# An experience on statistical machine translation between Spanish and the regional languages of Spain

**Mireia Farrús[1], Gonzalo Iglesias[2], Carlos Henríquez[1], Marc Poch[1], Roberto Muñoz[1], Nerea Ezeiza[3], Eduardo R. Banga[2], José B. Mariño[1]**

[1]Universitat Politècnica de Catalunya, TALP Research Center, Dep. Signal Theory and Communications, Barcelona, Spain
[2]Universidade de Vigo, Department of Signal Theory and Communications, Vigo, Spain
[3]Euskal Herriko Unibertsitatea, Department of Language and Computer Systems, Bilbao, Spain
{mfarrus,carloshq,mpoch,alrobems,canton}@gps.tsc.upc.edu, {giglesia,erbanga}@gts.tsc.uvigo.es, n.ezeiza@ehu.es

**Abstract**

Statistical machine translation systems between Spanish and other regional languages from Spain has become an interest of research during the last decade. However, regional languages are usually characterized by the lack of linguistic resources necessary to build such systems. This paper describes the development of three statistical machine translation systems between Spanish and three other languages: Galician, Catalan and Basque, focusing on the corpora used and the techniques applied in order to improve their performance.

## 1. Introduction

Several regions in Spain have two official languages: in addition to Spanish, another language shares the official status. These co-official languages are Galician (a romance language close to Portuguese and spoken in Galicia), Catalan (also a romance language spoken in the Mediterranean regions of Spain) and Basque (an ancient language spoken in the Basque Country). The interest in machine translation systems between Spanish and these regional languages is not new and several research projects have been devoted to develop such systems for text-to-text translation. Although some experience using the statistical machine approach has been carried out, the most known systems are the rules-based ones.

Within the framework of the AVIVAVOZ project, funded by the Spanish government under grant TEC2006-13694-C03, the statistical approach has been adopted to provide speech-to-speech translation between Spanish and Galician, Catalan and Basque. The main problem to deal with has been the lack of bilingual corpora. In this paper, a description of the corpora used in the text-to-text translation, the developed systems and their performance is provided.

## 2. A Spanish-Galician SMT system

Along the last decade, some previous efforts for Spanish-Galician machine translation (Diz-Gamallo, 2001; Gómez et al., 2003) can be found, such as the rule-based translation system *OpenTrad* (Corbí-Bellot et al., 2005).

Unlike rule-based systems, when working on statistical machine translation the real challenge is to obtain suitable parallel corpora, specially in minority languages like Galician. This section describes the parallel corpus extraction and the machine translation system between Galician and Spanish languages, whose core is the decoder *Marie*, developed at the Universitat Politècnica de Catalunya (Crego et al., 2005).

### 2.1. Corpus

The main problem in building a Galician-Spanish statistical machine translation is that very few resources are available. In this work, some translated sections coming from *El Correo Gallego*[1] newspaper and the *Eroski Consumer* web[2] are used, where news are translated by experts to all the official languages in Spain. The material from *El Correo Gallego* was used for training, whereas the development and testing was performed over a set of sentences extracted from the *Eroski Consumer* texts. The Official Galician Government Bulletin[3] is another available source for a legal domain.

#### 2.1.1. Extraction and preprocessing

Both materials required a web extraction, which included html filtering, sentence boundary detection, stripping off website names, e-mail addresses and similar pre-processing steps. For extraction and html filtering, the open-source tools *wget* and *lynx* were used, respectively.

#### 2.1.1. Sentence alignment

Sentences of these bilingual corpora have been extracted by using the well-known LCS algorithm. A variant of this algorithm provides a normalized score of similarity for a sentence pair candidate. This pair is accepted if the score exceeds a given threshold and has the highest value compared to other scores in the window of sentence pair candidates. Ideally, the complete bilingual corpora should be manually checked; however, due to the lack of human revisers, only and informal revision of small parts of the corpora was performed. Details about both corpora are showed in Table 1 and Table 2.

| | es | gl |
|---|---|---|
| sentences | 85073 | |
| average sentence size | 18.97 | 17.89 |
| vocabulary size | 72498 | 74343 |

Table 1. Training set from *El Correo Gallego* (1995-2002).

