# Joint efforts to further develop and incorporate Apertium into the document management flow at Universitat Oberta de Catalunya

**Luis Villarejo Muñoz**
Learning Technologies Office
Universitat Oberta de Catalunya
lvillarejo@uoc.edu

**Sergio Ortiz Rojas, Mireia Ginestí Rosell**
Prompsit Language Engineering
{sergio,mginesti}@prompsit.com

## Abstract

This article describes the needs of UOC regarding translation and how these needs are satisfied by Prompsit further developing a free rule-based machine translation system: Apertium. We initially describe the general framework regarding linguistic needs inside UOC. Then, section 2 introduces Apertium and outlines the development scenario that Prompsit executed. After that, section 3 outlines the specific needs of UOC and why Apertium was chosen as the machine translation engine. Then, section 4 describes some of the features specially developed in this project. Section 5 explains how the linguistic data was improved to increase the quality of the output in Catalan and Spanish. And, finally, we draw conclusions and outline further work originating from the project.

## 1 Introduction

Large institutions such as Universities must tackle important linguistic needs (press, marketing, web site . . . ). This is especially evident for virtual universities where written language is the basis of the everyday activity. In addition, Catalan universities are set in a multilingual environment where their administrative and learning materials must be translated into Spanish for non-Catalan speakers both within and outside Catalonia. Moreover, translations into English must also be performed not only for foreign students but also for the English versions of the university web site or the English version of university-sponsored journals.

Taking a look at some figures from 2008, UOC met correction and translations orders totalling 17,000 pages of text (4,590,000 words) -either in Catalan, Spanish or English- coming from more than 40 different groups within UOC. The amount of translated documents is so immense that the use of language technologies, particularly machine translation, is a must. Any other choice would result either in inability to meet the orders or a huge increase in bills for external language services. In addition, the use of a machine translation system enables the re-use of information, improves the homogeneity of the translation and optimizes the time invested by the language technicians freeing them from low value tasks.

Within this general framework, UOC has been developing free software resources (Villarejo et al. (2007)), such as a terminology extractor, that can aid the translation of academic and administrative documents. In order to take full advantage of these resources, they need to be organized around a machine translation engine that can adequately meet UOC's needs, which will be introduced in section 3.

## 2 Apertium

Apertium, presented in Ramírez-Sánchez et al. (2006), is a free rule-based machine translation system. It is being developed by a community of

J.A. Pérez-Ortiz, F. Sánchez-Martínez, F.M. Tyers (eds.)
*Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, p. 75–82
Alacant, Spain, November 2009

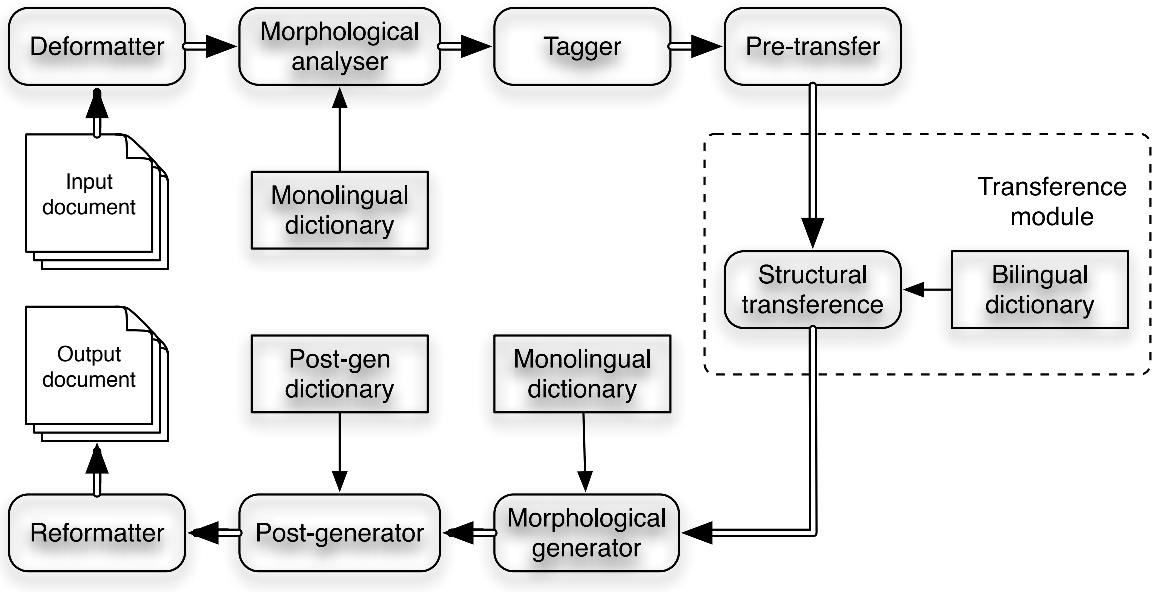

Figure 1: Apertium MT system

users worldwide. The system has a set of modules that are connected in a serial way to produce translations of texts in a diversity of formats.

