Opentrad Apertium open-source machine translation system: an opportunity for business and research

Gema Ramírez-Sánchez, Felipe Sánchez-Martínez,
Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz and Mikel L. Forcada

Transducens group, Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant. E-03071 Alacant. Spain
{gramirez,fsanchez,sortiz,japerez,mlf}@dlsi.ua.es

Prompsit Language Engineering, S.L.
Polígon Industrial de Canastell 77, Despatx 3
E-03690 Sant Vicent del Raspeig. Spain
info@prompsit.com

Abstract

Most successful machine translation systems built until now use proprietary software and data, and are either distributed as commercial products or are accessible on the net with some restrictions. This kind of machine translation systems are regarded by most professional translators and researchers as closed and static products which cannot be adapted or enhanced for a particular purpose. In contrast to these systems, we present Opentrad Apertium, an open-source shallow-transfer machine translation engine originally intended for related-language pairs but currently being extended to deal with not so similar pairs. The opportunities offered by open-source software are very interesting in new research projects but they are also promising in business and particularly as a business model for innovative companies. Questions like which is the most suitable license to release open-source software, or how to make it visible are discussed in this paper. Real opportunities for research and business derived from the Apertium machine translation system are presented as well.

1 Introduction

Most successful machine translation (MT) systems built until now use proprietary software and data, and are either distributed as commercial products or accessible on the net with some usage restrictions. Most professional translators and researchers view them as closed products since they cannot be easily adapted to particular purposes, integrated with other applications or used as resources in research or development projects. Besides, these MT systems mostly use *ad hoc* formats for linguistic data which are unreadable and very hard to maintain or extend.

All these aspects of commercial systems have a negative impact on the development of new techniques or the addition of new language pairs. A better concurrence between developers would had led to a positive motivation to improve existing MT systems, but making new systems from scratch is so costly that usually the primary goals are constricted by what has been already done. For this reason, it seems as if we were constantly reinventing MT and both the techniques and the resulting translations are often very similar to those of ten or even twenty years ago.

Fortunately, in the last decade, propelled by the globalization of the Internet, open-source strategies have established as a sound development practice allowing for reuse of code and data.[1] Under this new situation, developers can now focus on improving and extending available software and data. In order to ease collaborative development, open-source projects are managed on centralised websites which also act as source code repositories. Another fundamental aspect for open-source projects to success is the availability of complete documentation describing it.

In the last years, open-source programs and data have also appeared in the field of MT, coexisting with commercial alternatives and bringing new opportunities which are proving very positive on both research and business areas.

We present in this paper a real case of these positive effects achieved by Opentrad Apertium[2] (Apertium for short), an open-source shallow-transfer MT engine originally intended for related-language pairs but which is currently extended to deal with not so similar pairs. Apertium is part of a large, government-funded development project[3] involving four universities[4] and three companies[5], all located in Spain. For them, these new opportunities have become realities.

A brief introduction to open source as well as an analysis of the dichotomy open-source MT versus closed-source MT are presented in section 2; section 3 shortly introduces the Apertium open-source MT engine; section 4 summarises the possible chances of open-source MT for research and business. Also, some successful cases on both areas derived from the Apertium system are mentioned. Finally, section 5 ends the paper with a brief discussion.

2 Open-source versus closed-source machine translation

Free[6] software, according to the definition by the Free Software Foundation[7], is software that may be:

[1] In fact, sharing of code was a common practice in the first days of software development (scientific computing, Unix, etc.).

[2] The Opentrad platform (`http://www.opentrad.org`) is composed of two different open-source architectures: Apertium (shallow-transfer MT system) and Matxin (deep-transfer MT system).

[3] The Apertium project has been funded through project FIT-340101-2004-3 (Spanish Ministry of Industry, Commerce and Tourism), with additional support from project TIC2003-08681-C02-01 (Spanish Ministry of Science and Technology).

[4] Universitat d'Alacant, Universidade de Vigo, Universitat Politècnica de Catalunya, and Euskal Herriko Unibertsitatea.

