Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 525–530

ISBN 978-83-60810-27-9
ISSN 1896-7094

A web-based translation service at the UOC based on Apertium

Luis Villarejo, Mireia Farrús
Office of Learning Technologies
Universitat Oberta de Catalunya
Av.Tibidabo, 47. 08035. Barcelona (Spain)
Email: {lvillarejo,mfarrusc}@uoc.edu

Sergio Ortiz, Gema Ramírez
Prompsit Language Engineering, S.L.
Avinguda Sant Francesc, 74, 1L.
03195. L'Altet (Spain)
Email: {sergio,gema}@prompsit.com

***Abstract*—In this paper, we describe the adaptation process of Apertium, a free/open-source rule-based machine translation platform which is operating in a number of different real-life contexts, to the linguistic needs of the Universitat Oberta de Catalunya (Open University of Catalonia, UOC), a private e-learning university based in Barcelona where linguistic and cultural diversity is a crucial factor. This paper describes the main features of the Apertium platform and the practical developments required to fully adapt it to UOC's linguistic needs. The settting up of a translation service at UOC based on Apertium shows the growing interest of large institutions with translation needs for open-source solutions in which their investment is oriented toward adding value to the available features to offer the best possible adapted service to their user community.**

I. Introduction

Machine Translation (MT) is one of the classic tasks of Natural Language Processing (NLP) and still an ongoing problem. Since the first attempts, dating from the 1950s [1], the presence of MT in a multitude of assimilation (understanding) and dissemination (publishing) scenarios has increased, as has interest in producing and accessing multilingual content. Companies, public and private institutions, and individual users have looked for solutions to cover their needs and this increasing demand has led, in the last years, to a growing interest of big companies (such as Google[1]) or big public-funded projects (such as EuroMatrix[2]) for the field of MT. The number of MT initiatives has risen greatly in recent years, mainly in statistical MT, as a result of the availability of vast multilingual parallel texts, but also in rule-based MT, example-based MT or hybrid systems. Many of these efforts have been released in the last decade as free/open-source systems (FOSS)[3] making MT available to the whole user community and not just to restricted groups.

This is the case of the Apertium[4] FOSS MT platform presented in this paper. Apertium is a framework in which rule-based machine translation systems can be created. It was first released in 2005 with only two available language pairs while today there are more than 20 stable language pairs available. Apertium has, since then, been chosen by a range of users to meet a number of MT needs. Examples can be seen in a wide variety of scenarios:

- individual users interested in using or developing the Apertium platform,
- less-resourced language communities, interested in increasing language visibility,
- research groups interested in carrying out R&D projects related to MT or NLP,
- companies interested in using MT or in improving the platform for internationalization or to offer commercial services,
- companies from the translation or localization industry interested in increasing productivity,
- public administrations interested in promoting languages and language technologies or,
- academic institutions working in multilingual environments.

From the scenarios above, we can see that Apertium is used in many different real-life contexts[5]. Apertium is, for example, used to publish online bilingual versions of *La Voz de Galicia* newspapers in Spanish and Galician, to generate, via translation, book reviews in many languages on the online Casadellibro.com bookshop or, recently, to produce bilingual versions of the University of Alicante (UA) website in Spanish and Catalan. Beyond Spain, Apertium is being used in companies such as Autodesk, where it is used to produce rough translations from Spanish into Brazilian Portuguese [2] as part of its localization workflow.

One of the most successful industrial applications of Apertium is the online MT web service at the Universitat Oberta de Catalunya (Open University of Catalonia, UOC). A long-term project developed within the UOC in collaboration with Prompsit Language Engineering, as the Apertium service provider, was set up in July 2008. The aim of this project was to improve and adapt Apertium to the UOC's needs for a translation service on its virtual campus. All the adaptations and improvements made by the UOC are free and open-source, and available on the Apertium platform. The UOC has become a regular developer on the Apertium platform, making

[1] http://google.translate.com
[2] http://www.euromatrixplus.net/
[3] See FOSS MT systems at http://computing.dcu.ie/∼mforcada/fosmt.html
[4] http://www.apertium.org
[5] http://www.translationautomation.com/technology/open-mt-ready-for-business.html

978-83-60810-27-9/09/$25.00 © 2010 IEEE

contributions and benefiting from the improvements made by the whole Apertium community.