[1] http://www.elcorreogallego.es
[2] http://www.consumer.es
[3] http://www.xunta.es/diario-oficial

| | development | | test | |
|---|---|---|---|---|
| | es | gl | es | gl |
| sentences | 1550 | | 1651 | |
| avg. sent. size | 20.38 | 19.51 | 21.12 | 20.19 |
| vocab. size | 6668 | 6786 | 7203 | 7326 |

Table 2. Development and test sets from *Eroski Consumer* (2007).

### 2.2. Machine translation system

As it was stated above, the N-gram decoder Marie was used in this translation task. The translation was performed by using two different models: a bilingual language model and a target language model. The GIZA++ (Och, 1999), was used to estimate the word alignments –refined with the union of both alignment directions– from which we obtain the bilingual model based on tuples (Crego et al., 2005). Both language models are estimated using the SRILM tools (Stolcke, 2002). No reordering model is required. Out-of-vocabulary words are allowed to pass through from source to target: the closer two languages are, the bigger the probability of sharing common words.

By revising the translations manually, we discovered some issues that were solved for the improved version. For instance, the Galician language uses many contractions and clustered clitics. As a pre/post-processing stage for training and decoding, they were identified by a Galician POS tagger (Méndez et al., 2003), and split into smaller words. Interestingly, we also discovered systematic translation discrepancies on the *Eroski Consumer* sets (material from 2007). Just to put a simple example, the frequent expression *to the* is a contracted word in Galician, and used to be written years ago as *ó*. However, since recent times it is preferred to write it as *ao*. Many similar discrepancies were identified and sorted out manually. Other changes to the baseline system include a bilingual dictionary of over 40 000 words and exclusion of conflictive tuples from the bilingual language model.

### 2.3. Evaluation and results

Table 3 shows the performance of both baseline and improved systems –in terms of BLEU– over the *Eroski Consumer* development set in both directions of translation. On the other hand, Table 4 shows the performance of the improved system over the *Eroski Consumer* test set. In all cases, one reference is available.

| | baseline | improved |
|---|---|---|
| es2gl | 78.1 | 79.3 |
| gl2es | 70.1 | 75.8 |

Table 3. BLEU score for Spanish-Galician translation task over the *Eroski Consumer* development set.

| | es2gl | gl2es |
|---|---|---|
| BLEU | 77.7 | 80.6 |
| NIST | 12.66 | 13.00 |
| PER | 12.19 | 10.58 |
| WER | 12.85 | 11.26 |

Table 4. Performance obtained in the Galician-Spanish tasks over the *Eroski Consumer* test set.

As it was expected, the performance is reasonably high even with such a simple system, although it is worse for the Spanish-to-Galician task than for the Galician-to-Spanish task over the test set. An informal examination of translation hypotheses suggests that, although the output text is reasonably good and fairly understandable, the complex clitic clustering in the Galician language still remains an issue with such a small amount of training data.

## 3. A Spanish-Catalan SMT system

The Spanish-Catalan case has also been treated in rule-based machine translation in last years (Alonso and Thurmair, 2003; Alonso, 2005). This section introduces a statistical-based machine translation system between Catalan and Spanish (N-II[4]) developed at the Technical University of Catalonia (UPC). N-II is an N-gram-based statistical machine translation system (Mariño et al., 2006), trained with an aligned Spanish-Catalan parallel corpus taken from the bilingual editions of the *El Periódico*[5] newspaper.

Since the Catalan and Spanish languages belong to the same linguistic family, no serious errors are encountered during in the translation task. However, translations can be highly improved by using adequate techniques.

