Proceedings of ParlaCLARIN III @ LREC2022 pages 125–130
Marseille, 20 June 2022

ParlamentParla: A Speech Corpus of Catalan Parliamentary Sessions

Baybars Külebi[1,2], Carme Armentano-Oller[1], Carlos Rodríguez-Penagos[1], Marta Villegas[1]
[1]Barcelona Supercomputing Center - Centro Nacional de Supercomputación
[2]Col·lectivaT SCCL
{baybars.kulebi, carme.armentano, carlos.rodriguez1, marta.villegas}@bsc.es

Abstract

Recently, various end-to-end architectures of Automatic Speech Recognition (ASR) are being showcased as an important step towards providing language technologies to all languages instead of a select few such as English. However many languages are still suffering due to the "digital gap", lacking thousands of hours of transcribed speech data openly accessible that is necessary to train modern ASR architectures. Although Catalan already has access to various open speech corpora, these corpora lack diversity and are limited in total volume. In order to address this lack of resources for Catalan language, in this work we present ParlamentParla, a corpus of more than 600 hours of speech from Catalan Parliament sessions. This corpus has already been used in training of state-of-the-art ASR systems, and proof-of-concept text-to-speech (TTS) models. In this work we explain in detail the pipeline that allows the information publicly available on the parliamentary website to be converted to a speech corpus compatible with training of ASR and possibly TTS models.

Keywords: speech corpus, automatic speech recognition, data, found data

1. Introduction

Although Natural Language Processing is fast becoming a mainstream, readily-usable technology, for many of its core tasks it relies on a vast amount of data being available for development and training, especially in the age of neural networks-based Artificial Intelligence. Speech processing, be it recognition or synthesis, require many hours of recorded and transcribed data in order to be used reliably for electronic assistants, translation, voice operated interfaces, etc.

For some of the world's languages with many millions of speakers available, this data gathering is not especially challenging. For other less-resourced languages, it can be a matter of survival in an increasingly digital world.

Catalan is a romance European language spoken or understood by more than 9 million people, with deep roots in culture and history, and with a significant online presence, for example, in Wikipedia [1], popular media outlets and in print literature. Even so, it is not recognized as an official language in the EU, nor is it incorporated in many of the major apps and services from Big Tech (Amazon, Google, Apple, etc.) that people increasingly rely on for communication and even daily chores like using maps or scheduling appointments.

For some years now, a diverse and vibrant collaboration between regional government agencies, volunteer tech collectives and research institutions has promoted and supported projects that have produced state-of-the-art tools and resources, such as: translators, datasets, (Külebi and Öktem, 2018), dictionaries, correctors,[2] corpora,[3] pipelines,[4] massive Transformer-based language models and language benchmarks.[5]

These somewhat scattered efforts and the direct involvement of the Catalan government have led to the AINA initiative,[6] enabling the generation of high-quality corpora and datasets which, along with extensive language models, are being made available through various open platforms[7][8] in order to promote Natural Language Understanding (NLU) capabilities for any institution, organization, company or individual. The objective is for people to be able to engage in the digital world in Catalan to the same degree as speakers of a global language such as English, thus preventing the digital extinction of the language.

In this work, we explain the preparation of a Catalan speech corpus based on the Parliamentary recordings and metadata. The corpus is published with a CC-BY license and is fully downloadable from https://zenodo.org/record/5541827(Külebi, 2021).

2. Data Compilation

In order to contribute to the openly available speech resources for Catalan, we have compiled the parliamentary speeches and processed them with their corresponding transcriptions, as well as segmented them into a format that is compatible with ASR training pipelines.

[1]The Catalan version is the 20th largest language edition
[2]i.e. see the resources provided by the NGO Softcatalà
[3]Catalan has consistently been in the top 5 languages of the Mozilla Common Voice initiative
[4]CLIC, TALP and other university research labs.
[5]Barcelona Supercomputing Center, Text Mining Unit
[6]https://www.projecteaina.cat
[7]https://huggingface.co/projecte-aina
[8]https://github.com/projecte-aina

The use of parliamentary recordings for generating speech corpora is well established, with the earliest example for a limited resource language has been the creation of the Althingi, Icelandic parliamentary speech corpus (Helgadóttir et al., 2017). Parliamentary content has certain advantages, such as readily available transcripts, relatively natural speech and a controlled recording environment. Additionally, due to the transparency laws, it is customary to find parliamentary content with open and/or free licenses, hence facilitating the release of the final processed dataset with open licenses.