The modules inside Apertium are the following:

- **Format processing**: These are the modules that perform the text extraction in a way that keeps the original format in the translation result at the end of the processing. Deformatter and Reformatter are the modules that perform this processing.

- **Lexical processing**: Apertium provide dictionary-based text transformation modules using finite-state techniques like Beesley and Karttunen (2003). The modules are the morphological analysis, which provides all the possible interpretations of each word; the bilingual module of the transference module, that performs the word-by-word translation of each word; the morphological generation, that generates the correct surface form for each lexical form of a word coming from the transference module; and the post-generation module that performs some orthographical tasks as contractions and apostrophes.

- **Lexical disambiguator**: The lexical disambiguator (*tagger*) provides the right lexical interpretation of an ambiguous text coming from the morphological analysis. It is based on supervised or unsupervised hidden Markov model (HMM) processing (Rabiner (1989)).

- **Transference**: The transference module, that provides one or three pass transference operation depending on the pair of languages, do the word reordering needed by the translation process and some other tasks as agreement of genders and numbers between the words in each sentence.

We found that Apertium is specially suitable for its integration inside universities like UOC for various advantages:

- **Open source**: Apertium is licensed under the GPL (GNU General Public License). This implies that the source code is provided with the application, and this allows UOC to adapt both the MT engine and the linguistic data to its specific needs.

- **Free software**: GPL requires all derivative software to be also licensed under GPL; this promotes the availability of all new source code developed for Apertium by the user community. Therefore, anyone using the system automatically benefits from new developments made by third parties, both on the engine and the data.

- **Predictability**: Being Apertium a rule-based MT system, the obtained results when translating documents are highly predictable. We think that this is an advantage against other non rule-based MT technologies for several reasons. Firstly, many of the systematic mistakes made by the MT system can be corrected in a systematical way. Secondly, human post-editors, once accustomed to the system behavior, are able to reduce the amount of work checking the original when post-editing a document. This reduction, which can even be automatized, makes its work simpler and more productive.

## 3 Apertium for UOC

UOC is an university in constant growth. This is especially evident in terms of languages. UOC started in 1994 using just Catalan as the academic and administrative language. In the year 2000 it incorporated Spanish with the Ibero-American campus. And in 2009, UOC has opened its Global Campus adding English to its educational offer. Right now university degrees, postgraduate training, open programmes and PhD programmes can be taken in English. Likewise, French has also been added to UOC's educational offer through master's and postgraduate programmes such as *Études islamiques et arabes* or *Interprétation des Fondements de l'Islam*. There are some 1,167 courses on offer as part of various master's degree, postgraduate and extension programmes and there are more than 54,000 students enrolled on these courses. More than 1,475,000 articles, studies and teaching materials have been translated and downloaded from the Virtual Campus. This all leads to a general framework where translations into several languages have to be performed on a daily basis.

The Apertium engine had to meet UOC's requirements in terms of two different aspects. The first regards production needs; i.e. the number of pages, the languages and formats involved, response time, quality, traceability, scalability, potential for growth and a strong user and developer community. The second regards UOC's mission where it is clearly stated that *promoting free software, accessibility and interoperability* is one of the institution's core values.

Most of these requirements were met by Apertium with no need for special developments. But, there were some specific requirements, especially regarding formats, that were not initially met by Apertium and led us to produce specific developments to satisfy them. These specific developments, which are described in sections 4 and 5, are very diverse, but it is worth mentioning here the need for a translation service not usually available from most MT engines, i.e. a translation service that could translate whole web sites. It is very common to structure the contents of a web site in folders. Once the web site is built in one language, we then need to translate all the information it holds into a second language. Compiling the files separately and sending them one by one to the translation service would have required an enormous time investment. However, if we have the possibility of compressing the web site contents and maintaining the folder structure, we can translate the web site in a few seconds and have it ready for post-editing and publication with no extra folder manipulation.