First, this section describes the main characteristics of the corpus used, as well as the way it was processed to make it suitable for our MT system. Second, an error analysis and the techniques applied to solve the encountered errors are briefly introduced. Finally, an automatic evaluation allows showing the improvements achieved by using different corpora and by applying the presented improvement techniques.

### 3.1. Corpus

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. Therefore, one of the main objectives was to produce a new Catalan-Spanish parallel corpus as a basis for the development of a high quality SMT system for such pair of languages. Te parallel corpus used in our system was extracted from eight years (2000-2007) of the paper edition in pdf format. Then, such files were subject to the following processes in order to make them suitable for the MT system:

#### 3.1.1. *pdf* to *text* format conversion

The first step in processing the corpus was to convert each pdf file into the corresponding text file. This was done by using the *pdftotxt* open-source program. Once the data were converted, sentences were divided into different lines like in the newspaper columns. This sentence reconstruction allowed putting one whole sentence in each line.

[4] http://www.n-ii.org
[5] http://www.elperiodico.cat

#### 3.1.2. Filtering

Due to the complex layout of the paper edition, the recovery of the sentence from each pdf file may lead to incomplete words and sentences. Therefore, after extraction, a filtering process was followed in order to detect incorrect sentences. Additional content-based filtering procedures were also included to avoid duplicate newspaper sections and other parts that were considered of no interest for the final parallel corpus.

#### 3.1.3. Tokenization and normalization

The aim of tokenization and normalization is to reduce sparseness. Given a text, tokenization chops it into tokens, and this language-dependent task uses different tokenizations for Spanish and Catalan. On the other hand, the aim of normalization is to group characters into the same token that are, in fact, related to the same character but are written in a different way (e.g. the use of different symbols for the apostrophe: *l'arbre* and *l´arbre*, needs to be unified).

#### 3.1.4. Sentence alignment

In order to use the corpus as training material for corpus-based MT systems, it becomes necessary to align each line in the Spanish file with its corresponding translation in the Catalan file. For this purpose, the program *Bilingual Sentence Aligner* (Moore, 2002) was used, which searches paired translations and discards sentences without translations.

#### 3.1.5. Statistical bilingual sentence filtering

In this final process, parts of sentences that have statistically a low probability of having a translation correspondence are eliminated by using two different filtering approaches: IBM1 (Brown, 1993) and the contribution to the total probability from stopwords.

In an initial step, only the editions from years 2000-2003 where used. Afterwards, the complete training corpus from years 2000-2007 was used, so that it was possible to see how this corpus extension improved the system performance. Both corpora were split into two subsets: training, development and test. Table 5 and Table 6 show the statistics of both partial and complete corpora. Table 6 shows only the training corpus, since the development and test sets remained as in the partial corpus.

| | train | | development | | test | |
|---|---|---|---|---|---|---|
| | es | ca | es2ca | ca2es | es | ca |
| sentences | 2178796 | | 1986 | 1966 | 692 | 808 |
| avg.sent.size | 18.98 | 19.06 | 20.62 | 20.70 | 23.08 | 21.15 |
| vocab.size | 397766 | 430252 | 12518 | 12384 | 5705 | 5538 |

Table 5. Partial corpus statistics of N-II system.

| train | es | ca |
|---|---|---|
| sentences | 4656926 | |
| avg.sent.size | 20.80 | 20.82 |
| vocab.size | 1231358 | 1287234 |

Table 6. Complete train corpus statistics of N-II system.

### 3.2. Improvement techniques

In this section, the main translation errors encountered and the solutions applied to solve them are described (Farrús et al, 2008). The errors and solutions are classified according to the linguistic level they belong.

#### 3.2.1. Orthographic errors

The most common orthographic errors are the use of the apostrophe in Catalan and the modification of some conjunctions in Spanish in front of specific vowels. The correction of the apostrophe requires knowing the word gender, and thus, its corresponding grammatical-category tag. To solve the conjunction errors, a simple post-processing after the main translation is enough.

