Proceedings of ParlaCLARIN III @ LREC2022 pages 107–110
Marseille, 20 June 2022

Adding the Basque Parliament Corpus to ParlaMint Project

Jon Alkorta, Mikel Iruskieta
HiTZ Basque Center for Language Technologies - Ixa, University of the Basque Country (UPV/EHU)
jon.alkorta@ehu.eus, mikel.iruskieta@ehu.eus

Abstract

The aim of this work is to describe the collection created with transcript of the Basque parliamentary speeches. This corpus follows the constraints of the ParlaMint project. The Basque ParlaMint corpus consists of two versions: the first version stands for what was said in the Basque Parliament, that is, the original bilingual corpus in Basque and in Spanish to analyse what and how it was said, while the second is only in Basque with the original and translated passages to promote studies on the content of the parliament speeches.

Keywords: corpus, Basque, bilingualism, parliament

1. Introduction

There are three parliaments in the Basque Country and Navarre:

i) The Parliament of Navarre sited in Iruñea/Pamplona is the Navarre autonomous unicameral parliament.

ii) The Parliament of Navarre and Béarn is sited in Pau.

iii) The Basque Parliament is sited in Vitoria-Gasteiz (headquarters) and in Gernika (the symbolic town of Basque laws).

Elected Basque representatives have a representation in these three parliaments and with the aim to build a parliamentary data in Basque language, we decided to choose one of them where we think that Basque language is used most: the Basque Parliament (*Eusko Legebiltzarra*, in Basque).
The Basque Parliament is composed of seventy-five deputies elected from these three provinces: Araba, Biscay and Gipuzkoa and each province has twenty-five deputies. And the spokespersons from all the parties with a significant representation can speak Basque. This is the composition of the chamber after the last elections, held on September 26, 2016 and July 12, 2020 and the distribution of seats:

- Partido Nacionalista Vasco (EAJ-PNV): 28 deputies (37.36% votes) / 31 deputies (39.12% votes): Basque christian-democratic and conservative-liberal party.
- Euskal Herria Bildu (EH Bildu): 18 deputies (21.13% votes) / 21 deputies (27.84% votes): Basque left-wing political coalition.
- Partido Socialista de Euskadi-Euskadiko Ezkerra (PSE-EE / PSOE): 9 deputies (11.86% votes) / 10 deputies (13.64% votes). Spanish social-democratic political party.
- Podemos-Izquierda Unida (Podemos-IU): 11 deputies (14.76% votes) / 6 deputies (8.03% votes): Spanish left-wing electoral coalition.
- Partido Popular (PP) + Ciudadanos (Cs): 9 deputies (10.11% votes) of PP and 0 deputies (2.02% votes) of Ciudadanos / 6 deputies (6.75% votes): Spanish conservative and Christian-democratic political party.
- Vox (Vox): 0 deputy (0.07% votes) / 1 deputy (1.96% votes): Spanish right-wing conservative nationalist political party.

The Basque Government is composed through an agreement between EAJ-PNV and PSE-EE/PSOE, the two political parties that are in the government of the parliament from its creation in 1979 (with an exception where PSE-EE/PSOE governed with PP).
The official languages of this parliament are Basque and Castilian-Spanish and the speech transcripts are produced in two ways i) original transcript: as they were said (Basque and/or Spanish), and ii) reflected translation transcript: in another text column, Basque is translated to Spanish and Spanish is translated to Basque.
The chair of the Basque Parliament accepted to add the transcriptions to the ParlaMint: Towards Comparable Parliamentary Corpora project (CLARIN-ERIC) (Erjavec et al., 2021).
The aim of this paper is to describe the two versions of the Basque parliamentary corpus: the Basque version (with original and translated excerpts) and the original bilingual version (as it was stated in the parliament). We decide to compile two versions to promote research in Basque (a low resourced language) and offer NLP (Natural Language Processing) based tools for search. On the other hand, we build the original corpus to analyse language in use.
Other corpora compilations are possible, for example, we could create the Spanish corpus or a parallel Basque-Spanish translation corpus, in order to offer

data for machine translation studies, but this is out of our scope, and we leave this and other works for the future.

