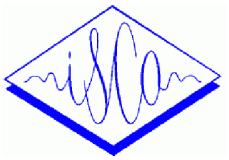

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

# EuskoParl: a speech and text Spanish-Basque parallel corpus

*Alicia Pérez*[1], *José M. Alcaide*[2], *María-Inés Torres*[2]

[1]Dpto. Lenguajes y Sistemas Informáticos, Universidad del País Vasco UPV/EHU, Bilbao (Spain)
[2]Dpto. Electricidad y Electrónica, Universidad del País Vasco UPV/EHU, Bilbao (Spain)

{alicia.perez, josemaria.alcaide, manes.torres}@ehu.es

## Abstract

The advances in corpus-based approaches and machine learning techniques have promoted the development of minority languages. The contribution of this work is to acquire a parallel corpus in Spanish and Basque with both text and speech data. In order to be able to compare the systems with those developed for other languages, Europarl corpus was taken as a reference in both domain and size. The acquisition process, carried out within the Basque Parliament reports and speeches, involved subtle differences to that of Europarl acquisition. The resulting corpus is described and a few preliminary experiments on machine translation with Moses are reported.

**Index Terms**: speech resources, statistical machine translation, under-resourced languages

10.21437/Interspeech.2012-619

## 1. Introduction

Language and speech technologies have evolved significantly in recent years boosted by machine-learning approaches based on data. An incessant development of speech recognition, language identification, machine translation, etc. is going on based on statistical methods. Yet, statistical models require huge amount of representative data for robust training of model-parameters. Moreover, for the sake of usability these enormous sets of data need to be processed in advance, indeed, we find a gap in the literature on contributions focusing on robust preprocessing strategies.

On the other hand, the availability of linguistic resources, speech and either monolingual or bilingual text, depends strongly on the socio-economic circumstances of the language itself. In this context, minority languages tend to be under-resourced, resulting in a slow development of speech technologies. This is the case of Basque, the official language, together with Spanish, in the Autonomous Community of the Basque Country (ACBC) holding 2.5 million inhabitants. Basque is also spoken in some areas of Navarre and Atlantic Pyrenees, and also in other communities abroad e.g. Nevada, Idaho, etc.

Even though Basque and Spanish coexist in the ACBC, these languages present great differences in lexicon origin, morphological structure and syntactic construction. Basque is a pre-IndoEuropean language considered to be the only isolated language in Europe without any relationship with other living languages. While the syntax is rich for both languages, and the phrases can be arranged in different ways, the expected syntax for Spanish is subject-verb-objects, while for Basque is subject-objects-verb. This aspect leads to long-distance alignments between Spanish and Basque. On the other hand, Basque is a highly inflected language yielding n-to-one relations between Spanish and Basque.

The final aim of this work is to develop a corpus as similar as possible in terms of both domain and size to Europarl in Spanish and Basque with text and speech resources. EuskoParl is named after Europarl, *Eusko-* meaning *Basque*. Specifically, EuskoParl consists of texts from the proceedings of the Basque Parliament separated into the two languages in monolingual parallel files; in addition, it counts on a subset of the original speeches (in either Spanish or Basque), along with their transcription and translation into the counterpart language.

Overall, both Europarl and EuskoParl are within the same domain, and the Spanish-side of EuskoParl is nearly the same dimensions of the Spanish-side of Europarl [1] $20 \times 10^6$ word-forms, (yet, current releases of Europarl are still much bigger). Not only have we stuck to the dimensions but we have also followed the acquisition process, yet, some differences arose and so have we described here. All in all, EuskoParl corpus will serve to boost the development of language technologies in Spanish and Basque and compare them with those developed for other European languages.

The arrangement of the paper is as follows: Section 2 is devoted to describe the corpus acquisition process; the characteristics of the obtained data-sets are shown in Section 3, in addition some preliminary machine translation results are presented. Finally, some concluding remarks and hints for future work are given in Section 4.

