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 postedited). 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/