[5] Eleka Ingeniaritza Linguistikoa S.L. as project coordinator, Imaxin Software, and Elhuyar Fundazioa.

[6] Free as in *freedom* not as in *free beer*.

[7] `http://www.fsf.org`

0. freely executed for any purpose;

1. freely examined to see how it works, and freely modified to adapt it to a new need or application;

2. freely redistributed to and by anyone anywhere; and

3. freely improved and released to the public so that the whole community of users benefits.

Note that the availability of source code is necessary for conditions 1 and 3 to hold. In fact, a closely related concept is that of *open-source software* as coined by the Open Source Initiative[8]; for the purposes of this paper, both concepts might be used interchangeably. A license compatible with the principles of open-source software must accompany a program in order to consider it as open-source.

Open-source software practices can be extended to items beyond software as long as the concept of end product's source is clear enough. For example, in many MT systems the original text-based form of the dictionary is converted (or *compiled*) into a much more efficient binary form which, however, is also more difficult to edit (when possible); in this case, the text-based entries may be considered as the dictionary's source code.

The terms *commercial software* or *proprietary software* are commonly used as opposed to *open-source* software. In this paper, and as far as MT software and data are concerned, we will follow the approach by Forcada (2006) and use the dichotomy *open-source* MT versus *closed-source* MT.[9]

In the next section, the conditions which should be met by a MT system to be considered as open source are stated.

2.1 Open-source machine translation

Two main approaches have been followed so far when building MT systems: rule-based and corpus-based; prominent approaches for the latter are statistical MT and example-based MT. Hybrid approaches can also be found in the literature; however, in order to simplify the following discussion we will focus on *pure* systems. The common components that may be distinguished in both rule-based and corpus based systems are:

- an engine (recombinator, decoder, etc.),
- data (either explicitly-coded linguistic data, or monolingual and bilingual corpora), and,
- optionally, tools to maintain these data and convert them into a form suitable for the engine.

[8] `http://opensource.org`

[9] The terminology in the field is very ambiguous: some proprietary programs have their source code available, but cannot be considered open-sourced, according to the definition by the Open Source Initiative, because of some restrictive clauses in the corresponding license.

On the one hand, for a rule-based MT system translating between a particular language pair to be considered open-source, source code for the engine should be distributed as well as the source code for all linguistic data. It is much more likely for users of an open-source MT system to introduce changes in the linguistic data than to modify the engine. Moreover, in order to ease the former task it would be desirable to a certain extent that the maintenance tools are open-sourced as well.

On the other hand, for a statistical MT system (a particular case of the corpus-based approach), source code for the programs learning the statistical translation models from parallel corpora, and source code for the decoders using these translation models to generate the most likely translations of new sentences should be opened. It is also necessary to distribute the original sentence-aligned parallel texts.[10]

2.1.1 Examples

Some examples of active open-source MT systems projects are:

Apertium: toolbox to build open-source MT systems, initially suited to related language pairs but currently being extended to translate between not so closed languages.[11] There is a demonstration website[12]; the currently open-source available pairs (both translation directions) are Spanish–Catalan, Spanish–Galician, Spanish–Portuguese, Catalan–Occitan, and Catalan–French. See section 3 for a longer description. By the end of 2006, data for Catalan-English will be released.

OpenLogos: open-source version of the Logos commercial MT system.[13] Last system update was performed in January 2006, and there is no demonstration site. Opened available pairs are German and English as source languages, and some European languages (French, Spanish, Italian and Portuguese) as target languages.

PSMT (Prolog Statistical MT): according to the author, "an unsophisticated statistical MT program written in Prolog".[14] It is platform-independent although it has only been tested on Linux. Currently it works for English and French. A prototype demonstration is available at the project's website.

