Computing and Informatics, Vol. 33, 2014, 907–920

A LARGE SPANISH-CATALAN PARALLEL CORPUS RELEASE FOR MACHINE TRANSLATION

Marta R. Costa-Jussà, José A. R. Fonollosa, José B. Mariño

TALP Research Center
Universitat Politècnica de Catalunya
e-mail: {marta.ruiz, jose.fonollosa, jose.marino}@upc.edu

Marc Poch

Institut Universitari de Lingüística Aplicada (IULA)
Universitat Pompeu Fabra
e-mail: marc.pochriera@upf.edu

Mireia Farrús

N-RAS Research Center
Universitat Pompeu Fabra
e-mail: mireia.farrus@upf.edu

Abstract. We present a large Spanish-Catalan parallel corpus extracted from ten years of the paper edition of a bilingual Catalan newspaper. The produced corpus of 7.5 M parallel sentences (around 180 M words per language) is useful for many natural language applications. We report excellent results when building a statistical machine translation system trained on this parallel corpus. The Spanish-Catalan corpus is partially available via ELDA (Evaluations and Language Resources Distribution Agency) in catalog number ELRA-W0053.

Keywords: Catalan-Spanish parallel corpus, machine translation

1 INTRODUCTION

The availability of large parallel corpora is usually the bottleneck in the development of statistical methods for multilingual natural language processing, especially for minority languages. One of the most important and direct applications of parallel corpora is the training of corpus-based machine translation systems such as statistical (SMT) or example-based (EBMT) machine translation [1, 2].

The main objective of the described work is to produce a new Spanish-Catalan parallel corpus as a basis for the development of a high quality SMT system for this pair of languages. However, the corpus will be also of great interest for other bilingual or monolingual language processing applications in Spanish and Catalan.

SMT has received increased interest from the research community since the publication of its theoretical foundations in 1993 by a group of IBM researchers [3]. In parallel, machine translation in general has also garnered interest because of the ubiquity of millions of multilingual web pages throughout the Internet. Moreover, in bilingual communities, such as Catalonia, the need for good translation engines is pervasive for institutions, companies and individuals seeking to generate and disseminate documentation in at least the two languages spoken in that community (in this case, Catalan and Spanish).

The newspaper *El Periódico de Catalunya*, one of the Catalan newspapers with the highest readership figures, is a good example of an institution that would benefit from machine translation. Every day, *El Periódico de Catalunya* has to generate a bilingual paper edition in a few hours, apart from keeping an updated version of the bilingual web edition. Fortunately, companies and institutions with this high level of translation need are usually good providers of parallel texts as well. However, the format of the original bilingual text is not always simple to process automatically.

The parallel corpus described in this article has been extracted from ten years of the paper edition of *El Periódico de Catalunya* (1997–2007) in PDF format. First, a significant part of our effort to transform the original files into a useful sentence-aligned parallel corpus was devoted to text extraction and content filtering. Second, we performed the tokenization and normalization of the bilingual corpus. Finally, we addressed the sentence alignment problem. Notice that the resulting text cannot be used for textual investigations; it is focused on obtaining a collection of individual sentences. That is, sentences are scrambled and thus any information in a unit larger than a sentence is not possible to obtain. Each intermediate step was carefully evaluated manually by a bilingual native speaker in Spanish/Catalan. Every 10 thousand sentences, the native speaker evaluated three parallel sentences. The parallel Spanish-Catalan corpus is already partially available via ELDA (Evaluations and Language Resources Distribution Agency) in catalog number ELRA-W0053.

The paper is organized as follows. Section 2 describes the original database, including its basic statistics. Section 3 reports some details of the database monolingual preparation. Section 4 describes the sentence alignment process, and Section 5 concentrates on the detection of poor alignment based on statistical confidence measures. New statistical sentence-filtering approaches are proposed and compared in terms of the quality of the obtained translations. In order to present these translation results, Section 6 shows the results of a statistical machine translation system trained with the presented parallel corpus. Finally, Section 7 presents the main conclusions of the paper.

2 DATABASE DESCRIPTION

El Periódico de Catalunya[1] is a Catalan newspaper that offers everyday news about politics, economics, culture, health and sports both in Catalan and Spanish languages. The bilingual edition of the newspaper started in October 1997. The news pieces are originally written in Spanish and then translated into Catalan using an internal machine translation (MT) system and three human post-editors, meaning each text is checked three times. Given that the newspaper readers are quite demanding about the quality of the written Catalan, *El Periódico de Catalunya* has to offer a very good translated text in Catalan to guarantee their competitiveness.