2. Related Works

Basque Parliament texts have been used on many occasions for the study of NLP. Mainly, two areas of study are distinguished: i) studies related to voice processing and ii) those related to text processing.

As long as the Basque Parliament offers the transcript, video, and audio of the sessions, the data has been exploited in some speech processing tasks (Bordel et al., 2011; Etchegoyhen et al., 2021; Pérez et al., 2012).

Bordel et al. (2011) present an automatic video subtitling system to subtitle the video recordings of the Plenary Sessions. Authors aligned the audio in Basque and Spanish, searching the minimum edit distance once they converted them to phonetic streams.

Etchegoyhen et al. (2021) present Mintzai-ST, the first publicly available corpus for speech translation in the pair Basque-Spanish. This is a Basque-Spanish parallel corpus compiled with both speech and text data. The corpus collects Session Diaries from 2011 to 2018. In total, the corpus consists of 370 videos (1,146.18 hours) and 217 PDF documents (Session Diaries and 18,625,252 words). The translations are bidirectional (Basque-Spanish and Spanish-Basque) and the corpus could be employed for research purposes.

Pérez et al. (2012) present Euskoparl, a parallel corpus in Spanish and Basque with both text and speech data. This corpus is aligned at sentence level and divided in train and test datasets. The train dataset consists of 741,780 pairs of sentences (22,668,478 words in Spanish and 18,161,805 words in Basque). The test dataset consists of 30,000 pairs of sentences (915,528 words in Spanish and 733,900 words in Basque).

Our work follows the ParlaMint project (Erjavec et al., 2022). The aim of this work is to build European parliamentary data into comparable, interpretable and highly communicative resource. During the first stage of the project (July 2020 – May 2021) corpora from 17 languages and parliaments were complied, analysed and made available (on GitHub) for research powered by CLARIN-ERIC. To mention some of the languages that participated in the first stage, we should mention Danish (Jongejan et al., 2021), Czech (Kopp et al., 2021) and Polish (Ogrodniczuk, 2018) among others.

In the second stage, more languages will be added to this project and, for example, the texts from the Basque parliament. Moreover, CLARIN-ERIC is promoting the use of parliamentary data in the university curricula (Fišer and de Maiti, 2020).

3. Methodology

3.1. Criteria

The aim of this work is to describe the creation of a bilingual Basque Parliament corpus for studies that aim to analyze what was said and how, and also the creation of an entirely Basque corpus. The Basque corpus will be very useful for the community to analyse the content of the Basque Parliament using Basque NLP tools. To do so, we use the excerpts of the original corpus that are in Basque, as well as the translated passages into Basque by the chamber.

3.2. Resources and Steps

To create the corpus, we follow the ParlaMint criteria, to include the Basque Parliament corpus in the project (`https://github.com/clarin-eric/ParlaMint`) which is labelled with ES-PV (Basque Country).

These are the steps in the corpus creation:

- Permission. Obtain permission from the parliament.
- Convert documents from DOC to TEI-XML format.
- Convert documents from TEI-XML to TEI-ParlaMintXML format.
- Check and validate TEI-XML.
- Metadata file. Describe the Basque Parliament, political parties and parliamentarians at the metadata file.
- Add morphosyntactic information with UDPipe analyser.

3.3. Development

i) **Collect the parliamentary data**.
The secretary of the senior lawyer send the transcripts from the Basque Parliament (and their translations, where possible), to include a corpus of Basque in the project ParlaMint. The request (2021/1887) was granted on March 21, 2021. The texts obtained are from the speeches between February, 2015 and February, 2021.

ii) **Create the original version**.
The parliamentary data is divided into several files, and each file corresponds to one parliamentary act. Sometimes, in a day, there is more than one parliamentary act.

In each DOC file, there are two columns. The left column contains the original speech, while the right column translated the original speech to the other official language. The original corpus contains only the text in the left column and we create TEI-XML format document using this text.

iii) **Create the Basque version**.
In this case, we choose the passages written in Basque from the left column (original text) and from the right column (translated text), by means of a script. The script first calculates how likely the paragraphs are to be in Basque or Spanish.