Open Source Toolkit for Statistical MT: as a result of the Johns Hopkins University Summer Workshop 2006,[15] a group of researchers "is refining and advancing state-of-the-art methods in an open-source toolkit for statistical MT".[16] The current system uses the Moses' decoder[17] and GIZA++[18] to learn statistical translation models from bilingual cor-

[10] This last requirement may sound strange, but it is actually the statistical MT analog of distributing linguistic data for a rule-based MT system.

[11] `http://apertium.sourceforge.net`

[12] `http://xixona.dlsi.ua.es/prototype`

[13] `http://logos-os.dfki.de`

[14] `http://psmt.sourceforge.net`

[15] `http://www.statmt.org/jhuws`

[16] `http://www.clsp.jhu.edu/ws06/application/opensource.shtml`

[17] `http://www.statmt.org/moses`

[18] `http://www.fjoch.com/GIZA++.html`

pora. The available baseline system is built from English–German, English–French and English–Spanish European Parliament corpora.[19]

Other attempts at open-source implementations of MT systems were initiated in the past, such as GPLTrans[20], Traduki[21] and Linguaphile[22], but their current level of activity is low or non-existing and they are far from reaching the usability levels of other systems.

2.2 Closed-source machine translation

2.2.1 Commercial machine translation

Most commercial MT systems are rule-based[23] although MT systems with a strong corpus-based component have started to appear.[24] Most of them have engines with proprietary technologies which are not completely disclosed (indeed, most companies view their proprietary technologies as their main competitive advantage). Linguistic data are not fully modifiable either; in most cases, one can only add new words or user glossaries to the system's dictionaries, and perhaps some simple rules, but it is not possible to build complete data for a new language pair and use it with the engine.

2.2.2 Non-commercial machine translation

One can also find non commercial but closed-source MT systems. For example, those systems on the web that may be freely used (with a varying range of restrictions); most of them are demonstration versions of commercial systems, but there are also some other freely-available systems that are not commercial. This is the case, for example, of interNOSTRUM (`http://www.internostrum.com`) a non-commercial but freely available MT system between Spanish and Catalan. Also Google's MT system (`http://www.google.com/translate_t`) is offered as non-commercial but freely available for translation.

2.3 Partially open/closed-source machine translation

Another possibility would be for the MT engine and tools not to be open-source (even using proprietary technologies) but just to be simply freely available and fully documented, with linguistic data being distributed openly (open-source linguistic data).

This kind of MT systems facilitates the improvement of existing data or the creation of new pairs of languages for translation. However, users are dependent of the engine and tools provided by the original development team and cannot control future support from them.

[19] `http://www.statmt.org/wmt06/shared-task/baseline.html`

[20] `http://www.translator.cx`

[21] `http://traduki.sourceforge.net`

[22] `http://linguaphile.sourceforge.net`

[23] Such as Reverso, `http://www.reverso.net`, or Babylon, `http://www.babylon.com`.

[24] Prominent examples of currently available corpus-based MT systems are AutomaticTrans, `http://www.automatictrans.es`, and Language Weaver, `http://www.languageweaver.com`.

2.4 Open-source licenses and software development management websites

After making up one's mind to implement an open-source MT system, a concrete license should be chosen for both the software and the linguistic data; although it is not essential, it is strongly recommended to decide how and where they will be published. Both choices are very important for the future of the project (Behlendorf 1999) and will be analysed next.

2.4.1 Software licenses

In most countries, copyright laws automatically impose severe restrictions (all rights reserved) on the distribution and modification of released software (and, in general, many other types of work); this way, unless otherwise stated, software is proprietary and the freedoms mentioned at the beginning of section 2 cannot be transferred to those who use the work. This implicit restrictions, however, may be explicitly modified by the author of the work via a copyright license.

In the late 80's, the Free Software Foundation created the *copyleft*, one of the most prominent examples of a copyright license tightening up the four freedoms of open-source. Besides those four freedoms, copyleft disallows people from changing the terms of copyright in derivative works. In addition, there are at least 60 different licenses in use for open-source software.[25] One of the main differences among them is their copyleft or non-copyleft nature; copyleft licenses force users to pass on their modifications, whereas non-copyleft licenses allow users to make their modifications private and distribute them as closed-source products. The most general and used licenses are:

The GNU General Public License (GNU GPL for short): a general copyleft license used by most GNU programs, and by more than half of all open-source software programs.