The company provided us with a large database, including all the Catalan and Spanish editions from October 1997 to February 2008 in PDF format. The database is organized in different levels of folders and subfolders. The first level splits the database in two main parts: Catalan and Spanish, which contain a different folder for each year. Every *year folder* is then divided into folders that correspond to different months. *Month folders* are divided into *day folders*, and a PDF file can be found inside each *day folder* for each newspaper page from that day edition. Every newspaper edition has approximately 90 pages per day, that is, 90 PDF files. In total, the original database has around 12 800 folders and approximately 850 000 files, and 163 GB of hard disk are needed to store the whole PDF database.

3 DATABASE PREPARATION

In general, each page of the newspaper editions is contained in a different PDF file in the original database. In order to obtain the necessary information from each page, the first step was to convert the PDF file into a corresponding text file. The program *pdftotxt* was used to accomplish this process. This program is an open source command-line utility to extract text data from PDF-encapsulated files, which is freely available and included by default with many Linux distributions, as well as being available on Windows. Such text extraction is complicated because PDF files are internally built on page drawing primitives, meaning the boundaries between words and paragraphs often must be inferred based on their position on the page. *pdftotxt* can operate in three different modes to recover the text: the layout mode, the normal mode and the raw mode. After a comparison of the three modes, the raw mode was selected as the best option to recover the complete sentences.

Once the data were converted into text format, sentences were still divided into different lines like in the newspaper columns. We performed a sentence reconstruc-

[1] `http://www.elperiodico.cat` and `http://www.elperiodico.es`

tion (i.e. put one whole sentence in every line) in order to join all sentence pieces in one line. Due to the complex column layout of the paper edition, the recovery of the sentences from each PDF file presents several problems and may lead to incomplete words and sentences. Therefore, after the extraction, a filtering process was followed in order to detect incorrect sentences. Moreover, additional content-based filtering procedures were included to avoid duplicate newspapers sections and other parts that were considered of no interest for the final parallel corpus.

The newspaper has many pages with advertising, television time tables or number games – such as sudoku – that must be filtered. After some research, two filters were developed. The first was used to identify undesirable content by analyzing the first lines of every page. This filter can be used to erase many useless pages, but it cannot remove all of them. The second filter reads the whole page searching for key lines used to identify undesirable content.

In order to reduce sparseness, we performed the usual tokenization and normalization. Given a text, tokenization chops it into pieces, called tokens. Thus, the task is language-dependent, and we used two different tokenizations for Spanish and Catalan. Next, some of the most relevant cases are described, and Table 1 shows an example of each one.

1. The apostrophe ('), which refers to the Catalan article in front of vowels, must be tokenised as a different word as *avi* (grandfather). Since the apostrophised article does not exist in Spanish, this rule does not apply.
2. Dashes and plus symbols are used to join initials that form a whole word, such as $R + D$ = *Recerca i Desenvolupament* (Research and Development), or the equivalent form in Spanish. In both cases (Catalan and Spanish), the three characters are not tokenised, since they are part of the same word.
3. The geminated l ($l \cdot l$) is a Catalan form that refers to a long-pronounced l, such as in *col·legi* (school, college), *il·lusió* (illusion), etc. As such, it must not be tokenised. Since the geminated l does not exist in Spanish, the rule is not applicable.
4. Even dots need to be tokenized in different ways depending on the situation, such as word abbreviations like name initials, which are different from those appearing at the end of the sentences. Unlike these ones, dots after initials are not treated as different tokens, since they are part of the initialized name. This rule applies to both Catalan and Spanish languages; however, since word abbreviations are different for Catalan and Spanish, two different lists were used by the tokenizer to avoid unnecessary tokenizations.

The aim of the normalization process is to group characters into the same token that are the same but written differently, e.g., using different symbols for an apostrophe. At this point, a deep analysis of the sentences is needed to find the remaining mistakes. Some words, for instance, are not separated by spaces, and some kinds of punctuation marks should be avoided. Table 2 shows the problems (input) found and their respective solutions and results. Note that the character average per word

Character	**Symbol**	**Catalan**		**Spanish**	
		before	**after**	**before**	**after**
apostrophe	’	l’avi	l’ avi	N/A	
dash and plus	− +	$R + D$	$R + D$	$I + D$	$I + D$
geminated l	$l \cdot l$	$l \cdot l$	$l \cdot l$	N/A	
initials’ dot	.	A. Pous	A. Pous	A. Pous	A. Pous

Table 1. Each row shows an example of situations with punctuation marks before and after preprocessing. N/A stands for Not Applicable

in Spanish is around 5.5. The longest word in Spanish has 27 characters, but there are very few used words above 20 characters. As an additional post-process, repetitive sentences that appear every day were detected and removed from the monthly files.

