when facing text that is not in a script that you are familiar with, working with toponymic lexicons, and making rough conversions to and from a writing system and the IPA.

```
<icu:transform type="Latin">
  # variables
  $gammaLike = [ΓΚΞΧγκξχϰ] ;
  ...
  # convert all to decomposed
  ::NFD (NFC) ;
  ...
  α ↔ a ;  Α ↔ A ;
  β ↔ v ;  Β ↔ V ;
  # contextual transforms
  γ } $gammaLike ↔ n } $egammaLike ;
  Γ } $gammaLike ↔ N } $egammaLike ;
  γ ↔ g ;  Γ ↔ G ;
  δ ↔ d ;  Δ ↔ D ;
  ε ↔ e ;  Ε ↔ E ;
  ζ ↔ z ;  Ζ ↔ Z ;
  # contextual transform
  Θ } $beforeLower ↔ Th ;
  θ ↔ th ; Θ ↔ TH ;
  ι ↔ i ;  Ι ↔ I ;
  κ ↔ k ;  Κ ↔ K ;
  λ ↔ l ;  Λ ↔ L ;
  μ ↔ m ;  Μ ↔ M ;
  # contextual transforms
  ν } $gammaLike → n\' ;
  Ν } $gammaLike ↔ N\' ;
  ν ↔ n ;  Ν ↔ N ;
  ...
  # convert back to composed
  ::NFC (NFD) ;
 </icu:transform>
```

Figure 1: Latin ↔ Greek Transliteration Sample in ICU

As provided, the ICU transliteration capability is promoted as supporting "50+" systems. New

[5] gbdfed Homepage: http://crl.nmsu.edu/~mleisher/gbdfed.html

[6] FontForge Homepage: http://fontforge.sf.net/

[7] ICU Homepage: http://icu-project.org/

transliteration systems can be added without requiring source code recompilation. Figure one presents a familiar example with the Latin and Greek alphabets:

A very valuable feature of transliteration in ICU is the extension to “compound transforms”. Under the concept of compound transformations defined transliteration systems may be chained together for special conversions. For example:

```
[:Lu:] Latin-Katakana; Latin-Hiragana;
```

Defines a compound transformation where uppercase Latin letters (identified with the Unicode character class [:Lu:]) are converted into Katakana. The output of the first transformation, terminated with the semicolon “;” symbol, becomes the input for the next. The remaining lowercase Latin characters will be converted into Hiragana. For example:

Input ⇒	[:Lu:] Latin-Katakana; ⇒	Latin-Hiragana;
Washington ⇒	ウashington ⇒	ウあしんぐとん

Reversible transliteration systems require strict adherence. The most common deviations from stringent transliteration come from the application of transcription rules. There are legitimate reasons for applying transliteration and transcription together, for example when a transcription is well established and the transliterated rendering would cause confusion. The meta-language for transcription should be capable of contextual processing to make some conversion of transcription and other phonological phenomena that manifest in an orthography possible. This capability will also be useful in validation and conformance work where conversion from transcription to stringent transliteration is the objective.

For example, a regular expression based rule for the elision of gemination characters:

```
([^aeiou]){2} ⇒ $1;
```

which replaces two occurrences of a non-vowel (i.e. a consonant) with a single occurrence. Contextual transliteration of Greek gamma:

```
{γ} [ ΓΚΧΞγκχξ ] > n;
γ > g;
```

which converts gamma into *n* if gamma is followed by any of: Γ, Κ, Χ, Ξ, γ, κ, χ or ξ. To enhance phonetic transcription it is advantageous to precondition an English string to better represent the spoken value of the word in the target language. For example “progeny” in American English spoken form is “pro–ǧə–ni” (IPA). Rules can be applied here, within the domain of American English orthography:

```
([^aeiou])y$ > $1i;
oge > oje;
```

thus the Cyrillic rendering, for example becomes “пройени” and not naïvely as “прогены”. These special rules should generally be applied prior to regular transliteration. At the end of a transliteration process the capability may also help further simplify symbol clusters that have occurred from the application of earlier rules.

2.7. Keyboards

When a U-RL requires a writing system which has little or no computer legacy it can be useful to develop a typing system for entering new in-language text. Adding an Input Method (a typing system, or “IM”) is very platform and API specific exercise. Unlike character encoding and fonts formats, there are no recognised standards for a keyboard implementation that is portable across operating systems or window environments. Unless you have hired typist to enter or compose text for your project, developing a keyboard system will likely not be worth the investment of time to learn the APIs and develop the IM software. When time is an issue it will be more efficient to develop a transliteration system, enter the text samples you need in regular Latin script, and convert them into the target script (still under Unicode encoding) with a transliterator.

If it is essential to have an IM for your project the best options on a Window system will be to work with the very robust Tavultesoft Keyman Developer[8] which has a small licensing fee. On Linux systems the trend has been to move towards the Smart Common Input Method (SCIM)[9] where an IM can be defined in configuration files without having to develop new source code. The SCIM based Keyboard Mappings For Linux (KMFL) also offers some Linux compatibility for Keyman IMs. Keyman and SCIM attempt to provide keyboard support for all applications within a window environment. Some applications do provide their own keyboard interpreters as a means to obtain greater independence from the window system and thus offer some portability between window and operating systems. Yudit[10] and Emacs are two such examples of editor applications that provide their own IM infrastructure. Adding an IM to Emacs[11] will take a little knowledge of the Lisp programming language and in Yudit it is a matter of writing a text based mapping file.

As a last resort, a platform portable IM can be developed in either Java or JavaScript but under the restricted contexts of where these languages can be used.

2.8. Text Processing

Unicode and ICU do not address analytical linguistic issues directly but do provide many of the language neutral facilities that you would build upon to address language specific problems. For instance ICU does not have support for stemming but will apply Unicode definitions for character properties to offer text segmentation, normalization, and highly advanced pattern matching.

A number of writing systems have numerous legacy encoding systems; for example Vietnamese had 43 systems (Erard) and Ethiopic over 70. Unicode offers a

[8] Tavultesoft Homepage: http://tavultesoft.com/

[9] SCIM Homepage: http://www.scim-im.org/

[10] Yudit Homepage: http://yudit.org/

[11] Emacs Home: http://www.gnu.org/software/emacs/

vendor neutral encoding that legacy systems can be converted into. Algorithms may then be developed to understand only a single system. ICU offers over 700 encoding system conversions and developers may add to it as needed. (Davis and Scherer)

ICU also provides character normalization services. A character in Unicode may still have for than one form, this is common with “composed characters”. For example ‘ä’ may be the singe character with address U+00E4 or may also be comprised of the two characters ‘a’ + ‘̈’ with the addresses U+0061 and U+0308. There are seemingly endless numbers of possible composed characters and ICU will know how to map components into a single character if available or possibly a Unicode defined named sequence; thus making text comparisons more successful.

Similar to normalization the notion of equivalence classes is supported in regular expressions languages. ICU extends the Java regular expressions support to account for all character classes in the “Unicode Character Database”[12]. Here every property of a character is defined, such as case, letter type, punctuation type, numeric type, and so on, there are a great number of these classes. In a very simple example of how they are applied in a regular expression, the following statement would match only letters with “uppercase” property but excluding (‘-’) those in the Latin alphabet:

```
/[\p{Lu}-\p{Latin}]/
```

Both the C/C++ and Java APIs support the Unicode style regular expressions extensions as defined in the Unicode Technical Report #18[13]. The Perl and C# programming languages also support Unicode style regular expressions without depending on ICU. While very powerful for complex pattern matching the Unicode regular expressions syntax does not always go far enough to support lesser understood properties of U-RL scripts. The Perl regular expressions support is very simple to extend and some modules have been developed to support pattern matching in syllabic scripts such as Cherokee and Ethiopic. For example with the Regexp::Ethiopic[14] Perl module the expression:

```
/[መ-ቀ]{#4,6#}/
```

matches any character in the range of መ through ቀ but only in the 4th or 6th orders, ie the set:

```
[ማምሣሥራርሳስሻሽቃቅ]
```

The module demonstrates overloading of the Perl regular expressions engine. The same can be done with the ICU classes, it is often faster however to prototype and experiment in Perl first and then follow with a Java or C/C++ implementation as needed.

3. Working *for* Unicode

As an NLP researcher working with an U-RL your objective will be to solve some very specific problem within a limited period of time and within the fiscal constraints of a budget. The native speaker community will benefit from the body of knowledge of their language having been expanded by your work. You will have made it easier for future researchers to enter into and further explore the language. You may solve the problem that lies before you and move on to others, you have no further obligation to work with the language beyond your original mandate. If you’re lucky, however, your experience with the U-RL may become a labour of love that you continue to seek funding to work on or take up in personal time.

If you are so able and wish to do so, working beyond the problem that brought you to the U-RL, you can expect to be drawn deeper and deeper into the community and engage in every broader computing resource problems. Indeed, you may be one of very few, or even the only, person actively developing software for the community for quite some time. As the resources that you develop and provide become utilised by the research and native-speaker community, you inherit some responsibility to maintain them and provide support services.

Gradually you will also become a technical bridge between the community and the greater software industry. At this level you have become a subject expert on the language and its computing resources. The best service that you can do for the language, and yourself, is to build upon your experience and community contacts to advance your work to the level of standards for the language. With standards available for the language software companies have in a sense the “legal basis” that they need to begin support for the requirements of the language. This will be a great benefit to the native speakers and with the software industry picking up support for a language you will be relieved of the burden of having to maintain and support your earlier work.

3.1. Character Encoding and The Script Encoding Initiative

If in the course of your work you have developed a font or experimental encoding system for the script, you will want to consider permanent encoding in the Unicode Standard. Doing so will ensure that the letters will live on for the life of the standard itself and give software companies the technical foundation needed to support the script and language.

The challenge of obtaining the standard however is not unlike trying to be your own trial lawyer in courtroom in a foreign environment where you do not know the laws. It is not for the faint of heart and will most definitely become a bigger undertaking than you had in mind.

It will be much better to have an experienced standardisation expert on your side that knows the system and can navigate the process for you while you provide the subject matter expertise. Launched in April 2002 at University of Berkeley, the Script Encoding Initiative (SEI)[15] is just such an expert.

[12] http://www.unicode.org/Public/UNIDATA/UCD.html

[13] http://www.unicode.org/reports/tr18/

[14] http://search.cpan.org/~dyacob/Regexp-Ethiopic/

[15] SEI Homepage: http://www.linguistics.berkeley.edu/sei/

The SEI has "…the goal of organizing and orchestrating the completion of the Unicode Standard. SEI involves other institutions than software companies, reaching out to academia and the public sector, drawing upon scholars around the world as a major resource and on their research results, existing publications and script descriptions. SEI also involves a small group of experts in script encoding to work with scholars to make finished, workable proposals and to move those proposals through the standardization process as soon as possible." (Anderson)

Even if you never anticipate working on a standard for the script, do collect as much cultural information about the writing system as you can while you have the opportunity in the field. If not you, a native speaker or another researcher wishing to work with the SEI, can apply the information in a proposal process. There is almost always more to letters than just the sounds they make.

3.2. Script Name and Language Codes

Along with the encoding of the written symbols of the script, the script and language identities likewise need to be encoded in the respective standards. In this case the standard relevant to language encoding is the ISO 639 family (parts 1, 2 and 3). It is in this standard where "en" is defined as a code for "English" and "lol" for "Mongo". These ISO 639 language codes are used by applications and operating systems to configure a language setting.

Parts one and two cover only 506 of the 7,300 main languages tracked by the Ethnologue. (SIL) Part three attempts to cover all of the world's languages and has been a draft standard for a number of years and should become a final standard shortly.

If the U-RL that you are working with is not covered by ISO 639-3, you will want to request its inclusion from the registrar[16]. Requests can also be made for the encoding of a dialect. It is still possible to request a code assignment from ISO 639-2 which may help gain software support faster. However, ISO 639-2 does require evidence of 50 in-language documents from 5 institutes that most U-RLs are unlikely to have[17].

To encode a script identity, such as "tfng" for "Tifinagh", ISO 15924 is the applicable standard. As with language name encodings the registrar should be contacted if the script the U-RL uses has not been assigned a code[18].

3.3. National Standards

Most every country will have some government recognized body such as a ministry, industrial consortium, or professional organization entrusted to define standards for the nation as a whole. Working with a national standards body should always be attempted before approaching an international body.

It may be the case that the national standards body does not have the interest, expertise, funds or other resources to develop a national standard for technology focused topics such as character encoding, collation, keyboard mappings, or more deeply linguistic areas such as lexicons.

On the other hand, a standards body may be grateful for any help it can get and will welcome your initiative particularly in information technology areas. Potentially, a standards body may fund and even take over the "defacto standard" that you have intrinsically developed for your resources and evolve it into a formal standard. Having a national standards body endorse and legalize a defacto standard for a system puts the highest level of clout behind it. This will make it much easier to achieve international recognition for the standard, particularly if the standards body is also the nation's representative to the ISO. Local software companies may also begin supporting the standard without waiting for the international recognition.

4. Conclusion

In uncharted territory there will be more to explore academically, and more pure knowledge build, but at the cost of moving a little slower while you tread the first roads.

The entry cost into working on a U-RL will always be higher relative to working on a well explored language. This greater cost is incurred for the lack of informational, computational and the lowest level of resources for language representation on a computer. With each passing year the growing availability of Unicode based tools helps lower the entry cost in both time and money required to undertake U-RL research. Indeed, it has never been easier.

While there can be a greater burden upon an NLP researcher to develop basic level resources to engage in U-RL research, there is also the opportunity to realize a greater and more meaningful impact from the research activity. By developing those resources and filling the resource void, you are helping the language and its culture make the leap into the information era and have a better chance at surviving into the future.

5. References

Anderson, Deborah, September 2003. The Script Encoding Initiative, Multilingual Computing, Volume 14, Issue 6, 34.

Davis, Mark, and Markus Scherer, 2005. Globalizing Software, Retrieved April 14, 2006, from http://icu.sourceforge.net/docs/papers/globalizing_software.ppt

Erard, Michael, 2003, September. Computers Learn New ABCs, Technology Review, 28.

International Components for Unicode. (n.d.). Retrieved April 14, 2006, from http://icu.sourceforge.net/

SIL International, 2006. Ethnologue, Retrieved April 14, 2006 from http://www.ethnologue.com/

[16] ISO 639-3 Registrar: http://www.sil.org/iso639-3/

[17] ISO 639-2 Reg. : http://www.loc.gov/standards/iso639-2/

[18] ISO 15924 Registrar: http://www.unicode.org/iso15924/

A Natural Language Generator for Minority Languages

Tod Allman, Stephen Beale

Linguistics Department
University of Texas at Arlington
todallman@sbcglobal.net

Computational Linguistics Department
University of Maryland at Baltimore
sbeale@cs.umbc.edu

Abstract

The Bible Translator's Assistant (TBTA) is a natural language generator (NLG) designed specifically for field linguists doing translation work in minority languages. In particular, TBTA is intended to generate drafts of the narrative portions of the Bible as well as numerous community development articles in a very wide range of languages. TBTA uses the rich interlingua approach. The semantic representations developed for TBTA consist of a controlled English based metalanguage augmented by a feature system designed specifically for minority languages. The grammar in TBTA has two sections: a restructuring grammar and a synthesizing grammar. The restructuring grammar restructures the semantic representations in order to produce a new underlying representation that is appropriate for a particular target language. Then the synthesizing grammar synthesizes the final surface forms. To date TBTA has been tested with four languages: English, Korean, Jula (Cote d'Ivoire) and Kewa (Papua New Guinea). Experiments with the Jula text indicate that TBTA triples the productivity of professional mother tongue translators without any loss of quality. A model of TBTA is shown below in Figure 1.

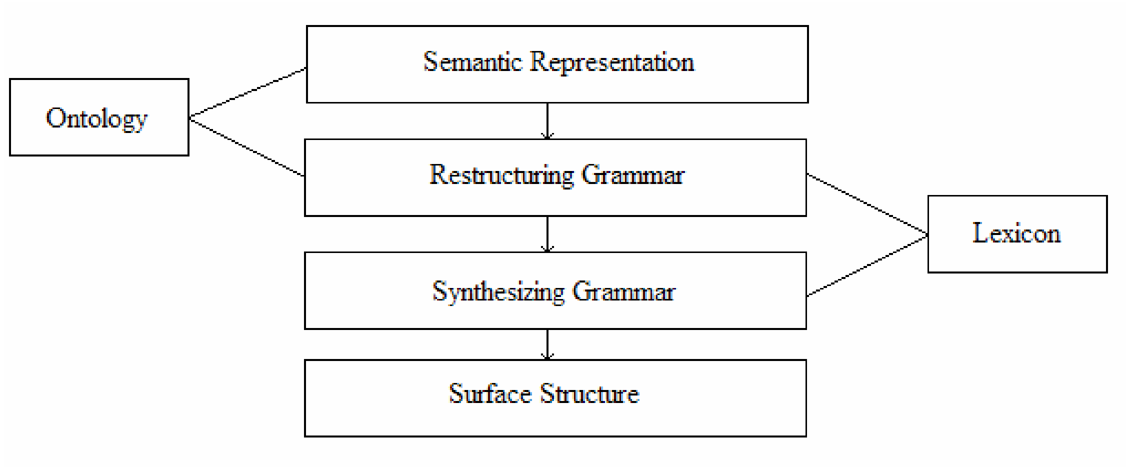


Figure 1. Underlying model of The Bible Translator's Assistant

1. The Semantic Representations

The development of an adequate method of meaning representation for TBTA's source texts proved to be a challenge. Formal semantics (Cann, 1993; Rosner, 1992), conceptual semantics (Jackendoff, 2002) and generative semantics (Lakoff, 1975) were each considered but found inadequate. Using the foundational principles of Natural Semantic Metalanguage theory, a set of semantically simple English molecules was identified in a principled manner (Wierzbicka, 1996; Goddard, 1998). These semantic molecules serve as the primary lexemes in TBTA's ontology. The ontology also includes semantically complex lexemes, but each of those lexemes has an associated expansion rule that automatically expands the complex concept in terms of the semantic molecules for those target languages that don't have a lexicalized semantic equivalent.

The feature set developed for TBTA encodes semantic, syntactic and discourse information. Each feature is an exhaustive etic list of the values pertinent to the world's languages. For example, each nominal is marked for Number, and the possible values are Singular, Dual, Trial, Quadrial and Plural. Each of these values is necessary because some languages morphologically distinguish all five of these categories. Examples of some of the features and their values are listed below in Tables 1 through 4.

Number	Singular, Dual, Trial, Quadrial, Plural
Participant Tracking	First Mention, Integration, Routine, Exiting, Offstage, Restaging, Generic, Interrogative, Frame Inferable
Polarity	Affirmative, Negative
Proximity	Near Speaker and Listener, Near Speaker, Near Listener, Remote within sight, Remote out of sight, Temporally Near, Temporally Remote, Contextually Near, Contextually Remote, Not Applicable
Person	First, Second, Third, First & Second, First & Third, Second & Third, First & Second & Third
Participant Status	Protagonist, Antagonist, Major Participant, Minor Participant, Major Prop, Minor Prop, Significant Location, Insignificant Location, Significant Time, Not Applicable

Table 1. Partial listing of the Features for Things (Nominals)

Time	Discourse, Present, Immediate Past, Earlier Today, Yesterday, 2 days ago, 3 days ago, a week ago, a month ago, a year ago, During Speaker's lifetime, Historic Past, Eternity Past, Unknown Past, Immediate Future, Later Today, Tomorrow, 2 days from now, 3 days from now, a week from now, a month from now, a year from now, Unknown Future, Timeless
Aspect	Discourse, Habitual, Imperfective, Progressive, Completive, Inceptive, Cessative, Continuative, Gnomic
Mood	Indicative, Definite Potential, Probable Potential, 'might' Potential, Unlikely Potential, Impossible Potential, 'must' Obligation, 'should' Obligation, 'should not' Obligation, Forbidden Obligation, 'may' (permissive)
Reflexivity	Not Applicable, Reflexive, Reciprocal
Polarity	Affirmative, Negative, Emphatic Affirmative, Emphatic Negative

Table 2. Partial listing of the Features for Events (Verbs)

Semantic Role	Participant, Patient, State, Source, Destination, Instrument, Addressee, Beneficiary, Not Applicable

Table 3. Partial listing of the Features for Thing Phrases (NPs)

Type	Independent, Coordinate Independent, Restrictive Thing Modifier, Descriptive Thing Modifier, Event Modifier, Participant, Patient, Attributive Patient
Illocutionary Force	Declarative, Imperative, Content Interrogative, Yes-No Interrogative
Topic NP	Participant, Patient, State, Source, Destination, Instrument, Beneficiary
Discourse Genre	Narrative, Expository, Hortatory, Procedural, Expressive, Descriptive, Epistolary, Dramatic Narrative, Dialog
Notional Structure Schema (Longacre, 1996)	Narrative-Exposition, Narrative-Inciting Incident, Narrative-Developing Conflict, Narrative-Climax, Narrative-Denouement, Narrative-Final Suspense, Narrative-Conclusion, Hortatory-Authority Establishment, Hortatory-Problem or Situation, etc.
Salience Band (Longacre, 1996)	Pivotal Storyline, Primary Storyline, Secondary Storyline, Script Predictable Actions, Backgrounded Actions, Flashback, Setting, Irrealis, Evaluation, Cohesive Material, Not Applicable
Direct Quote	Man to Woman, Woman to Man, Man to Man, Woman to Woman, Father to Child, Child to Father, Mother to Child, Child to Mother, Husband to Wife, Wife to Husband, Employer to hired Worker, Hired Worker to Employer, Teacher to Student, Student to Teacher, King to Man, Man to King, King to Woman, Woman to King, Queen to Man, Man to Queen, Queen to Woman, Woman to Queen, etc.

Table 4. Partial listing of the Features for Propositions

Because it's impossible to represent meaning in a completely language neutral way, it was decided that a subset of English sentence structures would be used. Taking all of the above into consideration, the semantic representation for the very simple sentence *John did not read those books* is shown below in Figure 2.

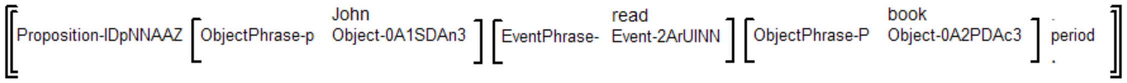


Figure 2. Semantic Representation of *John did not read those books.*

As seen in Figure 2, each lexeme has a set of features indicated by the numerals and letters immediately below it, each Object Phrase (NP) is marked for its semantic role, and the proposition is characterized by a set of features. The features associated with the event *read* in Figure 2 are expanded below in Figure 3.

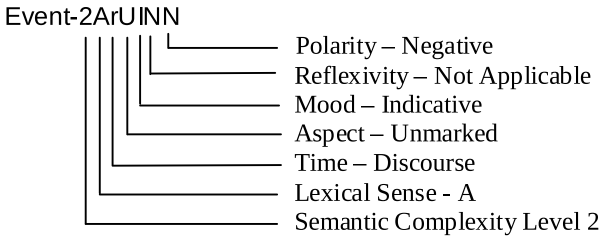


Figure 3. Expansion of Features associated with *read* shown in Figure 2

2. The Generator's Grammar

As was mentioned above, users of TBTA build a restructuring grammar and a synthesizing grammar for their target languages. The restructuring grammar restructures the semantic representations so that they contain the target language's structures, lexemes and features. The synthesizing grammar then synthesizes the final surface forms. The synthesizing grammar in TBTA has been designed to look as much as possible like the descriptive grammars that linguists routinely write. Therefore the synthesizing grammar includes phrase structure rules, constituent movement rules, clitic rules, spellout rules, morphophonemic rules, and feature copying rules. Figure 4 shows all of the types of rules in the synthesizing grammar and the sequence in which they're executed.

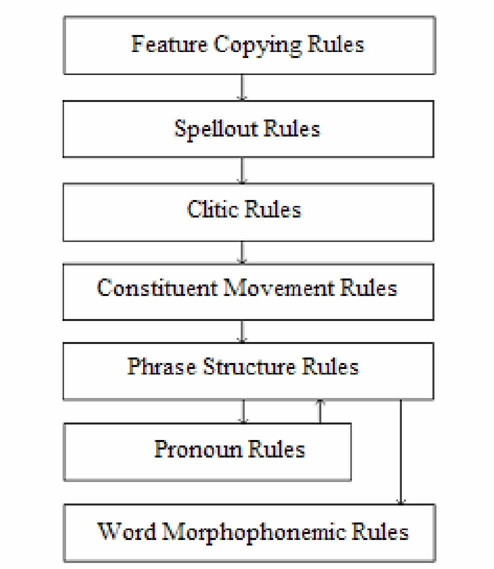
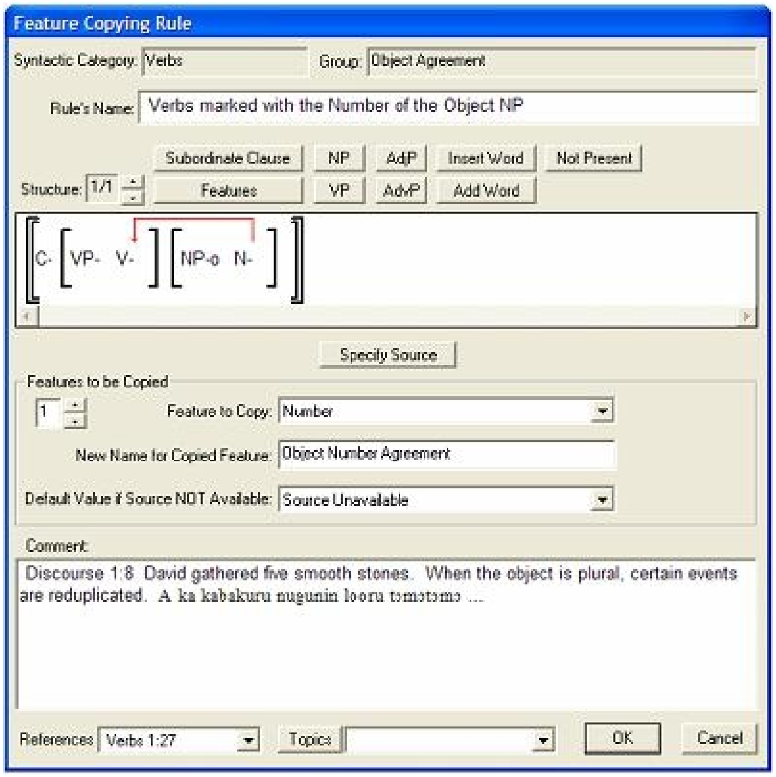


Figure 4. Overview of the Synthesizing Grammar in TBTA

Samples of some of these rules are shown below in Figures 5 through 7. Figure 5 shows a Feature Copying rule for Jula. Certain verbs in Jula are reduplicated when their objects are plural. Therefore a Feature Copying rule copies the number of the object nominals to the verb. If there are multiple object nominals, the system finds all of them and sums their number values (e.g., singular + singular = dual, singular + dual = trial, etc.).



Syntactic Category: Verbs

Group: Object Agreement

Rule's Name: Verbs marked with the Number of the Object NP

Structure: 1/1

Features to be Copied: 1

Feature to Copy: Number

New Name for Copied Feature: Object Number Agreement

Default Value if Source NOT Available: Source Unavailable

Comment:

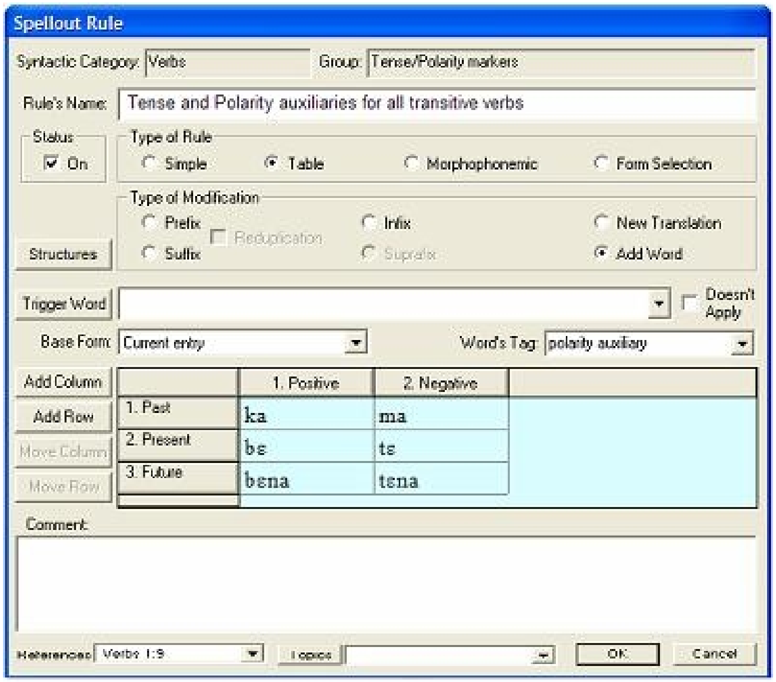
Discourse 1:8 David gathered five smooth stones. When the object is plural, certain events are reduplicated. A ka kabakuru nugunin looru tɔmɔtɔmɔ ...

References: Verbs 1:27

Topics:

Figure 5. Feature Copying rule for Jula

Figure 6 below shows a table spellout rule for Jula. All transitive verbs in Jula are marked with an auxiliary that indicates both tense and polarity. The table in this rule shows the six auxiliary verbs and their environments.



Syntactic Category: Verbs

Group: Tense/Polarity markers

Rule's Name: Tense and Polarity auxiliaries for all transitive verbs

Status: ☑ On

Type of Rule: ○ Simple ● Table ○ Morphophonemic ○ Form Selection

Type of Modification: ○ Prefix ○ Suffix □ Reduplication ○ Infix ○ Suprafix ○ New Translation ● Add Word

Trigger Word: □ Doesn't Apply

Base Form: Current entry

Word's Tag: polarity auxiliary

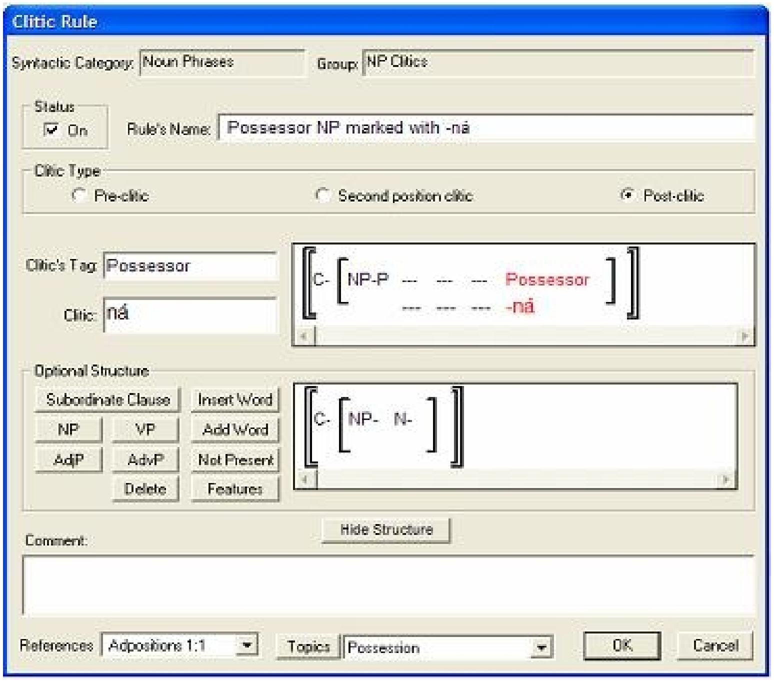
	1. Positive	2. Negative
1. Past	ka	ma
2. Present	bɛ	tɛ
3. Future	bɛna	tɛna

Comment:

References: Verbs 1:9

Topics:

Figure 6. Spellout Rule for Jula



Clitic Rule

Syntactic Category: Noun Phrases

Group: NP Clitics

Status: ☑ On

Rule's Name: Possessor NP marked with -ná

Clitic Type: ○ Pre-clitic ○ Second position clitic ● Post-clitic

Clitic's Tag: Possessor

Clitic: ná

Comment:

References: Adpositions 1:1

Topics: Possession

Figure 7. Clitic Rule for Kewa

Kewa marks many of its NPs with post-clitics which signal a variety of relationships. Figure 7 above shows a Clitic Rule for Kewa that inserts the post-clitic *–ná* which indicates possession.

3. Generating Target Text

As the linguist builds his lexicon and grammar, TBTA acquires knowledge of the target language and is able to generate target text; the more knowledge the linguist enters, the less assistance TBTA requires. Figures 8 and 9 shown below indicate that each subsequent chapter of text requires less effort by the linguist. Eventually TBTA acquires sufficient knowledge of the target language that it is able to generate drafts of all the analyzed source materials without any additional assistance from the linguist.

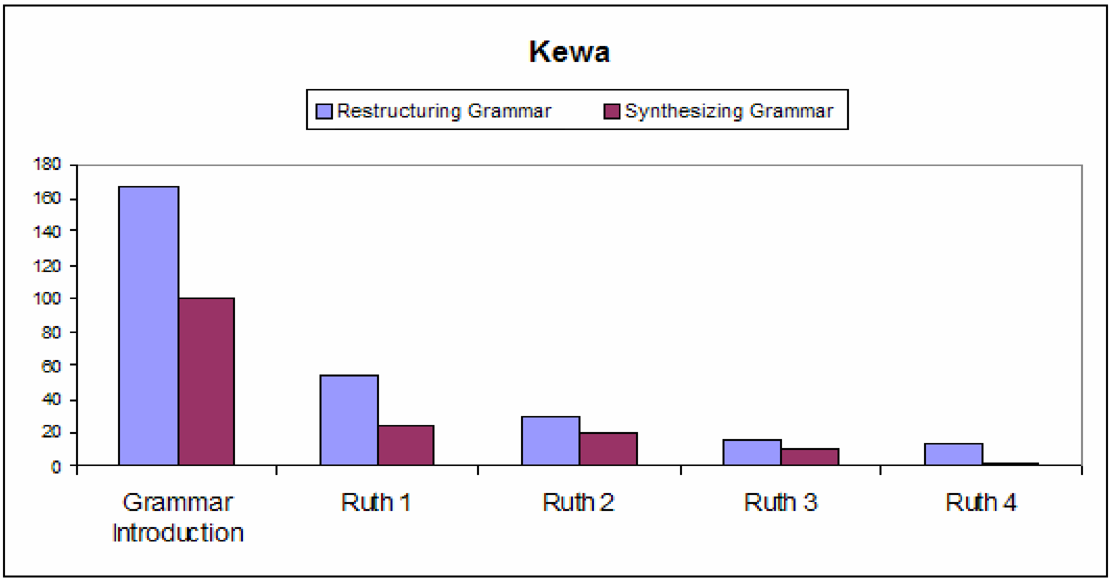


Figure 8. Number of new grammatical rules required for each chapter of Kewa text

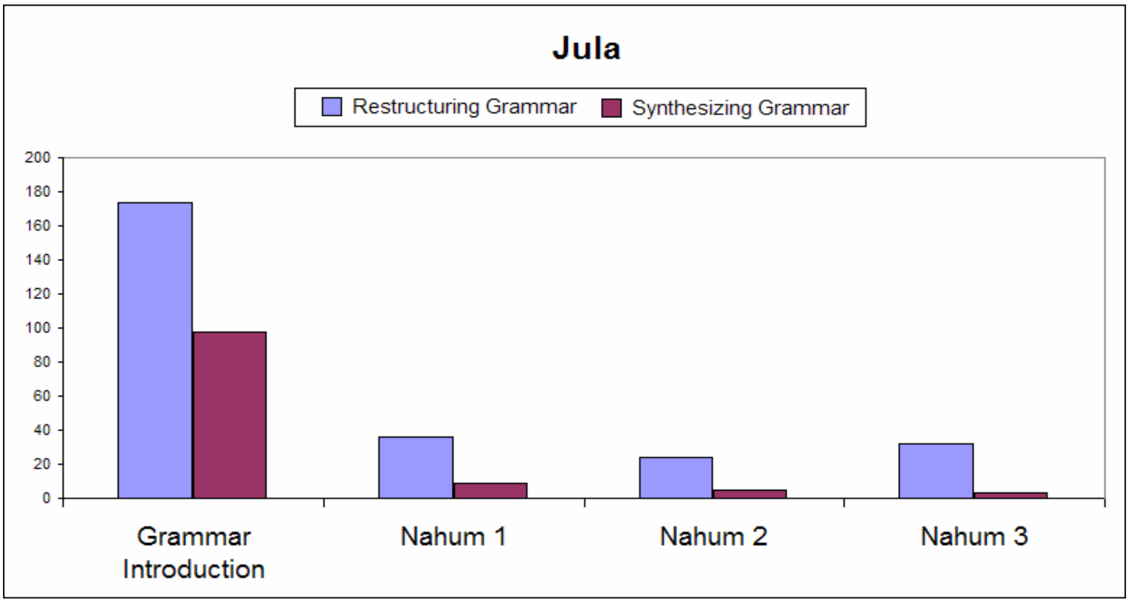


Figure 9. Number of new grammatical rules required for each chapter of Jula text

TBTA has been tested with four languages: English, Korean, Jula which is spoken in Cote d'Ivoire and Mali, and Kewa which is a clause chaining language with a switch reference system spoken in Papua New Guinea. In each of these four tests TBTA has produced text that is easily understandable, grammatically perfect and semantically equivalent to the source texts. However, the generated texts lack naturalness and need to be post-edited in order to produce presentable first drafts. Experiments with the Jula text indicate that TBTA triples the productivity of professional mother tongue translators without any loss of quality. Those experiments will be described in Section 4.

4. Evaluating the Generated Text

In order to determine whether or not the quality of the text generated by TBTA is sufficient so that it actually improves the productivity of a translator, several experiments were performed with the generated Jula text. As was shown above in Figure 9, a lexicon and grammar were developed for Jula so that TBTA could generate a draft of the biblical book of Nahum. Then eight professional mother tongue Jula translators in Mali were asked to participate in an experiment that was designed to determine the quality of the generated text. In particular, four of the translators were asked to edit the first half of the generated text and make it a presentable first draft. Then they were asked to manually translate the second half of Nahum from the French *La Bible en Français Courant*, again producing a presentable first draft. The other four translators were asked to perform the same two tasks, but they manually translated the first half of Nahum and then edited the second half of the generated text. All of the translators were told that they'd be timed during each of the two tasks. Table 5 below shows the results of this experiment. On average these eight professional mother tongue translators spent three times as much time translating as they did editing. These results were encouraging, but another experiment was considered necessary to determine whether or not the translators had actually done a thorough job of editing the generated text.

In the second experiment, the eight drafts of Nahum were evaluated by forty other Jula speakers in order to compare the quality of the two halves of each text. These other speakers had no idea how the texts had been produced or where the texts had come from. Each of the evaluators was given one text that consisted of two halves – one half had been manually translated and the other half had been generated by TBTA and then edited by the same translator. The evaluators were each asked just one question: Is the quality of either half significantly better than the quality of the other half, or are the two halves essentially equal in quality? The results of this experiment are also summarized below in Table 5.

Translator	Editing Time	Translating Time	Ratio	Evaluations
Translator #1	24 minutes	65 minutes	2.7:1	C1 - M1 - E3
Translator #2	51 minutes	89 minutes	1.7:1	C1 - M2 - E2
Translator #3	56 minutes	132 minutes	2.4:1	C4 - M1
Translator #4	40 minutes	150 minutes	3.8:1	C2 - M3
Translator #5	70 minutes	145 minutes	2.1:1	C1 - E4
Translator #6	52 minutes	120 minutes	2.3:1	E5
Translator #7	62 minutes	192 minutes	3.1:1	C2 - M1 - E2
Translator #8	20 minutes	296 minutes	14.8:1	C1 - M3 - E1

Table 5. Evaluating the Quality of the generated Jula text

Average translation time: 1189/8 = 149 minutes
Average editing time: 375/8 = 47 minutes
Ratio: 3.2:1

In the Evaluations column of Table 5, the numbers prefaced with a 'C' indicate the number of evaluators that chose the computer generated half as better, the numbers prefaced with an 'M' indicate the number of evaluators that considered the manually translated half to be better, and the numbers prefaced with an 'E' indicate the number of evaluators that said the two halves of the text were equal in quality. Considering all of the evaluations together, a total of twelve evaluators thought that the edited computer generated half was better, eleven evaluators chose the manually translated half as being better, and seventeen evaluators considered the two halves to be of equal quality. Therefore twenty-nine of the forty evaluators said that the halves that had been generated by TBTA and then manually edited were as good as or better than the halves that had been professionally translated. So this second experiment confirmed that the translators had done a thorough job of editing the generated text even though they had only spent a third as much time editing as translating. Therefore, in this particular case, TBTA tripled the productivity of professional mother tongue translators without any loss of quality.

5. Conclusions

TBTA is a tool that will help field linguists who are translating texts into a variety of languages. The information encoded in the semantic representations combined with the capabilities of the restructuring and synthesizing grammars enables this project to generate target text that is easily understandable, grammatically perfect, and semantically equivalent to the source texts. The generated texts lack naturalness, but this problem may be easily corrected with post-editing. Additional experiments are currently being performed to ascertain the quality of the generated texts in other languages. It is hoped that this project will help produce translations of many different documents into the world's minority languages.

6. References

Allman, T., Beale, S. (2004). An environment for quick ramp-up multi-lingual authoring. In *International Journal of Translation*, Vol. 16, No. 1.

Beale, S., Nirenburg, S., McShane, M., and Allman, T. (2005). Document Authoring the Bible for Minority Language Translation. In *Proceedings of MT-Summit.* Phuket, Thailand.