This paper describes Apertium and its application in the context of the UOC. Section II describes the Apertium platform and its main innovative features. Section III sets out the language needs at the UOC, and how Apertium has been integrated in response to these needs. Section IV presents the evaluation and user feedback obtained after integrating Apertium, and finally, section V details the main conclusions and work for the future.

II. Apertium

This section includes a description of the tool: the general architecture, a more specific technical description, and its historical evolution.

A. General Architecture

Apertium is a rule-based MT platform providing the engine, tools and data for a large number of languages under the GNU General Public License[6], and it is being developed by a community of users worldwide. The platform has been described in depth in papers such as [3]. Here is a brief overview of the platform, its history and the main innovative features.

B. Technical Description

The Apertium MT engine is a shallow-transfer system consisting of pipelined independent modules that intercommunicate using text streams (see the architectural diagram in Figure 1). Modules can be used in isolation and other modules can be added to the pipeline. Data and engine are fully decoupled to make the engine language-independent. During the translation, finite-state lexical processing, statistical disambiguation and shallow structural transfer based on finite-state pattern matching takes place. Linguistic data feeding the engine are coded in XML-based files which are compiled into binary format (finite-state letter-transducers [6]) to speed up the translation process (10,000 words can be processed per second on a basic desktop PC). An extensive documentation is also available[7].

The technology behind it is largely based on that of systems already developed by the Transducens group at the University of Alicante (UA), such as the Spanish–Catalan MT system interNOSTRUM[8] [4], and the Spanish–Portuguese translator Traductor Universia[9] [5].

Apertium-based systems consist of three different packages:

- `apertium`: MT engine modules for format management (deformatters and reformatters), part-of-speech tagging (tagger) and transfer tasks (structural transfer).
- `lttoolbox`: toolbox for processing and compiling letter transducers into which data are transformed (morphological analyzers and generators, lexical transfer and post-generators).
- `apertium-l1-l2`: language package containing XML-based data for translating between two languages, l1 and l2, (monolingual and bilingual dictionaries, post-generation dictionaries, data for part-of-speech tagging).

Moreover, the system has the following modules that are connected in a serial way to produce translations of texts in a diversity of formats:

- **Modules for format processing**: they are in charge of separating (without removing) the original format information from text to be translated and restoring the format at the end of the translation process. The **deformatter** and the **reformatter** are the modules that perform this processing.
- **Modules for lexical processing**: they use the information contained in the monolingual and bilingual dictionaries. These are:
  - the **morphological analyzer**, which provides all the possible lexical forms (consisting of lemmas and morphological information) for each word (surface form) in the original text;
  - the **lexical transfer** of the transfer module, which performs the word-by-word (or multiple-word-by-multiple-word) translation of each lexical form delivered by the morphological analyzer and, if needed, disambiguated by the lexical disambiguator;
  - the **morphological generator**, which generates the correct surface form for each lexical form of a word coming from the transfer module;
  - and the **post-generator** that performs some orthographical tasks such as contractions.
- **Lexical disambiguator**: it provides, based on probability estimates, one single lexical interpretation (and the most probable but not always the correct) of an ambiguous word for which the morphological analyzer delivers more than one lexical form.
- **Structural transfer module**: it performs one or three pass transfer operations (depending on the language pair) to apply structural changes between source and target language such as gender, number or case agreement, reorderings, changes in verb tenses (including clitics) or verbal structures, changes in prepositions, generation or deletion of partitives, articles, prepositions or subject pronouns for non-pro-drop languages, etc.

C. Historical Evolution

Since 2005, with the first release of the engine (which was aimed at dealing with closely related languages), tools and the Spanish↔Catalan and Spanish↔Galician pairs, the Apertium platform has been extended greatly:

[6] http://www.gnu.org/copyleft/gpl.html

[7] The Apertium Wiki provides documentation of a wide variety of development and usage scenarios (http://wiki.apertium.org/).