## 2. Corpus acquisition

In the acquisition process two data-sets can be distinguished: parallel-text and speech. Corpus acquisition process is often left aside in the literature. Yet, we find it important to give details for future references and better understanding of the forthcoming results.

### 2.1. Parallel text-data acquisition and processing

Parallel text-data consist of two sets of monolingual sentences (in Spanish or in Basque) aligned, as it is the typical case of a parallel corpus for machine translation (e.g. Europarl). For the data acquisition process we stuck to the procedure described in [1], yet subtle differences arose that are well-worth mentioning:

**1. Inventory:** the web was crawled in order to get an inventory of location of raw bilingual-data, source-language speeches, their transcription and translation. The inventory mapped the documents and their relations in our data-base and their content mirrored on the web. The mapping resulted of much benefit for the following steps, and deserved the time took.

**2. Raw data acquisition:** raw data was extracted following the mapping in the inventory. According to the Translation Service of the Basque Parliament, the translations were typically carried out at paragraph level. Nevertheless, files with different number of paragraphs were found. The disparities seemed to be due to file-formatting issues. On this regard, the files that did not contain the same number of paragraphs in both languages were discarded on this first approach but we meant to retrieve them on future works focusing explicitly on alignment strategies. A manual inspection of randomly selected 100 paragraphs showed that the 100% were correctly aligned. This inspection together with the directive by the Translation Service lead us to work on the assumption that the documents were aligned at paragraph level. The raw data-set is case-sensitive, holds tags derived from html format, and the paragraphs keep formatted with line-breaks, hyphens, printable and non-printable characters, etc.

**3. Tokenization:** the raw corpus would not be usable for machine learning. To begin with, 12,986 lines in the Spanish set and 13,005 lines in the Basque (out of 371,645) contained non-printable characters such as `['82', '89', 'ad', '91', '92', '93', '94', '96', '97']` (coded in Unicode/Latin1). It is of high importance to detect such characters. For example, while printable-hyphen mark allows to join words, not removing the soft-hyphen (`0xad`) causes fake word-splitting yielding fake word-forms in the vocabulary.

On the other hand, all the lines within a paragraph were joined into a single line (taking care of hyphens). Next, the punctuation marks and symbols were separated from words, in order to avoid artificially increasing the vocabulary. Finally, the corpus was lower-cased keeping a list of names for the sake of carrying out appropriate phonetic transcriptions. The corpus still shows a series of printable characters that should possibly be filtered as they seem to be derived from text-editing and formatting processes.

On this stage we tried to verify the assumption that the files were aligned at paragraphs level by means of a naive fertility-based-filter. The idea is as follows: given that the number of running words of the entire Spanish and Basque data-sets are approximately $2.3 \times 10^6$ and $1.8 \times 10^6$ respectively, the average fertility results in 1.3. The filter would allow to detect the pairs that are not within the allowed fertility threshold. Note that, individually, a Basque word could produce many more Spanish words than 1.3, for instance, the Basque word *amamarenean* (meaning *at granny's home*) would produce 6 words in Spanish *en la casa de la abuela*. Thus it is important to allow a margin bigger than the average for the fertility of the individual pairs. Figure 1 shows the number of pairs (on the ordinate) that were out of a fertility threshold (on the abscise) with the threshold ranging from 2 to 9 (GIZA's default fertility parameter is set to 9). For example, 2,494 paragraphs inconsistent with a fertility parameter 3 were detected. A manual inspection over 100 of those paragraphs showed that in 3 out of 100 the differences were due to differences in translation style. However the remaining 97 were associated to a bad document segmentation of the file into paragraphs.

That is, the prior hypothesis settling that the documents were aligned at paragraphs level has revealed not to be always true. By means of the fertility-based-filter (set to 3), it has been shown that 2,494 paragraphs (out of 743,214) might turn to be bad translation candidates, and thus, well worth of either manual or automatic inspection. Yet, in this preliminary version of the corpus we did not get rid of any of them.

