Machine Translation for Catalan↔Spanish: The real case for productive MT

Juan Alberto Alonso

Comprendium España S.L.
Plató 6, 1r, 1a, 08021 – Barcelona – Spain
juan.alonso@comprendium.es

It may come as a surprise to many people to know that Spain in general and Catalonia in particular are probably the places in the world where machine translation systems are most extensively used for productive applications. The peculiar position of Catalan and Spanish in Catalonia, both being official and therefore mandatory for Public Administration publications and websites, the fact that Catalan is used in everyday life, business and media, and the close linguistic relationship between both languages enabling an excellent translation quality in the Catalan↔Spanish language pair in our MT system, has made it possible for a number of Public Administration organisms and other private companies to use it in a productive way for massive translation. This paper presents the reality of MT for Catalan↔Spanish, together with two practical cases where our MT system is currently being used within a productive environment.

1. Introduction

Catalan is co-official with Spanish within the autonomous regions of Catalonia, Valencia and the Balearic Islands. This means that, quite often, all kind of written texts – Public Administration documents and websites, newspapers and magazines, books, etc. – are published in both languages. Moreover, there is a real need for translation of documentation, user-guides, brochures, etc. both by non-Catalan companies that establish themselves in Catalonia, or by Catalan companies that want to expand their business activities outside the Catalan-speaking territory. This is the perfect environment for the productive use of machine translation systems, especially when the close linguistic relationship enables a very good translation quality, at least in our MT system.

2. Some Facts about the Catalan Language

Catalan is a Romance language, linguistically closely related to other neighbor languages such as Occitan, Spanish and French. Catalan is widely spoken in the autonomous regions of Catalonia, Valencia (where the language is called "Valencià") and Balearic Islands, in Spain, where it is co-official with Spanish. It is also spoken in the department of Pyrénées Orientales, in France, in the small country of Andorra, where it is the only official language, and in the city of Alghero, on the island of Sardinia. All in all, some 7 million speakers speak it.

Catalan is widely present today in the public media, especially in Catalonia: radio and TV stations, daily and weekly press, books of all kinds, cinema, theater and Internet. It is also widely used in public announcements, official documents and websites, business, banks, schools and universities – where teaching usually takes place in Catalan – and, in general, in all kinds of everyday-life activities.

The current linguistic situation is thus one of bilingualism: virtually all Catalan native speakers can also speak and understand Spanish. The opposite (Spanish native speakers living in Catalonia who are fluent in Catalan) is also often the case, but not always.

3. The Uses of MT Systems

One of the factors that determines the appropriateness of a machine translation system is the specific use that it is given. Basically, there are

two kind of uses that can be given to MT systems:

- "Informative use": The MT system is used to extract (at least some) information out of a text written in a language that the user does not know. The aim is thus to use it to obtain information that otherwise would not be available to the user. A typical example of this case is the use of MT systems to translate web pages in foreign languages. In this case, the quality is not a key factor, but the number of language pairs available and the translation speed are.
- "Productive use": The MT system is used to reduce the time and effort needed to produce professional translations (i.e. meant to be published). The aim is thus to use it as an efficient tool within a productive chain. In this case, translation quality is of course the key factor, together with translation speed and customization of the system to the specific needs of the user.

4. MT for Catalan↔Spanish

4.1. The Outstanding Quality of Catalan↔Spanish MT

Catalan↔Spanish is, by far, the language pair that yields a better translation quality out of the 23 translation directions currently offered by Comprendium. According to Translation Quality Evaluations recently made in our company, it yields 93% of good and understandable sentences[1], compared to a 84% for Spanish→French or to a 67% for English→Spanish. This excellent translation quality is largely due, but not only, to the close linguistic relationship between both languages. Figure 1 illustrates these results.

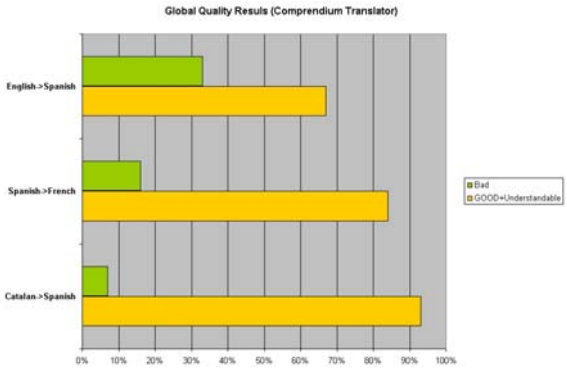

Figure 1

4.2. Typical Cases of Catalan↔Spanish MT Usage

Given the facts presented so far, we find ourselves with almost the perfect environment for MT productive use:

- There is a constant need for translation between Catalan and Spanish of big amounts of documents and websites.
- There is an MT system for this language pair that yields a degree of translation quality good enough to be used efficiently to produce professional translations, and that can actually save time and money in doing it.

Our Catalan↔Spanish system is currently being used in the following environments[2]:

- Newspapers: there is a need to publish daily editions in Spanish and Catalan for the same newspaper, but most of them are reluctant to do so because of the costs involved. However, among others, "Diari Segre", a newspaper from the Catalan town of Lleida, decided to do so and use our system as a means to reduce costs and save time. This case will be presented with more detail below in this paper
- Public Administration organisms and institutions . Some examples are
  - Generalitat de Catalunya: the autonomous Government of Catalunya. Our Catalan→Spanish system has been used there for years, for internal use within different departments.