[8] http://www.internostrum.com

[9] http://traductor.universia.net

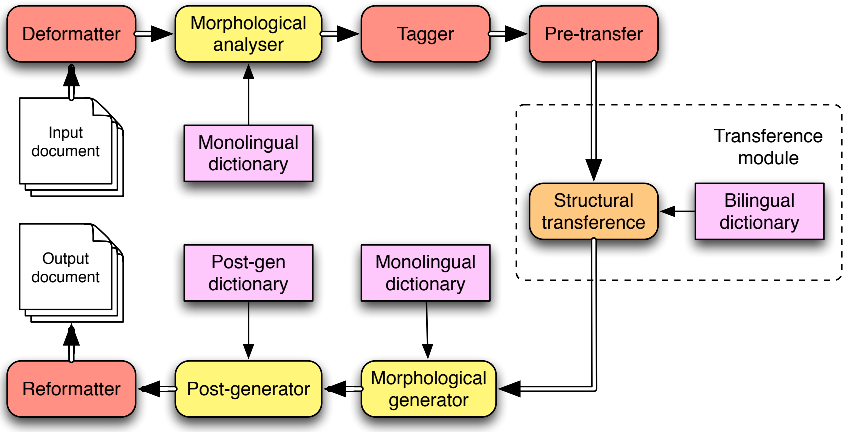


Fig. 1. Modules of the Apertium machine translation system

- more than 22 language pairs have been released[10] and many others have been started or are in development,
- the engine has been improved to deal with less related languages (thanks to the three pass transfer) and to be Unicode compliant
- file format support has been extended to all Office formats, Quark-Xpress, special XML-based formats, etc.,
- support for translation memories has been enabled (still experimental),
- language variants, polysemic or specific domain management has also been enabled,
- applications and tools have been developed for Apertium, such as Tinylex[11], a version of the bilingual dictionaries for mobile devices; `apertium-subtitles`, a tool for translating subtitles; user interfaces or add-ons for Firefox, and
- research on social MT development based on Apertium is currently being developed as part of a project entitled Tradubi [7].

All these efforts result in the continuous improvement of the Apertium platform and are carried out by a worldwide community of developers and committers acting individually or as groups in companies, organizations, research groups, etc.

We found that Apertium is especially 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**: Given that Apertium a rule-based MT system, the obtained results when translating documents are highly predictable. We believe that this is an advantage over 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 their work simpler and more productive.

Some of the improvements described above, as well as linguistic improvements in existing language pairs, were made by Prompsit as part of the development project designed by the UOC as described in the following section.

III. Practical Developments for Adaptation to the UOC

The UOC is a private Catalan online university whose mission is to provide people with lifelong learning and education. The UOC community is made up of more than 54,000 students, over 2,000 teaching counsellors and faculty working alongside an administrative staff of around 500 people. More than 1,475,000 documents have been downloaded from its Virtual Campus[12], including articles, studies and teaching materials. Given these figures and the fact that the university mainly works with three languages: Catalan, Spanish and English, there is the vital need to make intensive use of language technologies. Thus, the Office of Learning Technologies[13] and the Language Service[14] at the UOC have been developing several language tools [8] in order to exploit and reuse the

[10] http://wiki.apertium.org

[11] http:www.tinylex.org

[12] Data from academic year 2007-2008

[13] http://learningtechnologies.uoc.edu/

[14] http://www.uoc.edu/serveilinguistic/

language data generated by the University. In this context, several machine translation tools were evaluated. Apertium, thanks to its modular architecture, fully customizable nature and coverage of the languages used at the UOC, was adopted and improved on to meet the University's needs. Once Apertium was selected, it was integrated with the aforementioned language tools in order to create a document flow [9] for semi-automation of language processes. Below is a description of the translation service functionalities available and the user interface developed as part of the integration process.