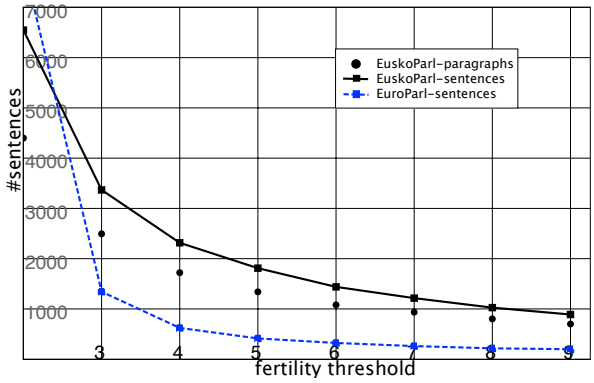


Figure 1: *Number of sentences above the fertility-threshold for the EuskoParl corpus in before and after segmentation (denoted as paragraphs and sentences respectively) and for Europarl Spanish-English corpora.*

**4. Segmentation:** so far, the corpus is aligned at paragraph-level, still, smaller translation pairs are desirable. Typically the paragraphs do not contain the same number of sentences, that is, the translation does not happen at sentence-level. Bilingual alignment at sentence level was a hectic field of translation at late nineties: there are methods based on sentence-lenght [2] (eg. Europarl corpus was segmented following this approach), others that focus on spotting cognates to serve as pivots [3] and methods that rely on word-alignment matrices to perform

recursive refine the alignment hypothesis or splitting [4]. In our case, an extension of Thot [5] toolkit was used to segment paragraphs into sentences. The extension implements alignment models based on IBM models [6] (from 1 to 4) and also mixtures of IBM-1. Specifically, it was use the module to segment parallel corpora based on recursive alignments (described in [4]). After this process the 371,607 paragraph-pairs were split yielding 771,780 sentence-pairs. Note that we are referring to sentences while often the pair consists of more than one sentence.

In order to quantitatively assess whether the segmentation has done good or harm to previously detected paragraphs over the fertility threshold, we run again the filter leading to the results shown in Figure 1 (denoted as EuskoParl-sentences). For the sake of comparisons, the same results were obtained with the Europarl Spanish-English release from Europarl. Note that in any case, the threshold for Europarl and EuskoParl should be different since the former represents Spanish-English and the latter Spanish-Basque. It is well worth mentioning that, by contrast to Europarl, the EuskoParl corpus does not contain any empty line that would potentially do harm to machine learning algorithms.

**5. Shuffling:** the sentence-pairs were shuffled and afterwards 30,000 pairs were extracted at random without replacement for testing purposes and the remaining was left to do training.

### 2.2. Speech-data acquisition

Speech-data, by contrast, consist of speeches in the source language (being it either Spanish or Basque) together with the transcription (in the source language) and its translation into the counterpart language. It often occurs the fact that one speaker in the same speech segment switches from one language to the other. Thus, both speech, transcription and translation are bilingual, meaning that both languages may appear together within a sample of the data.

These data represent a sub-set of the text-data, however, transcription and translation files are not monolingual. Transcription and translation files represent the source language of the speaker and its translation into the other language. Both, text transcription, translation and splitting languages into monolingual files was carried out by the linguist-staff from the Basque Parliament.

The transcription sums up to 1,001,147 running words, associated to approximately 189 hours of audio at 16KHz and uttered by 81 different speakers. The audio files contain long silences related to pauses, boos, claps, etc. that are not annotated in the associated transcription.