The Lesser General Public License: a weaker case of the GNU GPL used for program libraries which allow their use in closed-source products.

The BSD-like licenses: simple, permissive non-copyleft open-source software licenses. They come in many versions (FreeBSD License, Modified BSD License), whose permissiveness varies. It is used by the Apache server and the BSD version of the Unix operating system.

The Mozilla Public License (MPL): a weak copyleft license which may be seen as an hybrid between the BSD-like licenses and the GNU GPL.

2.4.2 Licenses for documentation

Documentation is crucial for open-source software projects to be successful. Just as special copyright licenses have been written for open-source software, so too, there are special ones to

[25] See the list made by the Free Software Foundation at `http://www.gnu.org/licenses/license-list.html` and the one made by the Open Source Initiative at `http://opensource.org/licenses`.

publish open-source documentation. These are a few of them:

The GNU Free Documentation License: a copyleft license specially indicated to documentation of software released under the GNU GPL License.

The FreeBSD Documentation License: a permissive non-copyleft license specially created for BSD-style software documentation.

Creative Commons Licenses: a set of licenses created in principle for artistic works (documentation included). Depending on the particular restrictions applied to each license (Attribution, Non-Commercial or Share-Alike) it may be considered as copyleft (Share-Alike) or non-copyleft (any other combination of the three restrictions).

Note that open-source software licenses such as the GNU GPL can be used for documentation as well. Linguistic data may be released in many cases either under software licenses or under documentation licenses.

2.5 Collaborative management websites for open-source software developers

An important step to help collaborative development of open-source software projects is to make their source code available on the net. Many projects are hosted on personal websites but there are other hosting alternatives presenting many advantages such as control of versions, increase of visibility, or developers management. These websites allow for centralised collaboration and distribution, and are known as *open-source development websites*. Two of the most commonly used are:

SourceForge.net is the world's largest open-source software development website, hosting more than $100\,000$ projects and over $1\,000\,000$ registered users with a centralised resource for managing projects, issues, communications, and code.[26]

Savannah is a central point for development, distribution and maintenance of open-source software that runs on free operating systems. It hosts more than $2\,500$ projects and has over $45\,000$ registered users. It includes issue tracking, project member management by roles and individual account maintenance.[27]

3 Apertium

In this section we briefly describe Apertium (Armentano-Oller et al. 2006; Corbí-Bellot et al. 2005), an open-source shallow-transfer MT engine, initially intended for related-language pairs

[26] `http://sourceforge.net`

[27] There is a close relationship between Savannah and the Free Software Foundation. Indeed, there are two different branches for GNU projects at `http://savannah.gnu.org` and non-GNU projects hosted at `http://savannah.nongnu.org`.

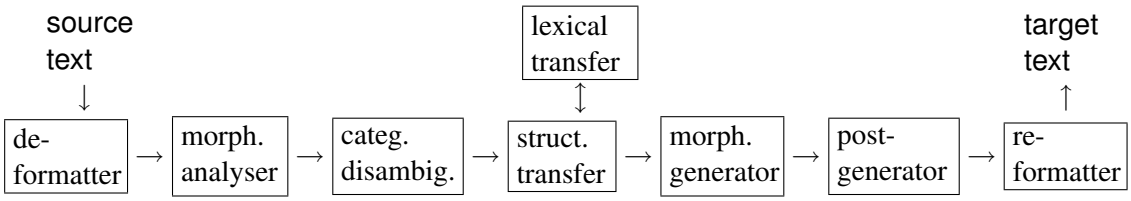

Figure 1: The eight open-source modules of the Apertium MT system (see section 3).