Next, the script takes the paragraphs most likely to be written in Basque from the left column or from the right column, if Spanish text was detected on the left column. After this, using the only text in Basque, we have created another TEI-XML document file of each parliamentary act.

iv) **Metadata file**.
Finally, we have created the metadata file that is valid for both versions of the corpus. The root file is in Basque, Spanish and English and contains information on: the title, the size, the date of creation of the corpus, as well as political parties, parliamentarians, sessions and positions of the Basque Parliament.

4. Corpus Description

	Basque corpus	Bilingual corpus
Basque	7.37	1.98
Spanish		7.35
Unidentified		0.05
Total	**7.37**	**9.38**

Table 1: Estimation of size of the corpus in words (in millions)

Table 1 shows the characteristics of both versions. The version in Basque has 7,37 million words. In contrast, in the bilingual version, 7.35 million words are in Spanish (%78.4), while 1.98 million words are in Basque (%21.39). Finally, 50 thousand words have not been identified[1] (%0.12). The Basque Parliament corpus is available on GitHub.

4.1. The Metadata File

The metadata file contains information on all aspects related to the parliamentary speeches, which is: the title, authors, project to which it belongs, the size of the corpus, licence, the taxonomy of the participants, organizations, and acts related to parliamentary speeches. Likewise, the data on the legislative periods (3 in total) and a governing body (the Basque Government) are detailed.
Secondly, the political parties are listed (8 in total): EAJ/PNV, EH Bildu, PSE-EE, Elkarrekin Podemos, Ezker Anitza, PP(+Cs), Vox and UPyD. The date on which the political parties were created, their acronym and their Wikipedia web page are also specified in the entry of each political party.
In the third place, parliamentarians are listed. In total, there are 176 parliamentarians. In each entry of the parliamentarian, the following information is detailed: name and two surnames, date and place of birth, gender, their political party affiliation, and their Wikipedia web page (if available).

[1]The unidentified words can be words in other languages, or Basque or Spanish words from short sentences that have been assigned the same probability of being in Spanish or Basque by the script.

Finally, the list of files containing parliamentary sessions is enumerated. Each file corresponds to a parliamentary session.

4.2. Basque and Original Versions

All files in the bilingual version have the same structure. At the beginning, there is a metadata section about the file. Then there is the body of the file. There, each paragraph is segmented. The Basque version maintains the same structure.

5. Hypothesis and Discussion

The characteristics of the corpus may be adequate to analyse some factors related to language and society.

5.1. Sociolinguistic Study

Diglossic situation (if there is a language considered to be of low variety usage and another used for high variety usage) between Basque and Spanish could be analysed.
The amount of words in Spanish and Basque in the corpus already shows that Spanish is used more in the Basque Parliament (which is considered as a high-level institution), which already reflects the diglossic situation in the Basque Parliament.
However, a more exhaustive analysis of this phenomenon can be done using NLP tools. Some interesting research questions arise for future work:

- Which language do parliamentarians speak when they have to say something important?
  - We hypothesize that parliamentarians change their language according to the importance of what they have to say.

    In other words, in our opinion, parliamentarians mainly use Basque when they greet or say something irrelevant, and they use Spanish when they have to say something important or when they have to address the other parliamentarians.

- Do parliamentarians use one language to express objective or narrative facts and another language to express subjectivity or their point of view?
  - We believe that parliamentarians use Basque when they have to say something objective or factual. However, parliamentarians use Spanish to express their opinion or address other parliamentarians. This is also related to the diglossic situation, since it would show that for parliamentarians, Basque is not useful to express opinions.

  In relation to our hypotheses, Gagnon (2006) studied the sociolinguistic situation in the Parliament of Canada with a similar approach.