The complete process, which takes various types of parliamentary content and converts them into a speech corpus, takes advantage of different types of tools and algorithms. In the following subsections, we first explain the details of the publicly available content, and follow with the specifics of the preprocessing steps applied first to textual data and metadata and later to the audio content.

2.1. Catalan Parliamentary Data

The Catalan Parliament (Parlament de Catalunya) consists of 135 elected representatives from an ideologically diverse group of political affiliations. In the last decade, the Catalan Parliament has witnessed lively and intense debates about topics such as social equality, national identity and statutes, language preservation, etc. In this work, we have profited from this diverse content to create a speech corpus.

In order to build our corpus we have first extracted the available content, directly from the Parliamentary website, taking advantage of an earlier easy to access version of it.[9] The audio segments were extracted from recordings of the Catalan Parliament plenary sessions, and the dates chosen were between 2007/07/11 and 2018/07/17, when the scraping was made.

Within the old version of the website, the video files were presented per plenary session and per speaker intervention. For each plenary session, the sequence of interventions per speakers were accessible in the DOM, in addition to a link to the complete transcripts in pdf format.

In short, the overall preparation of the corpus involves matching metadata from two different sources, namely the website and the transcripts in pdf format. Furthermore, the combined data is processed in order to create the ASR training ready corpus. The visual summary of the whole data processing can be found in Fig. 1. Apart from the matching of the metadata from both sources, the most time-consuming aspect of the data preprocessing has been the development of the pdf parser for the parliamentary transcripts. In addition, the existence of multiple official languages has been an extra inconvenience specific to our case.

[9]the old version of the site can be seen in web archive

2.2. Preparation of Session Metadata

The data processing pipeline starts with the scraping of the webpage of the Catalan Parliament, where a recording of a speaker during an intervention per session are available separately. Since the audiovisual content is speaker specific, the first important step was to associate each video file with the corresponding session, intervention and actual speaker and finally the corresponding text. Although the speaker list was provided in the DOM of the session video page, this information had to be aligned with the parsed pdf of the official transcript of the session, converted into structured data, showing the speaker and the corresponding text.

As a first step we have downloaded the list of interventions, the name of the corresponding speakers and the pdf of the full transcription per session, and the corresponding video files for the speaker. Furthermore we have repeated the whole process for each plenary session. As explained before, the intervention recordings are conveniently organized as per speaker per intervention, hence for the given time window we downloaded in total 12918 recordings, with lengths ranging between 10 seconds and up to 30 minutes. The pdfs or "diaris" in Catalan, include the interventions for each topic discussed during one specific day of a plenary session. Due to the content of the sessions, we have concentrated only on the regular sessions and not the extraordinary ones which might include constitutive sessions, with only the voting processes that are contentwise uninteresting for a speech corpus.

In addition to the structured metadata of the audiovisual content page, the pdf files also needed to be converted to structured data, comprising the sequential speakers with their corresponding scripts. For this task we used a two-step process, in which we first extracted the xml information from the pdf, where each line of text have their own coordinates and typographical information. And on a second step we implemented the template logic into a parsing script where the undesired parts of the text, like headers, footers and lists of contents were eliminated, and exploiting the typographical information, text versus speaker name were recognized. Since the structure of the official scripts changed only once during the chosen period, our parser needed to consider only two alternatives.

At the end of this process, we ended up with a structured data file which included the sequence of speakers with their associated speeches within a session.


```
{
  "_id" : "0fa8ea289797a5c40a9106",
  "value" : {
  "urls" : [
  [["Sra. Eulàlia Reguant i Cura
     (Membre)"],
    "4a20b06b2748a0060b3b.mp3"]
  ],
  "text" : [
  ["La presidenta",
```

