Using the Apertium Spanish-Brazilian Portuguese machine translation system for localization

François Masselot & Petra Ribiczey
Autodesk Development Sàrl
Rue Puits-Godet 6
CH-2002. Neuchatel. Switzerland
{Francois.Masselot
Petra.Ribiczey}@eur.autodesk.com

Gema Ramírez-Sánchez
Prompsit Language Engineering, S.L.
Av. Sant Francesc, 74, 1L
03195. L'Altet. Spain
gramirez@prompsit.com

Abstract

We present a user case of the free/open-source Spanish ↔ Brazilian Portuguese Apertium machine translation system inside the localization workflow of Autodesk. This system, initially developed to perform general-domain translations, has been customized by Prompsit to fit with Autodesk needs and by respecting the localization workflow as much as possible. This original scenario shows that post-edited machine translation can generate immediately significant productivity gains with publication-ready linguistic quality.

1 Introduction

Autodesk is a leader company in design software and localizes its products and their user documentation from English into 20 languages. Offering products in the user's language gives an obvious competitive advantage and helps gaining market share. However, when fully-fledged localization involves costs out of proportion with the sales forecasts in emerging markets, low-cost localization solutions must be explored, otherwise the localization might be not happen at all. One of the 3D-mechanical design flagship products of Autodesk was in this situation: the sales team requested in vain a Portuguese version for the Brazilian market for several years.

At Autodesk, low cost localization consists of localizing only the software user interface (UI), and not the user documentation, or the software UI and the *Getting Started* manuals, while leaving the reference, developers and other advanced manuals in English. This approach was chosen, resulting in 318,000 words of software UI and 111,000 words of the manuals (10% of the total documentation set) to be translated into Portuguese. In addition, machine translation (MT) combined with a full revision of the machine-translated material by professional translators was considered to help save on the translation costs and time.

Autodesk had been preparing the introduction of MT in its localization workflow for several months. The long-term MT solution was designed around statistical MT (Plitt and Masselot, 2010), but a combination of circumstances opened the door to an unusual, later proving to be successful, localization scenario for this project. Instead of translating from English into Brazilian Portuguese, we decided to use the Spanish version as the basis for the translation, resulting in chain translation between English and Spanish and then Spanish and Brazilian Portuguese. Once the Spanish version was completed, the plan was to machine-translate Spanish into Portuguese with the rule-based Apertium MT system, and perform the final revision between English and Portuguese, to cover for the mistakes resulting from chain translation. The use of standard CAT tools working at sentence level supported such a 3-way localization. The reasons for our choice were the following:

- Autodesk's in domain English → Portuguese data (600,000 words) appeared to be insufficient to obtain satisfactory results with statistical MT.
- An evaluation of several rule-based MT engines for the language pair English-Portuguese did not give satisfactory results for Autodesk's type of contents. We concluded that in case we used English as a source, post-editing of the user interface

strings would be unproductive due to the density of the text (very short strings with very specific vocabulary and meaning), and for documentation, to some extent, we would have to compromise on stylistic quality of output.

- The Apertium rule-based platform with the un-customized language pair was giving good translations from Spanish, whereas from English, the results were not as good. The proximity of Portuguese and Spanish in the Romance family of languages, a large vocabulary shared between the two, was making the so-called rule-based shallow-transfer approach an adequate solution. Chain translation, is generally a non-recommended practice, but in this particular case, it was combining advantages that no other scenario could offer.

The project goals for Prompsit, providing the service around the open-source platform Apertium, and to Alpha CRC, the translation agency in charge of the post-edition, were set like this:

- The linguistic quality of the final Portuguese translation should match publication quality as set by Autodesk for all its products and languages.

- The customization of the rule-based engine had to be performed to address problems in decreasing frequency, and stopped at a reasonable point in time, to avoid that further customization ended up being more expensive than manual corrections.

- The MT system had to produce translation of such a quality that post-editors should be able to review and correct between 4,000 and 6,000 words per person per day.

- The MT system had to be integrated with the CAT tools used at Autodesk. This goal was dropped latter, as it appeared that engineering costs involved were not in proportion with the rest of the project. The Apertium web application and the post-edition in Autodesk's standard tools actually reached the same result than a proper integration of the MT system.

Several features of Apertium made this project a success story where goals were met, as further analysis will show:

- No license cost (free software), all the investment goes into adding value to the language-pair with the appropriate domain customization.