Cann, R. (1993). *Formal Semantics*, Cambridge: Cambridge University Press.

Goddard, C. (1998). *Semantic Analysis: A Practical Introduction*, New York: Oxford University Press.

Jackendoff, R. (2002) *Foundations of Language*, New York: Oxford University Press.

Lakoff, G. (1975). *Pragmatics in Natural Logic.* In Keenan, E. (ed.) (1975) *Formal Semantics of Natural Language*, Cambridge: Cambridge University Press.

Longacre, R. (1996). *The Grammar of Discourse*, New York: Plenum Press.

Rosner, M., Johnson, R. (1992). *Computational Linguistics and Formal Semantics*, Cambridge: Cambridge University Press.

Wierzbicka, A. (1996) *Semantics: Primes and Universals*, New York: Oxford University Press.

Machine Translation for Amharic: Where we are

Saba Amsalu[*] **& Sisay Fissaha Adafre**[†]

[*]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
Kiskerstrasse 6, Bielefeld
Tel: 0049 (0)521 1063519
{saba@uni-bielefeld.de}
[†]Informatics Institute, University of Amsterdam
Tel: 0031205257545
{sfissaha@science.uva.nl}

Abstract

We describe some of the efforts in research and resource acquisition towards developing an Amharic machine translation system. A brief account of the challenges of unavailability of text corpora and linguistic tools is presented. In the direction of research, our survey shows that only limited attempts to integrate Amharic into a unification based machine translation system and to extract lexical information from bilingual corpora which are the basis for machine translation have been made. In order to fill the gap, significant research on different aspect of the language needs to be done. In this paper we identify and describe some of the tasks that need to be given due attention both in terms of resource acquisition and developing tools. This way we hope to inspire ideas and more discussions on the issue.

1. Introduction

Machine translation (MT) systems have shown to provide significant economic gains in a number of areas. Now that notable progress has been made in the development of MT systems for economically important languages of the world, the focus has shifted recently to the design of methods for the rapid development of MT system for minority languages which have no or very little natural language processing resources. Amharic is one of such languages for which limited research has been carried out in the area of MT or NLP in general.

In this paper, we provide a short summary of some of the works done in the areas of Amharic natural language processing, which are essential to the development of machine translation system for Amharic. Sections 2. provides background information on Amharic language. Section 3. briefs the need and challenges for developing MT system and the resources and tools available. Sections 4. & 5. describe relevant research towards having an MT system and future directions in corpus acquisition developing relevant tools.

2. Amharic Language

Amharic is a Semitic language which is widely spoken in Ethiopia. It has 17.4 million native and around 4 million non-native speakers and a long history of service as a medium in education and government activities (Wikipedia, 2006). It has its own script (Fidel) that is borrowed from Geez, another Ethiopian Semitic language (Leslau, 1995). The script is a syllabary writing system where each character represents an open CV syllable.

Amharic has a complex morphology. Word formation involves prefixation, suffixation, infixation, reduplication, and Semitic stem interdigitation, among others. Like other Semitic languages, Amharic verbs and their derivations constitute a significant part of the lexicon. In Semitic languages, words, especially verbs, are best viewed as consisting of discontinuous morphemes that are combined in a non-concatenative manner. Syntactically, Amharic is an SOV language.

3. Available Data and Tools to support MT

The need for automatic means of processing Amharic language has recently been recognised (Bayou, 2000; Bayu, 2002; Alemayehu and Willett, 2002; Getachew, 2001; Fissaha, 2004a; Adafre, 2004b; Amsalu and Gibbon, 2005; Alemu, 2005). Most of these studies are limited in scope and are far from meeting the demands of modern NLP applications such as MT. Unavailability of text in machine readable form, the limited number of researchers working in the area and the economic limitations to support projects in resource acquisition and developing tools are some of the major problems facing the development of Amharic MT. Despite these short-comings, the recent efforts coupled with the increasing availability of information technology for Amharic resulted in useful resources.

Amharic Script

Advances have been made with respect to the representation and manipultation of Amharic script using computers. Several fonts and encoding schemes have been proposed for the script in the past, which resulted in the proliferation of non-standard software packages. Currently, there are unicode compatible fonts, and companies are updating their packages to this standard.

The creation of the Unicode standard for Ethiopic was an important step for computational linguists in that certain problems of incompatibility to operating systems and the need to transliterate text from different packages to a common representation in Latin script has been alleviated.

Text Corpora

Historically, Amharic has served as the national language of Ethiopia. Currently it is being used as the working language of the Federal Government and several local governments. This has resulted in a large quantity of publications which range from news articles, working documents in organisations and novels to religious manuscripts and teaching materials for elementary schools. Among these documents, news articles, the Bible, government documents such as constitutional papers, judiciary documents, forms of Banks and Insurance and other companies are available in bilingual versions; and in most cases in machine readable form. The rest are mostly monolingual print documents.

Bilingual Dictionaries

There are different print bilingual dictionaries (Amharic-English). Some of these dictionaries are also available in machine readable form (Aklilu, 1998; Leslau, 1996; Davidovic Mladen, 1992). There are also online dictionaries (Eliab, ; Cain,). Though these dictionaries may have limited coverage both in the lexical entries they contain and the lexical data categories for the entries, they can be useful for bootstrapping large scale dictionaries or other linguistic resources such as wordlists and thesauri.

Linguistic Tools

Studies in the languages of Ethiopia, particularly Amharic, have been mainly motivated by pure linguistic interest (Baye, 1986; Leslau, 1995; Bender and Hailu, 1978). Most of the works are of descriptive nature and focus mainly on the morphology of the language. Some of the rare attempts at formalising Amharic grammar have been carried out within the framework of early theories of transformational grammar and are limited in scope. Regarding the computational aspect, it is only recently that works have begun to apply some of the techniques of formal linguistics (Bayou, 2000; Bayu, 2002; Alemayehu and Willett, 2002; Getachew, 2001; Fissaha, 2004a; Adafre, 2004b; Amsalu and Gibbon, 2005). The results of most of these works are prototypes with limited scope. To the best of our knowledge, there are no wide coverage parts-of-speech taggers, morphological analysers or syntactic parsers for Amharic. Recognising these problems, some recommendations have been made to develop resources required for the development of these tools such as Amharic treebank (Alemu et al., 2003).

4. Related Research

With respect to MT, one of the early attempt to integrate Amharic into a unification based machine translation system is work done by (Fissaha and Haller, 2003a). This work provided formal description of Amharic language borrowing ideas from contemporary linguistic theories, and applied different natural language tools and techniques for solving some of the problems of Amharic languages. Primarily, Xerox finite state tools (XFST) are used for modelling Amharic morphology. Amharic syntax is described using a unification-based grammar formalism. A fragment of the grammar thus developed have been used to develop a prototype transfer-based Amharic-English machine translation system. Corpus-based methods such as collocation extraction, clustering and classification techniques were applied for lexical development (Adafre and Haller, 2003b). However, as with other related researches, this work was severely limited by the inadequacy of monolingual linguistic researches and resources available for Amharic at the time.

Recently, there is increasing interest in the extraction of bilingual lexicon from parallel corpora which we briefly summarise below.

4.1. Bilingual Lexicon Extraction

Development of bilingual lexicon constitutes an important and achieveable short term goal that contributes greatly to machine translation research activities for Amharic. Bilingual lexicon acquisition from Amharic-English parallel corpora, using the Bible as a data source have been given due attention recently (Atelach Alemu and Eriksson, 2004; Amsalu, 2006a; Amsalu and Gibbon, 2006; Amsalu, 2006b).

Atelach Alemu and Eriksson (2004) devised a method for identifying noun translation equivalents from Amharic-English bilingual corpus. They used statistical method with and without the use of an affix stripper for Amharic.

Several modules that work independently and claim to have reasonable degree of success have also been developed by Amsalu and Gibbon (2006):

1. Analysis of term distribution in text for 1:1 alignment
2. Analysis of context of terms for m:n alignment
3. Use of relatively fixed realization of keywords as anchor for alignment
4. Use of syntactic location for parsing Amharic verbs

These modules use the distributional properties of lexical items in parallel corpora and characteristics of syntactically fixed expressions. Promising results have been achieved by modules 1, 2 and 3 (Amsalu, 2006a; Amsalu, 2006b). Some preliminary results have also been obtained by module 4.

5. Future Work

We believe that significant work needs to be done at all levels of Amharic NLP in order to bring about meaningful change to the current status. Essentially, due attention needs to be given to aspects that we broadly categorise as corpus acquisition and developing linguistic tools. Subsequently, we forward tractable approaches which we also believe are applicable for languages in similar situation.

5.1. Corpus Acquisition

Large text corpora form the basis of many monolingual and multilingual research in natural language processing, ranging from developing multilingual lexicons to statistical machine translation systems. Apparently, collecting text corpora written in different languages constitutes an important prerequisite for these research activities. Some of the tasks that can be done in this respect are:

1. Exploit the web: Automatic or semi-automatic acquisition of available corpora from the Web is an easy way to obtain free data. Mainly, newspaper archives are available in large quantities.
2. Collaborative content development (Wiki): Amharic is one of the languages for which a free encyclopedia, i.e. Wikipedia, is being created by Wikimedia foundation (Wikipedia, 2006). Though the current content of Amharic Wikipedia is very small, it has a potential of enabling a rapid development of Amharic corpus; provided that adequate awareness is created among

Amharic speaking community. Any native Amharic speaker with a working knowledge of English can translate English Wikipedia pages into Amharic versions. Wikis provide a general framework in which people on the Web can collaboratively develop content. Recognising this fact, Yacob (2006) has created an Amharic Wiki site where people can share ideas and contribute resources.

3. Exploiting data from other sources: An integrated approach of using OCR system for scanning print documents (there are some attempts to develop OCR systems (Cowell and Hussain, 2003; Alemu, 2005)), gathering print versions of anything available from authors or publishers, and organising projects for manually encoding documents into electronic format.

4. Developing programs for data conversion: To have data of longer period of time, tools that convert text written in non-standard packages need to be developed. This surely is a necessary task and less expensive.

5.2. Developing Tools

A strategic approach of developing tools that enable fast production of machine translation systems is of utmost importance. We propose some useful strategies as follows:

1. Adopting tools developed for related languages: Amharic shares a number of common linguistic properties with Arabic and Hebrew for which active research is being carried out. Use of resources developed for these languages may speed up some of the efforts on Amharic NLP.

2. Machine learning approaches to language modelling: Exploring application of unsupervised machine learning methods for Amharic needs to be given due attention.

3. Using one language as a pivot: Focusing into translating text in one language, namely English, and to perform the translation from other languages through this language.

4. Building domain specific translation machines: A phase by phase approach of addressing a sublangauge would be much easier instead of trying to create a general purpose MT system at once.

5. Fast production of unidirectional MT system: Not much is there to translate from Amharic to English, but the reverse is a lot, so developing a unidirectional MT system that translates from English to Amharic would be practical.

6. References

Sisay Fissaha Adafre and Johann Haller. 2003b. Application of corpus-based techniques to amharic texts. In *MT Summit IX Workshop Machine Translation for Semitic Languages: Issues and Approaches*.

Sisay Fissaha Adafre. 2004b. Adding amharic to a unification-based machine translation system.

Amsalu Aklilu. 1998. *English-Amharic Dictionary*. Oxford University Press.

Nega Alemayehu and Peter Willett. 2002. Stemming of amharic words for information retrieva. *Literary and Linguistic computing*, 17(1):1–17.

Atelach Alemu, Lars Asker, and Gunnar Eriksson. 2003. An empirical approach to building an amharic treebank. In *Proceedings of 2nd Workshop on Treebanks and Linguistic Theories*, Vaxjo University, Sweden.

Worku Alemu. 2005. *Handwritten Amharic Character Recognition Applied to Bank Checks*. Ph.D. thesis, Dresden University of Technology.

Saba Amsalu and Dafydd Gibbon. 2005. Finite state morphology of amharic. In *Proceedings of the International Conference on Recent Advances n Natuaral language processing*, pages 47–51, Borovets, Bulgaria.

Saba Amsalu and Dafydd Gibbon. 2006. Methods of bilingual lexicon extraction from amharic-english parallel corpora. In *Proceedings of The 5th World Congress of African Linguistics*, Addis Ababa. to appear.

Saba Amsalu. 2006a. Data-driven amharic-english bilingual lexicon acquisition. In *Proceedings LREC2006*, Genoa, Italy.

Saba Amsalu. 2006b. Scaling up from word to phrasal alignments of amharic-english parallel corpora. Submitted.

Lars Asker Atelach Alemu and Gunnar Eriksson. 2004. Building an amharic lexicon from parallel texts. In *Proceedings of: First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, a workshop at LREC*, Lisbon.

Abiyot Bayou. 2000. Developing automatic word parser for amharic verbs and their derivation. Master's thesis, Addis Ababa University, Addis Ababa.

Tesfaye Bayu. 2002. Automatic morphological analyzer for amharic: An experiment involving unsupervised learning and autosegmental analysis approaches. Master's thesis, Addis Ababa University, Addis Ababa.

Matthew Cain. Online dictionary of the official language of ethiopia. `http://www.amharicdictionary.com/`.

John Cowell and Fiaz Hussain. 2003. Amharic character recognition using a fast signature based algorithm. In *Proceedings of Seventh International Conference on Information Visualization (IV'03)*, page 384, Vaxjo University, Sweden.

A. Zekaria Davidovic Mladen. 1992. *Amharic-English / English-Amharic Dictionary*. Hippocrene Books Inc.

Eliab. Online amharic-english dictionary. `http://www.ethiopiandictionary.com/`.

Sisay Fissaha and Johann Haller. 2003a. Amharic verb lexicon in the context of machine translation. *TALN*.

Sisay Fissaha. 2004a. Formal analysis of some aspect of amharic noun phrases. In *EAMT 2004 Workshop*, Malta.

Mesfin Getachew. 2001. Automatic part of speech tagging for amahric language: An experiment using stochastic hidden markov model (hmm) approach. Master's thesis, School of Graduate Studies of Addis Ababa University.

Wolf Leslau. 1996. *Concise Amharic Dictionary*. University of California Press.

Wikipedia. 2006. Amharic language. http://en.wikipedia.org/wiki/Amharic_language.

Daniel Yacob. 2006. Welcome to amharic nlp. http://nlp.amharic.org/.

Open-source machine translation between small languages: Catalan and Aranese Occitan

Carme Armentano i Oller [1], Mikel L. Forcada [1,2]

[1] Transducens Group, Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
[2] Prompsit Language Engineering,
Polígon Industrial de Canastell, Ctra. d'Agost, 77, office 3, E-03690 Sant Vicent del Raspeig (Spain)

Abstract

We describe the use of an open-source shallow-transfer machine translation engine, Apertium, and existing open-source linguistic data to build a bidirectional machine translation system for a new pair of 'small' languages, Catalan (6 million speakers) and the Aranese variety (5000 speakers) of Occitan (about 1 million speakers), and discuss its possible uses and their effects on the linguistic normalization of the smaller language.

1. Introduction

Language technologies are pervasive in society: the use of spell-checkers, translators, search engines, etc. is growing. Among these tools, machine translation systems are proving their utility, especially when it comes to translating large amounts of text between related languages, both to produce texts that will be published after post-editing and to allow users to understand documents written in a language that they cannot read fluently. In the case of languages having a small number of speakers, machine translation systems may indeed be very useful to generate texts in the less-spoken language and to help speakers of the majority language get closer to the cultural reality of the minority. This is particularly interesting in the current context, in which cultural and linguistic differences are being perceived as a richness instead of as a trouble.

Unfortunately, however, small languages are usually ignored by enterprises (they do not consider them economically interesting) and, in absence of clear support by public institutions, may remain without linguistic resources. It is in these cases indeed where it is most important that linguistic tools and data are made available to the community, because this makes it easier and cheaper to generate new tools and data. Also, one has to take into account that many languages which are far from a status of officiality or normality have activist groups with people having the necessary linguistic skills and who are willing to volunteer to create and improve these resources.

The Transducens group and Prompsit Language Engineering are currently developing open-source linguistic data for use with an open-source shallow-transfer machine translations system called Apertium for a pair of small related languages: Catalan and Aranese, a subdialect of Occitan.

1.1. Catalan

Catalan (a medium-sized Romance language having about 6 million speakers) is spoken mainly in Spain, where has been recognized as co-official in some regions, but is also the official language of Andorra, and is spoken in South-Eastern France and in the Sardinian city of l'Alguer (Alghero), Italy, where it is basically non-official, but there exist groups that struggle for its normality, especially groups asking for Catalan schooling of children. Publicly-funded Catalan schooling is possible since the eighties in the areas of Spain where it is official (Catalonia, Valencia and Balearic Islands).

1.2. Aranese Occitan

Occitan is spoken mainly in France but also in parts of Italy and Spain. This language, one of the main literary languages in Medieval Europe, and usually called Provençal after one of its main dialects, is reported to still have about a million speakers, but has almost no legal existence in France and Italy and a limited status of co-officiality in a small valley of the Pyrenees in Catalonia, inside the territory of Spain, called Val d'Aran. In addition, standardization of Occitan as a single language faces still a number of open issues. Aranese, a subdialect of Gascon (another of the main dialects of Occitan) is the variety spoken in this valley. According to the Linguistic Census published by the Catalan Statistical Institute IDESCAT, out of 7500 inhabitants in the Val d'Aran, 4700 people can speak it, 4400 can read it and 2000 can write it; the Government of Catalonia has adopted an orthographical standard for Aranese (Comission de Còdi Lingüistic 1999). For more information on Aranese, see Generalitat de Catalunya and Govern de les Illes Balears (2001). As to Occitan taken as a whole, there are groups, mainly in France, who want to increase the legal recognition of Occitan, with people prepared to build linguistic data and who could collaborate with us to adapt our translator to a more general variety of Occitan and even to generate a French—Occitan translator.

1.3. Uses of Aranese-Catalan machine translation

Machine translation between Catalan and Aranese may serve two important groups of uses. On the one hand, most official and educational documents published in Catalonia are in Catalan; therefore Catalan—Aranese machine translation would be an important asset in what could be called the "linguistic normalization" of Aranese in the Val d'Aran (it would be used to produce Aranese versions of these documents which would have to be post-edited). On the other hand, it could also be useful for non-Aranese Catalan speakers who want to read, for example, approximate translations of Aranese-only web documents.

2. The Apertium machine translation toolbox

2.1. What is Apertium?

A brief description of Apertium follows; more details can be found in Corbí-Bellot et al. (2005) and at the project webpage http://www.apertium.com. Apertium is a machine translation toolbox born as part of a large government-funded project involving universities and linguistic technology companies.

Apertium is based on an intuitive approach: to produce fast, reasonably intelligible and easily correctable translations between related languages, it suffices to use a MT strategy which uses shallow parsing techniques to refine word-for-word machine translation. Apertium uses finite-state transducers for lexical processing (powerful enough to treat many kinds of multi-word expressions), hidden Markov models (HMM) for part-of-speech tagging (solving categorial lexical ambiguity), and finite-state-based chunking for structural transfer (local structural processing based on simple and well-formulated rules for some simple structural transformations such as word reordering, number and gender agreement, etc.).

The Apertium machine translation toolbox, whose components have been released under open-source licenses such as the GNU General Public License[1] and one of the Creative Commons licenses[2], includes:

1.

- the open-source engine itself, a modular shallow-transfer machine translation engine largely based upon that of systems we have already developed, such as interNOSTRUM (Canals-Marote et al. 2001) for Spanish—Catalan and Traductor Universia (Garrido-Alenda et al. 2004) for Spanish—Portuguese,
- extensive documentation (including document type declarations) specifying the XML format of all linguistic (dictionaries, rules) and document format management files,
- compilers converting these data into the high-speed (tens of thousands of words a second) formats used by the engine, and
- pilot linguistic data for Spanish—Catalan and Spanish—Galician and format management specifications for the HTML, RTF and plain text formats.

In addition to these language pairs and to the translator presented in this paper, the Transducens group has also created linguistic data for the Spanish—Portuguese language pair.

2.2. Why apertium?

To build the Aranese-Catalan translator we have chosen the Apertium architecture because it offered several advantages: on the one hand, Apertium may be seen as an open-source rewriting and improvement of the machine translation architecture which was successfully used by Transducens to build Spanish-Catalan (http://www.interNOSTRUM.com, Canals-Marote et al. 2001) and Spanish-Portuguese (http://traductor.universia.net, Garrido-Alenda et al. 2004) machine translation systems; the results encouraged us to use the architecture for this new pair of related languages.

On the other hand, since the toolbox and the linguistic data has an open-source license, we have been able to take advantage of the whole architecture and the linguistic data for Catalan, and it has only been necessary to create monolingual data for Aranese and bilingual data for Catalan-Aranese and adapt those of Catalan.

We have also found the way in which linguistic data are managed in the Apertium toolbox very convenient. On the one hand, linguistic data are found in independent files. This makes it very easy to develop or improve or a translator, since it frees those people responsible of building and maintaining the linguistic information of worrying about programming details, and makes it easy to recycle linguistic data from other language pairs. On the other hand, all modules in Apertium have a rather straightforward linguistic motivation (and many bear names based on those of well-defined linguistic operations such as morphological analyzer from morphological analysis). As a result, building an Apertium machine translation system for a pair of languages means just building the required linguistic data for each module in well-defined, XML-based formats. As a result of this, and of the intuitive approach to machine translation used in Apertium, the amount of linguistic knowledge necessary about the source and target language to build data for Apertium is kept to a minimum, and it may be easily learned on top of basic high-school grammar skills such as: morphological analysis of words: parts-of-speech or lexical categories (noun, verb, preposition, etc.) and basic morphology (number, gender, case, person, etc.; agreement (such as gender and number agreement between nouns and their modifiers: adjectives, determiners, etc.); main local structural differences between the source and target language: position of adjectives with respect to nouns (e.g, adjective after noun in Spanish, before noun in English), prepositional regime, etc.

2.3. How does Apertium work?

The engine is a classical shallow-transfer or transformer system consisting of an eight-module assembly line:

2. The **de-formatter** separates the text to be translated from the format information (RTF, HTML, etc.). Format information is encapsulated so that the rest of the modules treat it as blanks between words.
3. The **morphological analyser** tokenizes the text in *surface forms* (lexical units as they appear in texts) and delivers, for each surface form, one or more lexical forms consisting of *lemma*, *lexical category* and

[1] http://www.gnu.org/licenses/gpl.txt

[2] The Creative Commons Attribution—NonCommercial-ShareAlike 2.5 license: http://creativecommons.org/licenses/by-nc-sa/2.5/

morphological inflection information. The system is capable of dealing with contractions and fixed-length multi-word lexical units (either invariable or inflected).

4. **Part-of-speech tagger:** a sizeable fraction of surface forms (for instance, about 30% in Romance languages) are homographs, that is, ambiguous forms for which the morphological analyser delivers more than one lexical form. The part-of-speech tagger chooses one of them, according to the lexical forms of neighbouring words. When translating between related languages, ambiguous surface forms are one of the main sources of errors when incorrectly solved. The part-of-speech tagger reads in a file containing a hidden Markov model (HMM) which has been trained on representative source-language texts (using an open-source training program in the toolbox). The behaviour of both the part-of-speech tagger and the training program are both controlled by a tagger definition file.
5. The **structural transfer module** uses finite-state pattern matching to detect (in the usual left-to-right, longest-match way) fixed-length patterns of lexical forms (*chunks* or *phrases*) needing special processing due to grammatical divergences between the two languages (gender and number changes to ensure agreement in the target language, word reorderings, lexical changes such as changes in prepositions, etc.) and performs the corresponding transformations.
6. The **lexical transfer module** is called by the structural transfer module; it reads each source-language lexical form and delivers a corresponding target-language lexical form. The dictionary contains a single equivalent for each source-language entry; that is, no word-sense disambiguation is performed. For some words, however, multi-word entries are used to safely select the correct equivalent in frequently-occurring fixed context.
7. The **morphological generator** delivers a target-language surface form for each target-language lexical form, by suitably inflecting it.
8. The **post-generator** performs orthographical operations such as contractions and insertion of apostrophes.
9. Finally, the **re-formatter** restores the format information encapsulated by the de-formatter into the translated text and removes the encapsulation sequences used to protect certain characters in the source text.

To ease diagnosis and independent testing, modules communicate between them using text streams. This allows for some of the modules to be used in isolation, independently of the rest of the MT system, for other natural-language processing tasks.

These linguistic data are dictionaries (two monolingual dictionaries, two post-generation dictionaries, a bilingual dictionary), two structural transfer rules that perform grammatical and other transformations between the two languages involved in each direction, and control data for each one of the part-of-speech taggers; these data are XML files, whose format is governed by document-type definitions (DTD). Details may be found in the documentation posted in the web http://www.apertium.org.

3. The Aranese—Catalan machine translation system

To build linguistic data for the Aranese-Catalan translator, we have used existing open-source Spanish-Catalan data (package apertium-es-ca in http://www.apertium.org). We have been able to use the Catalan post-generation dictionary and the control files for the Catalan part-of-speech tagger. The Catalan morphological dictionary has been used with small changes (such as entries specially designed for the Catalan-Spanish language pair). As to structural transfer rules, we have been able to reuse rules already present in other language pairs (for example, gender and number agreement in noun phrases), but new rules have also been written such as the one that translates Aranese "*en*+*tot*+<infinitive>" into Catalan "<gerund>" (for instance, *en tot cantar* by *cantant* "singing"). Only those data involving Aranese have been built from zero: the morphological dictionary, the post-generator, and the bilingual dictionary. The lexical categories of Catalan have been preseved for Aranese; this eased the design of our first Aranese part-of-speech tagger: we have been able to use temporarily that for Catalan, in view of the fact that our Aranese dictionaries were too small to build a training corpus. The current tagger, however, has already been trained in an unsupervised way (using the Baum-Welch algorithm on a small Aranese corpus.

Having Catalan data available has made it possible for us to build prototypes in a very short time. The main problem we have found is the relative scarcity of Aranese resources such as Aranese grammars, dictionaries, or text. We have used an Aranese course (Ané Brito et al. 1987), a web grammar (González i Planas 2003), the official orthographic norms for Aranese (Comission de Còdi Lingüistic 1999), a children's Catalan-Aranese vocabulary (Oficina de Foment de l'Aranés and Associació Punt d'Intercanvi 2004), verb conjugation tables (Frías and Rius 2006), and corpora, such as the Aranese supplement *Aué* of the Catalan daily *Avui* (many issues may be found at http://www.occitania.org/aueoccitania.asp), the Aranese documents in the Government of Catalonia web (http://www.gencat.net), etc.

3.1. Current status of the translator and immediate work

At the time of writing these lines, and after less than 2 person-months of work, we have produced an Aranese—Catalan prototype with dictionaries having 2500 lemmas (in addition to 1500 proper names), and 33 structural transfer rules. The results of a quick evaluation on a short Aranese text (a mixture of government and newspaper texts having 2525 source words, 2700 target words) is shown in the following table.

A brief evaluation of the Aranese—Catalan system

	Correct		***Incorrect***		***Total***	
Known	2290	84.8%	96	3.6%	2386	88.4%
Unknown	151	5.6%	163	6.0%	314	11.6%
Total	2441	90.4%	259	9.6%	2700	100%

As may be seen, the current coverage (total known words) is 88.4% and the total error rate is 9.6% (total incorrectly translated words); these figures take into account the fact that the system leaves unknown Aranese words (11.6%) untranslated, some of which (5.6%) happen to be correct in Catalan too.

By the time the workshop takes place, we expect to have increased the coverage of the Aranese-Catalan prototype (which would have a native Aranese part-of-speech tagger instead of the Catalan one now in use) above 90% as well as to have an equivalent Catalan-Aranese system obtained by inverting and adapting the former. The error rates for these systems may be expected to be in the range 5—10%.

4. Concluding remarks

Language technologies offer an excellent opportunity that small languages have to be able to take advantage of. Apertium, both because of being free and because of the way it treats linguistic data, is an adequate toolbox which permits the development of new MT systems in little time. As learning to manage the linguistic information is easy, nonspecialists may learn in a short time to add vocabulary and to make small modifications, so that, if it were necessary, dictionary growth could be made by volunteers who would find it to be a very motivating task in view of the possible consequences of the availability of such a system.

We have also seen that, the more linguistic data are available, the easier it is to develop new data. This is the reason why it is so important that linguistic data developed have licenses that make it possible to adapt them to create new resources.

Acknowledgements: This work has been partially funded by the Spanish government through grants TIC2003-08681-C02-01, FIT-340101-2004-3 and FIT-340001-2005-2. We thank Gema Ramírez-Sánchez, Felipe Sánchez-Martínez and Sergio Ortiz-Rojas for their help during the development of the prototype.

5. References

Ané Brito, Manuela; Ané Sanz, Jovita, Sans Socasau, Jusèp Loís (1987) *Curs d'aranés.* Vielha: Centre de Normalisacion Lingüistica der Aranés.

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., 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, September 12-16, 2005, Phuket, Thailand.

Canals-Marote, R. Esteve-Guillen, A. Garrido-Alenda, A. Guardiola-Savall, M., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Pérez-Antón, P.M., Forcada, M.L. (2001) "The Spanish-Catalan machine translation system interNOSTRUM". In Proceedings of MT Summit VIII: Machine Translation in the Information Age, Santiago de Compostela, Spain, 18--22 July 2001.

Comission de Còdi Lingüistic (1999) Normes ortografiques der aranés. Tèxte aprovat en plen deth Conselh Generau d'Aran, 5 d'octobre de 1999. Vielha: Conselh Generau d'Aran (available at http://www6.gencat.net/llengcat/aran/docs/normes.pdf).

Corbí-Bellot, A.M. Forcada, M.L., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor, A., Sarasola, K. (2005) "An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain". In Proceedings of the Tenth Conference of the European Associtation for Machine Translation, p. 79-86, May 30-31, 2005, Budapest, Hungary.

Frías, X., Rius, R. (2006) "Es vèrbs der aranés", available at http://www.angelfire.com/falcon/ramonrius/verb.htm

Garrido-Alenda, A., Gilabert-Zarco, P., Pérez-Ortiz, J.A., Pertusa-Ibáñez, A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M.A., Forcada, M.L. (2004) "Shallow parsing for Portuguese-Spanish machine translation." In Branco, A., Mendes, A., and Ribeiro, R., eds., *Language technology for Portuguese: shallow processing tools and resources*, pages 135--144. Lisboa, 2004.

Generalitat de Catalunya, Departament de Cultura and Govern de les Illes Balears, Conselleria d'Educació i Cultura (2001) "Aranese, the language of the Aran Valley", in Catalan, Language of Europe, p. 26-27. Available at:http://membres.lycos.fr/aranes/ar-ang.pdf and http://www6.gencat.net/llengcat/publicacions/cle/docs/ecle9.pdf

González i Planas, Francesc (2003) Breu gramàtica aranesa (available at http://www.cesdonbosco.com/filologia/iberia/aranesa.htm and http://membres.lycos.fr/aranes/gramatica.pdf

Oficina de Foment de l'aranés and Asociació Punt d'Intercanvi (2004) "Català-Aranès: diccionari infantil", available at http://www.edu365.com/primaria/muds/aranes/dic/

Compilation and Structuring of a Spanish-Basque Parallel Corpus

A. Casillas *, **A. Díaz de Illarraza** *, **J. Igartua** *, **R. Martínez** †, **K. Sarasola** *

* Dpto. Electricidad y Electrónica & IXA Taldea
Euskal Herriko Unibertsitatea, Universidad del País Vasco (UPV-EHU)
{arantza.casillas, jipdisaa, webigigj, ksarasola}@ehu.es
† Dpto. Lenguajes y Sistemas Informticos
UNED - E.T.S.I. Informtica
raquel@lsi.uned.es

Abstract

In this paper we explain how we have compiled a Spanish-Basque parallel corpus. We also propose a corpus structure for containing: (1) translation units and the linguistic information for each unit, and (2) the whole documents with their linguistic information. The proposed corpus structure may be seen as composed of several XML documents and is based on stand off annotation model. This structure permits to work with the corpus from two points of view: as a annotated corpus with linguistic information, as well as a translation memory.

1. Introduction

There are two official languages in the Spanish side of the Basque Country: Basque (or Euskara) and Spanish. The latter is the third most spoken language of the world, and the former, Basque, is a minority language spoken in northern Spain and south-western France. There are 700,000 Basque speakers, and these comprise about 25% of the total population of the Basque Country - but they are not evenly distributed. There are six dialects, but since 1968 the Academy of the Basque Language has been involved in a standardization process. At present, morphology, which is very rich, is completely standardized, but the lexical standardization is still in progress. Most of the main public institutions such as Basque Government or universities try to publish official documents in the two official languages. In most of the cases, these type of documents are first written in Spanish and then translated manually into Basque.

A bilingual compiled corpus can be a helpful tool for different purposes: serving as training datasets for inductive programs; it can be used to learn models for machine translation, cross-lingual information retrieval; it could also be useful for automatic descriptor assignment, document classification, cross-lingual document similarity and other linguistic applications. This means that a Spanish-Basque corpus would be a very valuable resource for the research community. Once the corpus is compiled, it is possible to find different language resources inside it; for example translation memories or groups of classified documents. But nowadays the compiled Spanish-Basque corpus is poor, so there is not enough reference corpus to consult.

In this paper we explain how we have collected the bilingual corpus. We also propose a bilingual corpus structure that contains two types of informations: (1) translation units with their corresponding linguistic information, and (2) the whole documents with their linguistic information. We propose so a rich structure because our corpus resources are poor and we want them to be general and useful for different tasks in language technology research.

Similar works on compiling and representing bilingual corpus are: (Erjavec 2002), (Erjavec et. al. 2005) and (Tadie 2000). In all these three works one of the involved languages, at least, is a minority language. In (Tadie 2000) are presented procedures and formats used in building a newspaper bilingual corpus for Croatian-English. The author compares the two different ways to encode parallel corpus using XML: alignment by storing pointers in separate documents and translation memory (TMX) inspired encoding. One of the paper conclusions is to use the former due to the DTD's simplicity, because the original document keeps more unchanged, and because even with the stand-off way there is no problem to keep aligned sentences together in the same element while retaining upper levels of text encoding. In (Erjavec 2002) is used also a stand-off representation for bilingual corpus, so that linguistic information is in other separate files, that is, it is not included within the text. The authors in (Erjavec et. al. 2005) explain the compilation of massively multilingual corpora, the EU ACQUIS corpus, and the corpus annotation tool ”totale”. The EU ACQUIS corpus contents 8 to 82 million running words depending on the language. It contains EU law texts in all the languages of the current EU, and more, i.e. parallel texts in over twenty different languages. Unfortunately, we can not use Europarl (Koehn 2006) for Basque, the most useful corpus nowadays for research in MT.

Next section explains the characteristics of the bilingual corpus collected and the steps carried out to compile it. In Section 3 the structure of the bilingual corpus, which includes translation units and linguistic information, is explained. Finally, conclusions and future work are included.

2. Corpus Compilation

We have compiled a bilingual parallel corpus of 3 million words. This corpus is composed of two types of documents: official (about 2 million words) and not official documents (about 1 million words). The official documents are from local governments and from the University of the Basque Country. Mainly, they are edits, bulletins, letters or announcements. We have also collected some books, not official documents, that have been translated into Basque by this public university and they are about various subjects: fossils, music, education, etc.

Starting from the original plain text we are successively enriching the information contained in this corpus. The process consists of the following steps:

1. Obtaining the texts: we have downloaded the government official publications from (EHAA), in addition we have collected the available documents from the university. Actually, we continue with this collecting work and every day we download official publications from different websites. We also have got in touch with the editors of the public university to get more publications of this type.

2. Normalization of the texts into a common format: we have processed manually all the official publications because the documents were incomplete or there were some mistakes. On the contrary, there was no need of pre-processing the books. In both cases we have converted and saved all the documents into ASCII format.

3. Tokenization: involves linguistic analysis for the isolation of words.

4. Segmentation: to determine the boundaries of different types of units such as: paragraphs, sentences and entities (person, location, organization). Due to the differences between Spanish and Basque it was necessary to execute particular algorithms for each language in the detection process.

5. Alignment: the units detected in both languages were aligned. With the alignment process we have related the Spanish and Basque units of the same type that have the same meaning. Nevertheless, the alignment algorithms are independent of the language pair. The algorithms that we have executed to detect and align the different units are explained in more detail in (Martínez et al. 1998a) and (Martínez et al. 1998b).

6. Lematization and morpho-syntactic analysis: to know the lemma, number, gender and case of each word. FreeLing package (FreeLing) has been used for generating Spanish linguistic information. In the case of Basque, we have used a set of different linguistic processing tools. The parsing process starts with the outcome of the morphosyntactic analyzer MORFEUS (Aduriz et al., 2001). It deals with all the lexical units of a text, both simple words and multiword units, using the lexical database for Basque EDBL (Aldezabal et. al. 2001). This morphosyntactic analysis is an important step in our analysis process due to the agglutinative character of Basque. From the obtained results, grammatical categories and lemmas are disambiguated. The disambiguation process is carried out by means of linguistic rules (CG grammar) and stochastic rules based on Markovian models (Ezeiza et. al. 1998) with the aim of reduce the set of parsing tags for each word taking into account its context. Once morphosyntactic disambiguation has been performed, we have morphosyntactically fully disambiguated text. By the moment this is the deepest level we use to represent linguistic information in bilingual corpus, but we preview the inclusion of information about chunks, phrases and syntactic functions, in the same way we are doing for Basque monolingual corpora.

3. Bilingual Corpus Structure

The two main features that characterize the corpus structure are: (1) the richness of the linguistic information represented, and (2) the inclusion of relationships between units of the two languages which have the same meaning. The corpus structure proposed is based on the data model presented in (Artola et. al. 2005), which represents and manages monolingual corpus with linguistic annotations based on a stand off annotation and a typed feature structure. This representation may be seen as composed of several XML documents. Figure 1 shows the currently implemented document model for the bilingual corpus which includes: linguistic information, translation units (paragraphs, sentences and entities) and alignment relations. Next in this section, we will present the XML documents that constitute the proposed data model indicating their content.

With the corpus we have carried out two different processes: (1) detecting and aligning translation units, and (2) adding linguistic information to each subcorpus. With the proposed corpus structure, we have merged the output information of both processes. The final structure of the corpus is composed of the manuscript texts and of several files to define stand off annotations; these annotations contain the linguistic information and the delimitation of the units detected and aligned. The information to be exchanged among the different tools to manage this corpus is complex and diverse. Because of this complexity, we decided to use Feature Structures (FSs) to represent this information (Artola et. al. 2005). Feature structures are coded following the TEIs DTD for FSs (Sperberg-McQueen et al. 1994), and Feature Structure Definition descriptions (FSD) have been thoroughly defined for each document created. The documents created as input and output of the different tools are coded in XML. The use of XML for encoding the information flowing between programs forces us to describe each document in a formal way, with the advantages it offers to keep coherence, reliability and maintenance. This structure avoids unnecessary redundancies in the representation of linguistic features of repeated units.

The annotations which contain the linguistic information are saved into four XML documents:

- *eus.w.xml* and *cas.w.xml*: they contain single-word tokens in Basque and Spanish respectively.

- *eus.lem.xml* and *cas.lem.xml*: they keep for each single-word token of the two languages: its lemma, its syntactic function and some significant features of the morphological analysis. Words can be ambiguous and correspond to more than one lemma or syntactic function.

In order to represent the annotations that delimit translation units we have created six XML documents:

- *eus.par.xml* and *cas.par.xml*: these two documents are used to delimit the paragraphs detected in the bitext. Paragraphs are delimited with references to their first single-word token and their last single-word token.

- *eus.sen.xml* and *cas.sen.xml*: they contain the sentences of the parallel corpus by means of references to their first and last single-word token.

- *eus.nen.xml* and *cas.nen.xml*: they keep the name entities.

We have also created XML documents that relate units of the two languages with the same meaning:

- *alpar.xml*: this document is used to relate the paragraphs delimited in the files *cas.par.xml* and *eus.par.xml*. Each paragraph in one language is related with its corresponding paragraph (or paragraphs) in the other language, using the paragraph identifiers.

- *alsen.xml*: in this document are saved the relations between corresponding sentences from both languages. It is possible to set up 1-1 or N-M alignments.

- *alnen.xml*: name entities are aligned by means of this document. Relations of 1-1 and N-M are contemplated.

While translation memories take translation units as their primary "corpus", the corpus structure proposed contains the whole documents and the translations units detected and aligned. In the case of pure translation memories, only the units are saved, that is, the source text, the context from which the units come from, does not exist.

4. Conclusion and Future Work

In this paper we have explained how we have compiled a Spanish-Basque parallel corpus, the resultant language resources and its structure. The proposed structure supports linguistic information of the texts, as well as information of the alignment of the detected translation units.

The information contained in the resultant XML files is: (1) the whole document, (2) the linguistic information for each word, and (3) relations between translation units of both languages. This means that we have obtained mainly two resources: a translation memory and a morpho-syntactic tagged parallel corpus.

The main disadvantage of our proposal is that it needs more space than a translation memory or than a tagged corpus. Nevertheless, we think this representation will ensure the use of this "small" corpus in different tasks in language technology research. The compiled corpus, taking into account its structure, can be used as a translation memory for the automatic translation process or can be employed as a tagged parallel corpus for research in corpora based machine translation, machine learning, document clustering, cross-lingual information retrieval and other language applications.

Instead of repeating the same processing of the texts once and again for so different research lines, our representation makes easier and more efficient the use of parallel corpus, adding to the corpus structure to keep coherence, reliability and maintenance. Indeed, the work done so far confirms the scalability of our approach.