A. Features Installed and Improvements Developed

Apertium has been installed on the UOC's Virtual Campus and is available both for faculty and administrative staff. The translation service has been set up for Catalan, which is the language used most frequently at the University, and the supported language pairs are Catalan↔Spanish, Catalan↔English and Catalan↔French. The functionalities implemented include:

- text translation
- document translation
- translation as you browse
- advanced HTML treatment (with structure validation)
- compressed files
- TMX utilization creation

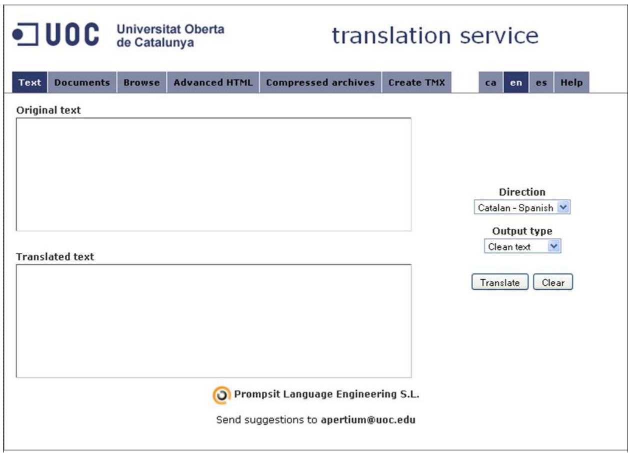
The integration in the Virtual Campus involved a series of specific actions in order to meet the UOC's needs. An upgrade was performed on the dictionaries using specific vocabulary extracted from the UOC's website. Regarding formats, specific support for Microsoft Excel and PowerPoint was developed. The marking of unknown and ambiguous words was adapted to ease post-editing.

Moreover, support for compressed files was developed so that users could send rar, zip, 7z, tar.bz or tar.gz files and receive an equivalent file with its contents translated.

In terms of the translation workflow, the use of translation memories prior to the machine translation phase was introduced in order to reuse the information generated within the university. This is especially useful in an institution where some departments generate similar documents, in which only small parts of them are modified, and many of the sentences translated could be reused in future translations. Moreover, since professors need to generate documentation in very specific domains, translation memories can help ensure more accurate and personalized translations.

B. User Interface

A special user interface was developed to integrate all the functionalities adopted or developed to create the translation service. This interface is available in Catalan, Spanish and English. Its different functionalities are structured by means of browser-like tabs as shown in Figure 2. Help is provided by means of HTML pages, and a downloadable PDF file, with detailed information on each functionality.



Direction: Catalan - Spanish

Output type: Clean text

Send suggestions to apertium@uoc.edu

Fig. 2. Apertium user interface

TABLE I
ERRORS OBSERVED WHEN TESTING THE FORMAT QUALITY OF ELEVEN DOCUMENTS.

Part of document	DOC	PPT	XLS
titles and subtitles	1	0	0
paragraph structure	0	1	0
bold and italics	0	1	1
figures	0	1	0
tables	0	1	1
headings and footnotes	3	0	0
lists	0	0	0
apostrophes	3	0	1
content (missing parts)	0	0	0
strange characters	2	0	0

IV. EVALUATIONS

Before releasing the first version of the translation system in December 2009, a series of tests were carried out in order to check its effectiveness in handling formats and the user interface's usability. To do so, we followed some of the ideas introduced in [10] and [11]. This section briefly outlines the design and the results obtained in both tests, performed by nine people from the university staff who volunteered to evaluate system quality for different formats and the usability of the interface.

After the release, the developers compiled user feedback from the suggestions and questions submitted to an address given for this purpose. The last part of this section summarizes the comments, frequently asked questions and suggestions made by users in order to improve interface usability and format handling.

A. Format quality test

The aim of the format quality test was to evaluate whether the use of different file formats led to problems in the translations or not, focusing on DOC, PPT and XLS formats. Table I shows the errors encountered by the nine evaluators in different parts of documents for each file format: titles and subtitles, figures, missing parts or strange characters in the translation, etc. A total of eleven documents were tested by the evaluators.