(such as Spanish–Catalan, Spanish–Portuguese, Czech–Slovak, Swedish–Danish, Kirwanda–Kiswahili, Bahasa Indonesia–Bahasa Melayu, etc.), but being currently extended to translate between not so related languages (such as Spanish–English); an early version of this extension is expected to be released by the end of 2006. Apertium's engine, linguistic data, and documentation can be found at the project's website at `http://apertium.sourceforge.net`.

3.1 Machine translation architecture

The open-source MT architecture Apertium is largely based upon that of systems already developed by the Transducens group at the Universitat d'Alacant, such as the Spanish–Catalan MT system interNOSTRUM[28] (Canals-Marote et al. 2001), and the Spanish–Portuguese translator Traductor Universia[29] (Garrido-Alenda et al. 2004). Both systems are not open-source; however, interNOSTRUM is publicly accessible through the net and used on a daily basis by thousands of users; Traductor Universia was also publicly accessible for some years until it was converted into a full commercial product.

The Apertium MT engine is a classical shallow-transfer or transformer system consisting of the following pipelined modules (see figure 1):

- A *de-formatter* which separates the text to be translated from the format information (RTF and HTML tags, white spaces, etc.). Format information is encapsulated so that the rest of the modules treat it as blanks between words.
- A *morphological analyser* which tokenizes the text in *surface forms* and delivers, for each surface form, one or more *lexical forms* consisting of *lemma*, *lexical category* and morphological inflection information.
- A *part-of-speech tagger* which chooses, using a first-order hidden Markov model (Cutting et al. 1992) (HMM), one of the lexical forms corresponding to an ambiguous surface form; this is the only statistical-centred module.
- A *lexical transfer* module which reads each source-language lexical form and delivers the corresponding target-language lexical form by looking it up in a bilingual dictionary.
- A *structural transfer* module (parallel to the lexical transfer) which uses a *finite-state chunker* to detect patterns of lexical forms which need to be processed for word reorderings, agreement, etc. and performs these operations.

[28] `http://www.internostrum.com`
[29] `http://traductor.universia.net`

- A *morphological generator* which delivers a target-language surface form for each target-language lexical form, by suitably inflecting it.
- A *post-generator* which performs orthographic operations such as contractions (e.g. Spanish *del*=*de*+*el*) and apostrophations (e.g. Catalan *l'institut*=*el*+*institut*).
- A *re-formatter* which restores the format information encapsulated by the de-formatter into the translated text.

The modules of the system communicate between them using text streams, which allows for easy diagnosis and independent testing. Furthermore, some modules can be used in isolation, independently from the rest of the MT engine, for other natural-language processing tasks. This extra application is also possible thanks to the full separation (or *decoupling*) of code and data.

3.2 Linguistic data

Apertium's linguistic data (which are fully decoupled from the translation engine) are coded using XML[30]-based formats; this allows for interoperability (that is, the possibility of using the XML data in a set of different scenarios) and for easy data transformation and maintenance. On the one hand, the success of the open-source MT engine heavily depends on the acceptance of these formats by other groups; this is indeed the mechanism by which *de facto* standards appear. Acceptance may be eased by the use of an interoperable XML-based format, and also by the availability of tools to manage linguistic data. But, on the other hand, acceptance of the formats also depends on the success of the translation engine itself. The XML formats for each type of linguistic data are defined through conveniently-designed XML document-type definitions (DTDs) which may be found inside the `apertium` package (see below).

Apertium contains *compilers* to convert the linguistic data into the corresponding efficient form used by each module. Two main compilers are used: one for the four lexical processing modules (morphological analyser, lexical transfer, morphological generator, and post-generator) and another one for the structural transfer. The first one uses a technique known as *finite-state letter transducers* (Garrido-Alenda et al. 2002) to efficiently code the lexical data; the last one uses *finite-state machines* to speed up pattern matching. The use of such efficient compiled data formats renders the engine able to translate tens of thousands of words per second in a current desktop computer.