In the future we preview the inclusion of a new level of alignment at phrase or chunk level. We also plan to extend the graphical web interface EULIA (Artola et. al. 2004) for creating, browsing and editing also parallel corpora.

5. References

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernndez G. and Lersundi M. "EDBL: a general lexical basis for the automatic processing of Basque". *IRCS Workshop on linguistic databases*, 2001.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., Artola X., Daz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., Urkia M. "A Framework for the Automatic Processing of Basque". *Proceedings of the First International Conference on Language Resources and Evaluation*, 1998.

Artola X., Díaz de Illarraza A., Ezeiza N.,Gojenola K., Labaka G., Salogaistoa A., Soroa A. "A framework for representing and managing linguistic annotations based on typed feature structures". *RANLP*, 2005.

Artola X., Daz de Ilarraza A., Ezeiza N., Gojenola K., Sologaistoa A., Soroa A. "EULIA: a graphical web interface for creating, browsing and editing linguistically annotated corpora". *LREC 2004, Lisbon*,2004.

"Euskal Herriko Agintaritzaren Ofiziala (EHAA)". *http://www.euskadi.net.*

Erjavec Tomaz: "Compiling and using the IJS-ELAN Parallel Corpus". *Informatica, 26* , 299-307,2002.

Erjavec T., Pouliquen C., Steinverger B., "Massive multilingual corpus compilation: Acquis Communautaire and totale". *Proceedings of the 2nd Language & Technology Conference*, 32-36, 2005.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. "Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. *Proceedings of COLING-ACL'98*, 1998.

"FreeLing 1.2 An Open Source Suite of Language Analyzers" *http://garraf.epsevg.upc.es/freeling/.*

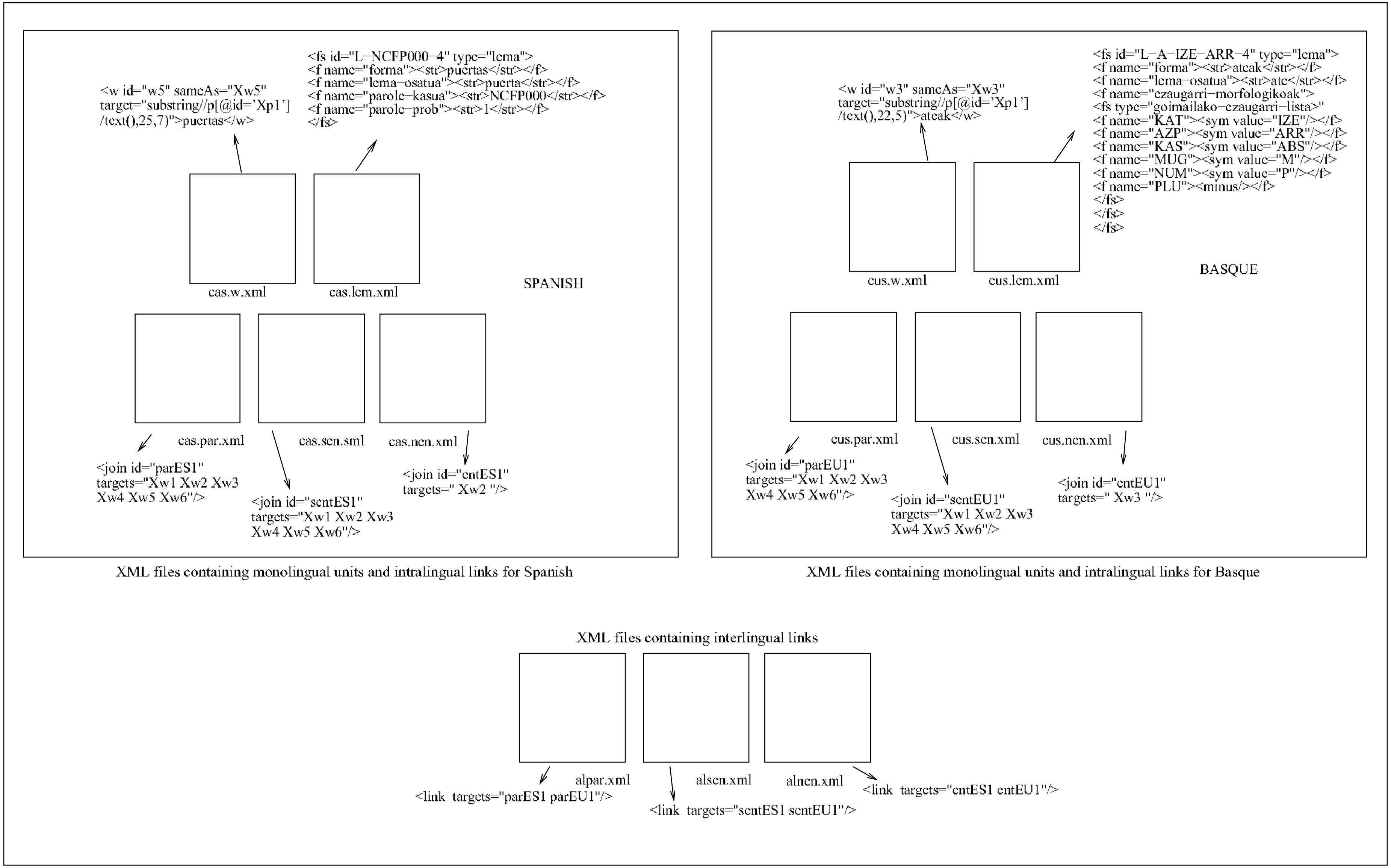
Koehn Philipp. "Europarl: A Multilingual Corpus for Evaluation of Machine Translation" *http://people.csail.mit.edu/koehn/publications/europarl/.*

Martínez R., Abaitua J., Casillas A. "Bitext Correspondences through Rich Mark-up". *Proceedings of the* 17^{th} *International Conference on Computational Linguistics (COLING'98) and 36th Annual Meeting of the Association for Computational Linguistics (ACL'98)*, 812-818, 1997.

Martínez R., Abaitua J., Casillas A. "Aligning tagged bitext". *Proceedings of the Sixth Workshop on Very Large Corpora*, 102-109, 1998.

MarSperberg-McQueen C.M., Burnard L. "Guidelines for Electronic Text Encoding and Interchange". *TEI P3 Text Encoding Initiative*, 1994.

Tadié Marko. "Building the Croatian-English Parallel Corpus". *LREC*, 2000.


SPANISH
cas.w.xml
cas.lem.xml
cas.par.xml
cas.sen.sml
cas.nen.xml
XML files containing monolingual units and intralingual links for Spanish
BASQUE
cus.w.xml
cus.lem.xml
cus.par.xml
cus.sen.xml
cus.nen.xml
XML files containing monolingual units and intralingual links for Basque
XML files containing interlingual links
alpar.xml
alsen.xml
alnen.xml

```
<w id="w5" sameAs="Xw5"
target="substring//p[@id='Xp1']
/text(),25,7)">puertas</w>
```

```
<fs id="L−NCFP000−4" type="lema">
<f name="forma"><str>puertas</str></f>
<f name="lema−osatua"><str>puerta</str></f>
<f name="parole−kasua"><str>NCFP000</str></f>
<f name="parole−prob"><str>1</str></f>
</fs>
```

```
<join id="parES1"
targets="Xw1 Xw2 Xw3
Xw4 Xw5 Xw6"/>
```

```
<join id="sentES1"
targets="Xw1 Xw2 Xw3
Xw4 Xw5 Xw6"/>
```

```
<join id="entES1"
targets=" Xw2 "/>
```

```
<w id="w3" sameAs="Xw3"
target="substring//p[@id='Xp1']
/text(),22,5)">ateak</w>
```

```
<fs id="L−A−IZE−ARR−4" type="lema">
<f name="forma"><str>ateak</str></f>
<f name="lema−osatua"><str>ate</str></f>
<f name="ezaugarri−morfologikoak">
<fs type="goimailako−ezaugarri−lista>"
<f name="KAT"><sym value="IZE"/></f>
<f name="AZP"><sym value="ARR"/></f>
<f name="KAS"><sym value="ABS"/></f>
<f name="MUG"><sym value="M"/></f>
<f name="NUM"><sym value="P"/></f>
<f name="PLU"><minus/></f>
</fs>
</fs>
</fs>
```

```
<join id="parEU1"
targets="Xw1 Xw2 Xw3
Xw4 Xw5 Xw6"/>
```

```
<join id="sentEU1"
targets="Xw1 Xw2 Xw3
Xw4 Xw5 Xw6"/>
```

```
<join id="entEU1"
targets=" Xw3 "/>
```

```
<link  targets="parES1 parEU1"/>
```

```
<link  targets="sentES1 sentEU1"/>
```

```
<link  targets="entES1 entEU1"/>
```

Figure 1: Corpus Structure: XML documents and their contents

Exploiting Similarity in the MT into a Minority Language

Boštjan Dvořák*, **Petr Homola**†, Vladislav Kuboň†

*Zentrum für allgemeine Sprachwissenschaft
Jägerstraße 10–11
10117 Berlin, Germany
dvorak@zas.gwz-berlin.de
† Institute of Formal and Applied Linguistics
Malostranské náměstí 25
110 00 Praha 1, Czech republic
{homola,vk}@ufal.mff.cuni.cz

Abstract

The paper presents a machine translation system from Czech to related languages, including 'small' (minority) languages Lower Sorbian and Macedonian. Lower Sorbian, a West Slavonic language, which is spoken by less than 20,000 people, preserves many acient features, it has supine, dual and some other grammatical forms which disappeared in most Slavonic languages. Macedonian is interesting for us in that its typology is quite different from other Slavonic languages (except Bulgarian). The paper presents the most important problems encountered during the implementation of the MT system *Česílko*.

1. Introduction

The problem of preservation of the cultural heritage of the mankind is very closely related to the problem of preservation of minority languages. A complete extinction of a language usually also to a large extent means the loss of most of the cultural heritage of the people speaking that language. The modern era has brought both positive and negative aspects into this problem. Among the negative ones is definitely the globalization bringing a very strong stress on unification, including the unification of languages. The more globalized our world becomes, the higher the number of people capable of expressing themselves in one or more of the "big" languages. This is of course also a positive factor decreasing the number of mutual misunderstandings among the people belonging to different nations. The negative effect is the decreasing number of people capable of speking or reading their own (minority) native language.

The positive effects of the modern world are the scientific and technological achievements which enable to preserve to some extent even the extinct languages in the form of various kinds of corpora of spoken or written language. This is of course the most obvious effect of natural language technology, but not the only one. In this paper we would like to show how a very simple machine translation system can help both to preserve the cultural heritage of a minority language by translating into some "bigger" language and to increase the number of translations in the opposite direction thus increasing the number of texts available in the particular minority language.

Democratic governments usually care about the minority languages, but in many cases a minority language is at the same time much less similar to a majority language than to a language spoken in a neighboring country. The MT from the similar language definitely can complement all other efforts preserving the minority language. This paper describes a particular case of a Lower Sorbian, a minority language spoken in Germany in the area around Cottbus. Lower Sorbian (and its neighbor Upper Sorbian) are Slavonic languages, typologically very different from the majority language, German, but very closely related to the languages spoken in a close geographical proximity, to Czech and Polish. Our paper presents an enrichment of an existing multilingual MT system to exploit the proximity of related languages for Lower Sorbian. The system is not new — it already exists for several language pairs (Czech-Slovak, Czech-Polish, Czech-Lithuanian), cf. (Hajič et al., 2003), and this paper describes the modifications of the original system allowing to insert a new module.

2. Česílko — a multilingual MT system for related languages

The system Česílko has been developed as an answer to a growing need of translation and localization from one source language to many target languages. It is quite clear that the independent translation or localization of the same document into several typologically similar target languages is a waste of effort and money. Our solution proposes to use one language from the target group as a pivot and to perform the translation through this language. It is quite true that applying the pivot language approach has a serious drawback — the translation quality, which needs to be very high, may deteriorate in this two-step process. A negligible shift of the meaning during the translation into a pivot language may be amplified by a subsequent translation from the pivot language to the actual target language. We focus on the 'second step' in the paper.

In order to overcome these problems we have suggested an approach combining the human-made translation

from the source language into a pivot language with a machine translation between a pivot and a (closely related) target language. The reviewer of the target language text may then review the translation against the original source language text and he thus can eliminate any problem caused by the translation from the source into the pivot language.
The system consists of the following steps:

1. Morphological analysis of Czech
2. Morphological disambiguation of Czech by means of a stochastic tagger
3. Search in the domain-related bilingual glossaries
4. Search in the general bilingual dictionary
5. Morphological synthesis of the target language

The need to account for phenomena which cannot be handled by this very simple architecture led us to the inclusion of a additional modules: a shallow parsing module for Czech for some of the language pairs which directly follows the morphological disambiguation of Czech (Homola and Rimkutė, 2003), and a named entity (NE) recognizion module (Homola and Piskorski, 2004).

3. Extending the system to new target languages

In this section, we describe the most significant problems encountered while adapting the system to new target languages — Lower Sorbian and Macedonian. The adaptation itself included, first of all, the creation of a bilingual glossary (with Czech as source language) and the implementation of syntactic and morphological synthesis. The most interesting part has been the modification of transfer. There are about twenty quite complex transfer rules that had to be rewritten according to the grammar of the target language to guarantee correct phrase structure and constituent order.

3.1. MT problems encountered with Lower Sorbian

Sorbian is a West Slavonic minority language spoken in Lusatia in Germany. It splits into many dialects which differ significantly from each other. Two written standards are used in the present, Upper Sorbian in Saxonia and Lower Sorbian in Brandenburgia. We have chosen Lower Sorbian for our experiments, mainly because there exists a morphological tool capable of generating inflected forms from many lemmas obtained as a result of the translation process.[1]
Both morphology and syntax of Lower Sorbian are very similar to Czech, nevertheless the grammar of Lower Sorbian is more complicated than the Czech one since Lower Sorbian is more conservative.[2] In the following text we describe some aspects of Lower Sorbian which are important with respect to MT from Czech.

- Lower Sorbian has *dual*, a special number used instead of plural for the amount 2, e.g. *dub* (1), *duba* (2), *duby* (more than 2) "oak(s)". We ignore this number because the number of persons or objects can only be decided with a proper understanding of the context. This may result in an translation error although the sentence as such is grammatical, but such a strategy is unavoidable if we want to keep the whole system as simple as possible.
- The *supine* is another grammatical form which is not present in Czech. It is an infinite verb form used to express a goal or decisions, usually together with a verb of movement, e.g., *źi spat* "go to sleep" (cf. the infinitive form *spaś*).
- The system of tenses is richer in Lower Sorbian. Whereas Czech only uses one periphrastic past form, Lower Sorbian also has synthetic past forms, *aorist* and *imperfect*. Nevertheless these forms are rarely used in contemporary texts, i.e., one can use the periphrastic form to translate past tense.
- Lower Sorbian does not drop the auxiliary verb *byś* in the third person of the past form (cf. Czech *převzala* "took over" vs. Lower Sorbian *jo pśiwzeła*). We ignore this difference in the current version of the system, since the participle forms are the same for all persons, therefore the shallow parser does not deliver the information about the person at all (in a full parse, the subject would contain the missing information; nevertheless the subject can be dropped as well, so the person may remain underspecified).
- The passive is constructed differently in some cases. There is the specific *bu*-pattern (e.g., *dom bu natwarjony* "the house has been built") and the colloquial *wordowaś* (e.g., *dom worduje twarjony* "the house is being built"), whereas Czech only has one equivalent with *být* "to be". Moreover, the reflexive passive is used more often (e.g., *drjewo se wót nana rubjo* "the tree is being cut by the father"). We use the reflexive pattern if there are more possibilities. The agent is expressed by a prepositional phrase with *wót* "from", whereas Czech uses the instrumental case.

One of the important things which really may substantially decrease the quality of output provided by our system is the word order. Due to the typological similarity of both languages and the fact that both Czech and Lower Sorbian use the word order to express topic-focus distribution, we can preserve the word order of the source (Czech) text.

3.2. MT problems encountered with Macedonian

Macedonian is a South Slavonic language spoken in the Republic of Macedonia and by national minori-

[1] We are very grateful to Gerat Nagora and Georg Müller who allowed us to use there morphology tool based on (Starosta, 1999).

[2] The transfer is based mostly on (Janaš, 1976).

ties in Albania, Bulgaria and Greece. It belongs to the South-East Slavonic Bulgarian-Macedonian dialect continuum, the written standard is based on South-West dialects. Macedonian is an interesting target language especially because its typology differs extremely from other Slavonic languages; it has a simplified nominal system with an analytical structure, but on the other hand, its verbal system is very complicated.
Although the vocabulary is similar to Czech, the sentence structure differs in many aspects, since synthetic construction have to be translated analytically in most cases and analytical constructions (e.g., past tense) have different syntactic structure too. For these reasons, the shallow parser and a deeper transfer are more important than for the other implemented langauge pairs as, e.g., for Czech-Polish.
For the first evaluation phase on small texts, we have developed our own morphological synthesizer with a limited word list based on (Koneski, 1952). The comparative analysis is partially based on (Koneski, 1965). In the following list, the most frequent discrepancies between Czech and Macedonian are presented.

- Macedonian has almost no cases except for pronouns, Czech cases have to be translated by analytic (prepositional) phrases, e.g.:

  (1) *hlavní* *město*
  main-**NEUT,SG,NOM** town-**NEUT,SG,NOM**
  Makedonie
  Macedonia-**FEM,SG,GEN**
  "the capital of Macedonia" (Cze)

  (2) главен град на
  main-**MASC,SG** town-**MASC,SG** on
  Македонија
  Macedonia-**FEM,SG**
  "the capital of Macedonia" (Mac)

  The assignment of prepositional cases to grammatical functions is quite straight-forward.

- There is object doubling in Macedonian, i.e., both direct and indirect objects get an additional pronoun in some cases, e.g.:

  (3) Му$_1$ ја$_2$ дадов
  him her-**ACC** gave-**1SG**
  книгата$_2$ на Стојан$_1$
  book-**FEM,SG,DEF** on Stojan
  "I gave the book to Stojan." (Mac)

  The Czech sentence would be:

  (4) *Dal* *jsem*
  gave-**RESPART,MASC,SG** am
  knihu *Stojanovi*
  book-**FEM,SG,ACC** Stojan-**SG,DAT**
  "I gave the book to Stojan." (Cze)

  The solution of this problem involves the decision, whether the enclitic pronoun has to be present in the sentence or not, and eventually the insertion of the pronoun at the right position.

- A complicated problem arises with the past tense. Czech has only one past tense — the compound perfect with a resultative (*l*-)participle, e.g.:

  (5) *On* *byl* *v*
  he was-**RESPART,MASC,SG** in
  Bitole
  Bitola-**FEM,SG,LOC**
  "He was in Bitola." (Cze)

  Unfortunately, an analogical construction cannot be used in Macedonian since these participles are used to express the renarrative (see below), so the English translation of the following example means "reportedly":

  (6) Тој бил во
  he was-**RESPART,MASC,SG** in
  Битола
  Bitola-**FEM,SG**
  "He reportedly was in Bitola." (Mac)

  Instead of that, one can use the compound past tense with има or the concise past tense (aorist or imperfect), e.g.:

  (7) Тој беше во Битола
  he was-**3SG** in Bitola-**FEM,SG**
  "He was in Bitola." (Mac)

  (8) Што имаш речено?
  what have-**PRES,2SG** said-**NEUT,SG**
  "What did you say?" (Mac)

- Verbal nouns are used in Macedonian more often since it has lost the infinive, e.g.:

  (9) Не треба седење
  not needed-**ADV** sitting-**NEUT,SG**
  треба работење
  needed-**ADV** working-**NEUT,SG**
  "One should not sit, one should work." (Mac)

  The other way to express the Czech infinitive is the embedded *da*-phrase, e.g.:

  (10) *Chci* *jít* *domů*
  want-**1SG** go-**INF** home
  "I want to go home." (Cze)

  (11) Сакам да одам дома
  want-**PRES,1SG** that go-**PRES,1SG** home
  "I want to go home." (Mac)

- Macedonian has a special verbal category which is not present in any other Slavonic language except Bulgarian, the renarrative. It is used to express facts which the speaker cannot verify, e.g.:

  (12) Тој кажа дека Стојан
  he says-**PRES,3SG** that Stojan-**SG**
  бил в куќи
  was-**3SG** in house-**FEM,LOC**
  "He says that Stojan was at home." (Mac)

- Existential propositions of the type *there is* are expressed in Macedonian using има+acc., whereas Czech uses *to be*, e.g.:

  (13) *V horách jsou*
  in mountains-**FEM,PL,LOC** are-**3PL**
  medvědi
  bears-**MASC,PL,NOM**
  "There are bears in the mountains." (Cze)

  (14) Има мечки во
  has-**PRES,3SG** bears-**FEM,PL** in
  планините
  mountains-**FEM,PL,DEF**
  "There are bears in the mountains." (Mac)

  The general pattern in Czech is a sentence with a subject and a locative phrase (with the auxiliar verb), thus the latter has to be transformed to accusative (this change is only relevant for pronouns), e.g.:

  (15) Него го нема
  Him him-**ENCL** has-not-**3SG**
  "He is not here." (Mac)

- The order of clitics is different. Basically, they are attached to the verb, often to the left, in Macedonian, whereas Czech usually places them at the second position in the clause (i.e., they follow the first (accented) phrase), e.g.:

  (16) *Nechce se mi číst*
  Not-want-**3SG** REFL me-**DAT** read-**INF**
  tu knihu
  that-**FEM,SG,ACC** book-**FEM,SG,ACC**
  "I do not feel like reading the book." (Cze)

  (17) Не ми се чита
  Not me-**DAT** REFL reads-**3SG**
  книгата
  book-**FEM,SG,DEF**
  "I do not feel like reading the book." (Mac)

As for the topic-focus articulation, we are trying to preserve the word (constituent) order given in the input sentence. Almost all changes concern enclitic elements which usually have a fixed position in the sentence or verbal phrase, e.g.:

(18) *Nemám rád politiku*
not-have-**1SG** like-**ADV** politics-**FEM,SG,ACC**
"I do not like politics." (Cze)

(19) Не ја сакам политиката
not her-**ACC** like-**1SG** politics-**FEM,SG,DEF**
"I do not like politics." (Mac)

(20) *V knize se netvrdí,*
in book-**FEM,SG,LOC** REFL not-says-**3SG**
že...
that
"One does not say in the book that..." (Cze)

(21) Не се тврди во книгата
not REFL say-**3SG** in book-**FEM,SG,DEF**
дека...
that
"One does not say in the book that..." (Mac)

Some changes concern noun phrases with embedded sentences, e.g.:

(22) *dívka, kterou jsem*
girl-**FEM,SG,NOM** which-**FEM,SG,ACC** am
viděl
saw-**RESPART,MASC,SG**
"the girl I have seen" (Cze)

(23) таа чупа што
that-**FEM,SG** girl-**FEM,SG** what
ја видов
her-**FEM,SG,ACC** saw-**1SG**
"the girl I have seen" (Mac)

Obviously, some elements can be dropped or added to the target syntactic structure.

4. Parser and transfer

4.1. Data structures

There are two essential data structures: a *multigraph*, which represents the input sentence and its structural analysis on different stages, and abstract *objects* that are embedded in an object-oriented hierarchy and contain properties and methods. These objects are widely autonomous in the parsing process, i.e. there are almost no global rules any more.

4.1.1. Objects and their instances

Objects are abstract entities which represent elements of the sentence and the result(s). Every object has a predefined template that defines which properties and methods the object has. Concrete realizations of objects are called *instances.*

Let us have a look at the following Czech noun phrase:

(24) *velmi staré auto*
very old-**NEUT,SG,NOM** car-**NEUT,SG,NOM**
"(a/the) very old car"

This NP is an abstract entity consisting of three words, or of other hierarchically organized entities which could be schematically described as follows: (((velmi) staré) auto)$_{NP}$

4.1.2. Properties

Each object can contain static data, called *properties*. Properties can be atomic values (strings, integers etc.) as well as complex entities (instances of other object). All properties of an object are accessible only to its instances or instances of its descendants; this feature is called *encapsulation*. In general, objects appear as black boxes that act autonomously. The behavior of the objects is defined in their methods (pieces of code). Each object defines an interface which is visible for others and allows invoking internal object methods which can manipulate object properties or decompose the action by using other objects' interface.

For example, we would define an object representing nouns. Its instance of the word *auto* from example (24) could be as follows:

$$\begin{bmatrix} \text{Noun} & \\ \text{LEMMA} & \text{`auto'} \\ \text{FORM} & \text{`auto'} \\ \text{NUMBER} & \text{sg} \\ \text{CASE} & \text{nom} \\ \text{GENDER} & \text{neut} \end{bmatrix}$$

4.1.3. Autonomous code

Besides properties, objects may have their own code (organized as methods) which can be invoked by other objects through a shared interface. This code is object specific, i.e. an object representing integers may provide methods for arithmetic functions, whereas an object representing a string or a text document may provide methods for searching for fragments, replacing pieces of text etc. In the described framework, most objects have a linguistic background, of course. Most of their methods are designed to build up the syntactic structure of parts of the sentence in an autonomous manner. For example, objects that represent nouns contain methods which assure building up noun phrases by incorporating adjectives and other attributes etc. The concrete implementation of such an object is, of course, language specific (e.g. there has to be a congruence in some morphological categories between a noun and its attributes, such as in case, gender and number).

4.1.4. Multigraph

The initial state of the multigraph is a chain of morphologically annotated objects. The sentence from example (24) would be represented by the following graph (the labels L/R denote immediate left/right neighbourhood) :

$$\begin{tikzcd} \text{velmi} \arrow[r, bend left, "R"] & \text{staré} \arrow[l, bend left, "L"] \arrow[r, bend left, "R"] & \text{auto} \arrow[l, bend left, "L"] \arrow[r, bend left, "R"] & \text{otce} \arrow[l, bend left, "L"] \end{tikzcd}$$

The parser would add dependencies, so the resulting syntactic structure may be:

$$\begin{tikzcd} \text{velmi} & \text{staré} \arrow[l, "D"'] & \text{auto} \arrow[l, "D"'] \arrow[r, "D"] & \text{otce} \end{tikzcd}$$

4.2. Parsing

The parsing process consists in our framework of two phases:

1. setting up hypotheses about relations,
2. transforming hypotheses to tectogrammatical structures (in general, acyclic directed graphs).

The first phase involves recognizing of relations between words and word groups, i.e., between nodes of the graph. We are working with four basic types of relations:

- dependencies,
- coordination,
- (co-)references,
- shackles.

Moreover, each relation (edge of the graph) may be labeled. The second phase involves the recognizing of tectogrammatical patterns in the graph.

The shallow parser that recognizes chunks is implemented as a set of Prolog rules. For example, the rule for combining an adjective with a noun is defined as follows:

```
rule(X1, Y1, X2, Y2, X, Y) :-
subType(X1, adjective), subType(X2, nounPhrase),
splitAvm(Y1, [gender, number, case], A1, B1),
splitAvm(Y2, [gender, number, case], A2, B2),
unifyAvm(A1, A2, A), unifyAvm(A, B2, Y0),
appendAttribute(Y0, attrAdj, B1, Y),
X = nounPhrase.
```

There are several auxiliar predicates. First of all, the type of the objects is checked (**subType**). For this rule to apply, the adjective has to agree with its governor in gender, number and case, thus we have to unify these attributes (**unifyAvm**). Finally, the adjective becomes a feature of its governor (**appendAttribute**).

4.3. Implementation

The code which integrates the independent modules of the system is written in Java (version 1.5), so that it is platform independent. Parser and transfer are written in Prolog.

5. A note on evaluation

The results have been evaluated using Trados Translators' Workbench. The measure gives the work amount of translator necessary to adapt the target sentence so that it would be grammatical (the method is described in more detail in (Hajič et al., 2003)).

Table 1 gives the results for implemented language pairs (the source language is Czech; (**P**) means that a shallow parser has been used).

We have no representative results for Macedonian yet (the preliminary result measured on a short text is about 88%).

target language	accuracy
English	*30%*
Slovak	90%
Polish	71.4%
Lithuanian (**P**)	87.6%
Lower Sorbian (**P**)	**93%**

Table 1: Evaluation of implemented target languages

6. Conclusions

Machine translation might become a very important tool for increasing the amount of texts available in minority languages. Although the work described in this paper has reached only an experimental stage for some of the language pairs mentioned in the paper, we believe that the experiments we have completed show the advantage of exploiting the language similarity among "smaller" languages may result in an good quality of the translation. Our paper shows that a thorough investigation of linguistic phenomena having a negative influence on the MT quality between two similar languages and the application of relatively simple but adequate means reflecting those phenomena is a relatively effective way how to add new language pairs to an existing simple MT system.

7. Acknowledgements

We are very grateful to Kiril Ribarov for his comments on Macedonian examples. This research was supported by the Ministry of Education of the Czech Republic, project MSM0021620838, and by the grant No. 1ET100300517. We would like to thank the anonymous reviewers for their valuable comments and recommendations.

8. References

Jan Hajič, Petr Homola, and Vladislav Kuboň. 2003. A simple multilingual machine translation system. In *Proceedings of the MT Summit IX*, New Orleans.

Petr Homola and Jakub Piskorski. 2004. How can shallow NLP help a machine translation system. In *Conference Human Language Technologies*, Riga, Latvia.

Petr Homola and Erika Rimkutė. 2003. Shallow machine translation — in between of two extremes. *In: Proceedings of the Tbilisi Symposium, Tbilisi.*

Pětr Janaš. 1976. *Niedersorbische Grammatik.* Domowina-Verlag, Bautzen.

Blaže Koneski. 1952. Граматика на македонскиот литературен јазик *[Grammar of the Macedonian literary language].* Skopje.

Blaže Koneski. 1965. Историја на македонскиот јазик *[History of the Macedonian languages].* Kočo Racin, Skopje.

Manfred Starosta. 1999. *Dolnoserbsko-nimski słownik.* Ludowe nakładnistwo Domowina, Bautzen.

Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish

Adrià de Gispert, José B. Mariño

TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
{agispert|canton}@gps.tsc.upc.edu

Abstract

This paper presents a full experiment on large-vocabulary Catalan-English statistical machine translation without an English-Catalan parallel corpus, in the context of the debates of the European Parliament. For this, we make use of an English-Spanish European Parliament Proceedings parallel corpus and a Spanish-Catalan general newspaper parallel corpus, both of which of more than 30 M words. Given the language proximity between Spanish and Catalan languages, we investigate the cost of using Spanish as a bridge towards large-vocabulary Catalan-English translation in a wholly automatical statistical machine translation framework. Experimental results are promising, as the achieved translation quality is nearly equivalent to that of the Spanish-English language pair, practically carrying SMT research for the Catalan language to the level of more prominent language, in terms of data availability.

1. Motivation

Catalan is a Romance language spoken or understood by as many as 12 million people who live mostly in Spain, where it is co-official in several regions, but also in Andorra (where it is the national language), and some parts of France and Italy.

Despite this, when it comes to the parallel corpora necessary to build statistical machine translation systems, it becomes nearly impossible to find freely-available large-vocabulary data in Catalan and other languages. Thanks to the bilingual nature of most of the Catalan society, one exception can be found in the Spanish language. In fact, much Catalan-Spanish parallel content is being generated on a regular basis (institutions, companies, newspapers, etc.), thus producing training material of which statistical machine translation models can make good use.

Belonging to the same family of languages, being very much influenced and sharing in many cases the same speakers, Spanish and Catalan languages exhibit a morphological and grammatical similarity favouring the quick deployment of high-quality machine translation tools. Since many more parallel corpora in Spanish are available, one can reasonably think of using Spanish as a bridge towards statistical machine translation from Catalan to other languages, and viceversa.

Under this particular situation, the question of whether the Spanish language can be used as a feasible bridge for building state-of-the-art SMT systems between Catalan and any other language is raised.

In this direction, this paper presents a full experiment on large-vocabulary Catalan-English statistical machine translation without an English-Catalan parallel corpus, in the context of the proceedings of the European Parliament debates[1]. For this, we make use of an English-Spanish European Parliament Proceedings parallel corpus and a Spanish-Catalan general newspaper parallel corpus, both of which of more than 30 M words, and we implement two strategies, namely the straight catenation of systems and a direct training between Catalan and English.

The organization of the paper is as follows. Section 2 details the experimental setup, presenting the copora we worked with, as well as the two approaches followed to produce English-Catalan translations, including the evaluation procedure. Section 3 introduces the statistical machine translation system used in all experiments, whereas section 4 reports results for all language pairs. Finally, conclusions are presented in section 5.

2. Experimental Setup

2.1. Parallel Corpora

In order to carry out the experiments, two parallel corpora have been used. On the one hand, a general newspaper Catalan-Spanish corpus has been used, whose main statistics are shown in Table 1, including number of sentences, running words, vocabulary size and average sentence length.

	General Newspaper	
	Catalan	Spanish
Sentences	2.18 M	
Running Words	43.28 M	41.51 M
Vocabulary	390.2 k	397.4 k
Avg. Sent. Length	19.86	19.05

Table 1: Catalan-Spanish corpus statistics

On the other hand, for English-Spanish a parallel corpus containing the proceedings of the European Parliament from 1996 to September 2004 has been used (see Table 2 for main statistics).

	European Parliament proceedings	
	English	Spanish
Sentences	1.22 M	
Running Words	33.37 M	34.96 M
Vocabulary	104.8 k	151.5 k
Avg. Sent. Length	27.28	28.60

Table 2: English-Spanish corpus statistics

[1] So far European Parliament debates are not being manually translated into Catalan.

As it can be seen, even though this is a large-vocabulary task, it proves much more domain-limited as the newspaper task, which includes politics, society, international and sports sections. Note that, whereas in the newspaper corpus a new English word occurs every 111 running words on average, this happens every 318 words for the EU Parliament corpus.

2.2. Bridging strategies

In order to carry out Catalan-English translation, we have implemented two strategies, namely sequential and direct.

Regarding the sequential strategy, it simply consists of catenating two independent statistical machine translation systems, one between Catalan and Spanish, and the other between English and Spanish. Therefore, this is an additive error approach, as errors from one system propagate to the input of the following system.

On the other hand, the direct approach consists of translating the whole Spanish side of the English-Spanish parallel text into Catalan by using the Spanish-to-Catalan system, which is of a more general domain. Then, an English-Catalan system is directly trained by using this automatically produced *noisy* Catalan text. With this we expect that some translation errors of the Spanish-Catalan system will not correlate with English text and may get very low probabilities when training English-Catalan translation models.

2.3. Task evaluation

For evaluating the Catalan-to-English task, we require clean Catalan source development and test data, as well as references for the English-to-Catalan direction. With the aim of minimising the human cost necessary to obtain these data, we automatically translated source and references from the Spanish-English task, and a human reviewer made the minimum corrections necessary to obtain a correct Catalan sentence carrying the same message as the Spanish and English sentences.

Apart from an effort reduction motivation, this has two additional advantages. Firstly, we end up with exactly the same development and test sets for Spanish-English and Catalan-English, which lets as compare both tasks and evaluate the quality loss from one task to the other. And furthermore, the corrected Catalan data can be used as a very accurate evaluation reference for the Spanish-Catalan system.

	Sent	Wrds	Vocab	AvLen	Refs
Cat à Eng					
Dev	504	15646	2627	31.0	3
Test	840	23140	3914	27.6	2
Eng à Cat/Spa					
Dev	504	15331	2300	30.4	3
Test	1094	26876	3975	24.6	2
Spa à Eng					
Dev	504	15415	2735	30.6	3
Test	840	22753	4085	30.4	2

Table 3: Statistics of European Parliament develop and test sets for each translation direction.

Table 3 shows the main statistics of the resultant Catalan develop and test sets, including number of sentences, runnings words, vocabulary size, average sentence length and number of reference translations. As this sets are produced from correcting the automatic translations of Spanish text, English references are identical both for Cat à Eng and Spa à Eng directions.

3. Baseline SMT system

The SMT system used for these experiments follows the maximum entropy framework (Berger, 1996), where we can define the translation hypothesis *t* given a source sentence *s* as the target sentence maximizing a log-linear combination of feature functions, as described in the following equation:

$$\hat{t}_1^I = \arg\max_{t_1^I} \left\{ \sum_{m=1}^{M} l_m h_m(s_1^J, t_1^I) \right\} \qquad (1)$$

where λ_m correspond to the weighting coefficients of the log-linear combination, and the feature functions h_m(s,t) to a logarithmic scaling of the probabilities of each model.

Following this approach, the translation system described in this paper implements a log-linear combination of one translation model and four additional feature models. In contrast with standard phrase-based approaches, our translation model is a bilingual Ngram model expressed in tuples as bilingual units (Mariño et al. 2005). These units are extracted from the automatical word alignment of a given parallel text generated by IBM Models (Brown et al., 1993) with the GIZA++ Toolkit (Och and Ney, 2003).

The tuple N-gram translation model is a language model of a particular language composed by bilingual units which are referred to as tuples. This model approximates the joint probability between source and target languages by using N-grams as described by the following equation:

$$h_m(s_1^J, t_1^I) = \prod_{i=1}^{K} p((s,t)_i \mid (s,t)_{i-N+1}, \ldots, (s,t)_{i-1}) \qquad (2)$$

where $(s,t)_i$ refers to the i^th tuple of a given bilingual sentence pair which is segmented into *K* units. It is important to notice that, since both languages are linked up in tuples, the context information provided by this translation model is bilingual.

As additional feature functions, the system includes the following models:

- a target language model
- a word bonus model
- a source-to-target lexicon model
- a target-to-source lexicon model

The first of these feature functions is a language model of the target language, estimated as an standard N-gram over the target words, as expressed by equation 3:

$$h_{LM}(t_k) = \prod_{n=1}^{k} p(w_n \mid w_{n-N+1}, \ldots, w_{ni-1}) \qquad (3)$$

where t_k refers to a partial hypothesis containing k target words, and w_n to the n^{th} target word in it.

Usually, this feature function is accompanied by a word bonus model in order to compensate the system preference for short target sentences caused by the presence of the previous target language model. This bonus depends on the total number of words contained in the partial translation hypothesis, and it is computed as follows:

$$h_{WB}(t_k) = e^{number\ \ of\ \ words\ \ in\ \ tk} \tag{4}$$

Finally, the third and fourth feature functions correspond to source-to-target and target-to-source lexicon models. These models use IBM model 1 translation probabilities to compute a lexical weight for each tuple, which accounts for the statistical consistency of the pairs of words inside the tuple. These lexicon models are computed according to equation 5, where word-to-word probabilities are obtained from IBM model 1:

$$h_{IBM1}((s,t)_n) = \frac{1}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(t_n^i \mid s_n^j) \tag{5}$$

Once the models were computed, sets of optimal log-linear coefficients are estimated on the development set for each translation direction using an in-house implementation of the widely-used **simplex** algorithm (Nelder and Mead, 1965). This system is proved to achieve state-of-the-art performance (Koehn and Monz, 2005; Eck and Hori, 2005).

As decoder, we used the freely-available Ngram-based decoder MARIE (Crego, 2005).

4. Results

In this section the translation results are presented and discussed. First, we evaluate separately the Catalan-Spanish and English-Spanish tasks. Then, translations for the Catalan-English task are evaluated. As automatic evaluation measures, we use Word Error Rate (WER), Position-independent word Error Rate (PER) and BLEU scores (Papineni et at., 2001).

	BLEU	WER	PER
Cat à Eng sequential	0.5147	36.31	27.08
Cat à Eng direct	**0.5217**	**35.79**	**26.79**
Spa à Eng	0.5470	34.41	25.45
Eng à Cat sequential	**0.4680**	40.66	32.24
Eng à Cat direct	0.4672	**40.50**	**32.11**
Eng à Spa	0.4714	40.22	31.41
Spa à Cat	0.8421	9.88	8.74
Cat à Spa	0.8334	10.08	8.86

Table 4: Translation results for all translation directions. Whereas all directions to and from English belong to the domain of EU Parlamentary Sessions, Spanish-Catalan results are in general newspaper task.

4.1. English-Spanish and Catalan-Spanish tasks

Results for the Spanish-to-English and English-to-Spanish translation directions (belonging to European Parliament task) are shown in the 3rd and 6th rows of Table 4, respectively. Results for the Catalan-to-English and English-to-Catalan translation directions are shown in the last two rows of Table 4.

Although these results correspond to a different task (general newspaper domain, evaluated on a 2000 sentences test with a single reference translation), the quality increase due to proximity between Spanish and Catalan is remarkable.

Interestingly, whereas the Ngram-based translation model takes advantage of the additional features in the Spanish-English tasks (obtaining a performance increase in development of several BLEU absolute points), this behaviour is not observed in the Catalan-Spanish tasks, where simply the ngram translation model suffices to generate high-quality translations. This behaviour tells about the grammatical similarity between Spanish and Catalan, which allows for a less-sparse estimation of the model even with much larger vocabulary sizes.

As mentioned in section 2.3., when generating development and test Catalan texts through correction of automatically-translated texts from Spanish to Catalan, we obtain references, which are ***adapted*** to the Spanish-to-Catalan European Parliament task. That is, references which evaluate only those mistakes committed by the Spanish-to-Catalan automatic translator (without including alternative translations which do not match the produced translation, if this is correct). When evaluating this task, we obtain the following results:

	BLEU	WER	PER
Spa à Cat *adapted ref*	0.9345	3.79	3.49

which reflect once again the high quality obtained with the Catalan-Spanish Ngram-based translation model trained on broad-domain newspaper data[2].