After this step, the quality of the process was verified. A native speaker checked three of every 10 000 sentences in order to verify that the output was correct. So, 0.03 % of the corpus (about 2 250 sentences) was checked manually with a successful result: just few and not significant errors were found after the normalization process.

Input: *Tras la revolución industrial, tenemos frente a nosotros la cuarta revolución: la revolución ambiental»*
Solution: Eliminate all punctuation marks at the beginning of the sentence, except for "¿" in Spanish. Erase some punctuation marks at the end.
Result: *Tras la revolución industrial, tenemos frente a nosotros la cuarta revolución: la revolución ambiental*

Input: *«some text A»* or *”some text B”* or *‘some text C’*
Solution: Have a unique character for quotes.
Result: *”some text A”* or *”some text B”* or *”some text C”*

Input: *l‘arbre* or *l’arbre*
Solution: Avoid using different symbols for the same meaning.
Result: always *l’arbre*

Input: *BruceWillistratadeevitarunataqueterrorista informático a Estados Unidos.*
Solution: Avoid joined words by erasing sentences with words longer than 20 characters
Result: (blank)

Input: *P r i m e r o f u e l a ofensiva contra Catalunya a raíz del Estatut.*
Solution: Avoid words consisting of characters separated by spaces (due to the PDF to text conversion) by deleting the whole sentence.
Result: (blank)

Table 2. Examples of normalization problems and the corresponding solutions and results

Finally, Figure 1 represents the block diagram corresponding to the database preparation, followed and briefly described in this section.

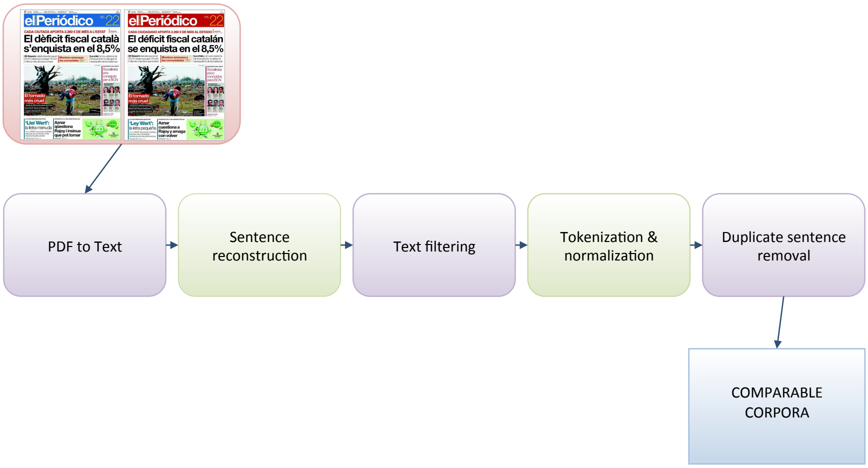
elPeriódico
El dèficit fiscal català s'enquista en el 8,5%
elPeriódico
El déficit fiscal catalán se enquista en el 8,5%
PDF to Text
Sentence reconstruction
Text filtering
Tokenization & normalization
Duplicate sentence removal
COMPARABLE CORPORA

Figure 1. Block diagram of the processes performed in the database preparation

4 SENTENCE ALIGNMENT

Up to this point, each language of the bilingual corpus is processed independently to generate a pair of comparable files. In order to be able to use this corpus as training material for corpus-based MT systems, it is now necessary to align every line in the Spanish file with its corresponding translation in the Catalan file. For this purpose, we selected the program Bilingual Sentence Aligner described in [4]. The program searches paired translations and discards sentences without translation.