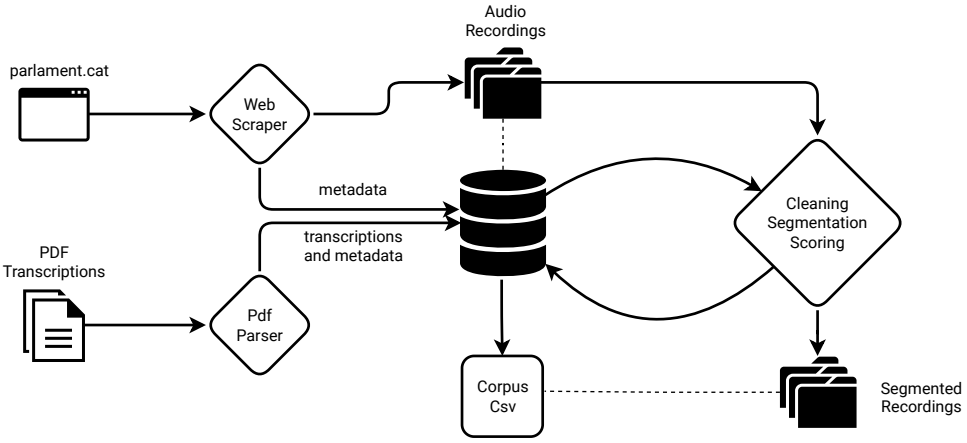

Figure 1: The flowchart of all the processes to generate the speech corpus. All the information persisted in a database, and as a final step a csv file is exported with the relative paths to the audio segments and their corresponding transcriptions.

```
    "Gràcies, president. Té la paraula
     la senyora Reguant."],
   ["Eulàlia Reguant i Cura",
    "Gràcies, president. Arrelada a
     Parets, amb el setanta-cinc per
     cent de la seu financera a
     irlanda:(...)"]
   ],
   "ple_code" : "2017_04_26_212616"
   }
}
```

Finally, to generate the input for the audio-transcript alignment process, we needed to merge this metadata coming from two separate sources, i.e. a sequence of speaker interventions with the corresponding video urls, and another sequence of speaker interventions with the corresponding text. However, the names of the speakers within these two different sources were not consistent: for example *H. Sr. Oriol Junqueras i Vies (Conseller)* vs *El vicepresident del Govern i conseller d'Economia i Hisenda (Oriol Junqueras i Vies)* from the pdf. Or in extreme cases, the names of the ministers did not appear corresponding to their appointment within the transcriptions, for example *H. Sr. Lluís Miquel Recoder i Miralles (Conseller)* vs *El conseller de Territori i Sostenibilitat*.

Furthermore, the audiovisual metadata did not include the minor interventions, neither from the President of the Parliament nor from the interrupting parliamentary members, whereas the data extracted from pdfs did include them. Hence, we used yet another step to align the speakers sources, using first a fuzzy match for the names within an intervention, and assigning them an index, and later sequence matching these indices using the Smith-Waterman Algorithm (Smith et al., 1981).

At the end of this process we ended up with a single database of metadata which include the session, its speaker intervention in the correct sequence with their corresponding text and video urls. The scripts implementing all these processes can be found at the repository of the project. [10]

2.3. Preparation of Audio Corpus

The metadata file we have prepared provided us with the audio file against its corresponding text, but the recording for each intervention had various lengths, up to 20 minutes, which needed to be segmented to 5-15 second clips in order to be applicable for ASR training pipelines. Thus to segment these long audio files into desired sizes, we executed a forced alignment process.

We initiated the process downloading all the content in video format and converting them into single channel wav files with a sampling rate of 16kHz. Furthermore, we processed the merged structured metadata file to eliminate all the non-Catalan speeches through the use of a basic language detector,[11] which gives a percentage of Catalan words for an intervention, based on a clean corpus. Finally, for the remaining set we applied a method very similar to the original LibriSpeech article (Panayotov et al., 2015).

We did a first pass of speech recognition using the Catalan CMU Sphinx models [12] trained with TV3, the public Catalan television corpus (Külebi and Öktem, 2018), with the language generated from the self-text of the intervention. This method gave us the word based timestamps for the text; however the resulting word se-

[10] https://github.com/gullabi/parlament-scrape

[11] The Catalan parliamentary discussions include Spanish as well as Aranese Occitan interventions, all three being the official languages of the autonomic region.

[12] https://cmusphinx.github.io/

quence did not correspond fully to the input text, partly due to the old ASR architecture and partly due to the non-literal transcription of the official records (omission of repetitions and disfluencies). However, we used the results of the ASR decoding, aligned them to the original text using the Smith-Waterman algorithm and used this information to define the text/audio "islands" (Panayotov et al., 2015) in order to segment the audio. For the segmentation, we have taken into account both the silences (minimum 300 ms) and the punctuation. We have used an algorithm similar to beam search, but with assigned probabilities, to find the most optimal segment. This way the algorithm prefer silences which coincide with punctuations (specifically comma, dot, question mark, colon, semicolon and exclamation mark).