3.3 Apertium-based working systems

Apertium's MT engine has been released in two open-source packages: `lttoolbox` (containing all the lexical processing modules and tools) and `apertium` itself (containing the rest of the engine); both are available under GNU GPL license. In addition to these programs, open-source data are already available for various language pairs:

[30] `http://www.w3.org/XML/`

- The Spanish–Catalan (packaged under the name `apertium-es-ca`) and Spanish–Galician (package `apertium-es-gl`) pairs developed under the Opentrad consortium and released under a Creative Commons (Attribution, Non-Commercial Share-Alike) license;
- The Spanish-Portuguese pair (package `apertium-es-pt`) developed at the Universitat d'Alacant and released under GNU GPL;
- Pilot data for Catalan–French (package `apertium-fr-ca`) and Catalan–Occitan/Aranese[31] (package `apertium-oc-ca`) released under GNU GPL;
- Pilot Catalan–English, expected to be released by the end of 2006.

Apertium gives a reasonably good translation quality between related languages (error rates around 5-10 per cent in general purpose translations). These results are obtained with the pilot open-source linguistic data already released (having around 10,000 lemmas and less than 80 shallow transfer rules) which might easily improve mainly through lexical contributions from the linguistic communities involved. The Apertium open-source engine itself is being actively developed and contributions to its design may enhance it to perform more advanced lexical and structural processing tasks.

For instance, Apertium's development team has been recently funded by the government of Catalonia so as to extend the translation engine to cope with non-related language pairs such as Catalan–English; the resulting extended system is expected to be released by the end of 2006.

All the available packages and documentation for Apertium are hosted at `http://www.sourceforge.net/projects/apertium`. Additional information may be found at `http://www.apertium.org`. Finally, web prototypes for MT systems for all the currently available pairs may be tested on plain texts, RTF and HTML at `http://xixona.dlsi.ua.es/prototype`.

4 Research and business opportunities

The Apertium open-source MT system, along with the already available documentation and tools for easing vocabulary management, offers good opportunities for researchers, professional translators and companies working on language engineering or linguistic services.

4.1 Opportunities for research

Individual users, researchers or linguistic communities (as well as private companies) can share their efforts by contributing to the development of Apertium technologies and data; anyone having the necessary computational and linguistic skills would be able to adapt or enhance Apertium to produce new or better MT systems, even for new language pairs.

[31] Aranese is a subdialect of Gascon, one of the main dialects of Occitan language. Occitan is reported to have about a million speakers, but it has almost no legal existence in France and Italy and a limited status of co-official language in a small valley of the Pyrenees in Catalonia, inside the territory of Spain, called Val d'Aran. Aranese is the variety spoken by about 6,000 people in this valley where it is, however, official.

Open-source development of software and data ensures open and free access to the created resources. And these resources need not to be built from scratch anymore, since the point of departure of new projects can be based on where previous works left.

Since Apertium code and data can be reused, researchers and developers can focus on improving them. Our experience shows that reusing data from existing language pairs speeds up the development of data for new ones; for example, the rules for gender and number agreement are basically the same for all the related language pairs already released by our team. Furthermore, monolingual dictionaries can be used in more than one language pair with little modifications.

Open-source software guarantees the reproducibility of research experiments. When reporting evaluation results of a MT system on a new technique, if the experiments conducted cannot be reproduced, authors are forcing readers to trust them. Open-source software and data make possible to reproduce experiments easily, to perform new ones, and to compare among them. *Reproducibility* is very important in science: it is necessary in order to check or verify experimental results constituting one of the basis of scientific progress: if a particular experiment cannot be replicated, it is not considered to provide *scientific evidence*.

Open-source software encourages collaborative interaction between research groups working on the same area. Governments and universities should support and fund this kind of shared work which is of special benefit to research groups with low resources. The open-source perspective stimulates and facilitates the presence of research groups in local, national and European programmes for technological development and research.