Defining the distance of alignment as $D = |N - M|$, where N is the line of a sentence in the Catalan file an M is the line of its translation in the Spanish file, it can be said that the aim of Bilingual Sentence Aligner is to reduce D to zero. The computational cost of the programmed alignment algorithm is very sensitive to the initial distance of alignment. If the input files have many lines and/or the initial distance of alignment is high, the program may require excessive resources (RAM memory) to be executed. For the sentence alignment experiments, a machine with 32 GB of RAM was used, and the distance of alignment was kept under control because of the data filtering previously described in this document. Moreover, the selected one-month files seemed to be the maximum practical size for each independent sentence-alignment run.

The data alignment took approximately 40 hours for each year. An additional loss of 10 % of the lines occurred due to the fact that the program discarded sentences without a corresponding translation or with a high distance of alignment. For each year of data, we finally obtained approximately 15 million words for each language, which means a total of 150 million words. The main statistics (sentences and running words) are shown in Table 3.

	Spanish	Catalan
Sentences	7.5 M	
Words	180.2 M	179.5 M

Table 3. Parallel corpus statistics

Again, the quality of the process was verified after the alignment step. As in the previous step, native speaker checked three of every 10 000 sentences in order to verify that the output was correct. So, 0.03 % of the corpus was checked manually, also with a successful result.

5 STATISTICAL BILINGUAL SENTENCE FILTERING

The parallel corpus built up to this point will be used to train a statistical machine translation (SMT) system. Therefore, we are interested in eliminating pairs of sentences that statistically have a low probability of having a translation correspondence. These sentences usually do not contribute to improvements in the translation quality, since they produce long translation units that are unlikely to reappear in a translation or because they produce translation units of low quality. Therefore, by detecting and discarding these sentences, the translation would be improved:

1. in efficiency, because the training corpus is reduced and the quantity of translation units is smaller; and
2. in quality, since the translation units, which are directly extracted from the bilingual corpus, would be more accurate.

In this section two bilingual sentence filtering approaches based on different criteria are presented. These methodologies have been shown to improve the quality of the parallel corpus at the level of sentence in our previous works [5]. In that work, we filtered an English/Spanish text of 1.3 milion sentences, filtered 12 % of the corpus, and gained 1+ point BLEU (from 43.33 to 44.52).

5.1 Filtering Using the IBM1 Approach

Here, a method based on IBM Model 1 [3] is proposed. This simple SMT model relies on establishing a translation probability at the word level. Given a bilingual corpus aligned at the sentence level, where each sentence is supposed to be the translation of its counterpart, IBM Model 1 finds the probability of a word being translated by another one. Basically, it uses information about co-occurrence. Therefore, a word is the translation of another word with high probability only if they have often appeared in the same bilingual sentences and rarely in different bilingual sentences. This word translation dictionary provided by the IBM Model 1 will be used to define the translation probability of a bilingual sentence. Then, a probability threshold must just be chosen, under which the bilingual sentences will be eliminated and over which they will be accepted.

A translation sentence is defined as $t_1^I = t_1, t_2 \ldots t_I$, and a source sentence is defined as $s_1^J = s_1, s_2 \ldots s_J$, where I and J are the number of target and source words, respectively. The sentence translation probability assigned using the IBM Model 1 dictionary is usually computed as follows:

$$P_{ibm1}(t_1^I, s_1^J) = \log \frac{1}{(I+1)^J \prod_{j=1}^{J} \sum_{i=0}^{I} p(s_j^n | t_i^n)} \tag{1}$$

where t_j^n and s_i^n are the j^{th} and i^{th} words in the source and target sentences, I being the number of words in the source sentence and J the number of words in the target sentence. That way $p_{IBM_1(t_j^n|s_i^n)}$ are the probabilities of translation in the source-target direction $p(t_k/s_k)$ asigned by the IBM Model 1.

Notice that this translation probability is asymmetric since the conditional probabilities given by the IBM Model 1 depend on the translation direction. In order to obtain a symmetric distance, the translation probability in the inverse direction is also considered. This inverse probability is given by:

$$P_{ibm1}(s_1^J, t_1^I) = \log \frac{1}{(J+1)^I \prod_{i=1}^{I} \sum_{j=0}^{J} p(t_i^n | s_j^n)}. \tag{2}$$

Thus, using the lexical probabilities the obtained IBM1 Model will calculate the probability that two parallel sentences in bilingual corpus are translations between them.

5.2 Stopwords

When computing the translation probability of a pair of sentences, we can usually have a high contribution to the total probability from stopwords (i.e. words which do not have a meaning by themselves). Many bilingual pairs of stopwords are so frequent that the probability of co-occurrence is very high, although they are not a reliable indication of parallelism since they have no specific meaning by themselves.