#### 3.2.2. Morphological errors

A common error between this pair of languages is the lack of gender concordance in those words where the corresponding translation has a different gender. In this case, a tag language model is created, which benefits those word sentences that maintain the gender coherence. However, this technique can only be used if the bilingual unit in question exists in the training corpus.

#### 3.2.3. Lexical errors

Lexical problems are mainly due to the fact that a word to translate is not included in the training corpus. To solve the existence of these unknown words, three strategies have been performed. First, a lexical categorization of numbers and time expressions; these are transformed into a code and generated a posteriori into the target language. Second, the addition of an external dictionary after the translation process in order to translate those words that remained as unknown. And finally, the inclusion of a spellchecker to help the user write correctly, since many of the unknown words come from misspellings.

#### 3.2.4. Semantic errors

When the same word has more than one meaning, it will probably have more than one translation in the target language. If the different meanings have different grammatical categories, these can be used to add a grammar tag to the word, so that the model will learn from the context and the different meanings will be disambiguated. In our system, this was applied to the Spanish element solo and to the Catalan sol, perquè and possessive pronouns and adjectives.

#### 3.2.5. Syntactic errors

The most common syntactic and relevant errors between Catalan and Spanish include the combination of the pronominal clitics with verbs, the use of the relative structure *cuyo*, and the non-systematic use of many prepositions, especially between *a* and *en*. The first one was solved by using some combination rules in both directions. In the second one, the *cuyo* pronoun was transformed into the equivalent structure in Catalan (*del cual*). In the case of the preposition errors, only a systematic rule could be found: in front of proper nouns, Spanish *en* is translated into *a*.

### 3.3. Evaluation and results

The evaluation of the system was performed by considering both the application of all these techniques and the extension of the corpus. Table 7 and Table 8 show the results in both directions of translation for the baseline system, the linguistically improved system, and

the improved system plus the extended corpus. The test corpus contained 692 Spanish sentences extracted from the *El País* and *La Vanguardia* newspapers and 808 Catalan sentences extracted from the *Avui* newspaper and transcriptions of the *Àgora* TV program. Two manual references were available.

| | baseline | +improved | +extended corpus |
|---|---|---|---|
| BLEU | 81.48 | 85.17 | 86.50 |
| NIST | 12.84 | 13.25 | 13.40 |
| WER | 14.27 | 11.93 | 10.91 |
| PER | 13.08 | 11.04 | 10.10 |

Table 7. Evaluation for the Spanish-Catalan translation.

| | baseline | +improved | +extended corpus |
|---|---|---|---|
| BLEU | 85.31 | 87.44 | 88.58 |
| NIST | 13.12 | 13.33 | 13.46 |
| WER | 10.30 | 9.12 | 7.92 |
| PER | 9.30 | 8.27 | 7.27 |

Table 8. Evaluation for the Catalan-Spanish translation.

# 4. A Spanish-Basque SMT system

Basque language has many particularities, which differences it from most European languages and make the translation between Spanish and Basque an interesting challenge that involves both morphological and syntactic features. Basque is an agglutinative language where morpho-syntactic information is expressed using suffixes (whereas in most of the European languages is used as separate words). Furthermore, there are also syntactic differences that affect the word order and have a negative impact on the translation (Díaz de Ilarraza et al., 2009).

In this section, a phrase-based statistical machine translation (Koehn et al., 2003) for Spanish and Basque languages is presented. This system aims to reach acceptable translation quality under conditions of smaller training material. It uses a POS language model and a lemmatized language model for Basque in order to improve the translations obtained by a baseline system, and it is based on the MOSES toolkit (Koehn et al., 2007).