- Brazilian-Portuguese was available as a target language in-the-box, where for some other MT systems customization of their continental Portuguese language would have been necessary.

- The customization proposed as a service included the addition of Autodesk's terminology, the new Portuguese orthography, and the addition of rules to comply Autodesk's style. Further localization projects could also take advantage of the MT system customization, as is the case nowadays. Once the customization was completed, the private web access allowed Autodesk's localization coordinators to manage the localization of components and updates autonomously according to internal schedule.

In section 2 we review the Apertium technology and the Spanish to Brazilian Portuguese language pair. Section 3 describes the customization process. In section 4, an evaluation of the output quality and post-edition team feedback will be reported and finally, in section 5 we describe ongoing and future work.

2 Apertium and its Spanish to Brazilian Portuguese system

2.1 Apertium overview

Apertium[1] is an free/open-source rule-based MT platform. It provides the necessary engine, tools and data for a large number of language pairs to build MT systems. All these components are distributed under the GNU General Public License.[2]

Building new systems or adapting an existing one in Apertium means simply writing the appropriate linguistic data for a particular language pair. Data and engine were fully decoupled in the original design to this aim.

[1] http://www.apertium.org
[2] http://www.gnu.org/copyleft/gpl.html

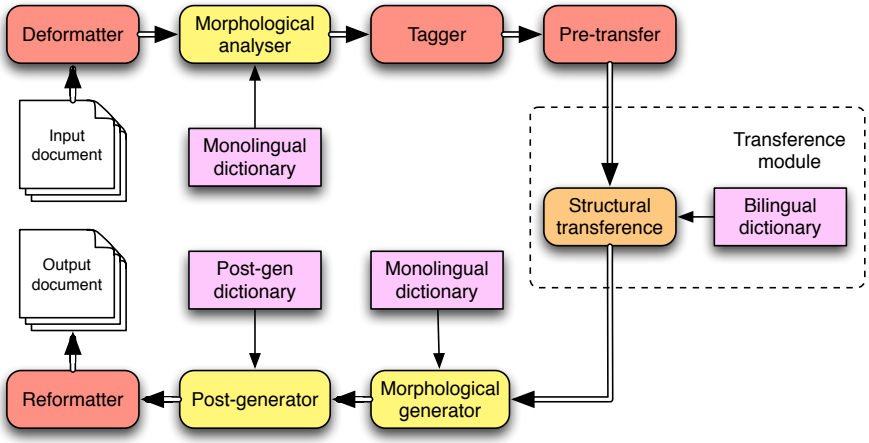

Figure 1: Apertium MT system

Linguistic data consist of monolingual and bilingual dictionaries, transfer rules and some other data useful for part-of-speech tagging (to help lexical disambiguation) and post-generation tasks (such as contractions or apostrophes). These components belong to the regular language pair package. It comes along with configuration and Makefile files to perform the compilation and installation of data to be used in binary format by the engine. This binary format (finite-state transducers (Garrido et al., 1999)) make Apertium very fast with very low hardware requirements (translation speed is around 10,000 words per second in a basic desktop PC).

All data are represented in XML-based formats to make them reusable and to ease interoperability. Both translation directions of a language pair can be represented in a single language package where normally the same monolingual and bilingual dictionaries (with restrictions depending on the translation direction) are shared. The other files are dependent on the translation direction.

The first two language pairs in Apertium (Spanish-Catalan and Spanish-Galician in both translation directions) and the first version of the engine and tools were released in 2005. The system was aimed at dealing with related languages and was inspired in the technology of two previous systems developed at the Universitat d'Alacant: interNOSTRUM [3] (Canals-Marote et al., 2001) and Traductor Universia [4] (Garrido-Alenda et al., 2004). Currently, more than 22 language pairs have been released and many others are started or in development.

The engine has been improved along the years to deal with not so related languages (3-level transfer), to be Unicode compliant and to give support to translation memories, new file formats, multiple translations of a given word, representation of language variants, polysemy or specific domain vocabulary.

A wide variety of applications and tools have also been developed for Apertium, such as a version of the bilingual dictionaries for mobile devices, a tool for translating subtitles, UI to ease diagnose, addons for Firefox, etc.

Engine, tools, data and other add-ons around Apertium are being developed by a worldwide community which involves individuals, research groups, public or private organizations and companies. All these efforts result in a continuous improvement of the Apertium platform.

2.2 The core

The system and the platform itself have been deeply described in many papers such as in (Forcada et al., 2009). We stress here only the features helping to understand the customization performed for Autodesk.