To solve this problem, a list of stopwords was computed for both languages and excluded when computing IBM Model 1 equations. In addition to the probability given by the IBM Model 1, other methods of computing the translation probability of a pair of sentences, such as the PER probability, were also considered. However, as shown in [5], the best results were obtained when considering the IBM Model 1 probability. Therefore, our analysis will be restricted to the results given by this sentence probability.

5.3 Filtering Statistics Using the IBM1 Approach

The filtering statistics are reported in this section. Both left and right graphics included in Figure 2 show the number of sentences (vertical axis) and their cost (probability expressed in negative logarithm) of being parallel in the direct and inverse directions, respectively. The peak in both graphics is around a cost of 1.5 and it contains around 90 000 sentences in both cases. By an experimental decision, the 30 most probable words (stopwords) of each language were not considered when computing this cost. We sorted words by frequency and considered that the first 30 could be removed.

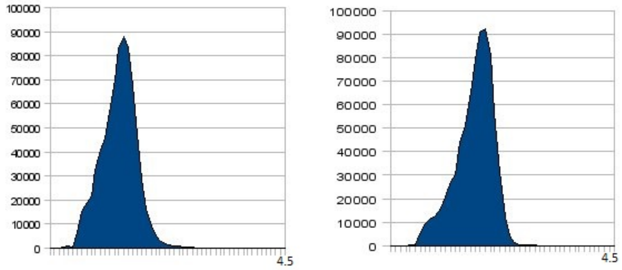

Figure 2. Cost distribution using IBM Model 1 direct (left) and inverse (right). The vertical axis shows the number of sentences and the horizontal axis shows the cost of being parallel

Given both tables, a threshold for each one is chosen and the results are combined with an AND or an OR operation. The former means that all sentences with a higher cost than the threshold (in Figure 2 left and right) will be discarded. The latter means that all sentences with a higher cost than the threshold (either in Figure 2 left or right) will be discarded. A comparison between these two logic operations is performed in the experiment section. The threshold is chosen to discard around 10 or 15 % of the corpus. Most of the discarded sentences were free-style translations that did not closely follow the form or organization of the original, very long sentences and/or sentences with many proper names. Some examples of discarded sentences can be seen in Table 4.

CA	*L'estrella va pagar 15 euros per un got de llimonada que els germans Carter, Chandler i Chase Fontaine, de 10, 6 i 5 anys, venien a 20 cèntims.*
ES	*El astro pagó con 15 euros un vaso de limonada que los hermanos Carter, Chandler y Chase Fontaine, de céntimos.*
CA	*Barcelona no podrà tenir l'organització dels Mundials d'atletisme de l'any 2013 i només organitzarà els Europeus del 2010.*
ES	*Barcelona no tendrá la organización de los Mundiales de atletismo del 2013 y se quedará solo con los Europeos del 2010.*
CA	*Sunyer: passeig de Sant Joan, 111 (Provença) 93 457 53 72.*
ES	*Laguna: Provença, 459 (Lepant) 93 455 12 07.*

Table 4. Some examples of discarded sentences

5.4 Filtering Non-Bilingual Pairs of Sentences

Non-bilingual pairs are other types of sentences that do not contribute to quality improvement of the translation system. In our corpus, we frequently observed situations where the aligned sentences were identical in both sides of the corpus (i.e. sentences with only proper names). It was found that these sentences represented a considerable increase in the size of the translation model without adding useful information. Therefore, we chose to discard these sentences in the translation table. In Table 5, some examples of discarded sentences are shown.

CA/ES	*Se ha reunido el comité de empresa/Y han decidido que se acabó la fiesta*
CA/ES	*Los catalanes llegan a esta decisiva cita con siete bajas de peso, un resultado a que se hace muchomás pesada teniendo en cuenta la situación desesperada del equipo.*
CA/ES	*Palmira Soler Bosch la seva família ho fa saber als seus amics i coneguts i els prega de voler-la tenir present en les seves oracions.*

Table 5. Some examples of non-bilingual discarded sentences

6 DEVELOPING AN MT SYSTEM

Machine translation systems can be built using the commonly known corpus-based approach; the knowledge is automatically extracted by analyzing translation examples from a parallel corpus (built by human experts). The advantage of this type of approach is that, once the required techniques have been developed for a given language pair, MT systems can be (in theory) developed very quickly for new language pairs using the provided training data. Recently, within the corpus-based approaches, statistical machine translation systems have been widely used. In statistical machine translation, parallel examples are used to train a statistical translation model. Thus, it relies on statistical parameters and a set of translation and language models, among other data-driven features. This approach worked initially on a word-by-word basis (thus it was classified as a direct method). Nevertheless, current systems attempt to introduce context, and the translation units are sequences of words.