## 4.1. Corpus

For the system design, the corpus used was provided by the Albayzin 2008 evaluation campaign[6]. This is a set of 61104 sentences, divided into three subsets: a training corpus (58202 sentences), a development corpus (1456) and a test corpus (1446 sentences). Only one reference for each set was supplied. The basic corpus statistics can be found in Table 9.

| | train | | development | | test | |
|---|---|---|---|---|---|---|
| | es. | eu | es | eu | se | eu |
| sentences | 58202 | | 1456 | | 1446 | |
| avg.sent.size | 19.77 | 15.20 | 23.33 | 18.70 | 22.32 | 17.85 |
| vocab. size | 97558 | 140931 | 7170 | 9031 | 6926 | 8691 |

Table 9. Basic Spanish-Basque corpus statistics.

The tokenization, Part-of-Speech tags and lemmatization for both languages was also provided although only the Basque information was used during development. According to the campaign documentation, the POS tags for Basque were obtained with the Estagger tool (Aduriz et al., 1996) and only categories and subcategories of the tags were provided.

An additional pre-processing consisted in changing the text encoding from iso88591 to utf8, by removing all the sentences longer than 100 words and lowercasing the entire corpus. The removing step was performed only over the training corpus due to a restriction implied by the alignment tool. The removed set was smaller than 1% of the training corpus.

## 4.2. Alignment

Although the baseline system used lowercased corpus to obtain the word-to-word alignment, the final system used a different approach. A segmentation tool was developed, which split the Basque words using the POS information and a suffix dictionary; wherever a verb, and adjective or a name was found, the word was checked with the dictionary, and if the word ended with any of the listed suffix, it was split into two (e.g. 'publikoen', a Basque adjective ended with the suffix 'en' listed in the dictionary, was split into 'publiko+ +en'.

By using the segmentation tool, the Basque lowercase corpus was segmented and the alignment was computed on this corpus. Once the alignment was completed, and using a developed desegmentation tool, which joins the words with their split suffixes, the corpus was desegmented to its original version and the links were properly relocated to the original words.

The Spanish lowercased corpus remained the same during the process. The alignment was automatically computed by the GIZA++ toolkit (Och and Ney, 2003). Finally, since all the design used a lowercased corpus, a final case-restoration was needed to establish a true-case translation. To this end, the *disambig* and *ngram-count* tools from the SRI Language Model toolkit were used.

## 4.3. Evaluation and results

For the Spanish-Basque task, four different systems were developed in a progressive fashion. Starting from a *baseline* with the default functions and parameters of MOSES, a *POS target language model* was added to the SMT system. In a third system, a modified *alignment* was performed as it was described in section 4.2.

The final system also included a *target language model based on lemmas*. As mentioned in section 4.1, the lemmas for Basque were computed automatically and were provided by the organizers of the evaluation campaign. Both language models (POS' and lemmas') were 7-gram models and the order of the surface words language model was 5. The maximum phrase size was set to 5 for all the systems.

Table 10 shows the different results obtained with all four developed systems. It can be clearly seen that the addition of the different features resulted in a final improvement of 0.5% BLEU points over the test set.

| | baseline | POS LM | segm. align. | lemma LM |
|---|---|---|---|---|
| BLEU | 12.66 | 12.76 | 13.01 | 13.26 |
| NIST | 4.77 | 4.72 | 4.80 | 4.90 |
| WER | 77.90 | 78.38 | 78.61 | 77.61 |
| PER | 60.15 | 60.61 | 60.36 | 59.37 |

Table 10. Results obtained on the test set.

[6] http://jth2008.ehu.es/en/albayzin.html

Another MT system applied to Basque (Díaz de Ilarraza et al., 2009) obtained slightly higher relative improvements in terms of BLEU, but using different techniques (morphological segmentation of words and word reordering). Those techniques could be integrated in our system to get even better results.

The tables have shown very low BLEU scores for Basque. They seem even lower compared to BLEU scores for the Romance languages that are extremely high (around 77-80%). Some reasons that could explain this fact are the following: first, BLEU scores are usually very low for agglutinative languages like Basque; second, Spanish, Catalan and Galician language are very close from each other, with almost no word order differences; and finally, in any case, it is commonly recognized that BLEU scores shouldn't be used to compare translation quality among systems for different languages (Callison-Burch et al., 2006).