It can be seen from the table that most of the errors encountered were related to Microsoft Word documents and headings, footnotes and the appearance of strange characters.

Nonetheless, when enquiring about system performance when translating compressed files (zip), no problems were encountered: the format and folder structure were maintained in the output file, and all the files included in the compressed folder were completely translated.

This test allowed us to identify and solve minor problems when dealing with formats. This was true in particular when translating Microsoft Office documents, due to their closed internal structure.

B. *Usability test*

The goal of the usability test was to evaluate the general level of user satisfaction regarding a number of aspects. We mainly wanted to detect possible inconveniences in the information layout and the user's ability to perform the intended tasks. Table II shows the results obtained where *NA* stands for *not answered*, *0* means *disagree entirely*, *1* means *disagree*, *2* means *agree*, and *3* means *agree entirely*.

TABLE II
DEGREE OF SATISFACTION IN THE INTERFACE USABILITY TEST

Concept evaluated	**NA**	**0**	**1**	**2**	**3**
easy-to-use interface	2	0	0	3	4
information easy to find	2	0	1	2	4
information clearly organized	2	0	1	2	4
submitted tasks completed	3	0	0	2	4
understandable help	3	0	1	3	2
help easy to find	3	1	0	2	3
easy to access interface	2	0	0	3	4

As can be seen from the table, generally speaking, the interface's usability was rated highly.

C. *User feedback*

Once the translation system was opened to the UOC community, users had the opportunity to send their questions and suggestions. More than 400 users visited the web every month, and among the most frequent errors made were forgetting to select the format type via a drop-down menu. This menu left room for errors when selecting the appropriate format and this situation could have been avoided from the start by incorporating the automatic format detector that it is used when translating compressed files. Another of the most frequent errors was to try to translate a type of document that differed from the default extension, or cutting and pasting XML files into Word documents instead of translating them in XML format. This error stems from use of the previous machine translator used by the UOC community, where XML documents were translated in this way.

Among the suggestions received, the most frequent were to include the translation of PDF documents or automatic selection of the file format, whereby the user does not need to be aware of the type of file they wish to translate.

The feedback we obtained led to a series of actions. First of all, we compiled the most frequently asked questions in a document, providing answers to these questions. This document was made available through the user interface so it could be easily accessed by users.

D. *Linguistic quality*

An evaluation of the linguistic quality of the three available language pairs was carried out after the integration of the Apertium system at the UOC. These three pairs are at different levels of development:

- Spanish↔Catalan is the most actively developed pair inside the Apertium platform, having, for example, more than 41,000 bilingual correspondences in its dictionaries.
- English↔Catalan has also been part of various development projects. It has around 34,000 bilingual correspondences.
- French↔Catalan is the least improved pair, still considered a prototype, with around 12,000 bilingual correspondences.

The evaluation was done using the tool *apertium-eval-translation*, which compares the marked Apertium output of a text (unknown words are marked with a star before the word) and a post-edited version of the same text to calculate some evaluation variables based on edit-distance techniques. During our evaluation, two variables were assessed: *Coverage*, i.e. the percentage of words for which Apertium returned at least one translation, and *Word Error Rate* (WER), i.e. the percentage of words being post-edited to convert the MT output into the post-edited file, as shown in Table III. These results are the ones obtained after improving for Spanish↔Catalan as part of the development project carried out by the UOC. Catalan↔English is also being improved as part of two projects led by the Linguamon-UOC Chair in Multingualism. Catalan↔French is almost at the same level of development as it was before starting the project.

TABLE III
SAMPLE APERTIUM MT SYSTEM OUTPUT QUALITY.