5.2. Other Possible Studies

The Basque corpus could be useful to answer the following question: Are translated paragraphs in Basque more complex or simpler if we compare with those original language forms in Basque?
The version in Basque may be suitable for the task called "text complexity". That is, taking into account some parameters, we can analyse if the translated texts differ a lot from the original texts in terms of complexity.
It can be assumed that the translated paragraphs are more complex, since they are texts created a posteriori. The syntactic structure can be more complex and the lexicon can be less repetitive, taking into account the characteristics of oral language.
Liu and Afzaal (2021) and Ausloos (2012) studied similar hypothesis with original and translated texts in English.

6. Conclusion and Future Work

In conclusion, we present the transcript and corrected written corpus of the speeches of the Basque Parliament. The corpus consists of two versions. One version is entirely in Basque and it is based on the original and translated texts. The second version is based on original texts and is bilingual, although most of the texts in Spanish. The two versions follow the format established in the ParlaMint project. For this reason, both versions are in XML format.
In this situation, the future works that we propose are the following works:

- To follow the ParlaMint document format since this corpus is in the xml format, but not in the TEI-ParlaMintXML format.
- To tag the paragraphs in both versions. We would like to tag paragraph indicating if the paragraphs are in Basque or Spanish (in the original and bilingual version) or if the paragraphs are original or translations (in the Basque version) in order to have more data for the different studies.
- To carry out the linguistic processing following the guidelines of the ParlaMint project. That is, we would add morphosyntactic information using Universal Dependencies framework and NER in both versions.
- Finally, we would like to study some hypotheses that we have raised in Section 5.

7. Acknowledgements

We would like to thank Kike Fernández (HiTZ Center-EHU/UPV) for his technical help in this work, INTELE network (Ministerio Ciencia, Innovación y Universidades de España RED2018-102797-E), and the ParlaMint project (CLARIN-ERIC) for the financial support.

8. Bibliographical References

Ausloos, M. (2012). Measuring complexity with multifractals in texts translation effects. *Chaos, Solitons & Fractals*, 45(11):1349–1357.

Bordel, G., Nieto, S., Penagarikano, M., Rodriguez-Fuentes, L. J., and Varona, A. (2011). Automatic subtitling of the basque parliament plenary sessions videos. In *Twelfth Annual Conference of the International Speech Communication Association*.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Pančur, A., Ljubešič, N., Agnoloni, T., Barkarson, S., Pérez, M. C., Çöltekin, Ç., Coole, M., et al. (2021). Parlamint: comparable corpora of european parliamentary data. In *Proceedings of CLARIN annual conference 2021, 27-29 September, 2021, virtual edition*, pages 20–25. Utrecht University.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., et al. (2022). The parlamint corpora of parliamentary proceedings. *Language resources and evaluation*, pages 1–34.

Etchegoyhen, T., Arzelus, H., Ugarte, H. G., Alvarez, A., González-Docasal, A., and Fernandez, E. B. (2021). Mintzai-ST: Corpus and baselines for Basque-Spanish speech translation. *Proc. IberSPEECH 2021*, pages 190–194.

Fišer, D. and de Maiti, K. P. (2020). Voices of the parliament. *Modern Languages Open*.

Gagnon, C. (2006). Language plurality as power struggle, or:: Translating politics in canada. *Target. International Journal of Translation Studies*, 18(1):69–90.

Jongejan, B., Hansen, D. H., and Navarretta, C. (2021). Enhancing CLARIN-DK resources while building the Danish ParlaMint corpus. In *CLARIN Annual Conference 2021*, page 73.

Kopp, M., Stankov, V., Kruza, J. O., Straňák, P., and Bojar, O. (2021). ParCzech 3.0: A large Czech speech corpus with rich metadata. In *International Conference on Text, Speech, and Dialogue*, pages 293–304. Springer.

Liu, K. and Afzaal, M. (2021). Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *Plos One*, 16(6):e0253454.

Ogrodniczuk, M. (2018). Polish parliamentary corpus. In *Proceedings of the LREC 2018 workshop ParlaCLARIN: creating and using parliamentary corpora*, pages 15–19.

Pérez, A., Alcaide, J. M., and Torres, M.-I. (2012). EuskoParl: a speech and text Spanish-Basque parallel corpus. In *Thirteenth Annual Conference of the International Speech Communication Association*.