## 5. Conclusions

Normally, regional languages are characterized by a lack of language resources in front of the main languages of the country. Spain is not an exception, and, in addition, Spanish is spoken by millions of people in other countries, so that the difference becomes substantially bigger. For this reason, a lot of effort is needed when trying to obtain language resources for minority languages such as Galician, Catalan and Basque. In this paper, all three languages share the same objective: to obtain parallel corpora between them and Spanish language for the statistical machine translation task.

Although some processes differ in minor details, the corpora extraction is successfully done in Galician, Catalan and Basque. However, corpora extraction is a necessary task, but not always sufficient to obtain a high quality system. To this end, each system applied suitable techniques with successful results, as it is shown in the evaluations performed.

As it was expected, the Galician-Spanish and Catalan-Spanish tasks present better results, since they all come from the same linguistic family, whereas in the Basque the translation task becomes more difficult. However, all the systems could be improved by applying the corresponding techniques and all of them outperformed their baselines.

## References


Aduriz, I., Ezeiza, N., Urizar, R. (1996). Euslem: A lemmatiser/tagger for Basque. In Proceedings of the EURALEX. Göteborg, Sweden, pp. 17--26.

Alonso J.A., Thurmair G. (2003). The Comprendium Translator System. In Proceedings of the MTS-IX, New Orleans, USA.

Alonso J.A. (2005) Machine Translation for Catalan-Spanish — The real case for productive MT. In Proceedings of the EAMT, Budapest, Hungary.

Callison-Burch, C., Osborne, M., Koehn, P. (2006). Re-evaluating the role of BLEU in Machine Translation research. In Proceedings of the EACL, Trento, Italy.

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 EAMT, Budapest, Hungary, pp. 79--86.

Crego, J.M., Mariño, J.B., de Gispert, A. (2005). An N-gram-based Statistical Machine Translation Decoder. In Proceedings of the Eurospeech -- Interspeech, Lisbon.

Díaz de Ilarraza, A., Labaka, G., Sarasola, K (2009). Reordering on Spanish-Basque SMT. (2009). In Proceedings of the MT Summit XII, pp. 207--213.

Diz-Gamallo I. (2001). The importance of MT for the survival of minority languages: Spanish-Galician MT system. Proc. MTS-IX, Santiago de Compostela.

Farrús, M., Costa-jussà, M., Poch, M., Hernández, A., Mariño, J.B. (2009). Improving a Catalan-Spanish Statistical Translation System using Morphosyntactic Knowledge. In Proceedings of the EAMT, Barcelona.

Gómez A.S., Conde E. V. (2003) The Functionality of a Tool Bar for Postedition in MachineTranslation between Languages with Linguistic Interference: the Spanish-Galician Case. In Proceedings of the MTS-IX, New Orleans, USA.

Koehn, P., Och, F.J., Marcu, D. (2003). Statistical phrase-based translation. In Proceedings of the HLT-NAACL pp. 48--54.

Koehn et al. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the ACL pp. 177--180.

Mariño, J.B., Banchs, R.E., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J.A.R., Costa-jussà, M.R (2006). N-gram-based Machine Translation. Computational Linguistics, 32(4):527--549.

Méndez, F., Campillo, F., Banga, E.R., Rei, E.F. (2003). Análisis morfológico estadístico en lengua gallega. In Proceedings of the SEPLN, 31:71--76.

Moore, R. (2002) Fast and Accurate Sentence Alignment of Bilingual Corpora. Springer-Verlag.

Och, F.J. (1999). An Efficient Method for Determining Bilingual Word Classes. In Proceedings of the EACL, pp. 71--76, Bergen, Norway.

Och, F.J., Ney, H. (2003). A systemantic comparison of various statistical alignment models. Computational Linguistics, 29(1):19--51.

Stolcke, A. (2002) SRILM - An extensible Language Modeling Toolkit. In Proceedings of the ICSLP -- Interspeech, Denver, Colorado.