4.2. Catalan-English task

As it can be seen in Table 4, results show in general that the automatic evaluation measures achieved in the Catalan-English task are pretty similar to those of the Spanish-English task, meaning that nearly no loss is found when bridging through Spanish to obtain a Catalan-English system. This is especially remarkable when English is the source language, which turns to be the most challenging case, basically due to the morphological richness of Romance languages in contrast to English.

At a more specific level, in Catalan-to-English, the performance loss due to the bridging through Spanish is higher (comparing the scores of the Catalan-to-English with the scores of the Spanish-to-English) than in the opposite direction. However, in this case the direct training with a noisy Catalan European Parliament corpus achieves slightly yet significantly improved scores in contrast to the sequential strategy.

This behaviour is not so clearly observed in the more difficult opposite direction, where both strategies achieve qualitatively the same performance, as automatic measures show discrepancies (best result is marked in bold). Remarkably, the scores are nearly as good as the English-to-Spanish experiment, although in this case the references are in different languages and therefore the results not strictly comparable.

[2] A demo of this Catalan-to-Spanish and Spanish-to-Catalan ngram-based SMT system can be found at http://www.n-ii.org

5. Conclusions

All in all, we believe that this is a very positive experience, whose implications are relevant, as one can conclude that it is indeed possible to build a large-scale statistical machine translation system between Catalan and any other language, so long as a large-vocabulary parallel corpus between the selected language and Spanish is found.

The achieved translation quality is nearly equal to that of the statistical translation system between Spanish and the given language, practically putting research in Catalan machine translation at the level of a major language, more powerful in terms of data availability.

On the other hand, and contrary to the catenation strategy, the approach implemented here has the advantage of producing a new corpus that can be improved and reused in further research.

Even though this corpus contains the translation mistakes generated by the Spanish-Catalan statistical system, this experiment proves that it is useful enough for training a state-of-the-art Catalan system. Besides, as these mistakes do not necessarily correlate with English data, the direct training achieves a small cancellation of some errors, slightly improving performance in one translation direction.

6. Further research

Unfortunately, at the moment the presented strategy seems to be valid only for those minority languages that are very closely related to other better-represented languages (for example, Galician). Therefore, it remains as a further research to investigate whether similar strategies could be devised for those minority languages which do not relate closely to any major language (for example, Basque).

Additionally, it will be interesting to investigate ways to detect consistent errors and clean them from the new noisy Catalan corpus, improving both the training of the Catalan-English models and the post-processing of the Spanish-Catalan translation system.

Finally, another further research line points towards multi-lingual machine translation, trying to take advantage of the information from both Spanish and Catalan data to improve translation into and from English.

7. Acknowledgements

The authors want to thank Prof. Climent Nadeu for his kind support. This work has been partly funded by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738), by Generalitat de Catalunya and by the European Social Fund.

8. References

Berger, A., Della Pietra, S.A. and Della Pietra, V.J. (1996) A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39-72, March.

Brown, P.F., Della Pietra, V.J., Della Pietra, S.A. and Mercer, R.L. (1993). The Mathematics of Statistical Machine Translation: parameter estimation. *Computational Linguistics*, 19(2):263-312, June.

Crego, J.M., Mariño, J.B. and de Gispert. A. (2005). An Ngram-based Statistical Machine Translation Decoder. In *Proceedings of the 9th European Conf. on Speech Communication and Technology*, Lisboa, Portugal, pp. 3193-96. September.

Eck, M. and Hori, Ch. (2005). Overview of the IWSLT 2005 Evaluation Campaign. In *Proceedings of the 2nd Int. Workshop on Spoken Language Translation*. Pittsburgh, Pennsylvania, pp. 11-32. October.

Koehn, P. and Monz, Ch. (2005). Shared Task: Statistical Machine Translation between European Languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Ann Arbor, Michigan, pp. 119-124, June.

Mariño, J.B., Banchs, R., Crego, J.M., de Gispert, A., Lambert, P., Costa-jussà, M.R. and Fonollosa, J.A.R. (2005). Bilingual N-gram statistical machine translation. In *Proceedings of the MT Summit X*. Pukhet, Thailand, pp. 275-282. September.

Nelder, J.A. and Mead, R. (1965) A simplex method for function minimization. *The Computer Journal*, 7:308-313.

Papineni, K., Roukos, S. Ward, T. and Zhu, W. (2001) Bleu: a method for automatic evaluation of machine translation. Tech. Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

SisHiTra: A Spanish-to-Catalan hybrid machine translation system

J. González*, A. L. Lagarda*, J. R. Navarro†, L. Eliodoro†, A. Giménez*, F. Casacuberta†, J. M. de Val‡, F. Fabregat‡

*Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
{jgonzalez, alagarda, agimenez}@dsic.upv.es

† Institut Tecnològic d'Informàtica
Universitat Politècnica de València
{jonacer, leliodoro, fcn}@iti.upv.es

‡ Servei de Normalització Lingüística
Universitat de València
{Joan.M.Val, Ferran.Fabregat}@uv.es

Abstract

In this paper, we describe how deductive and inductive techniques can be successfully combined in the framework of SisHiTra, a machine translation system from Spanish to Catalan with no semantic constraints. The translation process is based on finite-state machines and statistical models. Linguistic knowledge is appropriately incorporated as a database. Our results are compared with other systems.

1. Introduction

Machine translation (*MT*) is a challenging topic that engineers and scientists have been interested in for years. In addition to its importance for the study of human speech characteristics, MT is of social and economic interest because its development would allow for the preservation of the use of minority languages, such as Catalan, Basque or Galician in Spain. It could thus mean a reduction of linguistic barriers. This is particulary important in the access to some computer services. Another feature of Catalan is that it is a language that is spoken by more people than some official European languages.

Spanish and Catalan are languages that belong to the Romance language family, although they are from different linguistic branches: Catalan is a Gallo-Romance language, whereas Spanish is an Ibero-Romance one. Nonetheless, their resemblance is quite notable, since both of them are inflectively and morphosyntactically similar languages.

The approaches that have been traditionally used for MT can be classified into two families: *knowledge-based* and *corpus-based* methods. Knowledge-based techniques formalize expert linguistic knowledge in the form of rules, dictionaries, etc., in a computable way. Corpus-based methods use statistical pattern recognition techniques to automatically infer models from text samples without necessarily using a-priori linguistic knowledge.

Knowledge-based techniques are classical approaches for dealing with general scope MT systems. Nevertheless, inductive methods have achieved competitive results with semantically constrained tasks. On the other hand, finite-state transducers (Karttunen, 1993; Roche, 1999; Roche and Schabes, 1997) have been successfully used to implement both rule-based and corpus-based MT systems. Techniques based on finite-state models have also allowed for the development of useful tools for natural language processing (Mehryar, 1997; Mohri et al., 2000; Oflazer, 1996; Roche and Schabes, 1995; Casacuberta et al., 2004), which are interesting because of their simplicity and their adequate temporal complexity. SisHiTra makes use of finite-state models to combine knowledge-based and corpus-based techniques so as to produce a Spanish-to-Catalan MT system with no semantic constraints. Some other finite-state approaches to Spanish↔Catalan translation, such as *interNOSTRUM* (Forcada et al., 2001), confirm their adequateness to MT between these two languages.

SisHiTra's main aim is the achievement of high quality translations from Spanish to Catalan (and vice versa) for dissemination purposes. Of course, this is an ideal objective for any MT system; however, in our case, it is an especially important issue that has been taken into account in the design of each stage. Thus, we consider that *perfect* translations would be those that did not seem to be the result of a translation process, but that seemed as if they had been produced directly in the target language. This is not a problem for a human translator, but it is crucial for MT systems. For instance, semantic ambiguity is easily solved by a human speaker or reader, but it is usually a significant problem in MT. As a consequence of that, the evaluation of SisHiTra's performance is in terms of how far hypotheses are from a set of translation references, which experts have considered to be linguistically optimal.

The SisHiTra prototype is designed to be a serial process where every module performs a specific task. There is an online version running on the Internet[1] that is able to translate plain text, web pages, and LaTEX files.

Future versions of SisHiTra would be extended to other language pairs (Portuguese, French, Italian, etc.). In the fol-

This work was partially supported by the "Agència Valenciana de Ciència y Tecnologia" under grant GRUPOS03/031 and by the Spanish project ITEFTE (TIC2003-08681-C02-02).

[1] http://prhltdemos.iti.upv.es/˜taval/

lowing section, we will explain SisHiTra's architecture.

2. System architecture

SisHiTra is a general scope Spanish-to-Catalan translator with a wide vocabulary recall, so it is able to deal with all kinds of sentences. A previous version of the SisHiTra system can be found in (Navarro et al., 2004).

The methodologies to be used in the representation of the different knowledge sources are based on finite-state machines: on the one hand, stochastic transducers, which are employed as data structures for dictionary requests as well as for inter-module communication; on the other hand, Hidden Markov Models (HMM), which are applied in disambiguation processes (Sanchis et al., 2001). Finite states have proven to be adequate models for translation purposes. They can be easily inferred from corpora, and there are efficient algorithms for their manipulation (Viterbi, beam search, etc.). In addition, linguistic knowledge can be properly incorporated.

As previously stated, translation prototype modules are based on finite-state machines, providing a homogeneous and efficient framework. Engine modules process input text by means of a cascade of finite-state models that represent both linguistic and statistical knowledge. Finite-state models are also used to represent partial information during translation stages.

The SisHiTra system is structured in the following modules:

- **The preprocess module:** It divides the original text into sentences, thus allowing the translation process to be applied to each individual sentence. Let us introduce a simple example in order to better understand how SisHiTra performs. Figure 1 shows some Spanish text to be translated.

  La estudiante atendió.

  Figure 1: Translation text

  Moreover, sentences are split up as a sequence of translation units, where every translation unit is then identified and classified into one of the following groups: punctuation marks, numbers, abbreviations, proper names, or general words. Output is expressed in a *xml* format, in which every paragraph, sentence, translation unit, and case information, has been detected (see Figure 2).

  ```
  <doc>
   <p>
    <o>
     <ut ort="M">la</ut> <ut>estudiante</ut>
     <ut>atendió</ut> <ut uti="signo">.</ut>
    </o>
   </p>
  </doc>
  ```

  Figure 2: Preprocess

  As it can be deduced from Figure 1, the translation example is composed of only one sentence. Figure 2 shows how preprocess has segmented the whole text, thus identifying the most significant components. Xml tags stand for as follows:

  – <[/]doc> labels refer to the whole document.
  – <[/]p> labels point paragraphs out.
  – <[/]o> labels show sentence beginning/ending.
  – <[/]ut> labels identify translation units.

  Note that punctuation marks must be isolated from words in order to properly detect the right translation units. In the example, full stop is separated from last word *atendió*.

  In addition, upper case characters are identified, then lowered so as to be able to perform case-independent dictionary requests. Once the translation process has been carried out, case information can be restored to their original format. Figure 2 shows how uppercasing (in the example, initial word *La*) is handled by means of a translation unit feature, ort. Possible values for ort are 'M' (initial character), 'T' (all the characters) or 'U', which in conjunction with another feature, $mask$, take into account some particular configuration, just as it happens with *SisHiTra*.

- **The generation module:** A dictionary request produces a syntactic graph that represents all the possible analyses over the input sentence together with all their possible translations. For the proposed example, Figure 3 shows the result of this stage.

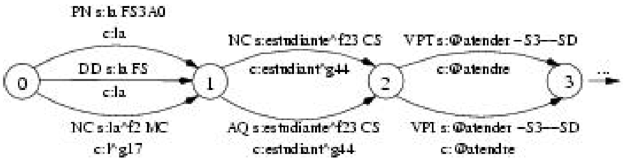
  

  Figure 3: Dictionary access

  Every edge represents a dictionary answer, which is directly related to, at least, one translation unit (note that set phrases would be modelled as a transition from state x to y, where $y - x > 1$). Therefore, transitions that share the same source/target states refer to the different readings that an input segment has, according to the dictionary. Information on edges is composed of a lexical category, together with a one-to-many relationship of Spanish and Catalan database entries. Because of space limitations, only one translation per edge appears, although it must be observed as if there could be several ones.

- **The disambiguation module:** Syntactic and semantic disambiguation is performed using statistical models. First, morphological and syntactic disambiguation is carried out through a tagging process, finding the most likely path over the analysis graph. This implies a segmentation of the the input sentence into tokens, selecting one lexical category per token. In a second step, semantic disambiguation is also performed by means of context-dependent methods. That means choosing one of the provided translations for every surviving edge from the previous morpho-syntactic disambiguation stage. This is accomplished through the very same statistical approach, namely the one that is based on HMMs, using a Catalan language model and a stochastic dictionary. Figure 4 shows the result of the disambiguation module over the proposed example.

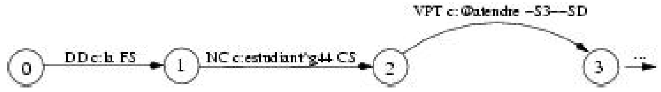


Figure 4: Statistical disambiguation

- **The postprocess module:** Here, several rule-based conversions are applied in order to transform output (that is not yet in natural language) into correct sentences from the target language. This includes noun phrase agreement, inflection, spelling, and format. Figure 5 shows the final output.

L'estudiant va atendre.

Figure 5: Translated text

MT needs to somehow semantically disambiguate source words before turning them into target language items. Semantic disambiguation methods try to determine the implicit meaning of a word in a surrounding context. SisHiTra makes use of statistical models for doing such a task.

Statistical models are becoming popular for several reasons. The most important reason is that they are cheaper and faster to generate than knowledge-based systems. Statistical techniques learn automatically from corpora, without the process of producing linguistic knowledge. Of course, obtaining corpora for model training is not a task that is free of effort.

Models for semantic disambiguation in SisHiTra need parallel corpora, that is, corpora where text segments (such as sentences or paragraphs) in a language are matched with their corresponding translations in the other language. These corpora have been obtained from different bilingual electronic publications (newspapers, official texts, etc.) and they have been paralleled through different alignment algorithms.

SisHiTra's modules were implemented following two different schemes: pre- and postprocess modules were coded in *flex* language, which is a very useful tool for generating programs that perform pattern-matching of regular expressions in text; and, the remaining modules were all written in *C*, following a Viterbi(Viterbi, 1967) searching strategy. In the next section, we will explain in detail SisHiTra's cornerstone: its dictionary.

3. Linguistic knowledge as a database

One of the advantages of our system is its convenient incorporation of linguistic knowledge, which seems to be essential to achieve *natural* translations. This knowledge is represented as a database where the main table registers refer to common dictionary entries, that is, words or set phrases from the source language (generally known as *tokens*).

Each possible translation is considered for every source token by means of a one-to-many relationship. Thus, in order to decide which translation is the most suitable, some extra information must be taken into account, such as surrounding context, the dialectal variety that is chosen, or the target language structure. All this information is reflected in our dictionary fields. These fields are:

- Spanish literal: the token's citation form. That is, masculine, if it has only gender inflection; singular masculine, if it also has number inflection; infinitive, if it is a verb, etc.

- Label of the Spanish token. It indicates the token's root and the way it is inflected.

- Morphotactics. Grammatical category of the Spanish token (noun, adjective, verb, adverb, etc.) together with some syntactic and lexical information (if a verb is transitive, intransitive, auxiliar, etc.; or if a conjunction is correlative, subordinating, etc.; and so on).

- Remission (only applied to verbs). Unlike nouns or adjectives, all the verb forms from the same infinitive do not necessarily have to share the same root. Therefore, we need to explicitely introduce every different pair (root, paradigm) as a separate entry. However, translation information is not cloned for every register but only stored in the one corresponding to the infinitive. This field links all these different entries towards their common translation information.

- Morphological unit class. Tokens are classified into one of the following 4 classes: common, infinitive verbs, prefixes, and suffixes. The usefulness of the first two groups has already been explained. Prefixes and suffixes are special tokens that are needed in order to be able to parse (and translate) unknown words, which are truly composed of:

  a) a well known prefix plus a dictionary entry

  or

  b) a dictionary entry plus a well known suffix

  Therefore, we can identify and successfully translate any compound or derivate word not explicitely included in the database.

- Abbreviation. It shows if tokens are abbreviations or acronyms.

- Case. It specifies if tokens are always written in upper case.

- Nominal inflection. Description of all the Spanish and Catalan inflectional paradigms.

- Extra. Additional information (literal or figurative sense, usual context, etc.).

- Number of meanings of the Spanish token.
- Senses. A set of different meanings for the current Spanish token. Each of them has the following information:
  - Thematic marks. Different topics where tokens might appear (technology, biology, sports, shows, chemistry, business, etc.). This field helps to do sense disambiguation in order to choose the right meaning according to the terminological sphere which tokens are referring to.
  - Semantic marks. Knowledge-based logical and semantic information.
  - Sense order. Priority over the set of meanings of the Spanish token. In addition, each sense has a set of Catalan equivalences, that is, there is a translation (or a set of synonymous ones) for each meaning of a Spanish token. Equivalences that belong to the first sense are preferred to the ones from the second sense; in turn, these are preferred to the ones from the third sense, and so on. These values have been manually established according to some linguistic criteria, taking into account both frequency of use and linguistic expressiveness.
  - Equivalences. A set of synonyms for a given sense of the Spanish token. Each equivalence has the following information:
    - Priority of a particular translation over a set of synonyms. As previously explained, a source token can have several meanings, and once one is chosen, there are several equivalent translations. We also set a range of priorities over these synonyms.
    - Catalan literal. See Spanish literal.
    - Catalan label. See Spanish label.
    - Catalan grammatical category. In general, given the similarity between both languages, Catalan tokens inherit their corresponding grammatical category of the Spanish token.
    - Preference. It refers to the linguistic dialect that is more likely to produce the Catalan item. Two of these Catalan dialects are taken into account: eastern and western. This distinction allows us to produce adequate expressions according to the user's linguistic area. In future reverse versions (from Catalan to Spanish), it will be essential to consider the whole Catalan vocabulary.

Although only a very superficial description of our dictionary has been presented here, it provides a general framework for building new dictionaries for other language pairs.

4. Evaluation

Several corpora were collected to assess the translation quality achieved. A comparison between SisHiTra and some other Spanish-to-Catalan systems has been made.

4.1. Corpora

In order to be able to make a statistical estimation of the different models used in the implemented version of the prototype, several corpora were collected.

Specific tools were developed to look for information through the web. The *LexEsp* corpus (Carmona et al., 1998), with nearly 90.000 running words, was used to estimate *syntactic disambiguation* model parameters. A label, from a set of approximately 70 categories, was manually assigned to each word.

Two other corpora (*El Periódico de Catalunya* and *Diari oficial de la Generalitat Valenciana*) were obtained by means of web tools. These corpora will be used in some system improvements such as training models for *semantic disambiguation*. These corpora consist of parallel texts that are aligned at the sentence level in a Spanish-to-Catalan translation framework without semantic constraints.

An evaluation corpus was created to perform the system assessment. This corpus is composed of 240 sentence pairs (4389 running words), which were extracted from different sources and published in both languages. Needless to say, they are not included in any training corpus.

- 120 sentence pairs from *El Periódico de Catalunya*, with no semantic constraints.
- 50 pairs from *Diari Oficial de la Generalitat Valenciana*, an official publication from the Valencian Community government.
- 50 pairs from technical software manuals.
- 20 pairs from websites (Valencia Polytechnical University, Valencia city council, etc.).

4.2. Results

Word error rate (WER[2]) is a translation quality measure that computes the edition distance between translation hypotheses and a predefined reference translation. The edition distance calculates the number of substitutions, insertions, and deletions that are needed to turned a translation hypothesis into the reference translation. The accumulated number of errors for all the test sentences is then divided by the number of running words, and the resulting percentage shows the average number of incorrect words. Since it can be automatically computed, it has become a very popular measure. The WER results for the SisHiTra system are similar to the ones achieved by other non-commercial systems (*interNOSTRUM*[3] and *SALT*[4]) as shown in Table 1.

interNOSTRUM is a realtime MT system that provides approximate translations from Spanish to Catalan. Texts can be processed in any of the following formats: ANSI, HTML, and RTF. *SALT* is a completely knowledge-based MT system that performs an interactive method that minimizes mistakes, thus providing naturalness to translations.

[2] Also known as Translation WER (TWER)

[3] See http://www.internostrum.com

[4] See http://www.cult.gva.es/salt/salt_programes_salt2.htm

Table 1: WER comparison between some MT systems

System	WER
interNOSTRUM	12.6
SisHiTra	12.5
SALT 3.0	12.2

A disadvantage of WER is that it only compares the translation hypothesis with a fixed reference translation. This does not offer any margin to possibly correct translations that are expressed in a different writing style. Therefore, to avoid this problem, we used the WER with multireferences (MWER[5]) to evaluate the prototype. MWER considers several reference translations for the same test sentence, then computes the edition distance with all of them, returning the minimum value as the error corresponding to that sentence. MWER offers a more realistic measure than WER because it allows for more variability in translation style. Other two more references were created by expert linguists, making variations to the original reference sentence. The MWER results for the SisHiTra system are the best in the three tested systems, as shown in Table 2.

Table 2: MWER comparison between some MT systems

System	MWER
interNOSTRUM	6.5
SisHiTra	**4.1**
SALT 3.0	6.1

With regard to the translation speed, SisHiTra is able to process more than 1000 words per second, which can be considered as realtime working.

5. Conclusions and future work

SisHiTra shows how deductive and inductive techniques can be successfully combined to produce a MT system with no semantic constraints from Spanish to Catalan, a *nearly* official European minority language that is spoken by an important number of the European people. The translation process is based on finite-state machines and statistical models that are automatically inferred from parallel corpora. The translation results are promising, but there are still several points that must be improved.

In addition, an appropriate representation of linguistic knowledge has been incorporated into a MT system as a database, which is essential for obtaining *natural* translations. Moreover, this database structure could be easily adapted to other language pairs.

The most relevant areas where the system could be improved are:

- Semantic disambiguation, where statistical models for ambiguous words could be trained in order to be able to choose the most appropriate context-dependent translations.
- Verb phrase agreement.

We also bear in mind a SisHiTra reversion in order to translate from Catalan to Spanish. A preliminary version of the needed linguistic dictionary can be automatically obtained from our current database.

Finally, SisHiTra's framework could be extended to other Romance languages (Portuguese, French, Italian, etc.).

[5] Multi-reference Word Error Rate

6. References

J. Carmona, S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. 1998. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC*, pages 915–922, Granada, Spain, May.

F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcia-Varea, C. Martinez D. Llorens, S. Molau, F. Nevado, M. Pastor, D. Pico, and A. Sanchis. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25–47.

M. Forcada, A. Garrido, R. Canals, A. Iturraspe, S. Montserrat-Buendia, A. Esteve, S. Ortiz Rojas, H. Pastor, and P.M. Pérez. 2001. The spanish-catalan machine translation system internostrum. *0922-6567 - Machine Translation*, VIII:73–76.

L. Karttunen. 1993. Citation of unpublished documents. Technical report, XEROX Palo Alto Research Center.

M. Mehryar. 1997. Finite-state transducers in language and speech processing.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2000. The design principles of a weighted finite-state transducer library. *Theoretical Computer Science*, 231(1):17–32.

José R. Navarro, Jorge González, David Picó, Francisco Casacuberta, Joan M. de Val, Ferran Fabregat, Ferran Pla, and Jesús Tomás. 2004. SisHiTra : A Hybrid Machine Translation System from Spanish to Catalan. In *EsTAL*, pages 349–359.

Kemal Oflazer. 1996. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. *Computational Linguistics*, 22(1):73–89.

Emmanuel Roche and Yves Schabes. 1995. Deterministic part-of-speech tagging with finite-state transducers. *Computational Linguistics*, 21(2):227–253.

Emmanuel Roche and Yves Schabes, editors. 1997. *Finite-State Language Processing*. Bradford Book. MIT Press, Cambridge, Massachusetts, USA.

Emmanuel Roche. 1999. Finite state transducers: parsing free and frozen sentences. pages 108–120.

A. Sanchis, D. Picó, J.M. del Val, F. Fabregat, J. Tomás, F. Casacuberta, and E. Vidal. 2001. A morphological analyser for machine translation based on finite-state transducers. In *MT Summit VIII*, pages 305–309, September.

A. Viterbi. 1967. Error bounds for convolutional codes and an asymtotically optimal decoding algorithm. *Annals of the New York Academy of Sciences*, IT-13:260–269.

Phrase-based Statistical Machine Translation between English and Welsh

Dafydd Jones, Andreas Eisele

Dept. of Computational Linguistics
Saarland University, PO Box 151150
D-66041 Saarbrücken, Germany
dafyddj@gmail.com, eisele@coli.uni-saarland.de

Abstract

This paper shows how a baseline phrase-based statistical machine translation (SMT) system can be set up for translation between English and Welsh, a UK language spoken by about 610,000 people, using well-documented and freely available tools and techniques. Our results indicate that the achievable performance for this language pair is among the better of those European languages reported in Koehn (2005). We argue that these preliminary results should be seen as a first step towards hybrid systems where linguistic knowledge from lexical and morphological resources are combined with additional terminology and stochastic preferences acquired from existing translations.

1. Introduction

As demonstrated in Koehn (2005), phrase-based statistical machine translation (SMT) can lead to interesting levels of translation quality for many language pairs if parallel corpora of sufficient size can be used for training. Results are particularly notable for target languages that do not make use of extensive morphology.

Compared with older models of SMT, where decoding (translation) methods were based directly on the mathematical models used for computing the word alignments, the phrase-based variant of SMT has the advantage of being flexible in the size of the building blocks it uses. Re-use of long subsequences of sentences leads to the high precision known from the use of translation memories (TMs). In contrast to TMs, where recall can fall dramatically when applied to texts written from scratch, phrase-based SMT shows a much softer degradation in these cases.

The availability of relevant tools (Koehn, 2006) made possible the organisation of a project seminar during the Summer semester, 2005 at Saarland University to investigate the use of these tools in the development of at least baseline SMT functionality for a number of different language pairs. Whereas most of these systems were built using the Europarl corpora (Koehn, 2002), and did not explore further than the 110 language pairs covered in Koehn (2005), our work involving the minority language of Welsh shows that this particular approach is equally applicable when one member of the language pair has little conventional language technology available to those wishing to conduct state-of-the-art computational linguistic research.

2. Corpus collection

Of crucial importance for any statistical analysis of language is the availability of a sufficiently large, high-quality corpus. Our corpus is collected from the online version of the Record of Proceedings of the Plenary Meetings of the National Assembly for Wales.

The Welsh Assembly is an elected body representing almost 3 million people and is responsible for developing policy and allocating funds made available to it from the UK Treasury. It consists of 60 members, who held their first meeting in May, 1999. Plenary meetings are held twice a week, and consist of questions and debate. Meetings are bilingual, and a speaker may speak in the language of their choice.

A verbatim report is made available to the public within 24 hours of any given meeting, with a 5-day, fully translated, official version made available on the Assembly website <http://www.wales.gov.uk/organipo/>. Though its publications are subject to Crown copyright protection, the Assembly has a policy of open access to information, and its records are free to use and can be reproduced under waiver conditions. The waiver conditions allow reproduction and distribution for any purpose and in any medium as long as the Record remains accurate and Crown copyright is acknowledged.

Collecting the source material for the corpus was simply a matter of crawling the relevant sections of the Assembly's website. The proceedings are presented as HTML in a 2-column table format, where each row of the table contains a turn of a speaker, one column for each language.

Of slight inconvenience is the fact that the left column always represents the speaker's utterance in its original language. A number of Assembly members choose to make their comments in Welsh, so the contents of the left column arbitrarily switches from English to Welsh. An ad-hoc language guesser was constructed, using data from the CPAN Perl module Language::Guess (Ceglowski, 2004), to identify Welsh and English paragraphs.

After sentence segmentation, corresponding Welsh/English paragraphs that differed in number of sentences were rejected. Inspection revealed that after this step corresponding sentences within speaker turns were aligned between languages. A final cleaning step was performed to remove a few garbage lines, convert HTML entities to ASCII equivalents and tokenize words and punctuation (apostrophes and hyphens within words were preserved). Table 1 shows some measurements of the final bilingual, sentence-aligned corpus.

	Welsh	English
Sentences	510,813	
Tokens	10,760,861	10,703,378
Types	65,219	49,719

Table 1: Corpus measurements after processing

Given the high-quality and size of this corpus, we are happy to make it available to other interested researchers.

Even in a monolingual form, it currently represents the largest freely-available Welsh corpus. Contact <dafyddj@gmail.com> for more details.

3. Software

Our work was made possible by the availability of software tools that cover the whole system development cycle from initial word-alignment and phrase table training to decoding and evaluation.

3.1. GIZA++

GIZA++, developed by Franz Josef Och (2003), is used to calculate word alignments between corresponding bilingual sentences according to refined statistical models. A detailed description of the design of the software can be found in Och and Ney (2003).

GIZA++ is itself an extension of the original GIZA software developed as part of the 1999 Summer workshop hosted by the Center for Language and Speech Processing at Johns Hopkins University (CLSP, 1999). The software is released under the GNU Public License.

3.2. Pharaoh

Phillip Koehn's Pharaoh (2004) system consists of scripts for extracting phrases from word-aligned sentences (provided by GIZA++, for example), and a decoder, which actually performs translation of an input sentence, given an appropriate translation model and a target language model. In this case, the translation model is a phrase-table that contains a set of phrase correspondences in the form of a foreign phrase, a native phrase, and a conditional probability that one could be the translation of the other.

Pharaoh is provided (Koehn, 2004b), without source code, for non-commercial purposes by the University of Southern California.

3.3. SRI Language Modeling Toolkit

SRI International provide a toolkit (SRI, 2006) for building and applying n-gram based language models under an open-source community licence that allows not-for-profit use and requires any code changes to be shared with other users. The Pharaoh decoder works with SRI language models.

4. Training and Decoding

Philipp Koehn provides a script, 'train-phrase-model.perl', as part of the Pharaoh package that automates the training process, using the above software. We split our corpus into a training set of 460,813 sentences, holding out 50,000 sentences for potential tuning and testing. Using this training data we generated bi-directional phrase tables of around 20 million phrases each (approx. 1.8 Gb in size on disk).

Due to the problem of loading and storing such a large amount of data in memory, sentence translation is accomplished via a filter script, that extracts a subset of phrases from the full phrase tables that account for actual phrases found in the source sentences. Nevertheless, this filtering incurs a time penalty, and data structures can still take over a minute to load into memory, with subsequent translation taking between 5 and 10 seconds per sentence.

5. Evaluation

Systematically measuring similarity between MT results and a human reference translation has become quite popular in the last five years (Papineni et al., 2002), but the metrics used for these comparisons, such as the BLEU score, are typically very superficial and do not allow qualified statements on absolute translation quality or comparison between systems across widely different architectures (Callison-Burch et al., 2006).

On the other hand, automatic evaluation has the advantage of being cheap, both in time and resources, and is therefore appropriate for measuring progress or regression during the development of a given system. In that sense, our measurements are based on BLEU, due to lack of time for more extensive manual investigations, and should only be viewed as a very first step towards a meaningful evaluation.

We picked a 5000 sentence test set, from previously held out data, as a basis for our measurements. Table 2 shows the BLEU scores measured on translations from Welsh to English and vice-versa. Figure 1 shows an example set of translations.

English to Welsh
Source ' iaith pawb ' clearly states that the availability of education through the medium of welsh has increased steadily in recent years , and that that is an aim your government wants to encourage
Translation mae ' iaith pawb ' yn datgan yn glir bod y ddarpariaeth o addysg drwy gyfrwng y gymraeg wedi cynyddu raddol yn ystod y blynyddoedd diwethaf , a bod eich llywodraeth yn dymuno annog nod
Welsh to English
Source mae ' iaith pawb ' yn datgan yn glir fod y gallu i gael addysg drwy gyfrwng y gymraeg wedi cynyddu'n gyson yn y blynyddoedd diwethaf , a bod eich llywodraeth yn dymuno hybu'r amcan hwnnw
Translation ' iaith pawb ' states clearly that the able to receive their education through the medium of welsh has steadily increased in recent years , and that your government wants to promote that objective

Figure 1: Examples of a spoken utterance and its corresponding translations.

From Welsh	Into Welsh
40.22	**36.17**

Table 2: BLEU scores for Welsh-English translation calculated over 5000 test sentences

To give some context to these measurements, the highest and lowest BLEU scores reported in Koehn (2005) are 40.2 for translating Spanish to French, and 10.3 for Dutch to Finnish. It should be noted that these systems were developed and tested on a different corpus from a different domain.

6. Other work

Our work is not the first reported instance of Welsh to English statistical machine translation. Phillips (2001) reports on the implementation of software to construct stochastic translation models from bilingual sources. In an interesting approach, he used the Bible as training text for building his system, extending this with a morphological component and a bilingual dictionary to compensate for the limited vocabulary of the Bible.

Due to a lack of an independent corpus, our Welsh language model was generated from our approx. 10 million word training corpus. Kevin P. Scannell's An Crúbadán project (Scannell, 2004) has collected a Welsh corpus of 95 million words. His software crawls the web, specifically collecting texts in minority languages, bootstrapping further crawls by using seed text as search terms to discover more web pages in the target language. We hope to investigate the use of this internet corpus as a basis for the language model required for translation into Welsh.

7. Conclusions

We see our SMT system as a first step towards machine translation for Welsh. We are convinced that better quality and coverage can be achieved when linguistic knowledge, such as rule-based morphological analysis and parsing is included in the process. Whereas this may look straightforward on the English side, a lack of language technology resources for Welsh is a hindrance.

However, we see the option to use the existence of large amounts of parallel texts and high-quality word alignments to project (parts of) linguistic analyses along these alignments to the other language, thus bootstrapping linguistic knowledge for Welsh in a manner that avoids the expense of treebanking but still promises higher quality than fully unsupervised approaches. We plan to use techniques and results from the Ptolemaios project (Kuhn, 2004) for a further exploration of this perspective.

We note that there is currently some interest in the commercial development of Welsh/English machine translation. A report commissioned by the Welsh Language Board (Somers, 2004) makes the recommendation that initial work in this area should be focused on the development of an SMT system capable of producing low-quality but usable translations. We hope our work shows that this is indeed a realistically achievable goal.

8. References

Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *EACL-2006 (to appear).*

Ceglowski, M. (2004). Language::Guess (version 0.01). CPAN, accessed April, 2006, <http://search.cpan.org/~mceglows/ Language-Guess-0.01/Guess.pm>.

Center for Language and Speech Processing (1999). The EGYPT Statistical Machine Translation Toolkit. Johns-Hopkins University, accessed April 2006, <http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/>.

Koehn, P. (2002). Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Unpublished draft, MIT, accessed April 2006, <http://people.csail.mit.edu/~koehn/publications/ europarl.ps>.

Koehn, P. (2004). Pharaoh: a beam search decoder for statistical machine translation. In *6th Conference of the Association for Machine Translation in the Americas*, Lecture Notes in Computer Science. AMTA, Springer.

Koehn, P. (2004b). Pharaoh: a beam search decoder for phrase-based statistical machine translation models. ISI, accessed April, 2006, <http://www.isi.edu/ licensed-sw/pharaoh/>.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*.

Koehn, P. (2006). Statistical Machine Translation. Accessed April 2006, <http://www.statmt.org/>.

Kuhn, J. (2004). Experiments in parallel-text based grammar induction. In *42nd Annual Meeting of the Association for Computational Linguistics*. ACL.

Och, F. J. (2003). GIZA++: Training of statistical translation models. Accessed April, 2006, <http://www.fjoch.com/GIZA++.html>.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics*. ACL.

Phillips, J. D. (2001). The Bible as a basis for machine translation. In *Proceedings of Pacling 2001*. Pacific Association for Computational Linguistics.

Scannell, K. P. (2004). Corpus building for minority languages. Saint Louis University, accessed April 2006, <http://borel.slu.edu/crubadan/>.

Somers, H. (2004). Machine translation and Welsh: The way forward. Technical report, The Welsh Language Board.

SRI International (2006). SRILM - The SRI Language Modeling Toolkit. Accessed April 2006, <http://www.speech.sri.com/projects/srilm/>.

BULGARIAN SENSE TAGGED CORPUS

Svetla Koeva, Svetlozara Lesseva, Maria Todorova

Department of Computational Linguistics, IBL - BAS
52 Shipchenski prohod, Bl. 17, Sofia 1113, Bulgaria
svetla@ibl.bas.bg, zara@ibl.bas.bg, maria@ibl.bas.bg

Abstract

The Bulgarian Sense Tagged Corpus is derived from the "Brown" Corpus of Bulgarian and annotated with word senses from the Bulgarian WordNet. The paper gives a brief account of the already available and currently developed language resources and tools which enabled the compilation and annotation of the Bulgarian Sense Tagged Corpus. We briefly describe the adopted methodology for constructing and preprocessing the source corpus of 63 440 words: all words were lemmatised, PoS-tagged and linked to the corresponding sets of senses in the Bulgarian WordNet. The paper also presents the annotation criteria underlying the sense selection process and outlines the general directions of expansion and modification of the Bulgarian WordNet. At the present stage 45 562 words (single words and multi-word expressions) are semantically annotated. The chief intended application of the Bulgarian Sense Tagged Corpus is to serve as a test and / or training dataset for word sense disambiguation with the further aim of developing a Bulgarian - English bi-directional machine translation system.

1. Introduction

The main objective of this paper is to present the Bulgarian Sense Tagged Corpus (BulSemCor) derived from the "Brown" Corpus of Bulgarian and annotated with word senses from the Bulgarian WordNet (BulNet)[1]. The chief intended application of the Bulgarian Sense Tagged Corpus is to serve as a test and / or training dataset for word sense disambiguation with the further aim of employing the results in the implementation of a Bulgarian-English bidirectional machine translation (MT) system. The paper also gives a brief account of the already available and currently developed language resources and tools which enabled the compilation and annotation of BulSemCor.

It is generally acknowledged that statistical approaches (completely or partially underlying any disambiguation process) can be efficiently combined with data derived from annotated corpora for testing and / or training as within the so-called supervised corpus-based methods. The statement that "a logical next step for the research community would be to direct efforts towards increasing the size of annotated training collections, while deemphasizing the focus on comparing different learning techniques trained only on small training corpora" (Banko & Brill, 2001) fully confirms our understanding of how part of speech and word sense disambiguation (WSD) should be handled. Therefore, effort should be concentrated in the devising of a balanced combination of the currently employed methods that will be able to yield a strong positive impact on the effectiveness of WSD.

In the compilation of BulSemCor we generally follow the methodology adopted for the English semantically annotated corpus – SemCor, created at the Princeton University (Fellbaum et al., 1998). The latter is a subset of the Brown Corpus of Standard American English containing almost 700 000 running words. All the words in SemCor are PoS-tagged, and more than 200 000 content words are additionally lemmatized and tagged with Princeton WordNet senses.

2. Bulgarian resources

Likewise, our target corpus for semantic annotation is a subset of the "Brown" Corpus of Bulgarian (BCB) (Koeva et al., 2005a). BCB consists of 500 corpus units of approximately 2000 words each, distributed proportionally to language use in 15 categories, thus forming an overall of 1 001 286 words. The methodology of the developing of Brown Corpus of Bulgarian is as close as possible to the original Brown corpus in terms of structure and content, but still it differs in some respects: some categories either partially or not at all represented in contemporary Bulgarian language use were replaced by more appropriate ones. The sub-corpus for sense annotation preserves the original structure of BCB by including a section of each BCB unit sampled according to the density of high frequency words.

The linguistic database which serves as a source for introducing and resolving ambiguity is the Bulgarian WordNet - BulNet (Koeva, 2004a). Synsets (as basic structural units of wordnet) are equivalence sets containing a number of obligatory elements: literals (single words and multi-word expressions (MWE) with the same referential meaning, expressed by an interpretative definition, usage examples and language notes. The synsets are interrelated in a lexical-semantic network – wordnet, by means of a set of semantic relations such as hyperonymy, antonymy, meronymy, etc. EuroWordNet (EWN) extended the Princeton WordNet (PWN) with cross-lingual relations, which were further adopted in BalkaNet (BWN) (Stamou et al., 2002) and by the Bulgarian WordNet as part of it. The equivalent synsets in the different languages are mapped to the same Inter-Lingual Index (ILI), thus connecting the individual wordnets in a global lexical-semantic network. The Inter-Lingual Index is based on PWN and is consecutively synchronized with new PWN versions.

At the moment BulNet consists of 27 045 synsets (synonym sets), containing 57 496 literals, and the average number of literals per synset is 2.12. The language-internal relations encoded in the Bulgarian WordNet are seventeen (following the Princeton WordNet), their occurrences are 48 371, the average number of relations per synset is 1.79.

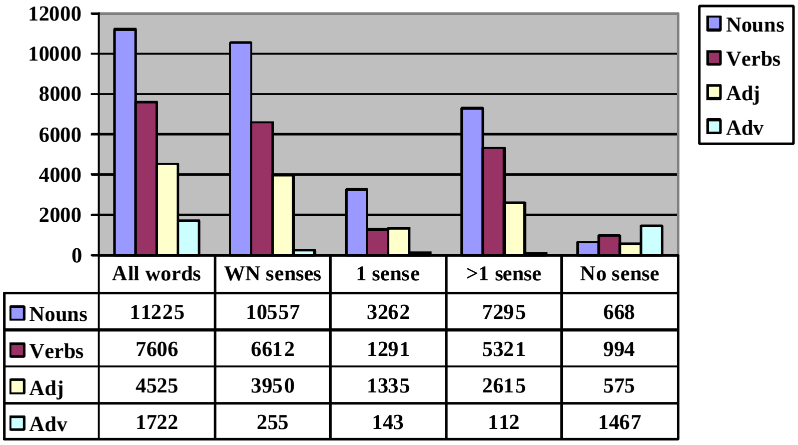
[1] The investigation is developed under the national funded project "*BulNet – Lexical-semantic Network of the Bulgarian Language*".