The parallel corpus developed in this project is a huge database (compared to others, like those offered in International Evaluation Campaigns as WMT for building an SMT system).

Using the parallel corpus and the N-gram-based approach [1], we built an SMT system for the Catalan-Spanish languages, which is freely available at `http://www.n-ii.org`. As shown later on, the size of the training corpus and the similarity between the two languages allowed very satisfying results to be obtained, according to the standard automatic measures, BLEU [6].

In order to show this corpus utility in MT, experiments are performed on a test set. The Spanish source test corpus consists of 711 sentences from *El País* and

La Vanguardia. The Catalan source test corpus consists of 813 sentences from *L'Avui* and transcriptions from the TV program *Àgora.* For both tests, there are two references available that were computed by linguistic experts. Table 6 shows the statistics of number of sentences, words and vocabulary. See further experimentation with this corpus in [7, 8, 9].

	Spanish	Catalan
Sentences	711	813
Words	15 974	17 099
Vocabulary	5 702	5 540

Table 6. Spanish and Catalan test corpora statistics

Table 7 shows the automatic translation results compared to the Catalan government official translation system (Translendium, `http://traductor.gencat.cat/`), which is a state-of-the-art system. Only the corpus corresponding to the year 2007 was used. Although results can be improved adding more years, these extra results are not currently available.

All in all, the results in Table 7 show that the system trained with our corpus is significantly better than the Translendium system in terms of BLEU. Significance was measured using "pair bootstrap resampling" method [10]. By performing a manual comparison of both translations, the UPC system shows a more fluent output.

	BLEU-Es2CA	BLEU-Ca2Es
Translendium	85.97	87.81
UPC	**86.54**	**88.58**

Table 7. Translation results. Best results in bold

7 CONCLUSIONS

This paper has presented a large Spanish-Catalan parallel corpus. The resulting database is a large Catalan-Spanish parallel corpus (of 7.5 M parallel sentences and 180 M words per language) of great interest for many natural language applications, including machine translation. The corpus has been successfully applied to train an SMT system which is capable to outperform the official MT system of the Catalan Government. The parallel Spanish-Catalan corpus is now partially available via ELDA in catalog number ELRA-W0053.

Acknowledgements

This work has been partially funded by the Seventh Framework Program of the European Commission through the International Outgoing Fellowship Marie Curie Action

(IMTraP-2011-29951). The authors also want to thank the Universitat Politècnica de Catalunya for its support and permission to publish this research.

REFERENCES

[1] Mariño, J. B.—Banchs, R. E.—Crego, J. M.—de Gispert, A.—Lambert, P.—Fonollosa, J. A. R.—Costa-Jussà, M. R.: *N*-Gram-Based Machine Translation. Computational Linguistics, 2006, pp. 527–549.

[2] Nagao, M.: A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle. Artificial and Human Intelligence, Elsevier Publishers 1984.

[3] Brown, P. F.—Della Pietra, S. A.—Della Pietra, V. J.—Mercer, R. L.: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 1993, pp. 263–311.

[4] Moore, R.: Fast and Accurate Sentence Alignment of Bilingual Corpora. Machine Translation: From Research to Real Users. LNCS 2002, Vol. 2499, pp. 135–144.

[5] Montolar, E.: Técnicas Para el Filtrado y la Categorización de un Corpus Bilingüe en la Traducción Automática Estadística. Final Degree Project. Universitat Politècnica de Catalunya, Barcelona 2008.

[6] Papineni, K.—Roukos, S.—Ward, T.—Zhu, W. J.: BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the ACL, Philadelphia 2002, pp. 311–318.

[7] Farrús, M.—Costa-Jussà, M. R.—Mariño, J. B.—Fonollosa, J. A. R.: Linguistic-Based Evaluation Criteria to Identify Statistical Machine Translation Errors. 14th Annual Meeting of the EAMT (European Association for Machine Translation), Saint-Raphael 2010.