## 3. Corpus description

As a result of the corpus acquisition process described in previous section, the corpus derived holds the features described in Table 1. The vocabulary refers to the different word-forms in the corpus (thus, singular and plural form of the same entry in a vocabulary would produce two entries in our vocabulary). The high declension of Basque makes the vocabulary grow (note that there are 2.3 times more words in the Basque vocabulary than in the Spanish one). The declension is also reflected in the differences in terms of running words (note that there are 1.2 times more running words in Spanish than in Basque).

| **EuskoParl** | | **Spanish** | **Basque** |
|---|---|---|---|
| **Train** | Pair of sentences | 741,780 | |
| | Running-words | 22,668,478 | 18,161,805 |
| | Vocabulary | 113,969 | 264,162 |
| | Avg. sentence-lenght | 30.5 | 24.4 |
| **Test** | Pair of sentences | 30,000 | |
| | Running-words | 915,528 | 733,900 |
| | OOV | 1,671 | 5,362 |
| | Avg. sentence-lenght | 30.5 | 24.4 |
| | PP (3grams) | 59.7 | 174.6 |
| **Speech** | Speakers | 81 | |
| | Hours | 189 | |
| | Running-words | 1,001,147 | |

Table 1: Main features of EuskoParl corpus.

The rich morphology of Basque (compared to the Spanish morphology) has an impact on the frequency of the words and n-grams of words in general. On average, the repetition ratio is 199 and 69 for Spanish and Basque respectively. Average values might result misleading, since the variance of the frequency of words is very high as the following statistics show:

- The most frequent word (the first word in the frequency rank) is the ' ,' symbol. It appeared 1,681,510 times in Spanish and 1,791,367 times in Basque, 7.4% and 9.9% of all words respectively.
- The number of singletons, that is, the number of words that appear once in the entire corpus (and thus, hold the last rank), is 42,448 for the Spanish set and 132,466 for the Basque set. Thus, singletons represent the 37% and 50% of the vocabulary in Spanish and Basque respectively. Singletons are typically due to misspellings, numbers, nouns, etc.

On account of frequencies, Figure 2 shows the frequency of Spanish and Basque words with respect to their rank in the vocabulary in logarithmic scale. The approximate linear trend of the figure out from the edges is an empiric evidence of Zipf's law [7]. Note that the Basque corpus shows higher sparsity than the Spanish one. The perplexity (denoted as PP in Table 1) of the test-set on 3grams (computed with SRILM-[8]) corroborates this hypothesis. The high ratio of singletons makes us expect that a high number of out-of-vocabulary words

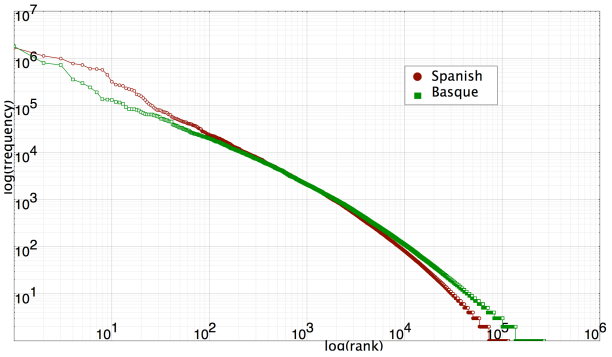


Figure 2: *Distribution of words in the corpus: log-frequency versus log-rank. Zipf's law is verified.*

(OOV) are likely to occur in a related but disjoint test set, as it is the case according to Table 1.

### 3.1. Preliminary machine-translation results

In order to validate the corpus we provide preliminary text translation results with Moses [9]. From the training set (referred in Table 1), 725,000 pairs were used for training the models and 16,780 were extracted to roughly tune the parameters of the model running 1-mert; finally the 30,000 sentences from the test set were translated yielding the results shown in Table 2.

| **EuskoParl** | **WER** | **BLEU** |
|---|---|---|
| **Spanish→Basque** | 71.3 | 11.8 |
| **Basque→Spanish** | 70.2 | 12.0 |

Table 2: *Machine-translation results with Moses.*

The difficulty of the task can be compared to that with Europarl when Finish language is involved (according to the results of the EuroMatrix project in `http://matrix.statmt.org`). Better results were expected translating into Spanish as its PP is much lower than the PP of Basque. A manual inspection suggests that the big amount of OOV words in Basque as input language represents a source of errors well worth to deal with for this pair of languages. Full mert-tuning and adding more factors would also improve the performance of the system.