3. Development and pre-processing of the source corpus

The annotation corpus consists of 500 excerpts (clippings) of approximately 100 words each, selected according to a criterion for well-balanced density of highest frequency Bulgarian open-class lemmas located in BCB. The calculation of the frequency list is based on the occurrences of content words in two Bulgarian POS disambiguated corpora – 71 876 words from Orwell's *1984* and a selection of 328 964 words from three thematic domains – economy, law and politics (400 840 words altogether).

The task of constructing the corpus for annotation consisted in the selection of a 100-word excerpt (clipping) from every file in the "Brown" Corpus of Bulgarian such that would contain the highest density of words from the frequency list. The selection procedure involved several experiments with frequency lists of different sizes derived from the original one by consecutively excluding words occurring one, two and three times. Relative weights were assigned to the lemmas featuring on the lists which were further modified proportionally to the frequency of the lemmas' occurrence both in BCB and in BulNet, so that less frequent words have greater weights. Further, additional weights were calculated according to part of speech as follows: 0.4 to nouns, 0.3 to verbs, 0.2 to adjectives and 0.1 to adverbs in order to provide a better balance in the proportion of nouns and verbs in comparison with adjectives and adverbs. After the clippings' selection the following statistics was made: nouns, verbs, adjectives and adverbs were divided in two groups depending on their occurrence on the frequency lists. For each group the number of words encountered in the wordnet and the number of multiple-sense words was calculated. Subsequently, the clippings that had the best lexical coverage estimated in terms of the greatest number of different lemmas in combination with the greatest number of corresponding wordnet senses were selected for the corpus.

The resulting corpus was further enlarged by expanding the clippings to the left and right sentence boundaries, thus amounting to a total of 63 440 words. The word forms in the source corpus were lemmatized, PoS-tagged and linked to the corresponding sets of senses in BulNet, if available. 6 031 lemmas were automatically linked to only one sense, 3 704 lemmas left without a sense matched in BulNet and 15 343 lemmas received more than one sense. Figure 1 below shows the distribution of open class lemmas in the resulting corpus across part of speech and the coverage of the same lemmas in the Bulgarian WordNet. Outside these figures remain the function words which had to be additionally encoded. In the course of annotation single-sense entries are subject to validation and possibly new senses for such lemmas are encoded where needed; for the lemmas not having a corresponding entry a new synset denoting the appropriate sense is to be included in BulNet (or the sense of an already existing synset has to be revised) and then associated with the word; for multiple sense lemmas the particular sense used in the context has to be picked up, or if not available - encoded.



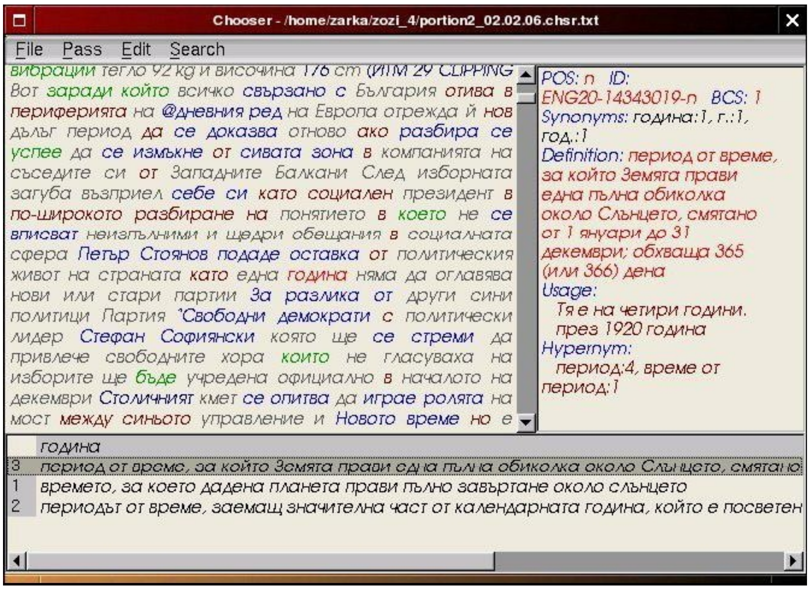
	All words	WN senses	1 sense	>1 sense	No sense
Nouns	11225	10557	3262	7295	668
Verbs	7606	6612	1291	5321	994
Adj	4525	3950	1335	2615	575
Adv	1722	255	143	112	1467

Figure 1: Distribution of content-word lemmas across POS and coverage in BulNet

4. Annotation tool

Sense annotation is conducted with the annotation tool Chooser developed at the Department of Computational Linguistics (DCL)[2]. Chooser was designed as multi-purpose multi-functional platform aimed at performing various tasks that require corpora annotation as well as at enabling automatic analysis and manual disambiguation of large volumes of text (Koeva et al., 2005a). Figure 2 below shows Chooser's layout. The application's visualisation and editing functionalities provide text display management and a number of other functions such as: text navigation according to various strategies, selection of particular options available for particular language units, group selection of adjacent or distant units (such as multi-words expressions, expressions whose constituents can be intervened by other words, etc.). The corpus for annotation is displayed in the left top window, the synchronization with the other windows is instantly initiated on navigating along the text. On selecting a current word (coloured red on the picture) the definitions of the senses available in BulNet for the word are displayed in the bottom window.

[2] Borislav Rizov from DCL has programmed the annotation tool.



вибрации тегло 92 kg и височина 176 cm (ИТМ 29 CLIPPING
Вот заради който всичко свързано с България отива в
периферията на @дневния ред на Европа отрежда й нов
дълъг период да се доказва отново ако разбира се
успее да се измъкне от сивата зона в компанията на
съседите си от Западните Балкани След изборната
загуба възприел себе си като социален президент в
по-широкото разбиране на понятието в което не се
вписват неизпълними и щедри обещания в социалната
сфера Петър Стоянов подаде оставка от политическия
живот на страната като една година няма да оглавява
нови или стари партии За разлика от други сини
политици Партия 'Свободни демократи с политически
лидер Стефан Софиянски която ще се стреми да
привлече свободните хора които не гласуваха на
изборите ще бъде учредена официално в началото на
декември Столичният кмет се опитва да играе ролята на
мост между синьото управление и Новото време но е

POS: n ID:
ENG20-14343019-n BCS: 1
Synonyms: година:1, г.:1,
год.:1
Definition: период от време,
за който Земята прави
една пълна обиколка
около Слънцето, смятано
от 1 януари до 31
декември; обхваща 365
(или 366) дена
Usage:
Тя е на четири години.
през 1920 година
Hypernym:
период:4, време от
период:1

година
3 период от време, за който Земята прави една пълна обиколка около Слънцето, смятано
1 времето, за което дадена планета прави пълно завъртане около слънцето
2 периодът от време, заемащ значителна част от календарната година, който е посветен

Figure 2: Layout of the annotation tool

The right upper window shows all the information for the sense corresponding to the selected item in the bottom window including the synonym set, definition, usage examples and the relations available in BulNet. The set of choices available for a given word can be ordered according to different criteria, presently the adopted one is the frequency of selection of the sense in the process of word sense annotation.

The application design envisages a number of pass strategies such as: passing all language units in the corpus, stopping at language units which are associated with certain information in the linguistic database, passing only ambiguous language units, language units that have been modified in the database since the last selection of the same item by the current user, or stopping at all occurrences of a particular item.

A major asset of the annotation framework is that it handles single as well as multiword expressions (MWE) referring to a single concept (i.e. *New York, nitric acid,* etc.) regardless of their structural or constituent variations, treating such expressions as single strings with spaces at certain places. The tool allows selection of adjacent as well as of distant MWE constituents, thus managing specific syntactic properties of multiword expressions - intervening words between constituents, varying constituents, different word order, syntactic transformations, etc. This has been pointed out as one of the main problems of encoding of MWE in WordNet (Fellbaum, 1998). The components of a MWE in the corpus are associated with their lemmas through which the tool makes correspondence to the relevant literal in BulNet.

Chooser is a multiple-user platform that performs dynamic interaction between the local users. User communication is implemented by means of a server that takes care of a number of activities in two principal directions:

- Interaction between the local users and the linguistic database;
- Interaction between the local users.

Chooser's database is permanently updated with enlarged BulNet versions contributed by the individual annotators. Thus, newly-entered and edited senses are imported for selection by the annotators, and passed along via the option of navigation along unselected items. By taking care of the frequency update and the respective reordering of the senses the tool keeps track of and stores valuable information about language data while facilitating the annotators' work.

5. Development of the Bulgarian Sense Tagged Corpus

After the processing of the source corpus, annotation of the language units in the corpus with the correct senses in the Bulgarian WordNet is performed.

For tagged words in the corpus the following outputs are produced:

- For nouns, verbs, adjectives and adverbs - Word, Lemma, Sense identification including the ID number and POS of the corresponding sense in BulNet.
- For multi-word expressions - MWE, Lemma, Sense identification including the ID number and POS of the corresponding sense in BulNet.
- For function words - Word, Lemma, Sense identification.

For example the Bulgarian sentence *Obichal mekiya kaliforniyski klimat* (He loved the soft Californian climate) will be annotated as follows:
Obichal{obicham#ENG20-01723774-v}
(love:have a great affection or liking for)
mekiya{mek#ENG20-00411931-a}
(mild: mild and pleasant)
kaliforniyski{kaliforniyski#ENG-02893758-a}
(Californian: of or related to or characteristic of California or its inhabitants)
klimat{klimat#ENG20-13692717-n}
(climate: the weather in some location averaged over some long period of time)

Sense Tagged Corpus	
New senses added in BulNet	5 328
Annotated units	45 562
Annotated single words	40 255
Annotated MWE	2 177
Words left for annotation	17 878

Table 1: Current state of BulSemCor

The current state of BulSemCor includes 45 562 semantically annotated single words and multi-word expressions (Table 1), of which 40 255 are single words and 2 177 - MWE. The average length of MWE is 2.19 words.

All the words initially assigned only one sense are considered annotated after the validation of the sense mapping. In the course of the annotation 5 328 new synsets have been added in BulNet so far. New senses are encoded where a corpus occurrence is not mapped in BulNet at all, or if the available candidates (single-sense or multiple-sense BulNet entries) do not match the meaning found in the corpus.

6. Annotation criteria

The annotation of the senses consists in the association of word occurrences in the corpus – single words and multiword expressions - with the appropriate senses in BulNet[3]. Coverage is ensured through the evaluation of the encoded data against the empirical evidence from the corpus and the respective revision and enlargement of BulNet with new senses. New literals and synsets are either such found in PWN or ones having no equivalent in PWN. In the latter case new BulNet-specific entries are created. Apart from this, optimisation of the encoded language data in BulNet is performed.

6.1. Selection of senses

Under this heading we discuss the implicit consistency criteria involved in the annotators' choice of a given sense of a graphical word (literal) from among the available candidates in BulNet. These procedures (or analogous ones) are extensively applied where a word in a language has a number of closely related senses:

6.1.1. Consistency with the other (if any) members of the synset

In deciding which is the most appropriate among the candidate senses, the first thing to be considered is the relation of equivalence defined between the members of a synset. This means that if an instance of a word in the corpus is semantically equivalent to an instance of another word in the same context, it is most likely that the correct sense is the one that corresponds to the synset where the two items appear as synonyms. Of course, cross-check with other criteria is performed even in this case, to avoid possible errors due to incompleteness in the database.

6.1.2. Consistency with the interpretative definition covering the general meaning of the synset

The interpretative definition (gloss) associated with the synset encodes the meaning of all the members of the synset in an explicit way, hence it is a principal clue in choosing between senses.

6.1.3. Consistency with the relative position of the synset in the overall wordnet structure

Unlike the previous criteria which establish the association between an instance of a word and a synset in BulNet according to the linguistic information contained in the synset, this one employs the degree of relatedness between pairs of synsets and is hence very helpful where a word has a number of closely related meanings. Similarity may well be signaled by identical or very close synonym sets and definitions. However, distinctness between similar lexical items will (or at least should) be observed in the different set of relations defined for a synset. Relatedness involves relations of similarity between semantically similar items, as well as other types of semantic relations (meronymy, antonymy, etc.) between dissimilar units (Budanitsky & Hirst, 2001). Hence, the exploration of the set of semantic relations encoded for the examined synset may provide helpful clues for the annotators.

The following examples illustrate the interaction of the three criteria:

Synonyms: *{nature:1}*

Definition: *the essential qualities or characteristics by which something is recognized*

Usage: *it is the nature of fire to burn*

Hypernym: *{quality:1}*

Synonyms: *{nature:3}*

Definition: *the natural physical world including plants and animals and landscapes etc.*

Usage: *they tried to preserve nature as they found it*

Hypernym: *{universe:1, existence:2, creation:6, world:2, cosmos:1, macrocosm:1}*

On looking at the Bulgarian counterparts one can see that the Bulgarian synset corresponding to *{nature:1}* has two members – *{estestvo:1, priroda:3},* and that corresponding to *{nature:3}* – one – *{priroda:2}*. The two senses are further distinguished by the glosses and the hyperonyms defined for the synsets. The usage examples also account for the distinction between the senses.

6.1.4. Consistency with the usage examples

Besides illustrating the context of use of a word, usage examples provide a quick way of scanning through and checking different senses of a word as well as of potential candidates for encoding. They are especially helpful in

[3] Ekaterina Tarpomanova and Hristina Kukova from the Department of Computational Linguistics (IBL-BAS) and Katya Alahverdzhieva and Nikolay Radnev, students at Sofia University, also worked in different capacity as annotators.

cases of similar synonym sets and / or unclear definitions as in the example given below:

Synonyms: *{disorder:1, upset:3}*

Definition: *condition in which there is a disturbance of normal functioning*

Usage: *the doctor prescribed some medicine for the disorder*

Hypernym: *{condition:1, status:2}*

The definition and the synonyms do not at first sight help to infer the meaning of the synset in this example. Scanning the usage examples, together with hyponyms, such as *{immunological disorder:1}*, *{cardiovascular disease:1}*, etc. help the annotator grasp the meaning at once.

6.1.5. Consistency with grammatical features accounting for sense distinctions

Certain sense distinctions may be suggested by grammatical differences. For example, the plural form of a noun signifying a member of a nation may stand for the relevant nation as well, as in *The Brits are a great nation* where the sense assigned to *the Brits* corresponds to:

Synset: *{British:1, British people:1, the British:1, Brits:1}*

Definition: *the people of Great Britain*

Hypernym: *{nation:2, land:8, country:3 a people:1}*
whereas the phrase *two Brits* in *Two Brits were rescued* is semantically equivalent to:

Synset: *{Britisher:1, Briton:1, Brit:1}*

Definition: *a native or inhabitant of Great Britain*

Hypernym: *{European:1}*

Hence, on coming across similar instances in the corpus, one should bear in mind this distinction and correctly assign the appropriate sense. It should be noted that since different lemmas (sg. and pl.) correspond to the considered literals, as well as to their Bulgarian counterparts, and lemmas are the mediator between the BulNet entries and the annotation tool, the appropriate lemmatization of the words in the corpus is a prerequisite for the generation of correct lists of choices. To be more particular, if the *Brits* in the first sentence is lemmatized as *Brit,* the synset featuring *Brits* will not be on the list of choices at all. The functionality of the annotation tool allows the system's update on manual corrections in the corpus if necessary and new set of choices is subsequently generated.

6.1.6. Appropriateness with respect to the available senses encoded in PWN

While the criteria (1-5) refer to the exploration of the senses already encoded in BulNet, this one applies mainly to the cases where the corpus occurrence might not be an instance of any of the senses present in the Bulgarian database.

For example, *nature* in the sentence *Nature has taken care of us for centuries and we are still discovering her many wonders* is not an instance of any of the senses encoded for *nature*, discussed above. On exploring PWN one can see that the sense *nature:2* corresponds precisely to the meaning of the word in the sentence:

Synonyms: *{nature:2}*

Definition: *a causal agent creating and controlling things in the universe*

Usage: *Nature has seen to it that men are stronger than women.*

Hypernym: *{causal agent:1, cause:4, causal agency:1}*

Appropriateness of the choice is also considered with respect to specific cases of language use. It is not infrequently the case that a more general and a terminological sense are overlapping. Therefore, special consideration to the type of annotated text should be involved in choosing between senses such as:

Synonyms: *{water:1, H2O:1}*

Definition: *binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid; freezes into ice below 0 degrees centigrade and boils above 100 degrees centigrade; widely used as a solvent*

Hypernym: *{binary compound:1}*

Hypernym: *{liquid:3}*

Synonyms: *{water:6}*

Definition: *a fluid necessary for the life of most animals and plants*

Usage: *he asked for a drink of water*

Hypernym: *{food:1, nutrient:1}*

6.2. Expanding the knowledge base - BulNet

The knowledge base BulNet is expanded in two principal directions: encoding of new entries found in PWN where a relevant occurrence in the corpus requires that in compliance with criterion 6.1.6, and encoding of BulNet-specific entries which fall into several categories:

6.2.1. Culture-specific concepts

In the development of the individual wordnets the BWN adopted the hierarchy of concepts and the structure of the relations established in the construction of the English WordNet. Hence, a strong rule for the preservation of the PWN structure has been strictly observed as a way of ensuring a proper cross-lingual correspondence and navigation via the ILI. Naturally, not all concepts stored in the ILI are lexicalized in all languages, and besides, there are language-specific concepts that might have no ILI equivalent. The structure preservation rule requires that in the first case empty synsets be created (called non-lexicalized synsets) in the wordnets of the languages that do not lexicalize the respective concepts. Thus, the non-lexicalized synsets preserve the hierarchy and cover the proper cross-lingual relations. In the second case culture-specific concepts not featuring in the English database are encoded such as:

Synonym:*{bogomilstvo:1}*

Definition: *an orthodox heretic sect founded by the Bulgarian priest Bogomil*

Hypernym: *{heresy:2 unorthodoxy:2}*

The adopted methodology for the incorporation of such concepts involved the further extension of the ILI with new records. The language-specific concepts shared among Balkan languages were linked via a BILI (BalkaNet ILI) index (Tufis et al., 2004). The initial set of common Balkan specific concepts consisted mainly of concepts reflecting the cultural specifics of the Balkans (family relations, religious objects and practices, traditional food, clothes, occupations, arts, important events, measures, etc).

6.2.2. Language-specific instances of lexicalization

Beside culture-specific concepts the semantic annotation involves the encoding of single words or

MWEs that are not lexicalized in English, for example: *stamva se* whose English counterpart is *get dark* and is not present in the PWN, as contrasted with its antonym *samva se - dawn* which is lexicalized in English. A systematically occurring case is presented by ingressive verbs in Bulgarian formed with a prefix which correspond to compositionally formed expressions in English.

There are four morpho-semantic relations included in PWN and mirrored in EWN and BWN, *Be in state*, *Derivative*, *Derived* and *Participle* (Koeva, 2004). These relations semantically link synsets although they can actually be encoded between pairs of literals (graphic and compound lemmas). There are systematic morpho-semantic differences between English and Slavic languages such as certain derivational mechanisms for forming classifying adjectives, gender pairs and diminutives. The Slavic languages possess rich derivational morphology which has to be incorporated into the strict one-to-one mapping with the ILI.

A productive derivational feature is the formation of classifying adjectives from Bulgarian nouns with the general meaning *'of or related to the noun'*. For example, the Bulgarian adjective

Synset: *{stomanen*:1}

Defintion: *made of or related to steel*

is expressed in English by the respective noun used attributively (rarely at the derivational level, consider *wooden ↔ wood, golden ↔ gold*), thus the concepts exist in English and the mirror nodes have to be envisaged.

The gender pairing is a systematic phenomenon in Bulgarian and other Slavonic languages that display binary morpho-semantic opposition: male ↔ female, and as a general rule there is no corresponding concept lexicalized in English. The derivation is applied mainly to nouns expressing professional occupations. For example, the masculine nouns in the Bulgarian synset *{prepodavatel:2, uchitel:1, instruktor:1}* corresponding to the English:

Synset: *{teacher:1, instructor:1}*

Definition: *a person whose occupation is teaching*

have their female gender counterparts *{prepodavatelka, uchitelka, instruktorka}* with a feasible definition '*a female person whose occupation is teaching*'.

Diminutives are another language-specific derivational class used to express concepts that relate to small things. The diminutives display a sort of morpho-semantic opposition: big ↔ small, however sometimes they may express an emotional attitude, too. Thus the following cases can be found with diminutives: standard relation big thing ↔ small thing ↔ small thing to which an emotional attitude is expressed, consider *{stol:1}* corresponding to English :

Synset: *{chair:1}*

Definition: *a seat for one person, with a support for the back*

and *{stolche}* with an feasible meaning '*a little seat for one person, with a support for the back*' and *{stolchence}* with a meaning '*a "dear" little seat for one person, with a support for the back*'.

6.2.3. Missing English senses and unaccounted systematic differences between senses

Cases where an English sense (attested in dictionaries and known to be the lexical equivalent of a particular Bulgarian sense) is not present in PWN fall into this category. A prominent example is presented by causative and inchoative verbs, which although in general are encoded in PWN and interrelated by means of the Causes relation, are sometimes either mingled or not represented at all:

Synset: *{modernize:2, modernize:1, develop:11}*

Definition: *become more technologically advanced*

The corresponding transitive verb used in *They modernized the cities* does not feature in PWN. In this case our approach is to encode the sense as a separate synset and link it through the Causes relation to its transitive or intransitive counterpart.

6.2.4. Closed word classes

For the purposes of WSD, BulNet is artificially being expanded to incorporate in a systematic way the classes of prepositions, conjunctions, pronouns, particles, modal verbs, etc. The distinction between the senses is based on the analysis of the syntactic evidence and the semantic features observed in the tagged corpus and the senses registered in different Bulgarian lexicographic and grammatical works.

The existing classifications of closed-word classes are sometimes overlapping, not precise enough or based on unclear criteria. This necessitated the elaboration of classifications for the different function word classes that give an adequate account for the sense distinctions found in language use. For example, high-granularity sense distinctions for the class of prepositions has been initiated, based on semantic roles, such as instrument, location, direction, addressee, etc. Thus, for example, one of the 22 senses encoded for one of the highly polysemous Bulgarian preposition *{na}* is defined in the following way:

Synset: *{na:4}*

Definition: *a preposition that introduces the receiver or addressee or beneficiary, etc. of the action*

For some of the closed-word classes, existing entries in PWN have to be considered, to ensure consistency between the Bulgarian and the English databases. The traditional classification of the Bulgarian pronominal system subsumes classes of words with adjectival or adverbial functions whose English equivalents are encoded as adjectives or adverbs, respectively. For example the senses of the Bulgarian demonstrative pronoun *takav* correspond to the synsets:

Synset: *{such: 2; such that: 1}*

Definition: *of a degree or quality specified (by the `that' clause)*

Sunset: *{such: 1; such as:1}*

Definition: *of a kind specified or understood*

Synset: *{such:3; so much:1}*

Definition: *of so extreme a degree or extent*

6.2.5. Proper names

Different types of proper nouns denoting unique entities are encountered in the corpus – person names, geographical names, names of institutions, companies, etc. Certain proper nouns, including anthroponyms signifying famous persons, are encoded in the English WordNet - for example:

Synset: *{Ploviv:1}*

Definition: *the second sized town in Bulgaria*

Regional and Bulgarian proper nouns of historical or social or political significance in case they are not included in PWN are encoded either as Balkan-specific concepts (BILI) or as Bulgarian-specific concepts (BUL) - for example:

Synset: *{Ivan Vazov:1; Iv. Vazov:1; Vazov:1; Ivan Minchov Vazov}*

Definition: *a famous Bulgarian writer, publicist and public figure.*

Otherwise, they are linked to the general term according to their referent e. g. *John* is connected to the synset *{first name:1, given name:1, forename:1}*.

6.2.6. Multi-word expressions

Multi-word expressions are linguistic units consisting of more than one distinct lexeme. They are incorporated in the Bulgarian WordNet in a similar way as single words, their POS having the same value as the head word of the expression. Decisions for the encoding of MWE in BulNet are taken according to the consistency criteria. The statistical data coming from the wordnets shows that the distribution of multiword expressions among natural languages is approximately equivalent and covers one forth of the lexis.

Following in part the existing literature, we adopt the following classification for phrases: free combinations of words, idioms and multi-word expressions. We consider a MWE a sequence of two or more words (including graphical words) that denotes a unique and constant concept. Idioms and idiomatic expressions are a lexicalized word group, whose meaning is not compositionally formed from the meanings of its components. Idioms can be part of the wordnet if they denote a unique concept, not a proposition.

An important class of syntactically-flexible MWEs are the so-called support (or light) verbs such as *do, give, have, take, etc.* which combine with certain nouns to express the same meaning as the corresponding lexical verb. They are either encoded as synonyms of the respective content verbs, for example *{ucahstvam:2, vzemam uchastie:1}*:

Synset: *{participate:1, take part:1}*

Definition: *share in something*

or annotated separately. Our approach is to follow the way the Princeton WordNet handles these expressions while at the same time considering factors such as the productivity (degree of collocativity) of the support verbs and taking into account whether it is the same or different support verbs that participate in the formation of semantically equivalent collocations in English and Bulgarian as a way to ensure correspondence between the senses of the support verbs in the two languages where appropriate.

An interesting case is presented by idioms. Some of them are the result of cultural interaction - the Bulgarian counterparts are loan translations of English expressions, for example *take the bull by the horns, close at hand, etc.* Others are functional equivalents and are therefore encoded in the synset representing the relevant meaning, for example *odera kozhata* and *svalyam rizata ot garba* are entered in the Bulgarian counterpart of the English:

Synset: *{overcharge:1, soak:2, surcharge:2, gazump:2, fleece:1, plume:1, pluck:3, rob:2, hook:2}*

Definition:*rip off; ask an unreasonable price*

Bularian idioms which have no idiomatic equivalents in English are encoded as hyponyms to an entry with roughly the same meaning. For example, the BulNet entry

Synset:*{med mi kape na sartseto:1}*

Definition: *be very delighted*

is encoded as a hyponym of *naslazhdavam se { delight:2, enjoy:5, revel:1}*

In the course of annotation the components of the multi-word expressions are grouped and linked to the corresponding wordnet synsets. The lemmas of the MWEs account for the grammatical features (e. g. adjective noun agreement) of the constituents and need not coincide with the lemmas of the individual words. For example: in *familna istoriya (family history)* the gender and person of the adjective *familna* agree with the feminine noun *istoriya* and is lemmatised both in the corpus and in the wordnet entry in its feminine singular form:

BulSemCor: familna{familna #ENG20-06112790-n 1144167285 5770 1} istoriya{istoriya#0 2000000000 5769 0}

Synset:Literal:*{familna istoriya:1}* Lemma:familna istoriya

The BulNet entries of MWEs reflect the neutral word order of the constituents where variations are possible as with idioms, collocations, etc. These features are handled at the stage of annotation.

6.2.7. Domain relations

With a view to modeling the tagged corpus into the Hidden Markov Model (HMM) WSD framework certain kinds of optimizations had to be implemented. A significant one was the association of adverbs with a semantic domain to which they pertain, following the PWN methodology of association of domain-specific words with the corresponding domain through the relation Category domain. All adverbs were linked to a synset corresponding to their semantic domain (such as time, location, manner, quantity, degree, frequency, etc.). For example the Bulgarian synset *{na zakrito}* corresponding to *{inside:1, indoors:1}* *'within a building'* is connected through the relation Category domain to the Bulgarian equivalent of the synset *{location:1}* - *'a point or extent in space'*. Further, grammatical peculiarities and syntactic function of certain items such as intensifiers, quantifiers, etc. are accounted through linking these items to the relevant domain synset by means of the relation Usage domain.

7. Evaluation

The evaluation of the Bulgarian Sense Tagged Corpus at this stage is performed manually by a second annotator. Further strategies of evaluation have to be developed in order for the consistency of the annotation to be guaranteed. The evaluation of BulSemCor is performed with respect to both the consistency and completeness of the corpus against the wordnet. The completeness check up has to take into account the following considerations:

There is still a large number of wordnet senses that are not mapped in BulSemCor, thus BulSemCor can be further enlarged with texts that include such words;

We may consider separately single-sense and multiple-sense words (as found in BulNet); this may reflect on the weights given to those categories.

Since the senses encoded in BulNet reflect largely the definition of senses in PWN, we may additionally perform experiments to estimate the number of senses attested in the existing lexicographic works, such as the Bulgarian explanatory dictionaries, that are mapped in BulSemCor.

8. Acquisition of multilingual Sense tagged corpora

The Bulgarian sense tagged corpus underlies an HMM formalism combined with additional operations over the wordnet (for the time being relatively low recall but high precision has been achieved) implemented for word sense disambiguation.

BulSemCor will provide an appropriate WSD foundation for a number of future purposes with a special focus on machine translation which is currently poorly explored for minority languages such as Bulgarian. For this purpose the Bulgarian Sense Tagged Corpus will be translated in English (this is also possible for any other language for which a WordNet is constructed). The resulting English corpus will be lemmatized and sentence aligned with the Bulgarian source corpus. Then, to every lemma from the corpus located in the English WordNet a corresponding identification number can be automatically assigned. This is one possible methodology among others (Bentivogli & Pianta, 2005) for obtaining a parallel sense-tagged corpus for Bulgarian and English (as well as for other language).

It has been noted that the sophistication of the statistical methods used in MT makes use of linguistic information (Hutchins, 1995) at different levels. An indispensable step in the further work is providing a proper basis for enhancement of the system in this direction. This will involve the encoding of different types of metalinguistic information in the corpus as well as the elaboration of approaches towards handling specific classes of words not encoded in dictionaries and units above the word level.

One major class to be considered is that of the named entities. Beside proper names (see section 6.2.5.) named entities subsume also locations (place names), company names, organizations names, etc. This is a heterogeneous group which will require different handling for the purposes of MT - transliteration, translation, etc. The task is even more challenging since named entities may incorporate units that require different type of rendering in another language, e.g. in *Емакар ООД*, the first part (*Емакар*), being the name of the company is transliterated (*Emakar*) while the second part of the name (*ООД*) denotes the type of company and is translated (*Ltd.*).

Another task whose relevance to MT has been acknowledged is "the use of syntactic transformations to bring source structures closer to those of the target language" (Hutchins, 1995). The task actually consists in finding functional equivalents of phrases and constructions and can be used in combination with the example-based approach where models are learned from actual expert translations of the same text.

9. Conclusions

The Bulgarian Sense Tagged Corpus contains 63 440 words, part of them linked to form MWE. Three-fourths of the corpus have been annotated and the results have been employed in the experiments on developing a WSD system. Our immediate goal is to complete the task of the annotation of the presented corpus, as well as to carry on enlarging it with more data. The next selection for annotation from BCB has to take into account not only the frequency, but also the already defined BulNet senses, especially those with more than one sense.

In the longer run, as noted in Section 1, our sense-annotated corpus will be employed as training and test dataset for a bidirectional machine translation system based on HHM.

Along with their immediate applications the MT platforms from and to minority languages will ensure these languages' equality at the international level. The experience gained in the elaboration of BulSemCor will be helpful to any future effort in this field and will further national and international cooperation in the creation of tools and resources for minority languages.

10. References

Banko, M., Brill, E. (2001). *Scaling to Very Very Large Corpora for Natural Language Disambiguation.* In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. *ACL*, pp. 26-33.

Bentivogli, L., Pianta, E. (2005). *Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus.* In Natural Language Engineering, Special Issue on Parallel Texts, Volume 11, Issue 03, September 2005, pp. 247-261.

Budanitsky, A., Hirst, G. (2001). *Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measure.* In Proceedings of the Workshop on WordNet and Other Lexical Resources, North American Chapter of the Association for Computational Linguistics. Pittsburgh, pp. 29-34.

Fellbaum, C. (1998). *Towards a representation of idioms in WordNet.* In Proceedings of the Workshop on the Use of WordNet in Natural Language Processing Systems (Coling-ACL 1998). Montreal, pp. 52-57.

Fellbaum et al. (1998). Fellbaum, C., Grabowski, J. and Landes, S. (1998). Performance and confidence in a semantic annotation task. In Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database.* Cambridge (Mass.): The MIT Press, pp. 217-237.

Hutchins, W. John (1995). Machine translation: a brief history. In E.F.K.Koerner and R.E.Asher (Eds.), *Concise history of the language sciences: from the Sumerians to the cognitivists.* Oxford: Pergamon Press, 1995. pp. 431-445

Koeva et al. (2004). Koeva, S., Tinchev. T., Mihov, S. *Bulgarian WordNet-Structure and Validation.* In Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004, pp. 61-78.

Koeva et al. (2005a). Koeva, S., Rizov, B., Leseva S. *Flexible Framework for Development of Annotated Corpora.*.In International Journal Information Theories & Applications, Sofia.. [In press].

Koeva et al. (2005b). Koeva, S., Krstev, C., Obradovic, I., Vitas, D. *Resources for Processing Bulgarian and Serbian.* In Proceedings from the International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries. Borovets, 2005, pp. 31-39.

Stamou et al. (2002). Stamou S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., Grigoriadou, M. *BALKANET: A Multilingual Semantic Network for the Balkan Languages.* In Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002, pp. 12-14.

Tufis et al. (2004). Tufis, D., Cristea, D., Stamou, S. *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview.* In Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004, pp. 1-32.

Statistical machine translation using the IJS-ELAN Corpus

Mirjam Sepesy Maučec and Zdravko Kačič

Faculty of electrical Engineering and Computer Science

Smetanova 17, 2000 Maribor, Slovenia

E-mail: mirjam.sepesy@uni-mb.si, kacic@uni-mb.si

Abstract

In this paper, we describe our experiments on statistical machine translation from Slovenian to English performed on IJS-ELAN bilingual corpus. Four different experiments are reported. They differ from each other in the type of information used in training the translation model. In first experiment word forms of sentence-aligned corpus are used as modelling units. Second experiment uses lemmas instead of word forms in Slovenian part of the corpus. In the third experiment word forms are replaced by lemmas and MSD codes attached to them. Fourth experiment tries to combine the advantages of previously performed ones. Only publicly available tools are used for training language model and translation model, as well as for decoding the test set. The results are evaluated by automatically calculated WER.

1. INTRODUCTION

The paper discusses corpus-based machine translation. The use of corpora of bilingual parallel texts seems to offer a promising tool for the future, thanks to the progress that has been made in the field of statistical machine translation. Various methods have been proposed for processing the different levels of correspondence between two texts, an original and its translation. In this paper well grounded methods of statistical machine translation are tested on the Slovenian-English language pair.

To our knowledge, "pure" statistical machine translation of Slovenian language has not been widely studied yet. There exist only one thesis (Vičič, 2002), which gave us fist impression about the topic

2. STATISTICAL MACHINE TRANSLATION

The problem is to find the optimal English translation $\hat{e}$ of a Slovenian sentence f . Statistical machine translation is referred to as the noisy channel model:

$$\hat{e} = \arg \max_{e} P(f|e) \cdot P(e) \qquad (0.1)$$

$P(f|e)$ denotes the translation model and $P(e)$ the language model of English language. The search for the optimal translation is encompassed by $\arg \max$. Language model is a conventional trigram model ^{e}with Katz back-off smoothing. This paper deals with translation model.

3. TRANSLATION MODELS

Translation model defines the correspondence between the words of the target sentence and the words of the source sentence. Translation model is a generative model, because it is a theory how Slovenian sentences are generated. First an English sentence is generated, and then it gets turned into a Slovenian one. Although we are building a Slovenian-to-English machine translation system, we reason in the opposite direction when training the translation model.

Translation models, defined by IBM in the early nineties, are used in our research. They are well documented (Brown at all, 1993) and the software for training them is publicly available (Och & Ney, 2003). The models are indexed from 1 to 4 according to their increasing complexity in training. The parameters are transferred from one model to another, for example from Model 2 to Model 3. It means that final parameter values of Model 2 are the initial parameter values of Model 3. Models 4 and 5 are the most sophisticated.

3. MODEL 4

Translation model is a word-for-word alignment model between two strings, an English string $e_1^I = e_1,...,e_I$ and a Slovenian string $f_1^J = f_1,...,f_J$. Because alignments are not known, the probability $P(a, f|e)$ for each particular alignment a is computed.

Model 4 is the final model, which is used by the decoder. It is composed of the following probabilities:

- $P(f_j|e_i)$ - translation probability. It is the probability of Slovenian word f_j being a translation of English word e_i.
- $P(\phi_k|e_i)$ - fertility probability. An English word can be translated into zero, one or more than one Slovenian word. This phenomenon is modeled by fertility. The fertility $\phi(e_i) = \phi_k$ of an English word e_i is the number of Slovenian words mapped to it. The probabilities of different fertility values ϕ_k for a given English word are trained.

- p_0, p_1 - fertility probability for e_0. The word e_0 is an invisible word in the initial position of an English sentence. It accounts for Slovenian words that have no counterpart in the English sentence. Instead of fertilities $\phi(e_0)$ one single parameter $p_1 = 1 - p_0$ is used. It is the probability of putting a translation of word e_0 onto some position in a Slovenian sentence.

  Between words being a translation of the same word we distinguish a head word and non-head words. Head word is the first word in the translation. All other words are non-head words.

- $P_{=1}\left(\Delta j \middle| \boldsymbol{A}(e_i), \boldsymbol{B}(f_j)\right)$ - distortion probabilities for the head word. Δj is the distance between the head of current translation and the previous translation. It may be either positive or negative. Distortion probabilities model different word order in the target language in comparison to the word order in the source language. Classes of words are used instead of words. $\boldsymbol{B}$ denotes mapping into classes for Slovenian words and $\boldsymbol{A}$ for English words.
- $P_{>1}\left(\Delta j \middle| \boldsymbol{B}(f_j)\right)$ - distortion probabilities for non-head words. In this case Δj denotes the distance between the head and non-head word.

Model 4 has some deficiencies. Several words can lie on top of one another and words can be placed before the first position or beyond the last position in the Slovenian string. An empty word also causes the problems. Training results in many words aligned to the empty word. Model 5 is a reformulation of Model 4 in order to overcome some problems. An additional parameter is trained. It denotes the number of vacant positions in the Slovenian string. It is added to the parameters of the distortion probabilities. In our experiments Models 4 and 5 will be trained, but only Model 4 will be used when decoding. Model 5 is not yet supported by the decoding program.

4. IJS-ELAN CORPUS

The translation system was tested on the IJS-ELAN corpus (Erjavec, 2002)[1]. The corpus has parts, which have a Slovenian origin and an English translation, and parts with origins in English and translations in Slovenian. The corpus is encoded in XLM/TEI P4. It is aligned at the sentence level, tokenised, and the words are annotated with disambiguated lemmas and morpho-syntactic descriptions (MSD). All annotations were done by the authors of the corpus.

We observed that the English part contains 18% more words than the Slovenian part. The average English sentence is 3 words longer than the average Slovenian sentence. One reason lies in determiners and pronouns. The subject pronouns in English (I, he, they) usually have a zero form in Slovenian.

The Slovenian corpus contains twice as many unique words than the English corpus; this is because of the highly inflectional nature of the Slovenian language. Almost half the words are singletons (they appeared only once in the training corpus). The data exposes the problem of data sparsity of the corpus and indicates the difficulty of the translation process.

5. EXPERIMENTS

5.1 TRAIN AND TEST SET

We discarded sentences longer than 15 words from the corpus because of the computational complexity. The rest of the corpus was split into training and test sets in the ratio 8:2. The test sentences were taken at regular intervals from the corpus (homogeneous partition). Some statistics of the training corpus are collected in Table 1. The training set contained 12,044 sentence pairs. Each appearance of a word or any other string of characters between two spaces is counted as one unit. The Slovenian part was 86,036 units long and the English part contained 97,062 units. The test set consisted of 3,069 sentences.

The vocabulary contained all those words (units), which appeared in the training set or in the test set. Almost half of the vocabulary units were singletons. Zerotons are units, which do not appear in the training corpus, but occur in the test set. These units not only remained untranslated but also "added noise" to the translation process of other words.

5.2 TOOLS

All experiments have been performed using only publicly available third-party tools. The language model was made by using the CMU-SLM toolkit (Rosenfeld, 1995). Classes of words were made by the tool, developed for language modelling (Maučec, 1997). The translation model training was performed using a program GIZA++ (Och, 2003). The decoding of test sentences was performed by an ISI ReWrite Decoder (Germann, 2003). Translations were evaluated using Word Error Rate metric (WER).

5.3 FIRST EXPERIMENT

In our first experiment all word forms appeared as unique tokens and were exposed as candidates for word-for-word alignments.

Before training, Slovenian words were mapped into 1000 classes and English words into 100 classes. The numbers of classes were predefined and were chosen to be the same as the number of different MSD codes in the corpus.

For English language a conventional trigram language model was built with Good-Turing discounting for bigrams and trigrams with counts lower than 7. No n-grams were discarded. Training corpus was the whole English part of IJS-ELAN corpus. It is relatively small, so there are a lot of singletons with significant information. The language model perplexity of the test set was 48.

[1] The authors are thankful for giving the corpus freely available.

	SLO part	ENG part
Sentences	12,044	
Units	86,036	97,062
Vocabulary	22,055	12,715
- singletons	11,355	5,611
- zerotons	2,566	1,274

Table 1: Training corpus for the first experiment.

For each translation model (Model 1 – Model 4) 10 iterations were performed.