[8] Costa-Jussà, M. R.—Farrús, M.—Fonollosa, J. A. R.—Mariño, J. B.: Study and Comparison of Rule-Based and Statistical Catalan-Spanish Machine Translation Systems. Computing and Informatics, Vol. 31, 2012, No. 2, pp. 245–270.

[9] Farrús, M.—Costa-Jussà, M. R.—Mariño, J. B.—Poch, M.—Hernández, A.—Henríquez, C.—Fonollosa, J. A. R.: Overcoming Statistical Machine Translation Limitations. Error Analysis and Proposed Solutions for the Catalan-Spanish Language Pair. Language Resources and Evaluation, Vol. 45, 2011, No. 2, pp. 181–208.

[10] Koehn, P.: Statistical Significance Tests For Machine Translation Evaluation. Proceedings of EMNLP, Vol. 4, 2004, pp. 388–395.

Marta R. Costa-Jussà is a Telecommunications Engineer at the Universitat Politécnica de Catalunya (UPC, Barcelona). She received her Ph. D. from the UPC in 2008. Her research experience is mainly in machine translation. She has worked at LIMSI-CNRS (Paris), Barcelona Media Innovation Center (Barcelona), the Universidade De Sao Paulo (Sao Paulo), Instituto Politécnico Nacional (Mexico) and Institute for Infocomm Research (Singapore). Since December 2012 she is implementing the IMTraP project (Integration of Machine Translation Paradigms on Hybrid Machine Translation), funded by the European Marie Curie International Outgoing European Fellowship Program. She has published around 70 papers in the field of machine translation and she has participated in 12 national and European projects.

José A. R. Fonollosa received the M. Sc. and Ph. D. degrees in electrical engineering from the Universitat Politècnica de Catalunya (UPC), Spain, in 1986 and 1989, respectively. Since 1986, he is with the Department of Signal Theory and Communications of the UPC, where he is now a Full Professor. From August 1991 to July 1992 he was visiting the Signal and Image Processing Institute, University of Southern California. He received the 1992 Marconi Young Scientist Award. He is co-founder (1999) of the company Verbio Technologies S. L. specialized in spoken language technology. Since 2006, he is the Director of the Center for Language and Speech Technologies and Applications (TALP Research Center). He has published more than 100 scientific papers in statistical signal analysis and their application in communication systems, speech processing and statistical machine translation. During the last 5 years, he has concentrated his efforts on (spoken) language translation participating in local as well as EU funded projects as FAME, LC-STAR, TC-STAR and FAUST.

José B. Mariño is Full Professor at UPC since 1981. He received the M. Sc. degree in telecommunication engineering from the Universidad Politécnica de Madrid, Madrid, Spain, in 1972 and the Ph. D. degree in signal processing from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1975. He is former Director of the Telecommunication Engineering School, Barcelona and is also former President of the Department of Signal Theory and Communications, both at UPC. Currently, he is heading a research team in speech-to-speech translation at UPC. He has been managing several speech-related projects in Spain, and he has participated in several European Union research projects. He has co-authored more than 200 technical papers in the fields of signal processing, speech recognition, acoustic modeling, confidence measures, spoken dialogue systems, speech-to-speech translation and statistical machine translation.

Marc Poch is a researcher at Institut de Linguística Aplicada (IULA) of Universitat Pompeu Fabra (UPF). He is a Telecommunications Engineer at Universitat Politècnica de Catalunya (UPC); he has worked in national and European projects. His research interests include NLP, machine translation, data mining, artificial intelligence and machine learning.

Mireia Farrús graduated in physics and in linguistics from the Universitat de Barcelona, and she received the degree of Doctor of Philosophy from the Universitat Politécnica de Catalunya (UPC, 2008), in Barcelona. She got a postdoctoral position for two years at the UPC to work in the machine translation field, where her main contribution was to use the linguistic knowledge to improve the state-of-the-art machine translation systems. From 2009 to 2011, she held a research position in the Office of Learning Technologies at the Universitat Oberta de Catalunya (UOC), where she worked on the UOCs Machine Translation System and on the development of e-learning applications related to natural language processing. She has also held researcher positions at DFKI (German Research Centre for Artificial Intelligence, Germany), Umeå Universitet (Sweden) and the National Centre for Biometric Studies at the University of Canberra (Australia). From January 2012 she is at the Universitat Pompeu Fabra, doing research in the Natural Language Processing group (TALN) and teaching at the Polytechnic School (ESUP).