The algorithm accepts minimum and maximum durations, but since it is probabilistic, in situations where it is not possible to segment the long audio (due to lack of silences for example) it allows for segmenting pieces that are longer or shorter than the given limits. This is preferable in order to ensure the quality of the resulting segments, and not introduce truncated speech in the corpus. The lengths of the resulting segments can be seen in Fig. 2. In the end, we have manually eliminated the segments longer and shorter than the desired limits, and ended up with corpora of bits between 5 and 15 seconds. The processing pipeline is available on github. [13]

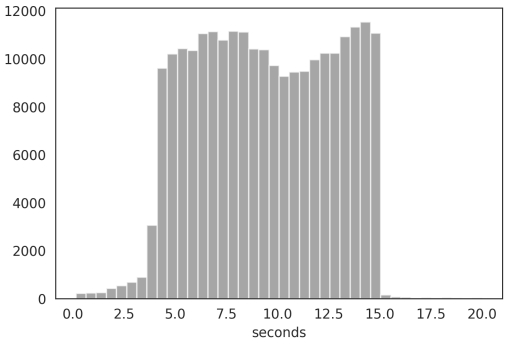

Figure 2: The histogram of segment lengths, the durations are in seconds. Although the duration interval is between 5 and 15 seconds, our method yields some segments with durations outside this interval in order not to generate segments with truncated utterances.

On a second pass, we have further processed the resulting segments in order to assess their quality. Namely, we applied once more the ASR models to the segmented audio and depending on the similarity of the results of the decoding versus the original text, we defined the quality thresholds for the corpus. The quality results were calculated according to Levenshtein distances with 100, complete equivalence of the two strings, and 0, no character overlap. Although we have ensured that our false negatives (i.e. segments with low score) are kept to a minimum through the use of finite-state-transducer based grammars built with the original text of the segment, similarly to Librispeech process (Panayotov et al., 2015), our scoring method was still prone to the quality of the ASR models. Specifically the fact that the ASR models that we used are biased against the non-central Catalan dialect, it is possible that these accents are also penalized by the scoring. However at this moment we don't have the dialectal metadata of the speakers to check this possibility.

In short, in order to eliminate the outliers, possible errors in transcription and/or segmentation pricess we have eliminated all segments below the score of 65 for the global corpus, and ended up with a total of 611 hours of total speech recordings.

2.4. Postprocessing ParlamentParla

For the final published speech corpus, we further processed the data in order to make it compatible with ASR pipelines. First, the text is all in lower-case, which is standard for the speech corpora. Furthermore, for the v2.0 we have included the speaker information, specifically anonymized ids and the corresponding gender. For detecting the genders, we have simply used the information provided in the metadata, using the gendered honorific titles in front of the name of the speaker. Specifically; Sra. (senyora) signifying female, and Sr. (senyor) signifying male.

Finally, using the quality values per segment, we have separated the corpus into clean and other. We have chosen all the segments above the quality score of 91 as clean and left the rest as other. This way we ended up with 211 hours of clean and 400 hours of other quality segments. Furthermore each subset was divided into three parts, training, dev and test, where dev and test datasets have 4 hours each and the rest goes to the training corpus. We have made sure that the speakers that are included in the test and dev subsets are not included in the training subset.

The final gender distribution of the corpus ended up being male dominant with female voice percentages of 28,7% for "other" and 39,3% "clean" subset. In total, the female voices comprise 32,4% percent of the total duration of ParlamentParla. The details of the total durations per subset can be seen in Table 1.

For the v2.0 of the corpus we have used a format similar to the Common Voice dataset since it became a *de facto* standard for most of the modern ASR training architectures during the recent years. Hence we released the dataset in a single file csv (comma separated values), with the speaker ids, the audio filename and the corresponding text, as well as the speaker gender and the duration of the utterance.