## 4 Developed features

In order to incorporate Apertium into the document management flow at UOC, the development of the following specific features, both in the machine translation engine and the user interface, was needed.

### 4.1 Machine translation engine

At the request of UOC, Prompsit developed the following new features in the Apertium MT System

- **Specific marking**: A new way to mark unknown words, and HTML tags to mark ambiguous words with colours. The ambiguous words marking feature was discarded and not considered useful.

- **Multiple meanings**: Apertium modes were implemented so it was possible to display the multiple meanings of a tranlated segment. To do this, Apertium uses the information present in some dictionaries or improving those dictionaries not having this kind of information available at development time. This feature allows one to display all the possible translations of a polysemous word in the target language. For example, the Spanish word *muñeca* can be translated into Catalan as *nina* ("doll") or *canell* ("wrist"). Text segments containing one or more words with multiple meanings are delimited by the patterns defined in the rules of the transference module. All possible translations are displayed, being the first one the one that contains the word marked as "default translation" in the bilingual dictionary. The output is presented as blocks of text between braces and the diverse meanings separated by vertical bars. For example, the phrase in Spanish *La dirección correcta* is translated in Catalan as {*La dirección correcta* | *L'adreça correcta*}. This work was implemented as the `apertium-multiple-translations` command and contributed to the trunk of the `apertium` module of the Apertium project.

### 4.2 User interface

- **Text translation**: The classic text box to allow translation of text pasted or typed in. It allows translation with and without information of unknown words and multiple meanings.

- **Document translation**: The purpose of this option is to allow translation of many different types of document, including HTML and ODT, and with the help of Openoffice, formats such DOC or XLS with acceptable quality. We also began the implementation of a filter to translate documents in PDF format generating ODT files using a commercial filter. This solution was not implemented finally due to technical problems but it will be in the future when these problems are resolved.

- **Translation-as-you-browse**: This translation service can translate web pages while browsing them, translating each link by modifying the source code of the links on each page.

- **Advanced HTML translation**: The purpose of this option is to provide more specific processing of HTML documents. In addition to the translation of these documents, they usually require some correction work and modification of certain parameters such as parts of the web addresses that change by changing the language of the document. For example if the document is translated into Catalan, the address `http://www.uoc.edu/masters/esp/web/index.html` would also have to have the link "translated" as `http://www.uoc.edu/masters/cat/web/index.html`, changing `esp` for Spanish (*español*) to `cat` for Catalan. This option also uses the Tidy HTML processing library to do some automatic correction in the HTML documents to fix systematic errors in the format.

- **Compressed archive translation**: The main use of this option is the translation of compressed archives. This kind of archive contains a series of documents of certain document types identified by known file extensions, and these files reside in directories that have some internal structure. The processing done by this option is the translation of all the files of known types respecting the internal structure of the subdirectories inside these files. The result is returned as a compressed archive of the same compression format as the original.

- **TMX creation**: This option allows the creation of bilingual TMX (Burns and Smith (1998)) memories with pairs of documents using the `apertium-tmxbuild` command. Every translation carried out inside the linguistic service is turned into a translation memory so it could be reused later for another translation order. This command was also included in the trunk of the Apertium project.

## 5 Linguistic data treatment

The linguistic work performed by Prompsit consisted of two separate tasks: on the one hand, preparing linguistic data needed by the Multiple Meanings feature explained in the previous section; on the other hand, improving the Spanish-Catalan translations of the engine as requested by UOC. By the time this improvement was designed, the Apertium es-ca translator reached a coverage of around 95% and a word error rate of approximately 5%.

The improvement of the linguistic data was conducted by means of three differentiated lines. The first line consisted of the improvement of the dictionary, the second the improvement of the lexical disambiguation module and the third the error correction regarding structural changes.

To be able to perform these improvements, a series of prior tasks were carried out. Firstly, a bilingual corpus (ca-es) was created from the UOC's web site. We used the translator that was available at that moment in Apertium to translate these texts in both directions (Spanish–Catalan and Catalan–Spanish) and, from this, we produced vocabulary lists containing frequent scientific-technical words and frequent words that were missing in the dictionaries. These lists would be used to add the most frequent unknown words to the system and to check the translation of the most frequent domain words.