After training some interesting observations were made (see Figure 1). Although the training-set perplexity continuously decreased, the test-set perplexity jumped at each transition point. At the transition point, the final estimates of one model initialized the estimates of the next model. In subsequent iterations after transition points, the test-set perplexity slowly increased, especially in Models 1 and 4. Each iteration of Model 4 made the test-set perplexity worse. The only exception was the transition to Model 5, although the estimates never improved the estimates obtained at the beginning of the training. The same observations have also been reported in Czech-English experiments (Al-Onaizan at all, 1999). We speculated that the reason was the small size (and consequently data sparsity) of the training corpus, so the translation probabilities become over-trained. Better alignments for training-set did not lead to better translations of previously unseen test-set.

In the first experiment only 37.5% of words got the correct translation. We obtained WER=71,6% (see first row in Table 4).

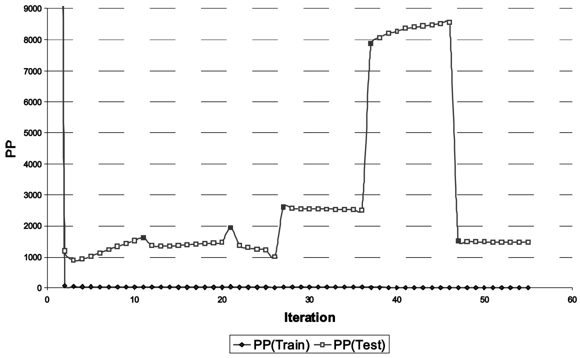


Figure 1: Train set and test set perplexities in first experiment

5.4 SECOND EXPERIMENT

The purpose of the second experiment was the reduction of data sparsity.
In the second experiment we used lemmatised Slovenian part of the corpus. English part remained unchanged. Lemmatising the Slovenian corpus reduced the data sparsity to a great extent (see Table 2).

	SLO part	ENG part
Sentences	12,044	
Units	86,036	97,062
Vocabulary	12,629	12,715
- singletons	5,654	5,611
- zerotons	1,292	1,274

Table 2: Training corpus for the second experiment.

New clustering of Slovenian lemmas was performed. The Slovenian lemmas were automatically clustered into 100 classes. Because Slovenian part of the corpus was lemmatised, there was no need to use extended set of classes.

The GIZA++ training was repeated. Comparison of train-set and test-set perplexities confirmed our assumptions about Model 4 over-training in first experiment (see Figure 2). In the second experiment each iteration of Model 4 training made a slight reduction of the test-set perplexity. Transition to Model 5 brought further improvements.

In the second experiment we obtained WER=68.5% (see second row in Table 4). Although some information was lost by lemmatization, the data sparsity reduction improved the results.

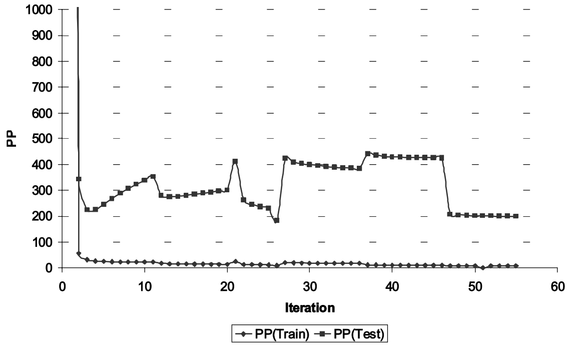


Figure 2: Train set and test set perplexities in second experiment

5.5 THIRD EXPERIMENT

In the third experiment we want to examine what is the influence of morpho-syntactic information on the translation success.

Slovenian words were replaced by lemmas and MSD codes attached to them. In this experiment we expose the problem of homographs. For example the word gori was replaced by goreti_[VMIP3S–N] and by gori_[RGP]. In this experiment we increase the data sparsity, so worsening of the translation results was expected. The data sparsity of third experiment is evident from Table 3.

	SLO part	ENG part
Sentences	12,044	
Units	86,036	97,062
Vocabulary	27,030	12,725
- singletons	14,809	5,645
- zerotons	3,409	1,272

Table 3: Training corpus for the third experiment.

Two versions of third experiment were performed. In the first one classes were made automatically. In this version of the experiment we obtained WER=75.5% (see third row in Table 4).

In the second version of third experiment classes of words were defined by MSD codes. We had 1100 different MSD codes. In this case we obtained WER=71.5% (see fourth row in Table 4).

The results of third experiment showed the problem of data sparsity. Data sparsity is even greater than shown at the beginning. The results obtained with MSD codes gave us the anticipation that MSD code is well correlated with the position of word in sentence.

5.6 FOURTH EXPERIMENT

In the fourth experiment the influence of MSD codes was further examined. We used them to improve distortion probabilities of our first experiment. The experimental setup remained unchanged, only automatic classes were replaced by MSD classes. Distortion probabilities of words depended directly on morpho-syntactic features of words. In this experiment we obtained WER=69.9% (see fifth row in Table 4).

One additional version of this experiment was performed. In this experiment translation probabilities $P\left(f_j \middle| e_i\right)$ were imported from the second experiment, where they were learned on lemmas. Each lemma-based probability value was assigned to all word forms, which belong to that lemma. The probabilities were normalized afterwards.

In the last experiment we obtained WER=68.9% (see last row in Table 4). This result is only slightly worse than the result obtained by lemmas. This experimental setup did not need the lemmatiser in decoding phase.

	CORR (%)	WER (%)
1. EXP.	37.5	71.6
2. EXP.	40.6	68.5
3. EXP. (a)	33.0	75.5
3. EXP. (b)	35.9	71.5
4.EXP. (a)	38.9	69.9
4.EXP. (b)	40.0	68.9

Table 4: Final results of all experiments

6. CONCLUSION

In this paper, we have discussed different types of translation model. The problem of data sparsity was outlined.

From the experiments we concluded that using lemmas is a good starting point for data sparsity reduction. On the other hand it has been shown that MSD codes are of great value. In the future we will examine how to use them more reasonable. Not all the information in MSD codes is important for translation. We would like to reduce its content just to the important parts.

7. References

Al-Onaizan, Y., Curin, J., Jahr M., Knight K., Lafferty, J., Melamed D., Och F. J., Purdy D., Smith N. A., Yarowsky D. (1999). Statistical Machine Translation. Final report, JHU Workshop.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J. And Mercer, R. L. (1993). The mathematic of statistical machine translation : Parameter estimation.. Computational Linguistic, 19(2):263:311.

Erjavec, T. (2002). Compiling and Using the IJS-ELAN Parallel Corpus. Informatica, Vol. 26.

Germann, U. (2003). Greedy Decoding for Statistical Machine Translation in Almost Linear Time. In Proceedings of HLT-NAACL-2003, Edmonton, AB, Canada.

Maučec, M. S. (1997). Statistical language modeling based on automatic classification of words, In Proceedings of workshop : Advances in Speech Technology, Maribor : Faculty of Electrical Engineering and Computer Science.

Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1).

Rosenfeld, R. (1995). The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation. In Proceedings of the ARPA SLT Workshop, Austin, TX.

Vičič, J. (2002). Statistično strojno prevajanje naravnih jezikov, Master thesis.

Lost in Translation: the Problems of Using Mainstream MT Evaluation Metrics for Sign Language Translation

Sara Morrissey[1], Andy Way

National Centre for Language Technology
School of Computing,
Dublin City University,
Dublin 9, Ireland.

IBM CAS Dublin
Mullhuddart,
Dublin 15, Ireland.

{smorri, away}@computing.dcu.ie

Abstract

In this paper we consider the problems of applying corpus-based techniques to minority languages that are neither politically recognised nor have a formally accepted writing system, namely sign languages. We discuss the adoption of an annotated form of sign language data as a suitable corpus for the development of a data-driven machine translation (MT) system, and deal with issues that arise from its use. Useful software tools that facilitate easy annotation of video data are also discussed. Furthermore, we address the problems of using traditional MT evaluation metrics for sign language translation. Based on the candidate translations produced from our example-based machine translation system, we discuss why standard metrics fall short of providing an accurate evaluation and suggest more suitable evaluation methods.

1. Introduction

Large amounts of money and resources are invested in dominant languages both in terms of linguistic analysis and their machine translation (MT). However, such investment serves to increase the prominence and power of these languages and ignores the less dominant, minority languages (Ó'Baoill & Matthews, 2000). Sign languages are the first languages (L1s) of the Deaf community worldwide and, just like other minority languages, are poorly resourced and in many cases lack political and social recognition.

As with other speakers of minority languages, Deaf people are often required to access documentation or communicate in a language that is not natural to them. In an attempt to alleviate this problem, we propose the development of an example-based machine translation (EBMT) system to allow Deaf people to access information in the language of their choice. The language of choice for us is Irish Sign Language (ISL). While corpus creation for English—ISL is ongoing, and we hope to avail of this data in the near future, in order to seed the development of our EBMT system, we have instead used a corpus of Dutch Sign/Language Nederlandse Gebarentaal (NGT) data created by the ECHO project. The annotation scheme used for NGT is the same as for ISL, so we anticipate that migration to ISL will be reasonably seamless.

In this paper we begin by introducing sign languages (SLs) in section two, briefly discussing their current status. Section three provides an overview of related work in the area of sign language machine translation. Section four reviews writing systems available for SLs. This is followed by a discussion on the annotated format we chose in section five. Section six gives a brief overview of EBMT before describing our own approach. We describe in section seven the experiments carried out on our system, and discuss both their evaluation and the problems with traditional evaluation metrics in section eight. Finally we conclude the paper.

2. Sign Language

SLs are the primary means of communication of the Deaf community worldwide. In SLs, the hands are the main articulators and non-manual features (NMFs) such as eyebrow movement, head tilt, and blinks add vital morphological and grammatical detail. SLs are fully formed natural languages that have developed in such a way that full articulatory use is made of the signing space, i.e. the area in front of the signer from waist to head and the extension of the arms in which discourse is articulated (Ó'Baoill & Matthews, 2000).

Most countries have their own native SL, although some are dialects of more widespread languages. Despite the use of these manual languages as the L1s of Deaf communities, in most cases they lack political recognition and are often not recognised as languages at all. As a result, SLs remain less resourced than spoken languages. This is apparent in the areas of SL linguistics and machine translation of SLs. Both are relatively new areas in comparison with their spoken language counterparts. Significant SL linguistic research began about 45 years ago with the work of (Stokoe, 1960) and the earliest papers on research into the machine translation of SLs date back only approximately 15 years.

In Ireland, Irish Sign Language (ISL) is the dominant language of the Deaf Community. As with other SLs, it is grammatically distinct from spoken languages. Despite being in use in Ireland since the 1800s, its status has remained low and a standardized form of the language is not taught to children in Deaf schools in the same way that English is in spoken language schools. The development of the language is

[1] This work is generously sponsored by a joint IBM-IRCSET scholarship.

slow as a result of "its users' lack of access to technical, scientific and political information" (Ó'Baoill & Matthews, 2000).

NGT is the SL of the corpora we use for translation. NGT is the primary language of the Deaf community in the Netherlands with a population of approximately 15,000 deaf. Similar to ISL, NGT was originally derived from French Sign Language and, as is the case in Ireland and many other countries, it has not attained recognition as an official language (Gordon, 2005).

We hope that the development of our MT system with first an NGT, then ISL corpus will help to raise the status of SLs in these countries and facilitate communication of information to the Deaf community in their preferred language.

3. Related Work

Many different approaches have been applied to sign language machine translation (SLMT). As might be expected, most approaches have concentrated on translating from spoken languages to SLs.

3.1. Traditional 'Rule-Based' Approaches

Given that SLMT has been tackled only quite recently, most approaches to date are 'second generation', namely transfer- or interlingual-based.

Many transfer-based translation methodologies have been used. (Grieve-Smith, 1999) uses the domain of weather reporting and uses a literal orthography to represent American Sign Language (ASL) for translation into English by mapping the syntactic structure of one on to the other. No evaluation methods have been used in his work.

Other transfer approaches have been applied in (Marshall & Sáfár, 2002; Sáfár & Marshall, 2002). Their work employs discourse representation structures to represent the internal structure of linguistic objects then uses HPSG semantic feature structures for the generation of ASL. No automatic or manual evaluation is discussed

A more syntax-based transfer approach is described in (Van Zijl & Barker, 2003) in their translation work from English to South African Sign Language. Their focus is on producing a signing avatar for manual evaluation at a later stage.

Interlingual SLMT methodologies have also been employed that use language-independent intermediate representations as the basis of their translation. (Veale et al., 1998) developed the ZARDOZ system, a multilingual sign translation system for English to Irish, American and Japanese Sign Languages using this approach. (Zhao et al., 2000) used an interlingual approach for translating English to ASL and employed synchronized tree-adjoining grammars. Evaluation metrics are not mentioned for either interlingual approach.

(Huenerfauth, 2005) attempts to combine the two previous approaches, transfer and interlingual, with a more simplistic direct approach to create a hybrid "multi-path" approach. This system translates English to ASL using first an interlingual method, then failing that, a transfer then direct approach. His work concentrates on the translation of classifier predicates and will be manually evaluated by native signers.

3.2. Corpus-Based Approaches

The first statistical approach that we are aware of was that of (Bauer et al., 1999), but this is the only model we have come across where translation is not from spoken to sign language. Their approach consists of a video-based recognition tool for a lexicon of 100 signer-dependent German Sign Language signs and a translation tool composed of a translation and language model, which is standard in the statistical MT (SMT) paradigm.

An SMT model for spoken to sign language is described in (Bungeroth & Ney, 2004, 2006) to translate German weather reports into German Sign Language using HamNoSys (Prillwitz, 1989) notation. Their initial experiments are automatically and manually evaluated and show promising results for a data-driven approach.

The first Example-based approach is our own model in (Morrissey & Way, 2005). Using an NGT corpus, we developed a prototype system for translating English and Dutch into NGT. Although traditional evaluation metrics were not employed, through manual analysis of a set of experiments we show that encouragingly good translations were obtained.

4. Writing Systems

When applying EBMT techniques to SLs, the lack of recognition and under-resourcing of SLs, together with their having no formal or widely used writing system, make SL corpora difficult to find. Attempts have been made to develop notation systems for these visual languages, examples of which include Stokoe Notation, HamNoSys and SignWriting.

4.1. Stokoe Notation

Stokoe notation (Stokoe, 1960) was developed in the 1960s for ASL and initially described three factors to be taken into account for SL description, namely *tabulation,* referring to the location of a sign; *designator,* referring to the handshape; and *signation,* referring to the type of movement articulated. SL-specific additions by international linguists over the years including the addition of a fourth factor *orientation,* describing the orientation of the handshape, have resulted in no universally accepted version of the Stokoe notation system (Ó'Baoill & Matthews, 2000). While this approach describes a comprehensive analysis of an SL, the method is data-heavy and not practical for use as a writing system for Deaf people. Furthermore there are no large corpora available in this format for use in a data-driven MT system.

4.2. HamNoSys

Another explicit notation system for SLs is the Hamburg Notation System (HamNoSys) (Prillwitz, 1989) that uses a set of language-independent symbols to iconically represent the phonological features of SLs (Ó'Baoill & Matthews, 2000). This system allows even more detail than that of the Stokoe system to be described including NMFs and information about the signing space. For reasons similar to those above, this system is not suitable for adoption by the Deaf community as a writing system and again, no large SL corpora are available in this format.

4.3. SignWriting

An alternative method was developed in (Sutton, 1995) called SignWriting.[2] This approach also describes SLs phonologically but, unlike the others, was developed as a handwriting system. Symbols that visually depict the articulators and their movements are used in this system, where NMFs articulated by the face (pursed lips, for example) are shown using a linear drawing of a face. These simple line drawings make the system easier to learn as they are more intuitively and visually connected to the signs themselves. The SignWriting system is now being taught to Deaf children and adults worldwide as a handwriting version of SLs. The system is not yet widely used but its usability and the rate at which it is being adopted suggests that corpora may be available in the near future on suitable topics for MT.

4.4. Manual Annotation

One way around the problems with writing systems is to manually annotate SL video data. This approach involves transcribing information taken from a video of signed data. It is a subjective process where a transcriber decides the level of detail at which the SL in the video will be described. These categories can include a gloss term of the sign being articulated by the right and left hands (e.g. HARE if the current sign being articulated is the sign for the animal *hare*), information on the corresponding NMFs, if there is repetition of the sign and its location. The annotations are time-aligned according to their articulation in the video. As the process is subjective, the annotation may be as detailed or as simple as the transcriber or project requires. On the one hand, this makes annotations suitable for use with corpus-based MT approaches as they are not loaded with linguistic detail and can provide gloss terms for signs that facilitate translation from and into spoken language. On the other, however, the problem of inter-annotator agreement remains; discrepancies in the training data will hinder the capacity of the corpus-based MT system to make the correct inferences.

5. Annotated Corpora for EBMT

A prerequisite for any data-driven approach is a large bilingual corpus aligned at sentence-level from which to extract training and testing data. For translation between major spoken languages, such data is available in large amounts: in the recent OpenLab[3] evaluation, we used almost 1 million aligned Spanish—English sentence-pairs from the Proceedings of the European Parliament (Koehn, 2005) to seed our MaTreX system (Armstrong et al., 2006). While this is the largest EBMT system published to date, many SMT systems use much larger training sets than this, e.g. the Chinese-English SMT system of (Vogel et al., 2003) is trained on 150 million words.

5.1. Dutch Sign Language (NGT) Corpora

As discussed above, finding corpora suitable for the task we are confronted with in this paper can be difficult. However, a collection of annotated SL data—albeit on a much smaller scale than the training sets typical of data-driven approaches—has been made available through the ECHO project.[4] This EU-funded scheme, based in the Netherlands, has made fully annotated digitised corpora for Dutch Sign Language (NGT: Nederlandse Gebarentaal) available on the Internet. The corpora have been annotated using the ELAN annotation software.[5]

ELAN provides a graphical user interface in which corpora can be viewed in video format with their corresponding aligned annotations (cf. Figure 1). The name of the annotation category tiers may be seen in the column on the left and the time-aligned annotations for each tier are displayed horizontally in line with the tiers.

Annotation has been included that displays a time-aligned translation in the native spoken language and in English. Further annotation groupings include a gloss in both spoken languages of the signs of both hands and various NMF descriptions. An example of some annotations used in the NGT corpus can be found in (1) (where numbers indicate time frame of annotation):

```
(1) 3:09:500 3:10:380
    (Gloss RH/LH English) TINY CURLS

    3:09:500 3:10:380
    (Gloss RH/LH) PIJPENKRULLEN

    3:09:500 3:10:380
    (Repetition) u

    3:09:740 1461310
    (Eye Gaze) l, d
```

Such suitably annotated corpora can be reasonably useful for an example-based approach to SLMT. Accompanying English and Dutch translation tiers and time-aligned annotations allow for easy alignment of corpora on a sentential level. The presence of time frames for each annotation also aids in the aligning of annotations from each annotation tier to form chunks that can then be aligned with chunks derived from the English/Dutch tier. As simultaneity (articulators signing two separate signs at the one time) and co-articulation (articulation of one sign being influenced by its neighbouring signs) are prevalent in natural signing, time-aligned annotations help tackle this issue by providing time boundaries to signs and NMFs so that each annotation remains complete and separate. As it is these annotations that are used in the translation output, once a satisfactory boundary width has been established, the issue of separating co-articulated words is removed automatically.

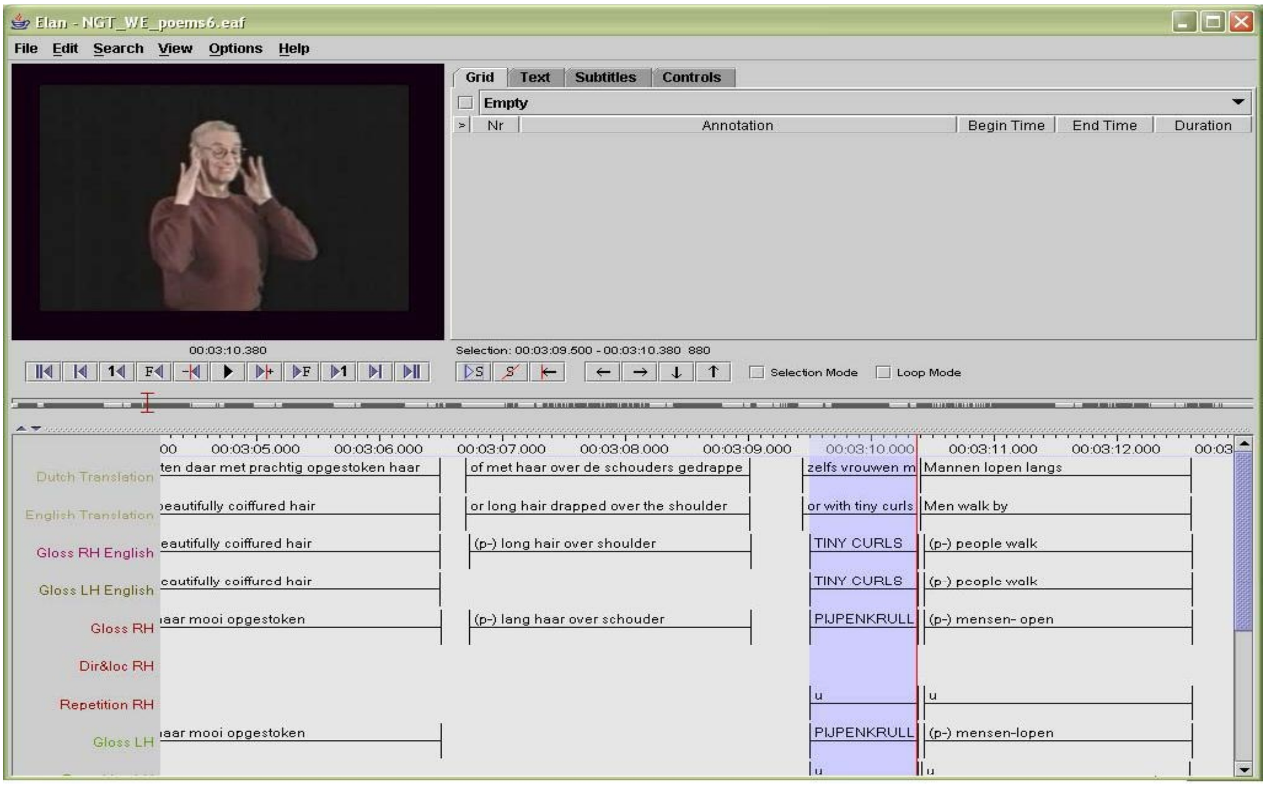
While we were grateful to avail of the ECHO data, there are two main problems with it: firstly, the data consists of annotated videos of two versions of Aesop's Fables and an NGT poetry file—this is hardly the most suitable genre for *any* MT system. Secondly, despite combining all NGT data files available, the corpus amounted to a mere 40 minutes of data, or just 561 sentences. This small corpus size obviously results in data sparseness; for any data-driven approach, the larger

[2] http://www.signwriting.org

[3] http://www.tc-star.org/openlab2006/

[4] http://www.let.kun.nl/sign-lang/echo/data/html

[5] http://www.mpi.nl/tools/elan.html



Dutch Translation: ten daar met prachtig opgestoken haar | of met haar over de schouders gedrappe | zelfs vrouwen m | Mannen lopen langs

English Translation: ɔeautifully coiffured hair | or long hair drapped over the shoulder | or with tiny curls | Men walk by

Gloss RH English: eautifully coiffured hair | (p-) long hair over shoulder | TINY CURLS | (p-) people walk

Gloss LH English: eautifully coiffured hair | TINY CURLS | (p-) people walk

Gloss RH: ıaar mooi opgestoken | (p-) lang haar over schouder | PIJPENKRULL | (p-) mensen- open

Dir&loc RH

Repetition RH: u | u

Gloss LH: ıaar mooi opgestoken | PIJPENKRULL | (p-) mensen-lopen

Figure 1 ELAN user interface

the amount of training data available, the greater the set of sub-sentential alignments that can be created. This provides a larger scope for finding translation matches for input string, which correspondingly increases the chances of improving system output. The ECHO project team have funding to increase their corpus creation by 2008, so we hope to increase our training data when this becomes available.

5.2. Irish Sign Language Corpora

Currently, a large annotated corpus of ISL data is under construction. The Centre for Deaf Studies[6] in Dublin is in the process of annotating a corpus of approximately 40 hours of ISL data. However, the subject matter of the ISL data is similar to that of the NGT data, namely stories and conversation. While the larger amount of training data will help increase the translation quality of our system, we will still be confronted with 'unsuitable' data.

A more suitable corpus which would have a practical use for the Deaf is, for example, the area of travel information. In airports and train stations, announcements of changes in travel information are usually announced over a PA system; often such information does not appear on the information screens until a later stage if at all. For this reason many Deaf people find themselves uninformed about changes to schedules etc. through no fault of their own. In many airports and train stations worldwide travel information is entered into a system that announces the changes in an electronic voice. It is quite possible that this system could be extended to accommodate SLs. The limited range of statements and information used in these circumstances could be compiled into a corpus and the information that is announced could be translated into sign language and displayed on video screens for the Deaf to view. We are negotiating with our local airport authority to obtain such data in order to compile a corpus of commonly used announcements suitable for the seeding of our EBMT system's memories.

6. Marker-Based EBMT

An example-based approach necessitates a set of sentences aligned in the source and target languages. Three processes are used to derive translation for an input string:

1. Searching the source side of the bitext for 'close' matches and their translations;
2. Determining the sub-sentential translation links in those retrieved examples;
3. Recombining relevant parts of the target translation links to derive the translation.

The methodology employed in our system is to make use of the 'Marker Hypothesis' (Green, 1979). Here closed class words are used to segment aligned source and target sentences and to derive an additional set of lexical and phrasal resources. In a pre-processing stage, (Gough & Way, 2004b) use 7 sets of marker (or closed class) words for English and French (e.g. determiners, quantifiers, conjunctions etc.) to segment the text into chunks which together with cognate matches and mutual information scores are used to derive three new data sources: sets of marker chunks, generalised templates and a lexicon.

[6] http://www.tcd.ie/Deaf_Studies/

Within our system, sentential alignments are extracted using the time-aligned borders of the English/Dutch translation tiers and grouping all annotations within those time frames together to form the corresponding SL sentence. The English/Dutch sentences are segmented on the basis of closed class words. As an example, consider the sentence in (2), from (Gough & Way, 2004a):

(2) The first part of the book describes the components of the desktop.

This string is automatically tagged with marker words, as in (3):

(3) <DET> The first part <PREP> of the book describes <DET> the components <PREP> of the desktop.

Given the tagged strings in (3), the marker chunks in (4) are automatically generated:

(4) (a) <DET> The first part
(b) <PREP> of the book describes
(c) <DET> the components
(d) <PREP> of the desktop

By generalising over the marker chunks we produce a set of marker templates. This is achieved by replacing the marker word by its relevant tag. From the examples in (4), we can produce the generalized templates in (5):

(5) (a) <DET> first part
(b) <PREP> the book describes
(c) <DET> components
(d) <PREP> the desktop

These templates increase the robustness of the system and make the matching process more flexible.

A different approach is used on the sign language side of the corpus. The annotations are segmented according to the NGT gloss time divisions and other corresponding annotations within the same time frame are grouped with that gloss to form a chunk. We therefore segment the sign language corpus into concept chunks to match the content of the English chunks. The example below demonstrates segments from both data sets (English (6) and NGT (7)) and their usability for chunk alignment:

(6) <CONJ> or with tiny curls

(7) <CHUNK>
(Gloss RH English) TINY CURLS
(Gloss LH English) TINY CURLS
(Repetition LH) u
(Repetition RH) u
(Eye gaze) l,d

Despite the different methods used, they are successful in forming potentially alignable chunks. Both chunks indicate the possession of "tiny curls", articulated by the words in the English chunk and the right and hand left hand in the sign chunk. Extra information is added in by the NMFs *repetition* and *eye gaze*. Repetition shows that the left and right hands signs are articulated a number of times, the 'u' indicates uncountable repetitions in a wiggling manner showing the plurality of the sign, i.e. many curls. The eye gaze 'l,d' indicates that the gaze of the signer goes from the left of the signing space (where the curls start at the signer's head) downwards, following the movements of the hands. This is an important feature in SLs. The gaze of the signer usually follows the movement of the main articulators. Eye gaze is also used to indicate distance of an object in relation to the signer. If eye gaze was not taken into account vital information on the location of objects in the signing space or the distance of objects from the signer would be lost.

7. Experiments and Results

We extracted 561 English— and Dutch—NGT sentence pairs. In order to provide an indication of data complexity, the English translations had an average sentence length of 7.89 words (min. 1 word per sentence, max. 53).

We began testing the system for translation of English and Dutch into NGT. The data was divided into an approximate 90:10 training-testing splits, with 55 randomly selected sentences withheld for testing purposes. Each test sentence is entered into the system and a translation produced based on best matches found at a sentential, sub-sentential (chunk) or word level.

Manual examination of the output showed that the system performed reasonably well and appeared to have correctly translated most of the central concepts in the sentences (Morrissey & Way, 2005). However, annotations can be complex and it is difficult for an untrained eye to discern the correctness of the output. Furthermore, due to the subjectivity and varying format of the annotations, there lacks a 'gold standard' against which they may be formally evaluated using traditional MT evaluation metrics.

In light of this issue, we chose to reverse the translation process taking in annotations as input and producing either English or Dutch output. Output into spoken language takes the form of written text and output in sign language takes the form of grouped annotations.

While reversing the directionality of translation enables automatic evaluation metrics to be used, the exercise is quite artificial in that there is little or no demand for translation from SL to spoken language. Of course, situations can be envisaged where this might be useful, e.g. in a post office, a Deaf customer could ask for stamps by signing into a camera and having it translated into text/speech for the hearing sales assistant, while the process could be reversed for communicating the information from the sales assistant to the Deaf person via a signing avatar.

From the change in direction we were able to obtain evaluation scores for the output as we had reference translations against which to measure the output. However, as SLs by their very nature do not contain closed class lexical items, the output was sparse in terms of lexical data and rich only with respect to content words. This resulted in decidedly low evaluation scores. In an attempt to improve these scores we experimented with inserting the most common marker word (*the* in English and *de* in Dutch) into the candidate translations in what we determined to be the most appropriate location, i.e. whenever an INDEX was found in the NGT annotations indicating a pointing sign to a specific location in the signing space that usually refers back to an object previously placed there. This was an attempt to make our translations resemble more closely the gold standard.

7.1. Automatic Evaluation Metrics

The system was evaluated for the language pair NGT—English using the traditional MT evaluation metrics BLEU (Papineni et al., 2002), SER, WER and PER. BLEU score is a precision-based metric that compares a system's translation output against reference translations by summing over the 4-grams, trigrams, bigrams and unigram matches found divided by the sum of those found in the reference translation set. It produces a score for the output translation of between 0 and 1. Sentence Error Rate (SER) computes the percentage of incorrect full sentence matches. Word Error Rate (WER) computes the distance between the reference and candidate translations based on the number insertions substitutions and deletions in the words of the candidate translations divided by the number of correct reference words. The Position-independent word Error Rate (PER) computes the same distance as the WER without taking word order into account. With all error rates, a lower percentage score indicates better candidate translations.

7.2. NGT—English Results

For the 55 test sentences, the system obtained a SER of 96%, a PER of 78% and a WER of 119%.[7] Due to the lack of closed class words produced in the output, no 4-gram matches were found, so the system obtained a BLEU score of zero. Ongoing experiments using the 'Add-One' ploy of (Lin & Och, 2004) will circumvent the 'Zero-BLEU' problem described here. An example of the candidate translation capturing the central content words of the sentence may be seen in (8) compared with its reference translation in (9).

(8) mouse promised help

(9) 'You see,' said the mouse, 'I promised to help you'.

Here it can be seen that our EBMT system includes the correct basic concepts in the target language translation, but for anyone with experience of using automatic evaluation metrics, the 'distance' between the output in (8) and the 'gold standard' in (9) will render the quality to be scored very poorly.

In the next section, we hypothesize whether a different evaluation metric might be more useful, both for SLMT, but also for MT as a whole.

8. Discussion of the Evaluation Process

As shown in the previous section, we struggled to use mainstream MT evaluation metrics such as BLEU, WER and PER, albeit in a rather artificial exercise. Of the related work mentioned in section 3, only the translations produced by (Bungeroth & Ney, 2005) have been evaluated using these metrics. The standard evaluation technique applied to SLMT seems to be a manual assessment by native and non-native signers.

We contend that in general, the traditional string-based metrics are inappropriate for the evaluation of SLMT systems, where the primary goal is translation from an oral to a non-oral language, as there is no 'gold standard' underlying sign language annotation available.

A typical annotation taken from our corpus was shown in (7). For our purposes, we concentrate mostly on the 'GLOSS' field, but other relevant information appears in other fields too, such as lip rounding, puffing of the cheeks etc. The absence of the semantic information provided by these NMFs affects the translation and thus the evaluation scores, so we intend to incorporate this information into the system in the next phase of development.

Our experiments were further hampered by the fact that we were generating root forms from the underlying GLOSS, so that a lexeme-based analysis of the gold standard and output translations via a morphological analysis tool might have had some positive impact. This remains an avenue for future research.

Subsequent experiments to (i) insert the most common marker word corresponding to the appropriate marker tag (to make our translations resemble more closely the 'gold standard', and (ii) delete marker words from the reference translations (to make them closer to the translations output by our system) had little effect on overall BLEU score.

In fact, we have come to the conclusion that rather than continue to attempt these 'transformations' in order to try to reconcile the differences between the reference and candidate translations, we would in fact be better off developing automatic MT evaluation metrics that were more suitable to sign language data

One measure that might have some promise is evaluation on the level of syntactic (or, even better, semantic) relations. For example, compare the (invented) reference and candidate translations in (10):

(10) *Reference*: I went to the shops yesterday.
Candidate: Yesterday I went to the shops.

Despite being a 'perfect' translation in many ways, the candidate translation in (10) obtains a BLEU score of just 0.669. However, at the level of syntactic relations, the sentences in (10) are identical.

One method of evaluating the 'goodness' of translations would be at the level of syntactic dependencies, rather than by measuring the distance between two strings. Dependency parsers for many languages exist already; two examples that one of the authors has been involved in are the LFG parsers of (Cahill et al., 2004) for English, and (Cahill et al., 2005) for German. New strings are parsed using a variety of PCFG-based LFG parsers, and LFG trees and f-structures are produced. Gold standard sets of f-structures exist (e.g. the PARC-700), and reference and

[7] It is possible to obtain a WER of more than 100% if there are fewer words in the reference translations than in the candidate translations.

system-generated f-structures can be compared and evaluated using F-score. An alternative means of evaluation would be to read the semantic forms (subcategorisation frames) off the f-structures generated using the method of (O'Donovan et al., 2005) and compare the 'predicate(filler, arg)' triples. Given examples such as (8), for example, it might be sufficient for our purposes to evaluate only the 'PRED' (or headword) triples.

All this remains for further research, and is outside the scope of this paper, but we are quite confident that such an evaluation would be more useful not only for sign language MT, but for all models of translation. Clearly, in addition, human evaluation remains crucial for all such approaches.

9. Conclusions

We have described ongoing work on our EBMT system to translate between oral and sign languages. Like other minority languages, SLMT suffers from the lack of suitable corpora for the training of corpus-based models of translation. In order to bootstrap the system, we have used 561 English— and Dutch—NGT sentence pairs from the ECHO corpus.

Despite the subjective nature of the corpus and its size, the availability and ease of use of the annotations facilitates speed of development of such an SLMT system. Were a larger corpus to be made available in another SL, the approach described above could easily be applied.

One disadvantage of a corpus-based approach, as discussed in this paper, is its evaluation. No 'gold standard' is available for evaluating candidate translations in SL and the metrics used for evaluating the English/Dutch output fall short of recognising that the candidate translations capture the essence of the sentence. We are confident that a syntactic- or semantic- based evaluation metric would better reflect the performance of an SLMT system while at he same time providing an improved evaluation approach for written languages.

References

Armstrong, S., D. Groves, M. Flanagan, Y. Graham, B. Mellebeek, S. Morrissey, N. Stroppa, and A. Way. (2006). The MaTreX System: Machine Translation Using Examples. Available at: http://www.tc-star.org/openlab2006/day1/Groves_openlab.pdf

Bauer, B., S. Nießen and H. Heinz. (1999). Towards an Automatic Sign Language Translation System. In *Proceedings of the International Workshop on Physicality and Tangibility in Interaction: Towards New Paradigms for Interaction Beyond the Desktop*, Siena, Italy.

Bungeroth, J. and H. Ney (2004). Statistical Sign Language Translation. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages (LREC 04)*, Lisbon, Portugal.

Cahill, A., M. Forst, M. Burke, M. McCarthy, R. O'Donovan, C. Rohrer, J. van Genabith and A. Way. (2005). Treebank-Based Acquisition of Multilingual Unification Grammar Resources. *Journal of Language and Computation: Special Issue on Shared Representations in Multilingual Grammar Engineering*, pp.247—279.

Cahill, A., M. Burke, R. O'Donovan, J. Van Genabith and A. Way. (2004). Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. In *ACL-04: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain, pp.319—326.

Gordon, R. G., Jr. (ed.), (2005). *Ethnologue: Languages of the World, Fifteenth edition*. Dallas, Tex.: SIL International.

Gough, N. and A. Way. (2004a). Example-Based Controlled Translation. In *Proceedings of 9th EAMT Workshop*, Valetta, Malta, pp.73—81.

Gough, N. and A. Way. (2004b). Robust Large-Scale EBMT with Marker-Based Segmentation. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04),* Baltimore, MD., pp.95—104.

Green, T. (1979). The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior* **18**:481—496.

Grieve-Smith, A.B. (1999). English to American Sign Language Machine Translation of Weather Reports. In D. Nordquist (ed.) *Proceedings of the Second High Desert Student Conference in Linguistics (HDSL2),* Albuquerque, NM., pp.23—30.

Huenerfauth, M. (2005). American Sign Language Generation: Multimodal NLG with Multiple Linguistic Channels. In *Proceedings of the ACL Student Research Workshop (ACL 2005)* Ann Arbor, MI., pp.37—42.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit X*, Phuket, Thailand, pp.79—86.

Lin, C-Y. and F.J. Och. (2004). ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp.501—507.

Marshall, I. and É. Sáfár. (2002). Sign Language Generation using HPSG. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-02)*, Keihanna, Japan, pp.105—114.

Morrissey, S. and A. Way. (2005). An Example-based Approach to Translating Sign Language. In *Proceedings of the Workshop in Example-Based Machine Translation (MT Summit X)* Phuket, Thailand, pp. 109—116.

Ó'Baoill, D. and P.A. Matthews. (2000). *The Irish Deaf Community (Volume 2): The Structure of Irish Sign Language*. The Linguistics Institute of Ireland, Dublin, Ireland.

O'Donovan, R., M. Burke, A. Cahill, J. van Genabith and A. Way. (2005). Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. *Computational Linguistics* **31**(3):329—365.

Papineni, K., S. Roukos, T. Ward and W. Zhu. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA., pp.311—318.

Prillwitz, S. (1989) *HamNoSys Version 2.0; Hamburg Notation System for Sign Language. An Introductory Guide.* Signum Verlag.

Sáfár, É. and I. Marshall. (2002). The Architecture of an English-Text-to-Sign-Languages Translation System. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-01)*, Tzigov Chark, Bulgaria, pp.223—228.

Stein, D., J. Bungeroth and H. Ney (2006). Morpho-Syntax Based Statistical Methods for Automatic Sign Language Translation. In *Proceedings of 11th EAMT Annual Conference,* Oslo, Norway.

Stokoe, W.C. (1960). An Ouline of the Visual Communication Systems of the American Deaf. In *Studies in Linguistics: Occasional papers, No. 8*, Department of Anthropology and Linguistics, University of Buffalo, Buffalo, NY., [Revised 1978 Lincoln Press].

Sutton, V. (1995). Lessons in Sign Writing, Textbook and Workbook (Second Edition). The Center for Sutton Movement Writing, Inc.

Van Zijl. L. and D. Barker. (2003). South African Sign Language Machine Translation System. In *Proceedings of the Second International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa (ACM SIGGRAPH)*, Cape Town, South Africa, pp.49—52.

Veale, T., A. Conway and B. Collins. (2000). The Challenges of Cross-Modal Translation: English to Sign Language Translation in the Zardoz System. *Machine Translation* **13**(1):81—106.

Vogel, S., Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao and A. Waibel. (2003). The CMU Statistical Machine Translation System. *MT Summit IX*, New Orleans, LA., pp.402—409.

Zhao, L., K. Kipper, W. Schuler, C. Vogler, N. Badler, and M. Palmer. (2000). A Machine Translation System from English to American Sign Language. In *Envisioning Machine Translation in the Information Future: Proceedings of the Fourth Conference of the Association for Machine Translation (AMTA-00)*, Cuernavaca, Mexico, pp.293—300.

A Spanish-Basque weather forecast corpus for probabilistic speech translation

Alicia Pérez[1], Inés Torres[1], Francisco Casacuberta[2], Víctor Guijarrubia[1]

[1] Elektrika eta elektronika saila
Euskal Herriko Unibertsitatea
{manes@we.lc.ehu.es}
[2] Departament de Sistemes Informàtics y Computació
Universitat Politècnica de València
{fcn@dsic.upv.es}

Abstract

The main goal of this work is to develop a bilingual corpus suitable for example based machine translation between Spanish and Basque. Spanish to Basque machine translation has proved to be a difficult task for statistical machine translation. Apart from the great linguistic differences between these two languages, the low performance of the few presently available statistical machine translation systems is likely due to the lack of adequate parallel corpora. Here we present the methodology involved to pick up and process bilingual data about weather forecast reports. We also present preliminary experiments carried out for speech and text input (statistical) machine translation.