Open-source may also facilitate the transference of new advances from the research community to companies interested in their application. This role of open-source products should be taken into account by public organisations promoting co-operation between researchers and companies in their socioeconomic environment.

In order to illustrate all these opportunities, we now overview some running research and development projects derived from the open-source Apertium project:

- Enlargement of Freeling[32] dictionaries used to perform dependence analysis of Spanish and Catalan. Apertium's XML-based format of dictionaries and some additional tools facilitate format conversion to Freeling dictionaries (conducted by the TALP Research Center at Universitat Politècnica de Catalunya).

- Improvement of the previously released Spanish–Galician MT system (see section 3.3). Dictionaries coverage, part-of-speech tagger module and transfer rules between both languages are being studied in order to improve translation quality (conducted by Seminario de Lingüística Informática at Universidade de Vigo).

- Enhancement of the translation engine to perform more complex structural transformations and to deal with the translation of polysemous words (conducted by Transducens research group at Universitat d'Alacant in collaboration with other research groups, such as Institut Universitari de Lingüística Aplicada at Universitat Pompeu Fabra).

[32] `http://www.lsi.upc.es/~nlp/freeling`

- Addition of the new language pairs Romanian–Spanish and Swedish–Danish (conducted by Transducens research group and Departament de Filologia Inglesa at Universitat d'Alacant).
- Use of information from the target language to train source language part-of-speech taggers to be used within the Apertium MT engine in an unsupervised way (Sánchez-Martínez et al. 2006) (conducted by Transducens research group at Universitat d'Alacant).
- Automatic inference of transfer rules to translate from Portuguese to Spanish and English (de Medeiros Caseli and Nunes 2006), and automatic extraction of vocabulary entries from parallel texts for the Apertium dictionaries (conducted by Núcleo Interinstitucional de Lingüística Computacional at Universidade de São Paulo).

4.2 Opportunities for business

Open-source software also brings new business models to private companies. Taking the Apertium MT system as an example, companies can offer a wide variety of services around it, such as follows:

- installing and supporting translation servers;
- maintaining, adapting and extending linguistic data;
- building data for new language pairs;
- integrating MT systems in multilingual documentation management systems;
- developing new tools for Apertium, etc.

Furthermore, companies and individual translators can adapt linguistic data to restricted language domains or to dialectical varieties in order to ease post-edition or better suit their client needs when offering translation services.

An illustrative example of companies benefiting from some of the previous profitable market segments is the case of the three companies participating in the Apertium project as well as the case of a new company named Prompsit Language Engineering,[33] created to exploit the challenges derived from the existence of Apertium. So far Apertium has been integrated in the edition of two of the most known Galician newspapers[34] originally published only in Spanish; now, the online version of these newspapers has a daily bilingual edition. Also, some institutions and public entities are juggling with the possibility of installing Apertium as their MT platform to offer online translation services. Banks which are present in different heterogeneous linguistic areas have also shown their interest in integrating Apertium-based MT systems in their documentation management systems.

However, it may be argued that all these possible markets would be even more productive if the systems were closed-source. Furthermore, it is not easy to change clients' mind to

[33] Visit `http://www.prompsit.com` to know about Prompsit's business model.

[34] *La Voz de Galicia*, `http://www.lavozdegalicia.es`, and *El Correo Gallego*, `http://www.elcorreogallego.es`.

encourage them to use open-source software: *open* and *free* terms are usually perceived as untrustworthy. However, there is a strong reason why clients would prefer open-source to closed-source software: clients who choose open-source software do not see companies distributing them as providers to whom they have a technological dependency, but as technological partners, since clients may feel free to contract services around the open-source system with any other company offering them; therefore, technological dependence, a typical feature associated with closed-source products, is strongly diminished.

Even more interesting for institutions, public entities and large companies is the social action they can contribute to by making open modifications, improving data or adding new functionalities to the open-source software specially developed for them. This gives them a very positive image before their clients and users; they are not only offering a better service, but also benefiting the whole community.