The system works as a pipeline of independent modules which produce raw translations of contents in various formats. The modules inside Apertium are the following (see figure 1):

[3] http://www.internostrum.com
[4] http://traductor.universia.net

- **Modules for format processing**: they are in charge of separating (without removing) the original format information from text to be translated and restore 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, that 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**, that 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 corresponding for which the morphological analyzer delivers to 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.

2.3 Linguistic package apertium-es-pt

At the time that Autodesk got interest in the Spanish-Portuguese pair of Apertium, `apertium-es-pt`, it had the the following characteristics:

- dictionaries had around 10,000 lemmata

- continental and Brazilian Portuguese variants, were represented in the package, although the Brazilian Portuguese variant did not comply with the recently adopted orthographical agreement in Brazil

- there were around 120 transfer rules in the dictionaries and structural transfer was a one-pass process

- vocabulary was complete for closed categories (such as determiners, pronouns, conjunctions or prepositions) and had the most frequent open categories (mainly noun, adjectives and verbs) according to general vocabulary extracted from various general domain resources

In this level of development, frequently considered a prototype by the size of the dictionaries, previous evaluations (Armentano-Oller et al., 2006) revealed that the system was ready to be used for dissemination purposes when dealing with general-domain texts.

3 Customization and setup

A development proposal to customize both engine and data to meet Autodesk's needs was presented and carried out by Prompsit. It included setting up a web service ready to be used by Autodesk's software engineers team. This section describes the customization process and setup.

3.1 Engine customization

The Apertium engine was modified to generate the mark @@@ instead of the default * in front of unknown words present wiht other meaning in Autodesk's contents. This behaviour was changed in the course of the project, based on the post-editor's feedback, to not mark unknown words at all.

New deformaters and reformaters were written to deal with Autodesk's exchange formats: TMX (Translation Memory eXchange) for documentation content, and CSV (Comma Separated Value) for software content:

- for TMX, Apertium receives a TMX having English and Spanish translation units variants

Original			Apertium MT output		
English	**Spanish**	**status**	**English**	**Brazilian Portuguese**	**status**
tahoma	tahoma	*Not Translated*	tahoma	tahoma	*Not Translated*
home view	vista pr&incipal		home view	vista &inicial	

Figure 2: Example on dealing with CSV special format.

for each translation unit and generates a trilingual TMX with an additional translation unit variant containing the Brazilian Portuguese, populated with the machine translation output.

- for CSV, besides having to convert UTF-16 into UTF-8, Apertium receives a CSV file with a fixed column number in English and Spanish and generates the same file in which the Spanish column has been replaced by the Brazilian Portuguese machine translation output. When a unit in the CSV is marked as *Not translated* or *Locked* or when the English and the Spanish column are identical, Apertium leaves them as in the original file. Two different CSV formats are currently supported by Apertium (see an example in 2).

Special attention was given to special characters, for instance Windows shortcuts represented by & inside words. As the engine cannot place them automatically in target words, Apertium was modified to dismiss them during translation and to place them in front of the translated word, which is most frequently the correct location.

3.1.1 Linguistic data customization

The linguistic customization was developed in different phases: compilation of resources and creation of suitable data for the Apertium MT system, term approval workflow, implementation and testing.

- **Compilation of resources and development strategy**

  Autodesk provided three different kinds of resources to Prompsit to customize the linguistic Spanish to Brazilian Portuguese language pair:

  - glossaries: bilingual Spanish ↔ English or English↔Portuguese glossaries, or multilingual (English, German, Italian, French, Spanish and Portuguese) glossaries from previously translated products inside Autodesk. They contained equivalences between terms in surface form without any further morphological information
  - translation memories from previously translated products from English to Spanish and English to Portuguese
  - the source texts (Spanish) to be translated into Brazilian Portuguese and that had recently been translated from English to Spanish

Monolingual and bilingual lists of domain adapted vocabulary were created from glossaries by semiautomatically converting them into a suitable format for the XML-based Apertium dictionaries.

Bitexts in Spanish and Brazilian Portuguese were created automatically by crossing English → Spanish and English → Portuguese translation memories. These texts were used to check equivalents and to infer transfer rules to be implemented according to Autodesk's style for Brazilian Portuguese.

A list of words unknown to the Apertium MT system sorted in decreasing frequency of occurrence was automatically created from the Spanish version of the text to be translated. As the number of entries was too large to be processed in the customization period, words appearing less than 15 times were not processed at all. The remaining words were contrasted with words in the glossaries to avoid double entries, which were removed from the list.