1. Introduction

In recent years, statistical methods have been successfully applied to machine translation. In this framework: stochastic translation models can be obtained for any pair of languages, whenever a representative number of examples is available. Thus, inductive approaches require a great deal of collection of bilingual data, that is, sentences from a (source) language and the corresponding translation from another (target) language (Brown et al., 1993). However, bilingual corpora for only some European languages are available.

The goals of this study are, on the one hand, to create a suitable bilingual corpus for a specific translation task from Spanish into Basque, for both text and speech-input purposes, and on the other hand to present the preliminary (speech and text) translation results using statistical machine translation (SMT) techniques. Another practical extension of the corpus is being created for translation from Basque into English.

1.1. The Basque language

Basque language is a minority but official language in the Basque Country. It is also spoken in some adjoining areas, such as Navarre, in Spain, and Atlantic Pyrenees, in France, as shown in Table 1.

Area	Total inhabitants	Basque speakers
Navarre	236,963	33.20%
Atlantic Pyrenees	515,989	10.15%
Basque Country	2,089,995	24.58%

Table 1: The percentage of Basque speakers in different areas (according to data published by the Basque Government).

Basque is a pre-Indoeuropean language of unknown origin. Thus, the etymology of words in Basque and Spanish is usually different. It also presents a different arrangement of the words within phrases, since, unlike Spanish, Basque has left recursion. These features are shown through the example in Figure 1. On the other hand, Basque is a highly inflected language, in both nouns and verbs.

1.2. The state of the art

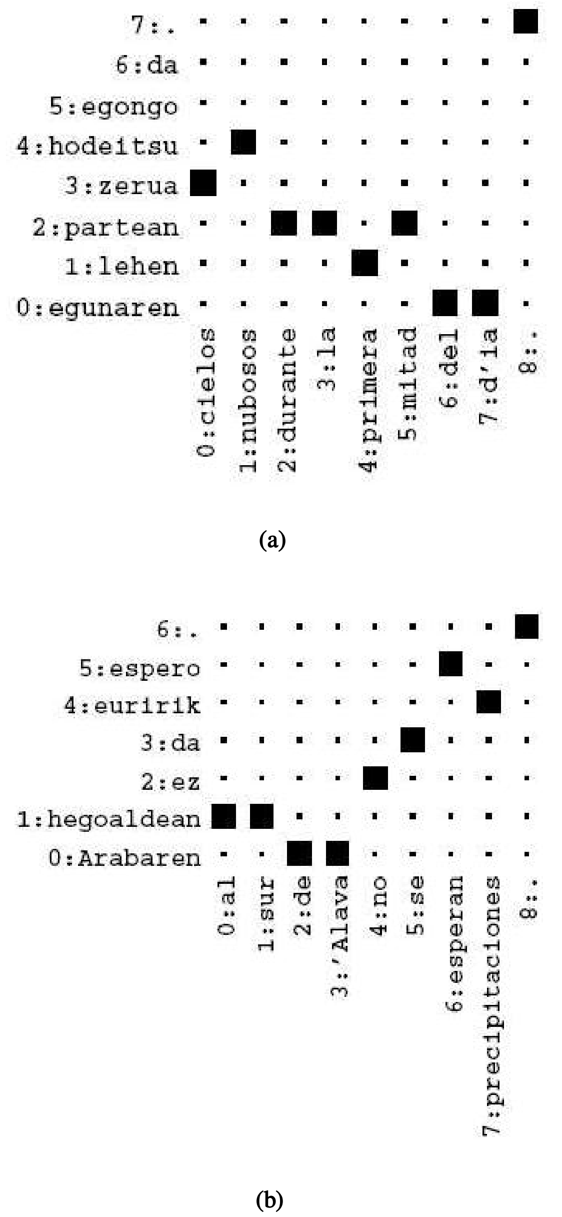
There are some papers related to transfer-based translation tools from Spanish to Basque (Alegria et al., 2005). Morpho-syntactic parsing has also been broadly studied (Arriola, 2004). However, there are few papers related to SMT between Spanish and Basque. Among them, we may highlight (Ortiz et al., 2003; González et al., 2004), where a corpus from the Basque Country's official government records was harvested. Nevertheless, low translation results were obtained. This low performance is mainly due to the scant bilingual corpora available. The lack of samples results in poorly trained statistical alignment models. We realize that those alignments were unable to capture long distance relationships (see Figure 1) between Spanish and Basque.

2. Stochastic finite-state transducers

Finite state transducers have proved to be useful in language processing and in automatic speech recognition (ASR) systems. In recent years they have also been proposed for SMT applications (Vidal, 1997; Casacuberta and Vidal, 2004). Moreover, FST can be easily integrated into an ASR system for speech translation application (Casacuberta et al., 2004).

Stochastic finite-state transducers (SFST) can be automatically learned from bilingual corpora by efficient grammar inference algorithms, such as GIATI (Grammar Inference and Alignments for Transducers Inference). Given a bilingual corpus, the GIATI algorithm provides a probabilistic finite-state transducer (Casacuberta and Vidal, 2004). This algorithm works as follows:

1. Given a bilingual corpus, find a monotone segmentation, and thereby, assign an output sequence to each



(a)

(b)

Figure 1: These two alignment matrices, extracted from the corpus, show the natural relationships between the words of a sentence in Spanish and their counterparts in Basque. Those relationships are not monotone.

input word, leading to the so called *extended corpus*.

2. Infer a probabilistic finite state automaton from the extended corpus. In this paper we propose the use of a *k-testable in the strict sense* (k-TSS) language model (Torres and Varona, 2001), rather than an n-gram model, since k-TSS models keep the syntactic constraints of the language.

3. Split the output sequence from the input word, on each edge of the automaton to obtain the finite state transducer.

The *stochastic translation* $\widehat{t} \in \Delta^*$, of an input sequence $\mathbf{s} \in \Sigma^+$, is the string which maximizes the joint probability described by the equation (1):

$$\widehat{t} = \arg\max_{\mathbf{t}} P(\mathbf{s},\mathbf{t}) = \arg\max_{\mathbf{t}} \sum_{d(\mathbf{s},\mathbf{t})} P(d(\mathbf{s},\mathbf{t})) \qquad (1)$$

where, $d(\mathbf{s},\mathbf{t})$, is a path in the SFST that deals with s and produces $\mathbf{t}$.
The resolution of the eq. (1) has proved to be a hard computational problem (Casacuberta and de la Higuera, 2000), but it can be efficiently computed by the *maximum approximation*, which replaces the sum by the maximum:

$$\widehat{t} \approx \arg\max_{\mathbf{t}} \max_{d(\mathbf{s},\mathbf{t})} P(d(\mathbf{s},\mathbf{t})) \qquad (2)$$

3. A weather forecast corpus

METEUS is the weather forecast corpus that we present here. It was composed from 28 months of daily weather forecast reports in the Spanish and Basque languages. These reports were picked from those published in Internet by the Basque Institute of Meteorology[1]. We obtained a first bilingual corpus where each report in Spanish was the translation of a report in Basque. Thus, *bilingual alignment* was assured at paragraph level.
At the end of the first corpus acquisition, there were 3,865 paragraph pairs, consisting of many sentences, with around 54 words per paragraph on average. Segmentation into sentences was solved by using statistical techniques, specifically *RECalign*, a greedy algorithm (Nevado and Casacuberta, 2004; García-Varea et al., 2005). A hundred paragraphs, randomly chosen, were checked and validated by experts, which assures the success of the algorithm.
After the segmentation process, the corpus was divided into training and testing sets (Table 2). Notice that the Basque language vocabulary size for this task is 1.6 times higher than the Spanish one. This is not unusual given the Basque language inflection mentioned in section 1.1.. In order to deal with this problem, a morphological analysis was carried out. As a result, we can work with word-forms or stems. The vocabulary size, in terms of stems, has been decreased to 462 units in Spanish and 578 in Basque.
The text-test set consists of 500 training independent pairs, all of them different. For speech input machine translation experiments, this test set was recorded by 36 bilingual speakers uttering 50 sentence-pairs each, resulting in around 3.25 hours of audio signal for each language.

4. Experimental results

In this section we present a preliminary evaluation of this corpus. We learned an SFST from the training set, and the test set was translated with the inferred models. Finally, the translations provided by the system were compared to the reference sentences.
Apart from text-to-text translation, we performed speech-input translation. There are many ways of building a speech input translation system, see (Vidal, 1997; Casacuberta et al., 2004). For these preliminary experiments, we simply chose the so called serial architecture, which consists of two steps. In the first the speech signal is decoded into a source

[1] http://www.euskalmet.net

		Spanish	Basque
Training	Pair of sentences	14,615	
	Different pairs	8,462	
	Different sentences	7,226	7,523
	Words	191,156	187,462
	Vocabulary	720	1147
	Average length	13.0	12.8
Test	Pair of sentences	500	
	Different sentences	500	500
	Words	8,706	8,274
	Average length	17.4	16.5
	Perplexity (3-grams)	4.8	6.7

Table 2: Features of the training and test sets.

sentence. In the second the decoded sentence is translated into a target sentence.

The speech signal database was parameterized into 12 Mel-frequency cepstral coefficients (MFCCs) with delta (ΔMFCC) and acceleration (Δ^2MFCC) coefficients, energy and delta-energy (E, ΔE), so four acoustic representations were defined (Rodríguez and Torres, 2003). For the speech recognition system, a total of 24 context-independent acoustic-units were used. Each phone-like unit was modeled by a typical left to right non-skipping self-loop three-state continuous hidden Markov model, with 32 Gaussians per state and acoustic representation. To train these models, a phonetically balanced Spanish database, called Albayzin (Moreno et al., 1993), was used. With regard to the language model, a 3-TSS was used, learned using the training corpus in Table 2.

In Table 3 we summarize some text and speech-input translation results for Spanish-to-Basque SMT. We have dealt with the following automatic evaluation measures:

WER: *Word Error Rate* is the string edit distance between the reference sentence and the system's output.

PER: *Position independent Error Rate* is similar to the WER but without taking into account the words-order inside the sentence.

BLEU: *BiLingual Evaluation Understudy* is based on the n-grams of the hypothesized translation that occur in the reference translations. The BLEU metric ranges from 0.0 (worst score) to 1.0 (best score) (Papineni et al., 2002).

	WER	PER	BLEU
Text input	46.5	37.6	0.46
Speech input	51.7	42.4	0.40

Table 3: Spanish to Basque speech-input and text-input machine translation scores.

Notice that the test set is completely independent of the training set (see Table 2). Therefore, our system is subjected to the worse case that could appear in practice.

4.1. Examples

Some translation examples are shown below, where **input** means "the input sentence", **reference** "the reference", and **system** "the system's output". Taking into account that an input sentence can be translated in more than one way, even though the output does not exactly match the reference, it could be translated correctly. Since we only have a single reference for each input, some outputs are unfairly penalized, therefore the reported error rates are quite pessimistic. In the following examples we emphasize the real errors in italics (in the third example the errors are related to the absence of some words).

- Test sentence #1.

  input las temperaturas máximas sin cambios o ligeramente más altas .

  reference tenperatura maximoak ez dira aldatuko edo gutxi igoko dira .

  system tenperatura maximoak egonkor mantendu edo gutxi igoko *da* .

- Test sentence #2.

  input por la tarde - noche cielos muy nubosos con precipitaciones débiles .

  reference arratsalde - gau aldera zerua oso hodeitsu egongo da eta euri arina egingo du .

  system arratsalde - gauean zerua oso hodeitsu egongo da eta euri arina egingo du .

- Test sentence #3.

  input por la tarde , los vientos girarán a componente sur , flojos a moderados , con intervalos fuertes en el litoral y zonas de montaña .

  reference arratsaldean , haizeak hegoaldera egin eta ahul - bizia ibiliko da , tarte gogorrekin kostaldean eta mendi inguruetan .

  system haizeak hegoaldera egin eta hegoaldeko haize ahul - bizia ibiliko da , kostaldean eta mendi inguruetan .

5. Concluding remarks and further work

The overall result of this work is a Spanish-Basque text and speech corpus, appropriate for building example based translation tools. It has been obtained using statistical techniques, which seem to be suitable from the practical point of view. Apart from that it has been enriched with morphological information. Professional translators are now translating this corpus into English, and it is also being recorded in the way mentioned in Section 3. This will lead us to a trilingual corpus in English-Spanish-Basque, along with speech resources. From this point onwards, we will complete the current work, evaluating the selected translation method under the same conditions for different language pairs. Due to the difficulties of the translation from Spanish to Basque, morpho-syntactic and phrase-based tags are being added to the corpus that will help the construction of SMT systems. Further work is needed in order to improve the statistical translation models used in this work. We suggest taking advantage of linguistic information such as word stems and declinations to enrich the corpus and, at the same time, to get more accurate translation models

6. Acknowledgements

This work has been partially supported by the Industry Department of the Basque Government and by The University of the Basque Country under grants INTEK CN02AD02 and 9/UPV 00224.310-15900/2004 respectively.

We would like to thank the *Ametzagaiña* group (`http://www.ametza.com`), and Josu Landa, in particular, for providing us with the morpho-syntactic tags.

7. References

I. Alegria, A. Diaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola, M. L. Forcada, S. Ortiz-Rojas, and L. Padr. 2005. An open architecture for transfer-based machine translation between spanish and basque. In *Proceedings of OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X (Phuket, Thailand, September 12–16).*

A. Aranzabe Arriola. 2004. A cascaded syntactic analyser for basque.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

F. Casacuberta and C. de la Higuera. 2000. Computational complexity of problems on probabilistic grammars and transducers. In Arlindo L. Oliveira, editor, *ICGI*, volume 1891 of *Lecture Notes in Computer Science*, pages 15–24. Springer-Verlag.

F. Casacuberta and E. Vidal. 2004. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225.

F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, and C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25–47, January.

I. García-Varea, D. Ortiz, F. Nevado, P. A. Gómez, and F. Casacuberta. 2005. Automatic segmentation of bilingual corpora: A comparison of different techniques. In *Iberian Conference on Pattern Recognition and Image Analysis*, volume 3523 of *Lecture Notes in Computer Science*, pages 614–621. Springer-Verlag, Estoril (Portugal), June.

J. González, D. Ortiz, J. Tomás, and F. Casacuberta. 2004. A comparison of different statistical machine translation approaches for spanish-to-basque translation. In *Actas de las III Jornadas de Tecnología del Habla*, Valencia, Spain, November.

A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Mario, and C. Nadeu. 1993. Albayzin speech database: Design of the phonetic corpus. In *Proc. of the European Conference on Speech Communications and Technology (EUROSPEECH)*, Berlín, Germany.

F. Nevado and F. Casacuberta. 2004. Bilingual corpora segmentation using bilingual recursive alignments. In *Actas de las III Jornadas en Tecnologías del Habla*, Valencia, Spain, November.

D. Ortiz, I. García-Varea, F. Casacuberta, A. Lagarda, and J. González. 2003. On the use of statistical machine translation techniques within a memory-based translation system (AMETRA). In *Proc. of Machine Translation Summit IX*, pages 115–120, New Orleans, USA, September.

K. Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association Computational Linguistics (ACL)*, pages 311–318, Philadelphia, July.

L. J. Rodríguez and I. Torres. 2003. Comparative study of the baum-welch and viterbi training algorithms applied to read and spontaneous speech recognition. In *1st Iberian Conference on Pattern Recognition and Image Analysis*, Puerto Andratx (Mallorca), Spain.

I. Torres and A. Varona. 2001. k-tss language models in a speech recognition systems. *Computer Speech and Language*, 15(2):127–149.

E. Vidal. 1997. Finite-state speech-to-speech translation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 111–114, Munich, Germany, April.

Machine translation for closely related language pairs

Kevin P. Scannell

Department of Mathematics and Computer Science
Saint Louis University
St. Louis, Missouri, USA 63103
scannell@slu.edu

Abstract

We exploit the close linguistic relationship between Irish and Scottish Gaelic to develop a robust machine translation system ga2gd, despite the lack of full parsing technology or pre-existing bilingual lexical resources.

1. Introduction

Irish (Gaeilge) and Scottish Gaelic (Gàidhlig) are close linguistic relatives, forming, together with Manx Gaelic, the Goidelic (or Q-Celtic) branch of the Indo-European family. For simplicity in what follows, we will refer to the two languages simply as "Irish" and "Scottish" (in spite of the fact that the latter term, when used in isolation, generally refers to the variety of English spoken in Scotland).

Both languages are spoken daily by small minorities (measured in the tens of thousands), primarily within geographic strongholds situated in each case on the margins of an overwhelmingly English-speaking country. Despite enthusiastic communities of learners scattered throughout the world, the numbers of native speakers have continued to decline over the past century.

In §2, we discuss in some detail the extent of the linguistic similarity between Irish and Scottish. For now we simply point out that the the languages are not, generally speaking, mutually intelligible. They have distinct orthographies, independent (and diverging) lexica, and a number of important structural differences in terms of syntax.

Nevertheless, the languages are close enough that high-quality machine translation can be achieved with a manageable number of syntactic transfer rules, at least when combined with robust, statistically-based word sense disambiguation. Furthermore, by leveraging the existing open source language resources developed by the author, a complete system was implemented without unreasonable effort.

We believe there are a number of similar under-resourced language pairs that could benefit from a comparably naive approach, (e.g. Zulu ↔ Xhosa, Hiligaynon ↔ Cebuano, or Tokelauan ↔ Tuvaluan). Perhaps also of interest are pairs in which one language is a global one (e.g. French ↔ Walloon, or Italian ↔ Sardinian), where robust MT in either direction would be immensely useful. This has already been achieved, for example, by the open source project *Apertium* (Corbí-Bellot et al, 2005), which uses a design quite similar to ours for Spanish ↔ Catalan and Spanish ↔ Galician MT.

We owe great thanks to Caoimhín Ó Donnaíle for providing a tremendous amount of machine-readable data for Scottish, and for sharing his substantial linguistic expertise (in both languages).

2. Linguistic comparison

Since the robustness of the translator depends to a great extent upon the close linguistic relationship between Irish and Scottish, we thought it important to include in this section an indication of just how close this relationship is. For simplicity, and because we are interested primarily in translating electronic documents, we focus on the written languages. More details on this subject can be found in (Ó Rathaille, 1932, Ch. XVI), (Mac Maoláin, 1962), and (McCone, 1994).

The main difficulty with making such a comparison is the fact that the languages in question are "moving targets" in the sense that a given text (in, say, Irish) is parameterized in a number of ways that influence its relationship with Scottish: e.g. the date it was written, the regional dialect in which it was written, or its linguistic register.

All three Goidelic languages share a common ancestor in *Middle Irish*, forms of which were spoken in Ireland, on the Isle of Man, and in the Scottish Highlands until roughly the 16th century. Even after the spoken languages had diverged, there was a shared literary tradition written in the so-called *Gaeilge Chlasaiceach* (Classical Gaelic) up through the 18th century; this was the language of most of the early printed books in both languages: from Carswell's translation of Knox's liturgy in 1567 to the Bible translations of the 17th century. Note, however, that by 1690 the languages had diverged to the extent that when the Irish translation of the Bible was reprinted in Roman characters for the benefit of Gaelic speakers in Scotland, there were complaints that the text was unreadable (Williams, 1986, pp. 101–102).

Geographically, there was at one time a continuum of dialects ranging from the far southwest of Ireland to the northernmost parts of Scotland. As a consequence, the Ulster dialect of Irish, spoken in northeastern Ireland, is by far the closest to Scottish. We note, however, that in practice these differences are of little importance to the translator: input texts are normalized in various ways that minimize any dialect differences that might be present; see §3.2.

The single greatest disaster in terms of mutual understanding between the languages was the introduction of the *Caighdeán Oifigiúil* (Official Standard) on the Irish side in the 1940's (Rannóg an Aistriúcháin, 1962). As an example, the Scottish Gaelic words *bàgh* (bay), *bàidh* (sympathy), and *bàthadh* (drowning) are immediately recognizable and distinguishable in pre-standard Irish (Dinneen, 1927) as *bádh*, *báidh*, and *bádhadh*, respectively, while the *Caighdeán* tragically conflates all three into the indescript "*bá*" (Ó Dónaill, 1977). Similar examples abound.

There was also a spelling reform on the Scottish side, put forward in 1981 with the publication of the "Gaelic Orthographic Conventions" (GOC) document[1]. These reforms were less sweeping, and offered more of a mixed bag in terms of the relationship with Irish. On the one hand, the GOC changed things like the shared acute accents on words like *mór* or *féin* to grave accents, but at the same time made other words look more Irish, by mimicking some of the *Caighdeán* reforms (e.g. replacing *sg* with *sc* and *sd* with *st*).

For the benefit of readers completely unfamiliar with these languages, we will attempt to make the preceding discussion a bit more concrete by offering a single sentence[2] rendered in each language; we will draw upon this example occasionally in the sections below.

> Cén fáth a bhfeiceann tú an cáithnín i súil do bhráthar agus nach n-airíonn tú an tsail i do shúil féin?
>
> Agus c'ar son a tha thu a' faicinn an smùirnein a tha 'an sùil do bhràthar, ach nach 'eil thu 'toirt fainear na sail' a tha ann do shùil féin?

3. Design and implementation

3.1. Overview

The `ga2gd` software is implemented, from the perspective of an end-user, as a standard Unix filter:

```
$ echo "lá breá éigin" | ga2gd
latha brèagha air choireigin
```

The same is true of the internal architecture; an input text is piped through a sequence of smaller standalone components which transform the text in various ways:

1. Irish standardization.
2. POS tagging, stemming, and chunking.
3. Word sense disambiguation.
4. Syntactic transfer.
5. Lexical transfer.
6. Scottish post-processing.

The sections below describe each of these components in brief, and, where appropriate, some indication is given of how they were assembled.

3.2. Irish standardizer

In recent work, the author created a web crawler and search engine tailored specifically to the Irish language documents on the web[3]. In contrast with general-purpose search engines, which tokenize and index the documents they find in exactly the form in which they appear on the web, our site also converts Irish documents to a form approximating the *Caighdeán Oifigiúil* and indexes them both ways. This allows access to all historical (or dialect) Irish documents on the web through a single, simple search interface.

The standardizer amounts to a finite state transducer that encodes the morphological rules of non-standard Irish together with mappings to standardized forms. These rules are augmented with a large database of non-standard/standard word pairs that was extracted in part from a parallel corpus of English and Irish texts (Scannell, 2005).

This phase is important in that it allows `ga2gd` to translate non-standard Irish as easily as standard Irish. It also has the advantage that, when constructing the bilingual lexicon, one need only provide Scottish translations for standard citation forms of Irish words.

3.3. Irish tagger, stemmer, and chunker

There is no full-scale parsing technology available for either Irish or Scottish, and, consequently, there are no treebanks one could use to implement a statistical MT system. There are, however, robust (rule-based) part-of-speech taggers built into the open source Gramadóir grammar checker[4] for each language. Irish, in addition, has a rule-based *chunker*, which delimits chunks in the spirit of (Ramshaw and Marcus, 1995). As it turns out, the syntactic differences between the languages are small enough that chunking is sufficient (in most cases) for finding accurate translations; this is discussed below in §3.5.

When this phase is completed, XML tags have been added to each word in the input stream that indicate the word's part of speech, its stem, and the stem's part of speech. For example, in our example sentence, *bhfeiceann* is transformed into:

```
<w>
 <t>
  <V p="y" t="láith">bhfeiceann</V>
 </t>
 <s>
  <V p="y" t="ord">feic</V>
 </s>
</w>
```

The stems and their POS tags are used in an essential way by the word sense disambiguation module; see the next subsection.

This component of the pipeline coincides almost exactly with the standalone Gramadóir grammar checker for Irish, and so we direct the reader to that project's documentation for further implementation details[5].

3.4. Irish word sense disambiguation

Because of the syntactic similarities between the languages, it turns out that the stickiest translation problems are, for the

[1] See `http://www.smo.uhi.ac.uk/gaidhlig/goc/`.

[2] Matthew 7:3: "Why do you see the speck that is in your brother's eye, but don't consider the beam that is in your own eye?"

[3] See `http://www.aimsigh.com/`.

[4] See `http://borel.slu.edu/gramadoir/`.

[5] Ibid.

most part, semantic. Solving these problems relies upon a robust word sense disambiguation (WSD) system.

For ga2gd, the WSD filter is implemented as a naive Bayes classifier. It takes as input a tagged and chunked Irish text, and begins by searching in each sentence for words that have more than one possible Scottish translation. When such a word is found, a *feature vector* is generated which is made up of the stemmed and tagged words from the sentence, plus features indicating whether or not the words adjacent to the ambiguous word have initial mutations[6]. Then the most probable sense for the ambiguous word is chosen, given the computed feature vector, and this sense is added to the text stream as an attribute within the <t> tag from the previous subsection, e.g. <t sense='1'>. The appropriate probabilities are computed using training data bootstrapped from a small, manually disambiguated corpus.

It is critical in a number of instances to consider the mutations on adjacent words. For example, the Irish adjective *céad* can mean "first" or "one hundred" and precedes the noun it modifies in each case. When it means "first", however, it causes lenition of the modified noun: *a céad cheacht* "her first lesson", but *céad bliain ó shin* "a hundred years ago". Without this clue, there is very little else (statistically speaking) that one might rely upon to perform this disambiguation.

The *bá* example mentioned above is a good one, though note that the part-of-speech tagger shares the responsibility for distinguishing the masculine "drowning" sense from the other (feminine) senses. A similar example is Ir. *fiach*, which can mean "a debt, obligation", "a raven", or "a hunt", and these senses are translated to Scottish as *fiach*, *fitheach*, *fiadhach*, respectively[7].

Ambiguous words are quite common. Looking up a random sample of words from our database in (Ó Dónaill, 1977) indicates that between 10% and 20% of words have multiple senses[8]. At present, the WSD module has been trained for less than 1000 ambiguous Irish words, though we hope to grow this to about 3000, or approximately 10% of the total lexicon. This should be more than adequate for accurate translation to Scottish since not infrequently an ambiguity in Irish is shared on the Scottish side, e.g. *bonn* can mean either "base, foundation" or "coin, medal" in both languages. This is another important way in which translation between closely related languages is substantially easier than the general case.

Even our simple example sentence from the end of §2 requires disambiguation of at least three words: *cáithnín*, *súil*, and *(t)sail*. The word *súil* illustrates a recurring issue for the language pair in question. Here it means "eye" but also functions as a so-called "verbal noun" after *ag*, e.g. *ag súil le* "hoping/waiting for", and the cognate translation *sùil* is not acceptable in the latter case. The same situation arises with many other Irish words such as *cnuasach*, *cruinniú*, *scrúdú*, etc.

In the sample sentence, the word *sail* means "beam, stick", but quite commonly means "dirt, dross" in the Irish corpus, and these senses translate to distinct Scottish words (*sail* (feminine) and *sal* (masculine) respectively). Similarly, the word *cáithnín* means something like English "speck, mote" in the present example, but in theoretical physics is used for things like subatomic particles. The WSD module has no difficulty distinguishing the latter sense since it often appears in sentences together with unambiguously technical terms such as *cosmach* "cosmic", *treoluas* "velocity", or *déacht* "duality" (though this is an instance in which disambiguation is less important – the single Scottish term *smùirnean* is probably a safe translation in either case).

3.5. Context-sensitive syntactic rewriting

As noted above, we do not have a full parse of the input sentence at this stage, but instead something resembling a parse tree of depth one.

An important example of a syntactic rewrite rule comes from the fact that there is no exact analogue of the present tense Irish verb in Scottish. This is why the phrase *(bh)feiceann tú* "you see" is translated as *tha thu a' faicinn* "you are a'seeing" in our example above. In cases like this, once the chunker has correctly identified the subject noun phrase of the present tense verb, then a simple syntactic transfer is sufficient:

(S (VBZ x) (NP y)) → tha t[y] a' t[x]

where the mapping $x \to t[x]$ means to recursively translate the given constituent, and in the special case of present tense Irish verbs we map to the appropriate verbal noun in Scottish.

The Irish imperfect tense is also not available in Scottish, and so we have the following similar rewrite rule:

(S (VBI x) (NP y)) → b'àbhaist do t[y] a bhith a' t[x]

The transfer rules are stored in a plain text input file and are expressed in a syntax similar to the examples above. Then, before the actual translation process begins, each rule is transformed into a finite state recognizer which can be compiled for exceptionally fast matching against the (tagged and chunked) input stream.

The current version has just under 100 transfer rules, though we expect this number to grow rapidly as we continue to add rules for handling additional multi-word phrases.

[6] An *initial mutation* in Irish or Scottish (or the other Celtic languages) is a phonological change that occurs at the beginning of a word in certain situations, usually depending on the syntactic relationship with, or some grammatical feature of, the preceding word. An important example in both languages is *lenition*, which is indicated orthographically by the insertion of an 'h" after an initial consonant. Irish has another consonant mutation called *eclipsis* that has no orthographic counterpart in Scottish.

[7] And note that, as with *bá*, the *Caighdeán Oifigiúil* is partially responsible for the ambiguity in this case: the 'hunting" sense is generally spelled *fiadhach* in pre-standard Irish

[8] This percentage is probably larger than what one would get by sampling the dictionary in full; our database omits a large number of (generally monosemous) rare or technical terms.

3.6. Bilingual lexicon

A number of different techniques were used to construct the Irish-Scottish lexicon required by the translator. Because of the scarcity of parallel texts in the two languages, we were unable to exploit mutual information techniques to any great extent (though a small number of word pairs were extracted by aligning the electronic Bible texts in the two languages).

At least 90% of the translations in the lexicon were extracted automatically from two existing electronic dictionaries, one Irish-English and one Scottish-English. The first of these was constructed by the present author while constructing a monolingual Irish thesaurus (Scannell, 2003). The Scottish-English data were provided by Caoimhín Ó Donnaíle, from among the many resources he manages at Sabhal Mòr Ostaig[9].

It was deemed desirable in constructing the bilingual lexicon to select cognate translations when possible, even at the risk of making the Scottish translations sound a bit "Irish", as a way of emphasizing (and maybe reinforcing) the common literary heritage of the two languages.

Finding cognates is straightforward. We first used a *fine-grained mode*, which applies a number of simple spelling changes to Scottish words to make them look as "Irish" as possible (grave accents made acute, *chd$* → *cht$*, *achadh$* → *ú$*, *sg* → *sc*, etc.). Pairs are then deemed to be cognates if the normalized Scottish word has edit distance zero or one from the Irish word, and if, in addition, they share at least one English translation. The *coarse mode* works similarly, but in this case both Irish and Scottish words are converted to a coarse phonetic encoding (originally used to implement Philips' metaphone algorithm in our Irish `aspell` spellchecker[10]); this approach generated pairs of cognates with fewer false positives than by merely increasing the allowable edit distance in the fine-grained mode.

The requirement that potential cognates share at least one English translation was sufficient to avoid all examples of *faux amis* known to us. In some instances we were saved by the limited size of the Irish-English and Scottish-English databases. For example, there were no English translations in common for *cuan* (Ir. "bay, harbor, port, inlet, haven", Sc. "ocean, sea"), though the "harbor" sense appears (with the qualification "rarely") in Dwelly's magnificent unabridged Scottish dictionary (Dwelly, 2001).

Finally, when no cognates were found for a given Irish word, a "best guess" translation was made automatically using a metric based on the number of shared English translations and the corpus frequencies of the Scottish words. In all, we were able to produce translations for 21,106 Irish lemmas with this approach; 8462 of these have been verified by hand against print dictionaries, and evaluated for potential disambiguation.

Note that the lexicon only pairs up Irish citation forms with their Scottish equivalents. A morphological generator for each language is then employed to pair up the corresponding inflected and mutated forms – at present this amounts to nearly 200,000 distinct Irish words that the translator can handle.

We have not as yet made any attempt at evaluating precision, but we have some preliminary results on the (word-for-word) recall of the translator. First of all, we should point out that our aim is to have the translator handle a very broad range of texts from different genres and literary registers (newspaper articles, government documents, email lists, newsgroups, blogs, etc.) since it will eventually be applied to the full web corpus of Irish (see §4). With this in mind, we performed an evaluation on a corpus of 1.89 million words of text crawled from the web. As a baseline, note that the Gramadóir spelling and grammar checker underlying `ga2gd` (and sharing the same Irish lexicon) recognized 91.14% of the words in the corpus (the others consisting mostly of English pollution, but also some misspellings and a few truly unrecognized words). The recall of the translator, as measured on the subset of 1.72 million known words, was 92.72%.

3.7. Scottish Post-processing

To this point we have not thought very carefully about generating grammatically correct Scottish sentences. For this, we use the nascent Scottish version of the Gramadóir grammar checker to automatically make any necessary local corrections when there are incorrect initial mutations, etc., in the naively generated output.

We have used a similar approach in English → Irish MT, where one can blithely translate a fragment like "the man" as "an fear" without considering the wider context, but then in post-processing, if the wider context happens to be "with the man", one obtains "le an fear" which is corrected to "leis an bhfear" by the grammar checker. This approach allows the complexity of the translator to be focused where it belongs (on global syntactic issues and WSD), and moves trivial post-processing to an independently useful (and independently developed) monolingual application.

4. Applications: Cross-language Information Retrieval

The `ga2gd` software will be integrated into the Irish language search engine mentioned above as a "Translate this page" feature, allowing Scottish speakers to browse and read the substantial amount of Irish language material on the web. The eventual aim is to combine all three Goidelic languages in a single search engine, where queries can be made in one language, with all relevant documents in any of the others being returned, and translated if necessary.

5. References

Patrick S. Dinneen, editor. 1927. *Foclóir Gaedhilge agus Béarla.* Irish Texts Society, Baile Átha Cliath.

Edward Dwelly, editor. 2001/1912. *Illustrated Gaelic-English Dictionary.* Birlinn Ltd., Dùn Èideann.

[9] This is a good illustration of the power of refactoring existing resources for minority languages (even resources designed and built with entirely different projects in mind), and argues for making all such data freely available under an open source license.

[10] See `http://aspell.net/metaphone/` and `http://borel.slu.edu/ispell/index-en.html` for more information.

Seán Mac Maoláin. 1962. *Gàidhlig agus Gaeilge*. An Gúm, Baile Átha Cliath.

Kim McCone, editor. 1994. *Stair na Gaeilge*. Roinn na Sean-Ghaeilge, Coláiste Phádraig, Maigh Nuad.

Niall Ó Dónaill, editor. 1977. *Foclóir Gaeilge-Béarla*. An Gúm, Baile Átha Cliath.

Tomás Ó Rathaille. 1932. *Irish dialects past and present.* Institiúid Árd-Léinn, Baile Átha Cliath.

Antonio M. Corbí-Bellot, Mikel L. Forcada, et al. 2005. An open-source shallow-transfer machine translation engine for the Romance languages of Spain In *Proceedings of the 10th Annual EAMT Conference*, Budapest, Hungary, 30-31 May 2005.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In D. Yarovsky and K. W. Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94. Association for Computational Linguistics, Somerset, New Jersey.

Rannóg an Aistriúcháin. 1962. *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil.* Oifig an tSoláthair, Baile Átha Cliath.

Kevin P. Scannell. 2003. Automatic thesaurus generation for minority languages: an Irish example. In *Actes de la 10e conférence TALN à Batz-sur-Mer*, volume 2, pages 203–212. ATALA.

Kevin P. Scannell. 2005. Applications of parallel corpora to the development of monolingual language technologies. Preprint.

Nicholas Williams. 1986. *I bPrionta i Leabhar*. An Clóchomhar, Baile Átha Cliath.

Models of Cooperation for the Development of NLP Resources: A Comparison of Institutional Coordinated Research and Voluntary Cooperation

Oliver Streiter (1), Mathias Stuflesser (2) and Qiu Lan Weng (3)

National University of Kaohsiung (1), Department of Western Languages and Literature;
European Academy Bozen Bolzano (2,3), Institute for Specialised Communication and Multilingualism
ostreiter@nuk.edu.tw (1), mstuflesser@eurac.edu (2), steph_weng@yahoo.com (3)

Abstract

The lack of freely available resources is still the main bottleneck in the processing of the word's estimated 6000 languages. To overcome the current limitations and to achieve the impossible, one might profit from studying the models of cooperation in free software projects or Wiki-projects. In this paper, we explore such models of voluntary cooperation for the collection and elaboration of free NLP-resources. We describe the database XNLRDF which has been set up for this purpose and how data can be collected in a model of voluntary cooperation. A comparison with BLARK concludes this paper.

1. Introduction

Resources for the processing of Natural Languages are scarce. While most languages have no NLP resources at all, for those which have, the resources may be incomplete, insufficient in quality, inaccessible, very expensive or protected under copyright. The 'smaller' the language, the bigger the gaps. Proposals which intend to overcome this situation, differ in their, e.g. which languages included, their degree of concreteness, their focus, e.g. spoken vs. written language and their approach. The general failure, however, to provide a solid infrastructure for the processing of the world's languages is not as much a consequence of lacking motivation or inadequate scientific methods, as it is a consequence of inadequate models of scientific cooperation. Scientists thus should not only reflect upon how to advance science internal topics, but how to improve the architecture of science in a way that data and minds can cooperate seamlessly.

Recently, new ideas on how scientific cooperation might be organized have sprung up. These new models share the common feature of a seemingly unstructured bottom-up cooperation between knowledgeable volunteers. More and more knowledge is thus created and managed by knowledgeable volunteers in domains previously restricted to affiliated specialist. For such a cooperation, a suitable software like CVS (http://www.nongnu.org/cvs/) or Wiki is used. It provides the platform for the communication and the management of the data. Sourceforge (http://sourceforge.net) and Wikipedia (http://www.wikipedia.org) are among the flagships of this movement, and as ongoing projects show, this movement does not stop before topics related to NLP. We thus can predict that this movement will influence the field of NLP in general.

The question of quality, frequently raised to object such models, seems no longer to be a principal concern, given the acknowledged quality of free software like Firefox or the quality of the articles in Wikipedia. What's more, these models have the advantage of a potentially unlimited number of people cooperating in a flat hierarchy, creating workforces which easily superate that of the most work-intensive academic NLP projects. Overall, it seems to us that understanding the potentials of these models of cooperation will be crucial to the question whether or not we will be able to overcome the current standstill in the creation of NLP-resources.

To substantiate our claim we will first sketch the main features of traditional coordinated research on the one hand and models of voluntary cooperation on the other. Then we will present XNLRDF, a project that tries to overcome the limitations in cooperation and the standstill in the creation of NLP resources. Contrasting XNLRDF to BLARK will reveal different but complementary conceptions and priorities. The two research models thus should find a way to interact and to share their respective advantages. As volunteer communities are expanding in all directions with or without the participation of affiliated researchers, it will be up to the latter to bring the two differently powered and differently paced dynamics together.

2. Institutional Coordinated Research

Institutional coordinated research is the established model of how science is organized in universities and research centers. This model of research is stamped by (a) its funding, (b) its limitation to time slices and (c) the definition of cooperation among partners.

Coordinated research is funded by a body which, more often than not, wants its money invested in what it perceives to be relevant to the financial resources of that body. Thus, research in France, paid by French tax payers is more likely to create NLP-resources for French than for Khamtanga. This creates a distorted relation between requirements and funding. As a consequence of this ego-centrism of the funding bodies, those languages, which have the smallest gaps in their infrastructure, receive most funding.

The second feature of institutional coordinated research is it's temporal limitation. Money is granted in well defined projects with starting and end points more likely to be determined by the accounting scheme of the funding body than related to inherent features of the project. Although this temporal limitation seems to be an organizational necessity, it is also obvious that topical or relevant research is an activity that should have started before the project and should continue after the project, otherwise research islands will be financed.

Finally, the cooperation between research units (e.g. univiersities or research centers) is organized in modules where the participating research units are autonomous within their modules and interact with other modules through specific interfaces, standards or protocols. In this model of cooperation, intellectual properties can be easily assigned to the research units. In addition, the consistency and coherence of the data within one module seem to be manageable. However, this model cannot take direct advantage of closely overlapping intellectual competences where a minor gap in one unit can be easily filled by another unit.

3. Models of Voluntary Cooperation

Funding is not a condition for a project to start. Volunteers cooperate on the realization of a content, be it software, lingware, translation, images or a new texts, despite the absence of any funding organization (c.f Bey et al. 2005). The only criteria for setting the research topic is the perceived relevance by the volunteers who, although not free of a ego-centric perception, can accommodate more easily to unbiased views. Thus, while in the institutional cooperation, no language resources are created for Khamtanga, except in Ethiopia itself, researchers from France and many other countries would contribute to the development of Khamtanga NLP-resource in the model of voluntary cooperation.

The shelf-life of such projects is much longer than the time span granted to institutional project, up to a decade or more. Although volunteers may drop out of a project defined within a voluntary cooperation, such projects tend to continue as long as the development seems to be relevant to the community. Thinothings temporal continuity is one of the principal assets which put such projects and not institutional project in state to improve, for example, the infrastructure for the processing of the world's languages.

Finally, the cooperation in these projects is less likely to be modular. This feature thus allows different people to work on the same data, so that overlapping but different knowledge resources can be merged. For this reason, this cooperation requires the support of a software which tries to minimize the friction between the cooperating units. A number of complementary techniques are used at these aims, depending mainly on the degree to which data are formally structured. A common feature is the fragmentation of data (e.g. in paragraphs in Wikipedia), so that each fragment can be worked on individually. Especially powerful are relational databases which provide maximal fragmentation. The overall coherence and consistency can be controlled through uniqueness constraints, references and triggers. Backup functions and human control complete the set of control mechanisms. Linking such a database to a web-interface, allows a person connected to the Internet to cooperate almost instantly without setting up a project design or a software.