[1] Out of the 7% of "bad" sentences, at least half of them could be considered as "understandable". In fact, curiously enough, when carrying out a Translation Quality Evaluation between closely related languages, the quality criteria tend to be stricter.

[2] The following is by no means a comprehensive list of our references, but a list with some example cases.

- Consell Consultiu de la Generalitat (Generalitat's Consulting Council)
- Government of Andorra.
- Consell de Mallorca (Mallorca Council)
- Town halls and other regional institutions: Diputació de Lleida and the town halls of Palma de Mallorca and L'Hospitalet del Llobregat. Below in this paper we will give more details about the installation of L'Hospitalet del Llobregat.

- Universities and Research Institutions: our Catalan→Spanish system has been installed in several universities around Catalonia, including, for example, the UOC (www.uoc.es), a e-University.
- Internet Portals: Vilaweb (www.vilaweb.com), a news portal in Catalan, offers translation through our system of their web contents from Catalan into Spanish and English for their registered users.

Most of these use the system to translate big amounts of documentation from Catalan into Spanish, or the other way round. However, some of them make a more sophisticated use of it, where they plug our system into their document management chain or their content management environment. Next, we are going to examine in more detail two of the cases listed above.

5. Two Real Cases

5.1. L'Hospitalet Town Hall

L'Hospitalet de Llobregat is one of the several "satellite" towns around Barcelona. It has some 300,000 inhabitants.

Due to the important economical and urban transformations that have taken place within the last years, and to the arrival of new citizens from other countries, the Town hall of L'Hospitalet has decided to publish the information in their Website (www.l-h.es) in several languages, apart from Catalan.

This is why they started a multilingual content management project. The aim of this project is to offer the website information in three languages – Catalan, Spanish and English – by developing a tool that, out of a pre-defined content model, extracts the relevant information – originally stored in Catalan – and, through the Comprendium Translator system, translates it into Spanish and English. Different working groups will then revise and validate the translation before posting it on the website.

5.2. "Diari Segre" Newspaper

"Diari Segre" (www.diarisegre.com/) is a daily newspaper published in the Catalan town of Lleida. Every day, both a Spanish and a Catalan edition are published. Since historically the newspaper was first published in Spanish, the Catalan edition is partially made out of the Spanish edition with the aid of the Comprendium Machine Translation system for Spanish→Catalan.

The general flow is illustrated in Figure 2 below.

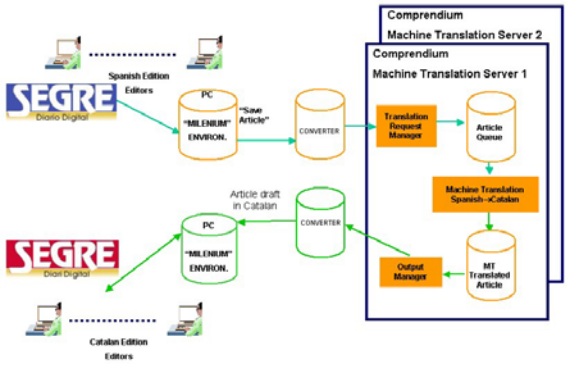

Figure 2

The newspaper is currently edited on the "Protec Milenium" edition environment. There, the Spanish Edition editors prepare the Spanish Edition articles. After saving the final version of every Spanish article, it is automatically fed to a converter specially designed to keep the original article layout and format and extract the actual article text. This text is then sent to the Machine Translation Environment. There are two Machine Translation Servers, one of them acting as a fallback in case of failure of the primary one.

The Machine Translation Environment interface with the Milenium Editor is an input/output queue. This queue is implemented in form of Windows file system folders, where the input folders (inbox) are being constantly scanned for incoming documents (text of articles to be translated). As soon as a new input text appears in the inbox, it is taken out from there and sent to the Machine Translation system queue. Texts in

this queue are then actually processed by the Spanish→Catalan Machine Translation Engine, and the corresponding translations are put into the output folders (outbox).

Another process is continuously scanning the outbox for the presence of translated texts, and as soon as they appear there, they are taken out from the outbox, they are processed by the converter, where the original layout and format is restored, and they are stored in the Catalan folders of the Milenium Editor. The corresponding Catalan Edition editors get a warning about a new Catalan article draft being available and finally they can correct it and store it in the final Catalan Edition.

It has to be stressed that all this process occurs in *real-time*, and that both newspaper editions, the Spanish one and the Catalan one, should be ready *on time* and at the *same time*. This can only be achieved if the Machine Translation system that is used is both fast and yields a very good translation quality, so that correction is kept to a minimum.

Some real data about the translationn process carried out at „Diari Segre" are:

- Around 32,000 words/day translated (which means some 12,000,000 words per year).
- MT System translation speed (Spanish→Catalan): around 1,500,000 words/hour
- Average MT result post-edition speed (Spanish→Catalan): 3,000 words/hour
- Average manual Spanish→Catalan translation speed: 500 words/hour

6. Conclusions

The availability of the fast, high-quality Machine Translation system between Spanish and Catalan from Comprendium has made it possible to use it in real-life applications as an efficient productive tool that saves time, effort and money. This, together with the real need for fast and massive translations between Spanish and Catalan, has promoted the use and expertise of machine translation for productive use in Catalonia during the past few years, with the current existence of many institutions and private companies that use it profitably on a daily basis.

7. Acknowledgements

- Ignasi Navarro (INCYTA Multilanguage S.L., official re-sellers of Comprendium products in Spain)
- Max Soria (Ajuntament de L'Hospitalet de Llobregat)
- Dori Llena and José Antonio Negredo (Diari Segre)