## 4. Concluding remarks and future work

A corpus acquisition process and the resulting data were described taking as a monitor procedure that of Europal acquisition [1]. The aim is to contribute to develop bilingual resources for Spanish and Basque. The corpus can be benefited by spell-checking [10] and morphological analysis [11]. EuskoParl seems to be a promising corpus to develop a variety of machine-learning applications: the text-parallel corpus will allow to infer machine-translation systems; the speeches are appropriate for speech recognition and also speech translation and acoustic modeling [12]; besides, the fact that speakers switch from one language to the other offers a challenge to language identification [13] within the translation loop.

For future work, the development of inferred-fertility-based is foreseen to refine segmentation and hopefully detect mis-aligned paragraphs in the raw-data. In addition, singleton and OOV are currently being tackled by means of tagging by virtue of FOMA [14] (a freely available FSM toolkit).

## 5. Acknowledgements

We are very grateful to Daniel Ortiz-Martínez (ITI, Valencia) for his kindly supplying us Thot's segmentation module.

We would like to thank the involvement and concern of the Administrative Services of the Basque Parliament in research and development on language technologies.

This work has been supported by the Spanish Ministry of Science and Innovation under both the Consolider Ingenio 2010 programme (MIPRCV CSD2007-00018) and the grant TIN2011-28169-C05-04, and also by the Basque Government under grant GIC10/158 IT375-10.

## 6. References

[1] P. Koehn, "Europarl: A multilingual corpus for evaluation of machine translation," http://people.csail.mit.edu/people/koehn/publications/europarl/, Tech. Rep., 2003.

[2] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," *Computational Linguistics*, vol. 19, no. 1, pp. 75–90, 1993.

[3] S. Bergsma and G. Kondrak, "Alignment-based discriminative string similarity," in *Proc. ACL*. 2007, pp. 656–663.

[4] F. Nevado, F. Casacuberta, and J. Landa, "Translation memories enrichment by statistical bilingual segmentation," in *Proc. LREC*, vol. 1. ELRA, 2004, pp. 335–338.

[5] D. Ortiz, I. García-Varea, and F. Casacuberta, "Thot: a toolkit to train phrase-based statistical translation models," in *MT-Summit: Asia-Pacific AMT*, 2005, pp. 141–148.

[6] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[7] W. Dahui, L. Menghui, and D. Zengru, "True reason for zipf's law in language," *Physica A: Statistical Mechanics and its Applications*, vol. 358, no. 2–4, pp. 545 – 550, 2005.

[8] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. Speech and Language Processing*, 2002, pp. 901–904.

[9] P. Koehn, H. Hoang, *et al.*, "Moses: Open source toolkit for statistical machine translation," *Proc. ACL*, pp. 177–180, 2007.

[10] E. Agirre, I. Alegría, X. Artola, A. Díaz de Ilarraza, M. Maritxalar, K. Sarasola, and M. Urkia, "XUXEN: A spelling checker/corrector for Basque based on two-level morphology," in *Proc. Applied Natural Language of the ACL*, 1992, pp. 119–125.

[11] I. Alegria, G. Aranbarri, K. Ceberio, G. Labaka, B. Laskurain, and R. Urizar, "A morphological processor based on foma for Biscayan (a basque dialect)," in *Proc. LREC*, 2010.

[12] J. M. Olaso, M. I. Torres, and R. Justo, "Representing phonological features through a two-level finite state model," in *INTERSPEECH*. ISCA, 2011, pp. 1733–1736.

[13] V. G. Guijarrubia and M. I. Torres, "Text- and speech-based phonotactic models for spoken language identification of basque and spanish," *Pattern Recognition Letters*, vol. 31, pp. 523–532, April 2010.

[14] M. Hulden, "Foma: a Finite-State Compiler and Library," *EACL 2009*. The Association for Computer Linguistics, 2009.