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.

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.

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

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 Sherematyeva 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*.

Sherematyeva, 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.