A similar work was done for bigrams and trigrams, i.e. two or three adjacent words, to create lists of multiple-word units having a different translation together than in isolation. A translation into Brazilian Portuguese for 1-2-3-grams in the final list was proposed by

Original Spanish sentence	Apertium output for Brazilian Portuguese	
	before customization	**after customization**
se puede mostrar	*pode-se exibir*	é possível exibir
pueden mostrarse	*podem exibir-se*	podem ser exibidos
se mostrarán	*exibirão-se*	serão exibidos

Figure 3: Examples of implemented structural rules from Spanish to Brazilian Portuguese.

looking in the bitexts or in other fonts and a final trilingual list (English, Spanish and Portuguese) was sent to Autodesk to double check them inside the regular term approval workflow.

- **Term approval workflow**

  As part of the usual localization preparation work, Autodesk uses a term approval process to ensure translators get the correct profession-related target language translation in the different products. We used this workflow for the Apertium engine customization as well.

  There were two phases:

  - In the first phase, in order to ensure consistency within all Brazilian Portuguese products, the initially extracted project-related key term list of 1709 terms was translated by the Autodesk's linguist and approved by the in-country Subject Matter Expert (SME) for the particular localized product.
  - In the second phase during the project the post-editing team provided feedback to the linguist and the SME, who communicated the necessary changes to the linguistic development team in Prompsit for analysis and implementation.

  There were 2070 terms approved in the process during the initial customization (first phase). Two minor updates were then performed based on the post-editors' feedback (second phase), adding or changing respectively 134 and 81 terms.

  This project being a pilot for machine translation, Autodesk increased the number of linguistic quality assesment tasks. These tasks consisted of several rounds of linguistic reviews of the newly localized content of the user interface and documentation. These checks revealed that (1) the linguistic quality is publication-ready and that (2) the linguistic issues found after post-editing were systematic minor corrections relating to terminology and its context.

- **Implementation and testing**

  Running in parallel with data preparation, the Apertium monolingual and bilingual dictionaries for Brazilian Portuguese were modified to meet the new orthographical agreement which came into force in Brazil in 2008.

  During double checking, Prompsit inserted into the Apertium dictionaries the terms extracted from glossaries and the bilingual list of frequent unknown words. As part of this task, all entries were classified according to a specific part-of-speech and inflection model (this information is required in the Apertium dictionaries). Two different strategies were followed:

  - to speed up the integration of glossaries (in surface form) into the system, it was decided to apply the correct part-of-speech but a general inflection model to each term for gender and number.
  - words in the list of unknown words were inserted with the correct part-of-speech and inflection model.

  After double checking, entries were modified (if needed) according to the SME's feedback.

  Also some transfer rules were implemented during customization, as the ones derived from the examples in figure 3.

  A test on random sentences of the Spanish files to be translated was carried out by Prompsit. The sentences were translated and reviewed by a native Brazilian Portuguese speaker to be sure that no frequent systematical contrastive phenomena were dismissed and to assess general output quality.

Vocabulary consistency checks and modifications were also performed inside dictionaries to be sure that domain-adapted terms were not interfering with general-domain terms.

3.2 Setup

Prompsit set up a web application with secure data transfer for Autodesk. This interface was used to translate from Spanish to Brazilian Portuguese the specific TMX or CSV files requested by Autodesk. An option to translate this kind of files in compressed format (.zip), was also set to speed-up the transfer time.

Software engineers from Autodesk can use the application at their convenience to generate or regenerate translations (for example, after implementation of changes provided by post-edition team). All improvements in engine and linguistic data are always available in their best version through the web application specially set up for Autodesk.

4 Evaluation

Two different kinds of texts were machine translated and post-edited: software user interface and documentation. The total amount of localized words was 429,000 words, 318,000 corresponding to software UI and 111,000 corresponding to documentation. This section is about the evaluation of the post-edited version and conclusions that can be drawn out of it.

4.1 MT output quality evaluation

In order to measure the MT output quality, Apertium output after customization was compared to the final post-edited files (see results in figure 4).

These figure allow to assess the impact of the different customization rounds. Different kind of scores were calculated:

The first three ones give an estimate of what was the post-edition effort done to convert the MT system output into the final post-edited version:

- Coverage, *naïve coverage* in this case, is the percentage of surface forms for which Apertium returned at least one analysis and one translation (not all possible ones, hence the word *naïve*).
- Word error rate (WER) is the percentage of words being post-edited to convert the MT output into the post-edited file.
- Edit-distance is the number of operations required to transform the MT output into the postedited text, edit operations being insertion, deletion, or substitution of a single character.