Then, a comparison was performed between texts translated directly by Apertium and texts translated by humans; this allowed us to detect wrong translations, which we tried to solve in two ways: on the one hand, by enlarging and correcting the dictionaries; on the other hand, by adding new transfer rules to the transference module. We also used this texts to build translation memories to be used in the pre-translation modules.

Finally, an independent bilingual corpus (ca-es) of about 5,000 words was compiled to be able to set a baseline regarding the performance of the system before the linguistic improvements.

### 5.1 Dictionary improvement

As a result of the vocabulary lists mentioned before, we managed to detect diverse vocabulary that was not present in the engine. This detection allowed us to reduce the percentage of unknown words in the system. Another source of improvements was the detection of errors in the translations comparison: some errors due to a mistranslated or a wrongly disambiguated word could be resolved by adding a multiword expression in the monolingual dictionaries. This phenomenon can be seen in table 1. For example, the wrong Catalan translation *arribar tarda* instead of *arribar tard* from the Spanish *llegar tarde* ("to arrive late") is due to a tagger error: the word *tarde* can be a noun or an adverb, and after a verb (*llegar*, "to arrive") both options are probable. The way we resolved this was adding the multiword expression *llegar tarde* to the Spanish monolingual dictionary only for the analysis (not for generation, since the multiword does not exist in the Catalan monolingual dictionary, where no ambiguity occurs). The way the morphological analyser works, in a left-to-right-longest-match manner, gives only one possible analysis for these two words and no ambiguity must be resolved by the tagger.

| Multiwords | | |
|---|---|---|
| Source lang | Previously translated as | New multiword |
| *llegar tarde* | **arribar tarda* | *arribar tard* |
| *m'obre* | **me obro* | *me abre* |
| *cuenta con* | **explica amb* | *compta amb* |

Table 1: Examples of new multiwords

The use of regular expressions in the monolingual dictionaries for web addresses allowed us to recognize and leave untranslated web page addresses like *www.sistemes.com/documents/arxiu.htm* that, before this modification, the translator segmented in single words and translated into the target language (the former web address was translated from Catalan to Spanish as *www.sistemas.como/documentos/archivo.htm*).

Postgeneration errors were also detected in the translated texts. For example, translating from

Catalan to Spanish, the Catalan conjunction *i* can be translated into Spanish either as *y* or *e*. In principle, the postgeneration module made the pertinent operations to change *y* to *e* following words beginning with *i*; but we detected that the engine was using *y* following words starting with *hi*, where it should use *e* (except those starting with *hie*). The same happened with the Spanish conjunctions *o* and *u* followed by *ho*. Some examples of these phenomena can be seen in table 2.

| Catalan | Spanish | Correction |
|---|---|---|
| *pares i fills* | **padres y hijos* | *padres e hijos* |
| *plom i ferro* | | *plomo y hierro* |
| *nen o home* | **niño o hombre* | *niño u hombre* |

Table 2: Post generation corrections in the output.

To sum up, table 3 shows the figures on the number of lemmas added to each dictionary and the resulting total figures.

| Dictionaries | es | ca | es-ca |
|---|---|---|---|
| New lemmas | 4,337 | 4,371 | 4,477 |
| Total lemmas | 24,735 | 24,660 | 26,662 |

Table 3: Figures on the addition of vocabulary

### 5.2 Lexical disambiguation

The better the dictionary coverage, the more information available to improve the statistical module in charge of text disambiguation for the source language. As a part of this work, the Spanish and Catalan disambiguation modules were retrained with the updated dictionaries. This retraining gave as a result the improvement of several errors. For example, take the sentence *se debe a que los primeros sistemas* and its erroneous translation as **s'ha d'al fet que els primers sistemes* where the use of *deber* as a modal verb was translated as *haver de* while it should have been translated as *es deu al fet que els primers sistemes*.

### 5.3 Structural transference

Regarding the structural transference module, the comparison of documents translated by the engine with documents post-edited by humans shed light on a series of linguistic errors both in the ca-es and es-ca directions. We solved many of these problems by defining new transference rules in the transference modules of both translation directions. Figure 2 shows some examples of its application.

### 5.4 Linguistic evaluation

The evaluation of the linguistic data improvement was carried out by means of the previously mentioned corpus made up of 5,000 words coming from different sources. This corpus was translated and post-edited both in the ca-es and es-ca directions. The two elements to carry out this evaluation were dictionary coverage and word error rate (WER).