The corpus is currently available through the Zenodo

[13] https://github.com/gullabi/long-audio-aligner

Subcorpus	Gender	Duration (h)
other_test	F	2.516
other_dev	F	2.701
other_train	F	109.68
other_test	M	2.631
other_dev	M	2.513
other_train	M	280.196
other total		400.239
clean_test	F	2.707
clean_dev	F	2.576
clean_train	F	77.905
clean_test	M	2.516
clean_dev	M	2.614
clean_train	M	123.162
clean total		211.48
Total		611.719

Table 1: The total duration of all the subsets, with gender distribution.

platform (Külebi, 2021) and Huggingface datasets,[14] with the latter allowing for a convenient interface for inspection of the published data.

3. Current and Future Work

The dataset was already used to train state-of-the-art ASR models, fine-tuning the pretrained multilingual models of wav2vec2.0 with 300m[15], 1b[16] parameters. For the training, ParlamentParla was used in addition to TV3Parla and the Common Voice Catalan v8.0 dataset. The achieved WER (word-error-rate) for these models on the Common Voice test set is 6,8% and 6,1%. Although these models do not solely rely on ParlamentParla, these wav2vec2.0 results are a good showcase of the importance of the corpus in achieving high quality ASR systems for resource-limited languages. We are currently making experiments training wav2vec2.0 models with Common Voice corpus vs Common Voice plus ParlamentParla corpus, and our results will be published in the Huggingface, within the AINA project[17] models.

Due to the innovative segmentation method introduced in the process, and the clear delineation of individual speakers, it is possible to use the corpus for training of text-to-speech (TTS) systems. In order to test this hypothesis, we have used 15 hours of speech from former Catalan President Artur Mas to train the NVIDIA implementation [18] of Tacotron2 (Shen et al., 2018). For privacy considerations, we did not publish the trained model, but synthesized speech snippets can be found on the webpage of the Catalan TTS project Catotron v1.0 [19]. The details of the architecture and part of the experiment are also explained in the original Catotron paper (Külebi et al., 2020).

The most important future work underway is the development of a data pipeline which the ParlamentParla speech corpus can be easily updated and published regularly. Currently, the biggest obstacle is the new structure of the website in which the assets are loaded dynamically; moreover, the connection between the session videos and the official transcripts is broken. However, there is development underway by the Departament d'Informàtica i Telecomunicacions (Department of Informatics and Telecommunications) to provide the session and intervention information via a RESTful API (application programming interface) as part of the improvement of the transparency standards of the Catalan Parliament. Thanks to this effort, the necessary information will be available by a simple API call without having to scrape the website.

Additionally, there is work underway to apply the methodology developed for ParlamentParla to process other parliamentary recordings, most importantly the Valencian Parliament (Corts Valencianes) and possibly also the Parliament of the Balearic Islands (Parlament de les Illes Balears), both Catalan-speaking territories.

3.1. Acknowledgments

The initial preparation of this corpus was partly supported by the Department of Culture of the Catalan autonomous government, and further development with the preparation of v2.0 was supported by the Barcelona Supercomputing Center, within the framework of the project AINA funded by the Generalitat de Catalunya, Departament de la Vicepresidència i de Polítiques Digitals i Territori, through the projects ECIA PDAD14/20/00001, AINA PDAD46/21/000001

4. Bibliographical References

Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Guðnason, J. (2017). Building an asr corpus using althingi's parliamentary speeches. In *INTERSPEECH*, pages 2163–2167.

Külebi, B. and Öktem, A. (2018). Building an Open Source Automatic Speech Recognition System for Catalan. In *Proc. IberSPEECH 2018*, pages 25–29.

Külebi, B., Öktem, A., Peiró-Lilja, A., Pascual, S., and Farrús, M. (2020). CATOTRON — A Neural Text-to-Speech System in Catalan. In *Proc. Interspeech 2020*, pages 490–491.

Külebi, B. (2021). ParlamentParla - Speech corpus of Catalan Parliamentary sessions, doi:10.5281/zenodo.5541827, October.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international*

[14] https://huggingface.co/datasets/projecte-aina/parlament_parla

[15] https://huggingface.co/PereLluis13/wav2vec2-xls-r-300m-ca-lm

[16] https://huggingface.co/PereLluis13/wav2vec2-xls-r-1b-ca-lm

[17] https://huggingface.co/projecte-aina

[18] https://github.com/NVIDIA/tacotron2

[19] https://collectivat.cat/catotron

conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Smith, T. F., Waterman, M. S., et al. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197.