It is also very difficult to convince tech companies to make their software open-source. The point is the change of the business model from a product-selling centred model to a service-offering one. Innovative services around a good open-source software are the main competitive advantages of this business model. Besides that, contributions to the open project coming from elsewhere are also contributions that companies can benefit from in order to offer better products services. This non-controllable aspect of the development makes heavy demands on those companies offering services based on open-source software but, despite this effort, in the current world it is crucial for tech companies to remain constantly updated.

Open-source software pose business challenges for those researchers working on new methods and techniques. Indeed, the number of technological-based *spin-offs* (here, companies created by researches as a result of a particular research activity) has increased in the last few years. These companies have not only a product and a catalogue of related services to offer, but also the *know-how* developed during the research work of their members, and, being half-way, they can offer the best services in collaboration with universities and companies. This is the case of Prompsit Language Engineering,[35] a spin-off startup company created by researchers from the Transducens group at the Universitat d'Alacant, that offers services around the open-source Apertium MT engine.

5 Discussion

With the recent trends in open-source software development, new challenges raise for both research institutions and companies. Open-source practices have recently reached the MT arena, therefore introducing new perspectives on MT system development. A new business model, which focuses on the services around translation engines and linguistic data more than on the programs and data themselves, is possible. In this paper we have presented Apertium, a full open-source MT system with a lot of potentials and introduced the main aspects around the new business model inspired by the Apertium system.

[35] `http://www.prompsit.com`

Acknowledgements

Work partially funded by Spanish Government through grants FIT-340101-2004-3, FIT-340001-2005-2 and TIC2003-08681-C02-01. Felipe Sánchez-Martínez is supported by the Spanish Government and the European Social Fund through research grant BES-2004-4711. The extension of Apertium is being funded by the Generalitat de Catalunya.

References

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. A. (2006). Open-source Portuguese-Spanish machine translation. In *Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006*, volume 3960 of *Lecture Notes in Computer Science*, pages 50–59. Springer-Verlag.

Behlendorf, B. (1999). *Open Sources: Voices from the Open Source Revolution*, chapter Open Source as a Business Strategy. O'reilly. Edited by Chris DiBona, Sam Ockman, Mark Stone.

Canals-Marote, R., Esteve-Guillen, A., Garrido-Alenda, A., Guardiola-Savall, M., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Perez-Antón, P. M., and Forcada, M. L. (2001). The Spanish-Catalan machine translation system interNOSTRUM. In *Proceedings of MT Summit VIII: Machine Translation in the Information Age*, pages 73–76. Santiago de Compostela, Spain, July.

Corbí-Bellot, A. M., Forcada, M. L., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor, A., and Sarasola, K. (2005). An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In *Proceedings of the 10th European Associtation for Machine Translation Conference*, pages 79–86, Budapest, Hungary.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing. Association for Computational Linguistics. Proceedings of the Conference.*, pages 133–140, Trento, Italy.

de Medeiros Caseli, H. and Nunes, M. G. V. (2006). Automatic transfer rule induction from parallel corpora. In *Proceedings of III Workshop on MSc dissertation and PhD thesis in Artificial Inteligence (WTDIA'2006), SBIA'2006*. (to be published).

Forcada, M. L. (2006). Open-source machine translation: an opportunity for minor languages. In *Proceedings of Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*.

Garrido-Alenda, A., Forcada, M. L., and Carrasco, R. C. (2002). Incremental construction and maintenance of morphological analysers based on augmented letter transducers. In *Proceedings of TMI 2002 (Theoretical and Methodological Issues in Machine Translation)*, pages 53–62.

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

Sánchez-Martínez, F., Pérez-Ortiz, J. A., and Forcada, M. L. (2006). Speeding up target-language driven part-of-speech tagger training for machine translation. In *Advances in Artificial Intelligence, Proceedings of the 5th MICAI*, volume 4293 of *Lecture Notes in Computer Science*, pages 844–854. Springer-Verlag. Accepted.