As can be seen in table 4, the coverage was pretty high before the improvement was carried out. Above 90-95%, it is difficult to achieve a remarkable improvement since the unknown vocabulary depends greatly on the domain of the translation. Most of the vocabulary added came from UOC's web site while the texts used for the evaluation come from newspapers. Regarding word error rates, the percentage improvement has outperformed the one achieved by dictionary coverage. Achieving a reduction of almost 1 percentage point in the es-ca direction and half a percentage point in the ca-es direction. This is explained by the improvement in the structural changes and high frequency linguistic phenomena.

| | es-ca | | ca-es | |
|---|---|---|---|---|
| | Before | After | Before | After |
| Coverage | 97.3% | 97.5% | 95.7% | 96.2% |
| WER | 4.86% | 3.93% | 5.52% | 4.94% |

Table 4: Results on the evaluation of the linguistic data improvement

## 6 Conclusions

The collaboration between UOC and Prompsit has given us a valuable experience about the problems that arise when integrating machine translation in the document management of an organization. The new features developed for this system have allowed us to explore the possibilities of automatic translation from an user

| **Spanish to Catalan** |
|---|
| *trabajaba **en** casa → treballava **a** casa*<br>(*trabajaba **en** Traducción Automática → treballava **en** Traducció Automàtica*) |
| *lo digo **como** amigo → t'ho dic **com a** amic*<br>(*vestìa **como** su amigo → vestia **com** el seu amic*) |
| *el ciudadano **medio** → el ciutadà **mitjà***<br>(***medio** ciudadano → **mig** ciutadà*) |
| **Catalan to Spanish** |
| *la informació que **cal** fer arribar als veïns → la información que **hay que** hacer llegar a los vecinos*<br>(*ens **cal** ajuda → nos **hace falta** ayuda*) |
| ***tots** quatre amics → **los** cuatro amigos* |
| ***tot** cantant → cantando* |
| *no **és al** domicili → no **está en** domicilio*<br>(*no **és a** tu → no **es a** ti*) |

Figure 2: Application of transference rules to solve translation problems between Spanish and Catalan

standpoint. Moreover, the developments made, both linguistic and in the translation engine, have been contributed to the Apertium user community so that the whole community can benefit from new improvements and expand the new functionalities in the future.

The system has been installed in UOC's infrastructure and is accessible through its web interface. Up to now, only selected users have been requested to use the system in order to detect further errors and improve the web interface before opening the system to the bulk of UOC's users. To be able to improve the web interface, we completed a user satisfaction survey involving 30 people who where asked about the interface functionality and usability. The results were very encouraging as the system achieved an overall mark of 4.3 out of 5 possible points.

## 7 Further work

Regarding formats, one of the needs that users at UOC have demanded is the possibility of translating files in Portable Document Format (PDF). Very frequently the translation of a PDF file is needed either for informational purposes or for further modification. As further work we have designed the implementation of a translation service that would accept such files as input for the translation process. The service would extract the text and the structure of the PDF file, translate it and re-build the file. The output of such a service would be either the PDF file translated into the desired language or a file in OpenOffice format which could be post-edited.

Regarding language pairs, the recent opening of the global campus marks a milestone in translation needs. As we have explained before, now English and French will play a central role in the translation activity of the university. And it is very likely that other languages like, for example, Arabic, will be incorporated in the university educational offer. With this general framework in mind, the translation needs of the university are in constant grow and the improvement of the MT engine regarding French, English and other languages will constitute a central goal in the following years.

## References

Beesley, K. R. and Karttunen, L. (2003). *Finite state morphology*. CSLI Publications, Stanford, Calif.

Burns, W. and Smith, W. (1998). Opentag and tmx: Xml in the localization industry. In *SIGDOC*, pages 137–142. ACM.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

Ramírez-Sánchez, G., Sánchez-Martínez, F., Ortiz-Rojas, S., Pérez-Ortiz, J. A., and Forcada, M. L. (2006). Opentrad apertium open-source machine translation system: an opportunity for business and research. In *Proceeding of Translating and the Computer 28 Conference*.

Villarejo, L., Moré, J., and Vàzquez, M. (2007). Proyecto restad: Herramientas de código libre para la traducción y postedición de documentos. *Proceedings of the Free/Libre/Open Source Systems International Conference*, pages 262–268.