These three scores were calculated using the tool *apertium-eval-translator*.

The fourth one is the BLEU score (Papineni et al., 2002) calculated using the non-marked MT output and its post-edited counterpart used as a reference.

The figure also allows to compare the two content types (software and documentation), and how they behave with MT. The content types feature the following characteristics:

	Software	**Documentation**
Translation units	69,466	9,134
Words	318,000	111,000
Distinct words	18,000	5,600
Words/translation unit	4.57	12.16

Table 1: Software and documentation texts characteristics

Translators do report that the software material was easier to post-edit than documentation. But the edit distance score tell that the documentation was slightly less edited than software. This apparent contradiction tells that the post-editing time does not only go into making the necessary changes, but also into non-editing time (thinking, reference look-up, etc.) that is obviously not captured by edit distance scores.

The initial objective of post-editing was between 4,000 and 6,000 words per person per day. The best speed reached by the post-edition (PE) team was 6,175 words and the overall post-editing throughput must have been around 4,500 words. This project did not implement an accurate time tracking, so we cannot give more precise reports on productivity.

The overall project conclusion shows that the customization of Apertium is fully covered by the productivity gains generated by MT, and that the margin remains significantly positive. In this project, the turn-around time also benefits from MT: the gain in throughput compensates fully the three week spent in the customization of Apertium.

Stage	Coverage	WER	edit-distance	BLEU
before customization (SW)	*84.2%*	*31.4%*	*100873*	*50.99%*
after initial customization (SW)	*93.0%*	*25.3%*	*81136*	*58.83%*
after final customization (SW)	*93.4%*	*21.9%*	*70305*	*64.73%*
after final customization (DOC)	*98.5%*	*20.4%*	*23012*	*68.31%*

Figure 4: Evolution of the output quality of the MT system for software (SW) and documentation (DOC).

5 Ongoing and future work

The use of MT and the use of the Apertium MT system to translate from Spanish to Brazilian Portuguese is being continued in two other of Autodesk's localization projects. An analysis of whether it is recommended or not to customize the system for a particular product is done before setting up the translation. The more the new contents are similar to previsouly translated ones, the lower the amount of customization is needed.

Previously translated and post-edited strings from software UI have been used to automatically feed Apertium and, mainly on short strings, this has proved to be very efficient for further similar projects.

It is presumed that, for later products the customization costs will be lowering and, at a given point, it will be not necessary at all considering that the most frequent vocabulary will already be covered by Apertium.

Autodesk has contributed to the Apertium MT system by agreeing to free the vocabulary created during customization that Prompsit considered suitable to be inserted in the publicly available linguistic package. Prompsit will make it available along with the adaptation to the new orthography for Brazilian Portuguese in the next release of the main `apertium-es-pt` language pair.

6 Acknowledgement

Autodesk is thankful to the partner involved in this project: Alpha CRC team providing translation and post-editing service into Brazilian Portuguese.

References

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, and Miriam A. Scalco. 2006. Open-source Portuguese–Spanish machine translation. *Lecture Notes in Computer Science*, 3960:50–59.

Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Perez-Antøn, and M.L. Forcada. 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.

Forcada, Mikel L., Francis M. Tyers, and Gema Ramírez-Sánchez. 2009. The free/open-source machine translation platform Apertium: Five years on. In *Proceedings of theFirst International Workshop on Free/Open-Source Rule-Based Machine Translation FreeRBMT'09*, pages 3–10, November.

Garrido-Alenda, Alicia, Patrícia Gilabert Zarco, Juan Antonio Pérez-Ortiz, Antonio Pertusa-Ibáñez, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Miriam A. Scalco, and Mikel L. Forcada. 2004. Shallow parsing for Portuguese-Spanish machine translation. In Branco, A., A. Mendes, and R. Ribeiro, editors, *Language technology for Portuguese: shallow processing tools and resources*, pages 135–144. Edições Colibri, Lisboa.

Garrido, A., A. Iturraspe, S. Montserrat, H. Pastor, and M. L. Forcada. 1999. A compiler for morphological analysers and generators based on finite-state transducers. *Procesamiento del Lenguaje Natural*, (25):93–98.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Plitt, M. and F. Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.