4. XNLRDF, Exploring Voluntary Cooperation

In order to bridge the gaps between a) the needs of languages users, b) affiliated research and c) the potential contributions of non-affiliated voluntary researchers, we started to create an environment for the cooperation through the Internet with the aim to collect and elaborate NLP-resources for the world's writing systems (Streiter & Stuflesser 2005). Simply the scope of the project, with currently 23.000 writing systems, 8000 languages, textual examples of 600 writing systems and 150 scripts, makes it obvious that models of traditional institutional cooperation do not provide a straight solution for the management of such an endeavor. Funding of such a project by any national body is unlikely given its multi-national/multi-regional scope. In addition, the required intellectual contribution (e.g. language expertise and access to resources) is beyond what a group of people can achieve in their life-time. Therefore, models of collaborative work are explored to achieve a Wikipedia-like cooperation of researchers.

Currently, the discussion of data structures and the collection of the first data is done by a small circle of volunteers and a few affiliated scientists. Within one year however, we succeeded in creating a basic architecture for the development of fundamental NLP-resources and to populate the database with data from the writing systems of the world. The potential of these resources starts to get visible with an automatic language recognizer and a basic spell checker working for currently about 600 languages.

The created NLP-resources are available in hourly builds under the GNU public license at http://140.127.211.214/research/nlrdf_download.html and intellectual insights related to the development of the resources are available under the Creative Commons License at http://xnlrdf.wikispaces.com. We hope that the circle of interested people might gradually enlarge, to open up finally for a free Wikipedia-like cooperation.

To prepare the project for a larger group of cooperators a number of design features have been proposed and are currently being implemented.

4.1 Relational database and XML

Not XML, but a relational database serves as backbone for data development. Modifications in the relational database affect individual data cells and thus provide a maximal fragmentation. This contrasts with XML which is normally organized in large text files.

Within the project, XML is used only for the exchange of data in RDF (cf. Manola and Miller 2004), hence the name XNLRDF. The database dump and a one-to-one representation of dump in XML can be already downloaded. An RDF will be designed which, due to the size of the download, will allow for extracts for single languages/writing systems.

The relational database has the advantage to integrate a client-server architecture for global collaborative work.

In addition to the standard clients for the management of the database and the data, there has been created a web interface that will allow volunteers to enter and check data. This interface, called the XNLRDF-browser, is accessible under http://140.127.211.214/cgi-bin/gz-cgi/browse.pl and allows its user to get an insight into the nature and wealth of the data. While currently the database still requires passwords for data modification, the system is gradually being opened up for voluntary collaboration as the appropriate checks have been tested.

Through the integration of simple tools into the XNLRDF-browser, which among others test the potential of XNLRDF, people should become motivated to enter data, e.g. to insert open licence-texts for a language to download seconds later a simple spell-checker.

4.2 Assuring quality

Quality checks are being installed at three levels: at database management level, at expert validation level and at the level of the voluntary user interface.

4.2.1 Coherence and consistency checks

Checks of coherence and consistency are being installed at the level of the database itself. These checks will thus apply to all kinds of servers that connect to the database. The checks can be defined to any level of complexity using triggers and functions.

Creating ambiguous data becomes impossible through uniqueness constraints.

References make it impossible to delete central data, e.g. a language referred to by a writing system.

The inclusion of false positives, e.g. pejorative language names, marked as deleted, makes it impossible to insert or inherit the same value again through the effect of uniqueness constraints.

4.2.2 Expert validation of data

The linguistic validity of inserted data is checked at the second level by a group of registered experts which delete unwanted data, freeze good data and wait for other data to be improved. At this level, the experts have to be informed on who created which data and how the data relate to the other data. Expert valdiation is to be done by two kinds of experts: by proficient speakers and by experts in formal or computational linguistics.

Letting linguists and native speaker experts declare an ever growing number of data in a network of data as unchangeable (freezing) will make the space for incorrect modifications smaller and smaller.

4.2.3 Guidance of voluntary users

At the level of the interface, voluntary users are guided to create as valid data as possible. On the one hand, this is done by making data fields obligatory and using picklists instead of free input. On the other hand, the database shows immediately the results of insertions, and the volunteers can see and check their input. Additionally, the setting up of the helpfile page for XNLRDF (http://xnlrdf.wikispaces.com/) has contributed to the usability of the database.

5. BLARK

BLARK, the Basic LAnguage Resource Kit tries to initiate coordinated actions to fill the gaps observed in the infrastructure for the processing of European languages. The project's aim is thus fully compatible with that of XNLRDF. The formal limitation of BLARK to European languages is not an inherent feature of the project but maybe necessary only to become 'fundable' and thus realizable in a traditional research framework. In fact, a BLARK-matrix for Arabic has been created (www.nemlar.org) beside BLARKs for European languages.

The idea of BLARK has been born in the Netherlands (Krauwer 1998), to be proposed as a project for the Fifth Framework Program of the European Commission. Unlike a project of voluntary cooperation this project idea didn't take off before any official funding was obtained. Revitalized in the Enabler Network, the concept gained the status of a reference according to which the development of NLP resources for a language can be measured and actions can be motivated. The amount of new language data however created under the direct influence of BLARK are relatively limited. Thus, although the Enabler Network shares our criticism of current funding policies and their incapacity to provide NLP-resources, the products of the network are basically theoretical in nature. Depending on national funding, BLARK thus lead mainly in the Netherlands and France to the production of new NLP resources (Mapelli & Choukri 2003). This minor impact on not more than 0.025% of the languages, falls back behind the ultimate goal of XNLRDF and BLARK which is to provide NLP resources to a great number of languages.

6. Conclusion

Thus, although BLARK and XNLRDF have similar goals and share a similar skepticism concerning the potential of current funding schemes to overcome the general misery in language resources, the contribution of BLARK is mainly in the development of concepts, schemes and metadata. Actually, nothing else could be expected, as the traditional models of scientific cooperation are not overcome but repeated within the project. XNLRDF however, without any funding and without a strong theoretical overhead, created very elementary language data (currently texts and wordlists, list of number words, function words etc) and very elementary tools (Liu et al. 2006) for about 600 languages through the participation of volunteers within one year. It is thus more than obvious, that bringing together the two research models, institutional coordinated research and voluntary cooperation, in one cooperative project design would provide the volunteers with sound conceptual features, whereas volunteers and language activists are willing and capable to provide urgently needed language data, especially for languages that start to be explored in electronic media.

To sum up, using BLARK and XNLRDF as examples, we tried to show that traditional models of scientific cooperation are unlikely to bridge the most urgent gaps in the development of NLP resources, while models of volunteer cooperation have the potential to do so. Researchers acquainted only with the first model of research should become aware of the enormous potential in voluntary cooperation and try to bring the two traditions together, profiting from the two motors the two traditions are powered by. Funding organizations should also acknowledge the potential of voluntary cooperation and support frameworks in which both traditions are alive.

7. References

Bey Y., Kageura K. & Boitet. Ch. (2005). A Framework for Data Management for the Online Volunteer

Translators' Aid System QRLex. In *Proceedings of 'PACLIC 19, the 19th Asia-Pacific Conference on Language, Information and Computation'*, 1st-3rd December 2005, Taipei, Taiwan.

Krauwer, S. (1998). ELSNET and ELRA: A common past and a common future, in *ELRA Newsletter* Vol. 3 N. 2. 1998.

Liu, D., Su, S., Lai L, Sung, E., Hsu, J. & Hsieh, S. & Streiter, O. Collaborative Development of African Language Resources. In *Proceedings of 'Networking the development of language resources for African languages'*. LREC Workshop Genoa, Italy, 22 May 2006.

Mapelli, V. & Choukri, Kh. (2003). ENABLER, European National Activities for Basic Language Resources, Deliverable D5.1, Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps, June 2003.

Streiter, O. & Stuflesser, M. (2005). XNLRDF, the Open Source Framework for Multilingual Computing, in: I. Ties (ed.) *Lesser Used Languages & Computer Linguistics*, European Academy Bozen Bolzano, Italy, 27th-28th October 2005. http://140.127.211.214/publs/files/nlrdf_lulcl_10.pdf

Developing a Spoken Corpus and a Synthesiser for Irish (Gaelic)

**John Wogan, Brian Ó Raghallaigh, Áine Ní Bhriain,
Eric Zoerner, Harald Berthelsen, Ailbhe Ní Chasaide, Christer Gobl**

Phonetics and Speech Laboratory, School of Linguistics, Speech and Communication Sciences
Trinity College Dublin, Ireland

{woganj, oraghalb, anibhri, zoernee, berthelh, anichsid, cegobl}@tcd.ie

Abstract

As one of the ‘lesser resourced’ languages, Irish (Gaelic) shares the disadvantage of other minority languages in lacking many of the resources that would be needed for the development of speech and language technology. This paper describes some of the linguistic, technical and practical difficulties presented in trying to put in place annotated corpora and resources for speech synthesis, and outlines how the developments took account of these challenges. Although the research is focussed on the provision of specific resources for a single dialect, the need for the long-term perspective is emphasised and the need to ensure maximum reusability of resources.

1. Introduction

Irish (Gaelic) shares the disadvantage of other minority languages in having almost no provision for the emerging speech and language technology, such as machine translation, speech synthesis and recognition, etc. Although these technologies are rapidly advancing and increasingly well understood, harnessing them for a new language requires many prior resources that are often lacking.

This paper deals with speech technology development, and thus relates to this workshop in the broader setting of what would be required for speech-to-speech machine translation. An important first step in the provision of speech technology for Irish is to develop a text-to-speech system. In principle, it should be a relatively straightforward matter to deliver such a system. In practice however, there are many perquisites that need to be in place and linguistic and practical issues that have to be resolved to enable the development of a synthesis system. Today’s commercial-grade synthesis systems for the ‘major’ languages are based on prior research and extensive spoken corpora, which have been fully annotated at the phonetic level, as well as pronunciation lexica, letter-to-sound rules, and often also sophisticated models of the prosodic and segmental features of the language. Thus, if one is not to sacrifice the quality of the eventual system, one needs to invest considerable time and effort to put these resources in place, either from scratch or by or adapting and upgrading existing materials.

These considerations informed our targets for Irish, within the Welsh-Irish project WISPR, funded by the EU-Interreg IIIA programme (www.tcd.ie/CLCS/phonetics/projects/prosody.html). Our objective was to develop a spoken corpus of Irish as a basis for synthesis, and to develop in parallel some of the other prerequisites for a text-to-speech system. Within the project, parallel work has also been carried out on Welsh, but in this paper we will describe the problematic issues needed to be considered for the Irish developments, and how during the development work the strategies adopted were geared towards extenuating these difficulties. Throughout, the emphasis has been less on the single application at hand, as on building a solid basis for the long-term development.

2. The situation of Irish

Irish is a Celtic (Goidelic) language spoken in Ireland. The Constitution of Ireland accords Irish the status as the first national language of the State. Under the Good Friday Agreement, Irish received formal recognition in Northern Ireland. As of 2007, Irish will also be recognised as an official language of the European Union. Despite the status afforded the language, the population of native speakers is small and decreasing, and there is little commercial incentive to develop speech technology resources.

Speech technology is particularly crucial for minority languages such as Irish whose future is precarious, and could contribute importantly to their preservation. Furthermore, as this technology is becoming increasingly crucial for education and access, particularly for people with disabilities, speakers of minority languages are becoming even more marginalised. Blind users of Irish have no access to electronic material in the language, and there is currently no way of providing synthesis-based communication devices to those that require them. In the pedagogical domain, speech technology could increasingly play a role in enabling the teaching and learning of the language. In an increasingly technological society, it would facilitate its widespread use among the broader Irish population. The lack of speech technology facilities in a minority language given their increasing availability and widespread use in the competing majority language will undoubtedly impact negatively on the sustainability of the former.

The difficulties that present for speech technology development may stem from a variety of sources. Many of the issues are specifically linguistic, such as the inherent complexity of the sound structure, the opaque orthographic system. Some others have to do with the deficit in the provision of necessary prerequisites: e.g., the lack of prior corpus collection, unavailability of resources such as pronunciation lexica (or in our case, lexica appropriate to the task), the availability of suitable prosodic and segmental analyses of the dialect, which are adapted to the needs of the envisaged developments.

Other issues arise from the present or historical context of the language. The fact that there is no standard dialect of modern Irish obliges us to consider long-term multi-dialect provision (see next section). Code switching between Irish and English also needs to be borne in mind: as most speakers of Irish are bilingual, and as English has such a dominant presence, code switching is a fairly typical feature for many native speakers, and embedded English words and phrases may be common in certain kinds of texts. Thus to deliver eventual Irish synthesis, we considered it wise to anticipate a future need for our synthesiser to also be able to 'speak' Irish English. Provision was therefore made for this in planning the corpus collection.

In the following sections we outline some of the major challenges to the intended development, and describe how the eventual research attempted to take account of them.

3. Reusability of existing resources: the challenge of dialect

There is no standard spoken dialect of Irish. There are approximately 150,000 native speakers living in pockets spread largely on the western fringes of the island (see Figure 1). The Gaeltachts (Irish speaking areas) can be broadly divided into three dialects which largely coincide with the provinces, i.e. Ulster (Donegal), Connaught (Connemara and Mayo) and Munster (Kerry, Cork). Any of the three main dialects would be an equally acceptable choice in the initial development of resources.

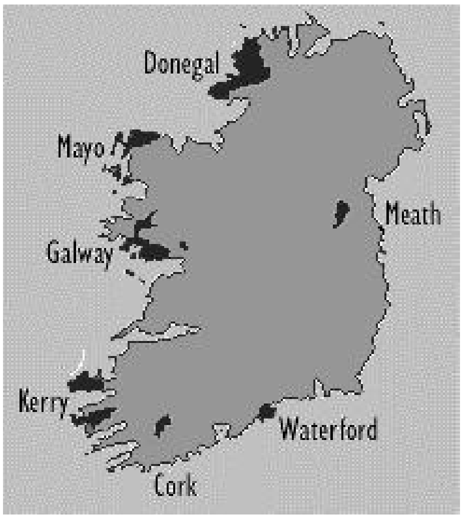


Figure 1: Map showing in black, the main regions where Irish is spoken.

The Donegal dialect was chosen for this project. From the outset, it was clear that whatever the initial dialect chosen, our long-term programme must encompass all three main dialects. Indeed, the Welsh-Irish grouping envisage a future extension of this research to the other Celtic languages and would furthermore strive to streamline methodologies, to enable them to be of wider use for similar developments with other minority languages. Thus, while the work is centred on a single dialect, many of the decisions made were at least partially guided by these long-term aspirations, and aimed to ensure reusability of resources and the establishment of an infrastructure for future developments.

4. Phonetic/phonological issues

For the task of speech synthesis, it is necessary to phonetically annotate a speech corpus. Before this activity can be undertaken, one must establish the system of contrasts used by the chosen speaker. While there is a long traditional of dialect studies of Irish, many of the studies based their analyses on the speech of older informants. As such, these resources may not represent the present day facts of the particular dialect. Questions arise in Irish dialects particularly about the inventory of laterals and nasals, and sometimes about the precise number of vocalic contrasts. This necessitates careful analysis of the chosen dialect as well as of the idiolect of the chosen informant.

The complexity of the phonetic and phonological systems of Irish presents particular challenges for corpus collection. Irish is complex from both phonetic (Ní Chasaide, 1999) and phonological (Ní Chiosáin, 1991) perspectives. Irish, possessing between 55 and up to 65 contrasts, depending on dialect and on the phonological approach adopted. A spoken corpus of any language needs to contain adequate coverage of all possible diphones, in as varied environments as possible. Consequently, for Irish, the corpus needs to be quite large if one is to be reasonably confident of ensuring adequate diphone coverage.

The large phoneset required for Irish is mainly due to the contrast of palatalised and velarised segments throughout the consonantal system. The secondary articulation of consonants has a major effect on the realisation of adjacent vowels. For example, Figure 2 shows a spectrogram of the word *buí* 'yellow' [b$^{\gamma}$i:].

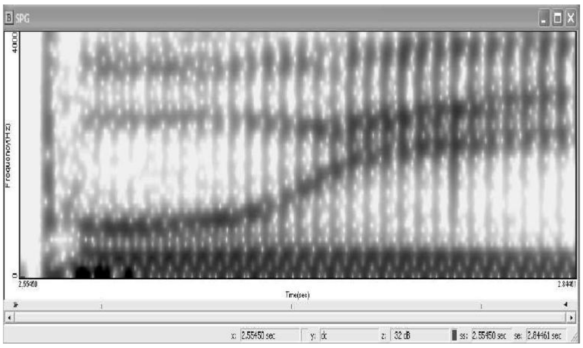


Figure 2: Spectrogram of the Irish word *buí* [b$^{\gamma}$i:] (yellow).

The *F*2 glide from the plosive release of the velarised [b$^{\gamma}$] to the steady state of the vowel [i:] is particularly noticeable, and this glide needs to be appropriately segmented as being part of the vowel. Such transitions are common when the vowel quality is particularly different from the (secondary) consonant quality, i.e. transitions from velarised consonants to front vowels and from palatalised consonants to back vowels. Thus, nearly all monothongs are potentially realised as diphthongs in this dialect (the degree of diphthongisation is somewhat variable across dialects). This feature of Irish presents quite a challenge to ensure full coverage of variants (see discussion below on the diphone corpus). It also poses a potential difficulty to the concatenation process in synthesis: as pointed out by Syrdal (2001), discontinuities are perceived more often in diphthongs than in monothongs.

Another issue that impinges on our efforts concerns particular gaps in prior linguistic analysis of the language. Despite the strong tradition of dialectology, all past descriptions have focussed on the segmental level and there has been virtually no coverage of the suprasegmental level. Having a good prosody model is vital to the provision of high quality text-to-speech. This gap in knowledge of prosody is being addressed in a concurrent project, Prosody of Irish Dialects (www.tcd.ie/CLCS/phonetics/projects/prosody.html). Within WISPR, while the lack of a prior prosodic model meant that we were not in a position to prepare even a demo diphone synthesiser, we have prepared the ground for such a development in the future, when the prosodic model is sufficiently elaborated.

5. A dialect-specific lexicon

The development of a lexicon for a minority language can also prove to be problematic. The only previously existing pronunciation dictionary was *An Foclóir Póca* (Rialtas na hÉireann, 1986). Since this dictionary provides some 15,000 words phonemically transcribed, one might expect that *An Foclóir Póca* could be used to bootstrap an automatic annotation process as well as the development of letter-to-sound rules. However, the forms and pronunciations in *An Foclóir Póca* do not correspond exactly to any one of the three spoken dialects, but rather represent an attempt at establishing a *Lárchanúint,* a 'Middialect' or a standard form, which compromises between the forms of all three.

In order to develop a dialect-specific lexicon for Donegal Irish, it was decided to adapt *An Foclóir Póca* to Donegal forms. To begin with, a short (20 min) corpus was carefully hand transcribed, and the words (orthographic and phonetic forms) were used to form a mini-lexicon for the dialect. This mini-lexicon was then compared to the forms in *An Focláir Póca*, and *sound-to-sound* rules were mapped, using the WAGON tool (www.cstr.ed.ac.uk/projects/speech_tools). Normally used to generate statistically based letter-to-sound rules, WAGON was in this case used to map between two sets of phonetic forms. The output rules were then applied to *An Focláir Póca*, in stages, beginning with the most common 500 headwords, to produce a Donegal lexicon. Before being added to the Donegal lexicon, the rule-predicted pronunciations were corrected by hand, to ensure that they were indeed in conformity with Donegal pronunciations. The Donegal-lexicon eventually included all 15,000 words of *An Focláir Póca*, plus 1,000 additional words gleaned from the 20-minute corpus.

Once this process was completed, lexicon development progressed alongside the automatic segmentation of the unit selection corpus and the development of letter-to-sound rules. Automatic segmentation involves forced alignment, and to do this the lexicon must contain all the words found in that corpus. Thus all words of the corpus, not yet in the lexicon needed to be added, along with their pronunciations. The Donegal letter-to-sound rules were therefore run on these new words, hand checked for accuracy and then added to the lexicon. This process was, reiterated until all the words of the corpus were entered, giving a total of 24,000 entries.

6. Letter-to-sound rules

Different methodologies can be used to generate letter-to-sound rules. As just mentioned, statistically based rules may be generated using the WAGON tool to map the correspondences between orthographic forms and the pronunciations in the Donegal-adapted lexicon. Through reiteration, with successive versions of the lexicon, one would expect increasingly accurate letter-to-sound mapping. The results using this approach turned out to be disappointing, yielding an unacceptably high error rate in the phonetic forms predicted. It may be that this approach may not in be well suited to the orthography of Irish.

A problem with statistically based letter-to-sound rules is that the rules themselves are not available in any explicit form. They can therefore not be scrutinised and corrected. Quite apart from this, as our long-term aspiration is to develop multi-dialect synthesis, we would in principle want to develop explicit letter-to-sound rules at some stage. Ideally this should be done in such a way as to differentiate between the common core of rules that hold across all dialects, and those that pertain to particular dialects, which might be viewed as a further layer of rules. This should in principle be the most interesting approach, not only as it yields important new linguistic information (both synchronic and diachronic), but also because it maximises the reusability of the resources we develop. As discussed earlier, though the quickest route to achieving a particular short-term goal is not necessarily be the best one: strategically 'reusable' approaches are important to our long-term interests.

For all of those reasons, handwritten letter-to-sound rules were subsequently encoded for use within Festival (www.cstr.ed.ac.uk/projects/festival). These rules were based in the first instance on *An Foclóir Póca* and Ó Baoill (1986), with rule adjustment for the Donegal dialect. The results of these handwritten rules have been encouraging, and the accuracy of their output has been much better than that achieved using the statistical approach. Consequently, the handwritten rules have replaced the latter, and now form the basis of the transcribed corpora and of the demo voice that have been developed. It is our expectation that only relatively minor further adjustments of these rules will allow us to begin the transcription of other dialects.

7. Developing annotated spoken corpora

Three corpora were collected to cater not only our immediate, but also our long-term objectives. They included a large corpus aimed at the eventual provision of a unit selection concatenative synthesis, an extended diphone corpus which will eventually enable the development of a diphone concatenative synthesis system, and finally, an Irish English recording of the ARCTIC (http://festvox.org/cmu_arctic/cmu) corpus, with a view to allow code switching in our synthetic voice. All corpora were recorded with the same speaker, a young female speaker of the Gaoth Dobhair dialect of Donegal. The recording conditions were also the same throughout, ensuring compatibility of corpora for eventual use within a single synthesis system.

The primary (unit selection) corpus involved 15 hours of recorded speech. Ideally, such a corpus should be designed to provide full coverage of all possible diphones of the language, in as many as possible environments, in the minimum of recording time. Clearly it was not possible at the outset to design such a corpus for Irish. Although we can figure out what coverage we need of distinct phones and can calculate how many diphones we need to cover, in the absence of any prior annotated data, or automatic segmentation facility for Irish, we could not calculate their frequency in any given text, or ascertain which were lacking. Thus, a large recording was in fact required: as the inventory of sounds is large and their interaction complex, the larger the corpus the greater the likelihood of all necessary sounds being captured along with their combinations in different contexts.

The texts for the unit-selection corpus were intentionally chosen from writings from this locality. This was deemed important, as it was feared that texts with forms and structures from other dialects might trigger dialect switching on the part of the reader, something that would introduce inconsistencies into our corpus. The novels of the Donegal author Séamus Mac Grianna (Máire) were the primary source. These were not available in electronic form, and were therefore scanned. However, as there is no optical character recogniser for Irish, the scanning resulted in numerous errors, and hand-correction was required.

The diphone corpus was recorded with a view to the future development of a diphone synthesis system, when a full prosodic model for the dialect has been elaborated. As the diphone corpus was recorded with the same voice and recording conditions as the unit selection corpus, it was also the intention to use it to extend the latter. By combining them, we wished to ensure complete coverage of all occurring sound sequences. This should prevent catastrophic failures of the eventual unit selection synthesis system that could conceivably arise from gaps in the coverage of the larger corpus. It should also be noted that for the purposes of building a diphone synthesiser, the large unit selection corpus will allow us to extend the diphone corpus further, by extracting sound combinations in more prosodic contexts.

Traditionally, diphones are recorded using nonsense words containing the diphones of interest, e.g. [daʃaʃa], is intended to capture the [ʃa] diphone and the [aʃ] diphone. These are elicited in citation form. We decided however against a ‘citation form’ diphone recording, given our interest in combining the two corpora. If citation-elicited diphones were concatenated with materials originally from the unit-selection corpus, it was feared that they would be temporally and prosodically out of kilter. To minimise this potential problem, diphones in the enhanced diphone corpus were elicited in sentence frames.

The diphone corpus recorded in WISPR was substantially enriched beyond what is usually included in a diphone set, to take account of the complexity of the sound system of Irish. It includes, to begin with, a full set of cross word-boundary diphones. It also includes Consonant$_1$ – Vowel – Consonant$_2$ sequences, where all C_1V sequences were elicited in contexts where C_2 was either palatalised or velarised, so as to allow for the very different vowel allophones that arise in these contexts. Likewise for VC_2 diphones, they were also elicited to allow versions where C_1 was either palatalised or velarised. Syllables were also recorded to include clusters of the form CCV, CCCV, VCC. Although a minimalist approach would suggest 55 phonemes and about 3,000 diphones for Irish, the enriched diphone corpus amounted to over 11,500 units.

The third corpus was an Irish English recording of the ARCTIC corpus, a compact corpus designed to yield coverage of the phonemes of English. This was to allow for code switching which is common in Irish speech. Although less prevalent in texts, it is nonetheless frequent enough in texts which purport to be representations of true daily conversational styles.

8. Conclusions

Using the corpora and the resources developed under WISPR we have put together a first demo of an Irish speaking synthetic voice, using the Multisyn voice building facilities in Festival. To this is added an Irish English voice, based on the ARCTIC corpus.

We would emphasise that neither the corpora nor the synthetic voices are in any way complete. Many aspects of a full Irish text-to-speech system remain to be done (e.g., tokenisation, implementation of a prosody model). The Irish corpora will also require more work to eliminate errors in the segmentation. These segmentation errors are clearly highlighted in the demo Irish voice.

Nonetheless, the demo voice also amply shows that high quality text-to-speech is well within our reach. Clearly, the work we report on here is, we hope, the beginning of an extended programme to provide speech technology facilities for the dialects of Irish. By extending the Welsh collaborations fostered by WISPR, we aspire to collaborations involving other Celtic languages, including Scottish Gaelic and Manx (near relatives of Irish) as well as Breton and Cornish (near relatives of Welsh). It is also our hope that our experiences and solutions to the linguistic and resource difficulties encountered, will be of use to others who envisage similar work on their languages. The most important message, perhaps, is to take the long view. To maximise the return on our efforts, we need to ensure that the research done provides not necessarily the *shortest* route towards the development of a specific application, but prioritises rather the establishment of an infrastructure for further developments.

9. Acknowledgments

This research was funded by the EU Interreg IIIA Community Initiative Programme.

10. References

ARCTIC, http://festvox.org/cmu_arctic/cmu

Festival, http://www.cstr.ed.ac.uk/projects/festival

Ní Chasaide, A. (1999). Irish. In *The Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press, pp. 111–16.

Ní Chiosáin, M. (1991). Topics in the Phonology of Irish. Ph. D. dissertation, University of Massachusetts.

Ó Baoill, D. P. (1986). *Lárchanúint donGhaeilge*. Dublin: Institiúid Teangeolaíochta Éireann.

Prosody of Irish Dialects, IRCHSS-funded, http://www.tcd.ie/CLCS/phonetics/projects/prosody.html.

Rialtas na hÉireann (1986). *An Foclóir Póca*, Dublin: An Gúm.

Syrdal, A. (2001). Phonetic effects on listener detection of vowel discontinuities. In *Proc. of Eurospeech* 2001, pp. 979-982.

WAGON, http://www.cstr.ed.ac.uk/projects/speechtools

WISPR, Welsh and Irish Speech Processing Resources, EU/Interreg IIIA project, 2004-2005, http//:www.tcd.ie/CLCS/phonetics/projects/prosody.html

An Attribute-Sample Database System for Describing Chuvash Affixes

Pavel V. Zheltov

Chuvash State University,
Russia, 428025, Gouzovskogo 21/37,
tchouvachie@narod.ru

Abstract

In the paper is described **“**KÜLEPEK” **–** a database system created by the author for description of Chuvash word-changing and word-forming affixes. The system is based on the attribute-sample model of affixes, which allows to describe affixes and their phonological, morphotactic and orthography rules. The system uses DBase IV database engine and was created in Borland Delphi 7.0 software environment, has a user friendly interface and can run under Windows 98/200/ME/XP. It is applicable for a large number of agglutinative languages with a finite-state morphology and can be used as a part of a morphological parser, and as an independent reference tool as well.

1. The Attribute-Sample Database System’s Common Structure and Description

The attribute-sample model of morphology (Zheltov, 2003) is based on following principles:

1) division of affixes to types;
2) conforming to each type a number of patterns, which describe phonological, morphotactic and spelling rules.

Chuvash language[11], being an agglutinative one, is mainly based on affixes and their interaction with stems and each other. In Chuvash morphemics are widely present such phonological variations (Sergeyev, 1992) as:

1) quality synharmonism - “soft” (front vowel) stems agglutinate “soft” allomorphs, while “hard” (back vowel) stems agglutinate “hard” allomorphs. As a rule each affix has minimum two allomorphs – a “soft” one and a “hard” one:

anne “my mother” + e (dat.-gen. affix) = annen**e** “to my mother”;
laşa “horse” + a (dat.gen. affix’s allomorph) = laşan**a** “to horse”.

But some affixes have only “soft” allomorphs (like 3-rd pers. sing. possessive affix -ĕ/-i “his/her”), thus a distorting of. synharmonism is observed sometimes:

šırăvĕ “his/her letter”.

2) interphonemes insertion:

anne “my mother” +e (dat.-gen affix) = anne**n**e “to my mother/my mother”.

3) vowels reduction (elisia)

a) in word formation:
vat šın < vată šın – “old man”,
purnăš < purănăš – “life”, from purăn – “to live”
b) in word changing:
vula (“to read”) + ăp (future tense 1-st pers. sing. affix) = vulăp “ (I) shall read”.

4) consonants reduction:
in ten verb stems, ending on -r, *r* is falling out in some verb forms:

pır (“to come”) + -t- (past tense affix) + -ăm (1-st pers. sing. affix) = pıtăm “(I) have come”.

5) consonants duplication in noun stems, ending on ă/ĕ, combined with final vowel reduction, when placed to dative-genitive case:

tulă “wheat” + -a (dat.-gen. case affix) = tulla.
sĕlĕ “oats” + -e (dat.-gen. case affix) = sĕlle.

6) final vowels alternation:

u – ăv,
ü – ĕv.

šıр**u** “letter” + ĕ (3-rd pers. sing. possessive affix) = šır**ăv**ĕ “his/her letter”,
vĕren**ü** “studying” + ĕ (3-rd pers. sing. possessive affix) = vĕren**ĕv**ĕ “his/her studying”.

The ***n*** in the example above as well as in ***laşana*** is an interphoneme, placed when a stem ends on a vowel and the agglutinating to it affix also begins on a vowel. Interphonemes are also being used in Chuvash in some other cases.

From this point of view Tatar language (a neighbour Turkic language of Kipchak group) has also two allomorphs of dative case: -a, -ə, while its others allomorphs -ga, -gə, -ka, -kə, -na, -nə are compound ones, decomposed to the allomorphs -a, -ə and interphonemes -g-, -k-, -n-. But from the formal point of view the representation accepted by Chuvash linguists is more comfortable, especially for computer analysis.

The database system has an interface structure, consisting of 6 tables that can be optionally filled for each affix:

Affix	Allomorphs	Morphologic feature	Type
A	a	Case (dative-genitive)	1
	e		

Table 1: Affixes.

[1] Chuvash language belongs to the Bulgar group of Turkic languages, together with extinguished Bulgar and Hazar and counts near 1,6 millions of speakers, their main part lives in Chuvash Republic and Volga region of Russian Federation. It is considered an endangered one and you can read more about the situation in (Zheltov, 2005).

Type	Stem – Affix	Stem˜ – Affix	Affix˜– Affix	Exception
1	1	2	3	4

Table 2: Type – Application rules.

Left context	Allomorph	Transformation	Result	Example
BC	a	+a	BCa	курак-а
FC	e	+e	Все	кĕрĕк-е
Ca	a	+н+а	Сана	лаша-на
Ce	e	+н+е	Сене	пике-не
Си	e	+е	Сие	пăри-е

Table 3: Stem–Affix.

Left context	Allomorph	Transformation	Result	Example	Exception
Сӳ	e	-ӳ ,+ĕв,+е	СĕВе	пĕлӳ–пĕлĕве	–
Су	a	-у,+ăв,+а	СăВа	çыру–çырăва	–
BCă	a	-ă,+с,+а	CCa	пулă–пулла	–
FCĕ	e	-ĕ,+с,+е	Cce	сĕлĕ – сĕлле	–
CCă	a	-ă,+а	Cca	карланкă – карланка	пуртă

Table 4: Stem˜– Affix.

Left context	Allomorph	Transformation		Result	Example
		Left context	Transformation		
ӳ	e	F	+н+е	ӳне	аннӳ–аннӳне
		FC	-ӳ, +н+е	Сне	кинӳ– кинне
		BCь	-ӳ,+ь,+н+е	Сьне	мăкăнӳ–мăкăньне
у	a	B	+н+а	уна	хулу– хулуна
		BC	-у,+н+а	Сна	арăму–арăмна

Table 5: Affix˜ – Affix.

Exception	Allomorph	Transformation	Result
пуртă	a	-ă, +с, +а	пуртта

Table 6: Exceptions.

The table “Affixes” contains the affixes (their lexical representation)[2], their allomorphs (surface representation of affixes), their morphological feature (in Chuvash there is usually 1 : 1 relation between an affix and morphological features it expresses) and the optional field “Type”. The field “Type” is used when in our database there exist already patterns of rules corresponding to the newly inputted affix, so we can just fill the type and the tables below will be filled automatically from existing patterns, with which the type value is related in the table “Type – Application rules”.

The table “Type – Application rules” is filled automatically by the system for each affix, after the user has filled phonological rules in the tables “Stem – Affix”, “Stem˜ – Affix”, “Affix˜ – Affix”, “Exception”.

The table “Stem – Affix”, describes contexts in which the current affix, when being glued to a stem, doesn’t cause any changes in the last one. The field “Left context” contains the finals of the stem, to which the current allomorph can be glued. In the field “Allomorph” the user enters an allomorph of the current affix, which can be glued in this context. The field “Transformation” contains the transformations, which have to be done to glue the allomorph. The sign “-” means reduction of a symbol before it, while “+” means addition. The Latin capital letter “B” means back vowel, “F” means front vowel and “C” consonant. In the field “Exception” are listed stems, the interaction of which with the current allomorph is exception to these rules. The field “Example” illustrates the application of each phonological rule.

The table “Stem˜ – Affix” describes contexts in which the current affix causes the change of a stem it is glued to.

“Affix˜ – Affix” is a table describing contexts when an interaction between the current affix and the affix that has been glued to the stem before it causes phonological changes in the last one.

While describing morphonological rules for Chuvash affixes, we have also encountered the recursion phenomenon.

The phenomenon of recursion on the morphology level is present both in Chuvash and Tatar, as well as in other Turkick languages. They are formed in Chuvash by:

1) relative affixes -ti/-çi.

Such recursive structures are translated into English with the means of relative pronouns – “who”, “which”, “what” and with the means of demonstrative pronouns “that”, “those”, “these”.

a) Yal+ti – “that, who/which is in the village”.
Village+ti.

b) Yaltisene – “to those who/which are in the village”
Village+ti+plural affix+dat-dir.

c) Yaltisençine –“to that, which/who is by those, who are in the village”.
Village+ti+plural+çi +dat-dir.

d) Yaltisençisene – “to those, who/which are by those, who/which are in the village”.
Village+ti+plural+çi +plural+dat-dir.

2) by the possesivity affix *-Ăn* (*-ăn*/*-ĕn*/*-n*). This affix is closely related with the possesivity’s category.

Let us have a hierarchy of possesivity’s relations: A⊂B⊂C… Then a following structure is possible: A-*Ăn*⊂B-*Ăn*⊂C-*Ăn*…:

*Tăvan appă**şĕn** ıvă**lĕn** a**çin** mănukĕ* – “the grandson of his/her elder sister’s son’s son (the grandson of cousin’s son)”;

3) by the affix of dative-accusative case **-***A* (-a/-e). Let us have an hierarchy of spatial relations. Then is possible a recursive structure A-*NA*⊂B-*NA*⊂C-*NA*…: *Aytar Pariş**a** universitet**a** texnika fakultetn**e** vĕrenme kayrĕ*. – “Aydar has gone to study **to** Paris, **to** university, **to** the technical department”.

4) by the locative case affix -*TA* (-*ta*/-*te*/-*ra*/-*re*/-çe).

*Aytar Pariş**ra** universitet**ra** texnika facultetĕn**çe** vĕrenet*. – “Aydar studies **in** Paris, **at** university, **at** the nical department”.

5) by ablative case affix -*TAn* (-*tan*/-*ten*/-*ran*/-*ren*/-*çen*). *Francin**çen** Pariş**ran** universitet**ran** Aytar**tan** şıru kilçĕ*. – “A letter has come **from** France, **from** Paris, **from** university, **from** Aydar”.

As it can be seen from the examples above, the recursion phenomenon is an important one while parsing. In our database we describe it by adding the word “recursion” into the morphologic feature of these affixes.

Affix	Allomorphs	Morphologic feature	Type
Ти (Ti)	ти (ti)	Relative affix (recursion)	3
	чи (çi)		
Ăн (Ăn)	ăн (ăn)	Possesivity affix (recursion)	4
	ĕн (ĕn)		
A	a	Dative-accusative case affix (recursion)	4
	e		
TA	та (ta)	Locative affix (recursion)	4
	те (te)		
	ра (ra)		
	ре (re)		
	че (çe)		
TAн (TAn)	тан (tan)	Ablative case affix (recursion)	4
	тен (ten)		
	ран (ran)		
	рен (ren)		
	чен (çen)		

Table 7: Affixes.

Using our system we have created a database, which describes over 120 word-forming and over 50 word-changing Chuvash affixes, and are planning to create on its basis a morphological parser of Chuvash language. We have also compiled on its basis a reference tool (Table 8).

[2] The lexical representation is an abstract uniting entry for a group of allomorphs, expressing the same morphologic features. For example we can lexically represent the allomorphs’ set {-a,-e} of dative-genitive case like -A. The concrete allomorphs {-a,-e} are surface representations of the abstract affix –A.

Lexical representation	Allomorph	Context		
		Vowel quality	Vowel / Consonant	Rejection /Addition
A	e (кĕнеке+н+е)	F	e	The interphoneme ‘н’ is glued after the stem.
	e (пĕрч+е, сĕлл+е)	F	ĕ	Ĕ is rejected. If the word ends on. Ĕ preceded by a consonant this consonant is duplicated in two syllable words.
	a (карланк+а, пулл+а, пуртт+а)	B	ă	ă falls out. If the word ends on ă preceded by a consonant this consonant is duplicated in two syllable words. The exception is the word пуртă.
	a (ача+н+а)	B	a	‘н’ is glued after the stem. The exceptions are loanwords from Russian ending on ‘a’. In them ‘a’ is rejected and changed on ‘ă’ (машина – машинăна).
	e (пĕлĕв+е, анне – аннÿ – аннÿ+н+е, ĕне – ĕнÿ – ĕнÿ+н+е)	F	ÿ	‘ÿ’ is rejected. Exceptions are words, where ‘ÿ’ is the possessive affix of the 2-nd person singular. In them ‘ÿ’ is not rejected and the interphoneme ‘н’ is glued after the stem, before the allomorph ‘e’.
	a (ывăл – ывăлу – ывăл-н-а) the possessive affix of the 2-nd person singular	B	the stem ends on a consonant	‘у’ falls out: ‘н’ is added after the stem.
	e (кин – кинÿ – кин+н+е) the possessive affix of the 2-nd person singular	F	the stem ends on a consonant	‘ÿ’ falls out: ‘н’ is added after the stem.
	a (кино+н+а)	B	o	‘н’ is added after the stem
	e (медаль – медал-е, мăкăнь– мăкăн-е, тетрадь – тетрад+е)	B	дь, ль. нь	‘ь’ falls out (медаль - медале)
	e (мăнукĕ – мăнук+н+е) the possessive affix of the 3-rd person singular	B	ĕ	‘ĕ’/’и’ fall out, ‘н’ is added after the stem.
	e (çулçи – çулçи+н+е) the possessive affix of the 3-rd person singular	B	и	‘н’ is added after the stem.
	e (тетрачĕ – тетрадь+н+е, турачĕ – турат+н+е) the possessive affix of the 3-rd person singular	B	ĕ	If the original stem ends on ‘т’/‘ть’/ ‘д’/‘дь’, then ‘ч’ which appeared due to the possessive affix of 3-rd person singular is rejected. The stem is transacted back and the interphoneme ‘н’ appears.

Table 8: The table of a reference tool for Chuvash affixes.

2. References

Zheltov, P.V. (2003). Attribute-Sample Model of Lexics in a Comparative-Correlative Aspect. *Newsletter of Chuvash State University, Natural and Technical science,* № 2. Cheboksary: Chuvash State University Press, pp.131-136.

Zheltov, P.V. (2005). Minority Languages and Computerization. Situation in Russian Federation. *Ogmios Newsletter,*3.03 (#27), pp. 8-11.

Sergeyev, V.I. (1992). *Modern Chuvash Language. A Handbook of Morphology.* Cheboksary: Chuvash State University Press.

Author Index