Language pair	**Coverage**	**WER**
Spanish-Catalan	97.1%	4.3%
Catalan-English	94.1%	29.3%
Catalan-French	92.4%	21.9%

V. CONCLUSIONS AND FUTURE WORK

We have presented the details of the integration of a web-based translation service based on Apertium at the UOC. This integration was conducted together with a series of evaluations that lead us to make minor changes in the overall project in order to better meet UOC users' needs. The translation service has been running since December 2009 and the number of unique visitors in the first five months adds up to more than 600. Given that figure and the positive feedback obtained from the users via e-mail and the different evaluations made, we can say that the translation service has had a general good acceptance. Once we covered an internal testing period, and as a conseqence of the general good acceptance of the system, we decided to provide the translation service for the general public and it is now publicly available at http://apertium.uoc.edu.

In terms of future work, and in order to obtain more effective translations, we plan to include semantic domains that will allow for the disambiguation of polysemic and homonymous words. The semantic domains will be classified according to the subjects linked to studies at the UOC. For instance, the word *table* that can be translated as *tabla* (table of results) or *mesa* (dining table), depending on whether the word is related to mathematics or general vocabulary, would be clearly disambiguated through the use of semantic domains.

In addition to that, and inside the Google Summer of Code program, we are currently developing an online post-editing tool which will provide the user with a smooth integration of spell checker, grammar checker and dictionaries together with the Apertium platform.

Another direction for future work is improving the accessibility of the interface. In order to achieve this goal experts in accessibility are currently running a series of tests on the web service to be able to identify potential problems from this point of view.

In more general terms, user feedback will guide future work which will take into consideration improvements to the Apertium platform regarding new functionalities or increasing linguistic quality.

ACKNOWLEDGMENT

The authors would like to thank the Linguistic Service technician's fruitful suggestions which helped us to improve the overall system not only in terms of the linguistic information but also in terms of the user experience.

REFERENCES

[1] W. J. Hutchins and H. L. Somers, *An Introduction to Machine Translation*. Academic Press, London, UK, 1992.

[2] F. Masselot and P. Ribiczey and G. Ramírez-Sánchez, *Using the Apertium Spanish-Brazilian Portuguese machine translation system for localization*. Proceedings of the EAMT Conference, Sain-Raphaël, France, 2010.

[3] M. L. Forcada and F. M. Tyers and G. Ramírez-Sánchez, *The free/open-source machine translation platform Apertium: Five years on*. Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation FreeRBMT'09, Alacant, Spain, 2009.

[4] R. Canals-Marote and A. Esteve-Guillén and A. Garrido-Alenda and M. Guardiola-Savall and A. Iturraspe-Bellver and S. Montserrat-Buendia and S. Ortiz-Rojas and H. Pastor-Pina and P. M. Pérez-Antón and M. L. Forcada, *The Spanish-Catalan machine translation system interNOSTRUM*. Proceedings of MT Summit VIII: Machine Translation in the Information Age, Santiago de Compostela, Spain, 2001.

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

[6] A. Garrido-Alenda and M. L. Forcada and R. C. Carrasco, *Incremental construction and maintenance of morphological analysers based on augmented letter transducers*. Proceedings of the TMI, Keihanna/Kyoto, Japan, 2002.

[7] V. M. Sánchez-Cartagena and J. A. Pérez-Ortiz, *Tradubi: open-source social translation for the Apertium machine translation platform*. The Prague Bulletin of Mathematical Linguistics, 2010.

[8] L. Villarejo and J. Moré and M. Vázquez., *Proyecto RESTAD - Herramientas de código libre para la traducción y postedición de documentos*. In Proceedings of the FLOSS (Free/Libre/Open Source Systems) International Conference, 2007.

[9] L. Villarejo and D. Cullen and A. Corral, *La integració de les tecnologies de la llengua en el flux de treball del Servei Lingüístic de la UOC*. Llengua i ús, revista tècnica de política lingüística 46, 2009.

[10] G. Letnikova, *Developing a Standardized List of Questions for the Usability Testing of an Academic Library Web Site*. Journal of Web Librarianship, vol. 2(2–3), pages 381–415, 2008.

[11] P. Gore and H. G. Sandra., *Planning Your Way to a More Usable Web Site*. Online, vol. 27(3), pages 20–27, 2003.