# The MAPLE Parliamentary Data Set

Full-text and annotated corpora of parliamentary speeches in the legislatures of six European democracies.

**YANI KARTALIS**
Institute of Social Sciences, University of Lisbon
eikartalis@campus.ul.pt

**MARINA COSTA LOBO**
Institute of Social Sciences, University of Lisbon
marina.costalobo@ics.ulisboa.pt

11 December, 2020

## Project Goals and Legislatures Selected

This dataset was created within the context of the MAPLE ERC (Consolidator Grant #682125) and aims at fulfilling its research goals. Namely, it aims at answering a bigger descriptive question with regards to the evolution of the politicization (salience and contestation) of the European Union as an issue in the public spheres of European countries before, during and after the recent Eurozone crisis. Not just if said politicization had occurred, but also how and at what pace, and what were the consequences it may have had on citizens' vote calculus in national elections[1].

To do that, the project collected data in six European countries: Belgium, Germany, Greece, Ireland, Portugal and Spain. For each country, data were collected from three arenas: The media, legislatures but also citizens' attitudes. This document describes the dataset collected regarding the parliaments of these countries. It is complementary to previous parliamentary text data-building efforts like the ParlSpeech datasets (V1:[3] and V2:[4]). In the following text, the data sources, variables included and a short description/summary of the data will be presented.

## Data Collection/Sources

Having selected the relevant countries along the original project research goals, data collection focussed on the national parliamentary chamber with most competences in legislation and government formation. As a result, the MAPLE parliamentary corpus covers the Belgian *Chambre de Representants* (or *Kamer van Volksvertegenwoordigers* in Flemish), the Greek *Hellenic Parliament (Vouli ton Ellinon)*, the German *Bundestag*, the Spanish *Congreso de los Diputados*, the Portuguese *Assembleia da República* and the Irish *Dáil Éireann*.

For each of these parliaments we were aiming at getting about 20 years of parliamentary plenary debates, covering the pre, during and post-crisis times roughly covering the entire 2000s. The majority of the data comes from transcripts of the debates which are readily available on each parliament's website and are mostly in a machine-readable (.docx or .pdf or even .xml in the Irish case) format for the period under observation. Table 1, provides information with links to the sources used to get the raw data on plenary sessions. These online sources were scraped with scripts (built in R and in Python depending on the case) customized to the structures and formats of the respective database. For the cases where there were no machine-readable texts, an extra step of using Optical Character Recognition (OCR) software had to be applied to extract the textual information. We relied on Finereader 14 by ABBYY.

The entire German corpus and some parts of the Spanish (2000 - 2018) corpus were covered by The ParlSpeech datasets and the efforts by Rauh and Schwalbach (2020)[4], while the entire corpus on Portugal was taken from a combination of De Giorgi & Dias (2019)[2] and Almeida,

[1] More info at: https://maple.ics.ulisboa.pt

Marques-Pita & Gonçalves-Sá (2020)[1]. For the rest, a set of scripts (customized to the specific country case) were applied that split the debate transcripts into individual speeches by identifying each utterance from a given speaker. These scripts also annotated the resulting speech observations with the speaker's name, the date of the debate, the parliamentary group (party) membership and the specific parliamentary role and agenda topic if the information was provided. A grand total of 2.501.833 speeches is included in the entire MAPLE parliamentary corpus

| Country | Parliament | Full Name | Legislature Type | URL | Scrape Date |
|---|---|---|---|---|---|
| Belgium | Chambre/ Kamer | Chambre des représentants/ Kamer van Volksvertegenwo-ordigers | Lower House | https://www.lachambre.be/kvvcr/index.cfm | 20/11/2020 |
| Germany | Bundestag | Bundestag | Lower House | https://www.bundestag.de/en/documents/minutes_neu | Rauh & Schwalbach (2020) |
| Spain | Congreso | Congreso de los Diputados of the Cortes Generales | Lower House | https://www.congreso.es/en/web/guest/datos-abiertos | Rauh & Schwalbach (2020) + 30/10/2020 |
| Greece | Vouli | Hellenic Parliament - Vouli ton Ellinon | Unicameral | https://www.hellenicparliament.gr/Praktika/Synedriaseis-Olomeleias | 10/10/2019 |
| Ireland | Oireachtas | Dáil Éireann of the Houses of the Oireachtas | Lower House | https://www.oireachtas.ie/en/open-data/ | 10/11/2020 |
| Portugal | Parlamento | Assembleia da República | Unicameral | https://www.parlamento.pt/Cidadania/Paginas/DadosAbertos.aspx | De Giorgi & Dias (2019) and Almeida, Marques-Pita & Gonçalves-Sá (2020). |

Table 1: Parliamentary Online Data Sources.

# 1 Variables

All six corpora have an identical column structure with ten (10) variables. We use a similar scheme to Rauh & Schwalbach (2020 [4] in order to facilitate further analyses by researchers.

The variable **Unique_ID** is a string variable that contains a unique id for each speech in the entire MAPLE parliamentary corpus. It was used for internal purposes, in order to match speeches with further analyses but is generally useful so it was kept in these versions of the corpora.

The variable **Speech** is a string variable that stores the raw text of the speech as given in the original plenary transcript. It contains normally rather long strings which also result in rather large files that are hard to render with regular text processors and without the use of some scripting language, but allows researchers as much text-as-data methodological choice as possible.

The variable **Speaker** contains a string vector with the full name (as given in the transcripts) of the person delivering the respective speech/utterance. It is worth mentioning here that researchers wishing to use this variable to match individual level data should consider that speaker names in almost all cases diverge from external lists, e.g. with regards to middle names, father's name, nicknames, initials or titles.

The variable **Date** stores the day of when the speech was delivered in a character vector that has the format: (YYYY-MM-DD).

The variable **Party** contains a string vector with the parliamentary group of the speaker as provided in the original transcript or as matched by external, time-stamped MP lists. In some cases, due to how the data was provided by the transcripts, some cleaning of the party names might be required (e.g. Instances of the same group spelled differently: VLAAMS BLOK, Vlaams Blok, Vl. BLOK). Furthermore, and for very small parts of each dataset, there are missing party values for speakers due to how the text was structured. Identifying the parties of those speakers might require some further manual from the user.

The variable **Role** is a string variable indicating whether the person delivering the speech had some parliamentary role in the specific legislature. Roles were not available in all cases and across years. However, the variable will always contain the information needed to distinguish between a speaker of the legislature and the acting chairperson during that session. When the variable takes the value "Chair" it specifies speeches given by the acting chairperson of the respective parliament or session. These speeches almost always serve for debate organisation only, no information on parliamentary group or even speaker name is normally provided and it is often useful to exclude them from analyses.

The variable **Session** is a string variable that contains the name of the session. It mostly contains numeric representations of a running count of sessions in each legislative period but due

to some exceptions it was kept as a string.

The variable **Period** is a numeric variable that contains the number of the legislative period during which the speech was given.

The variable **Terms** is a numeric variable and contains the overall length of the speech text calculated as the number of continuous chains of symbols separated by whitespaces.

Finally, the variable **Parliament** is a string variable that simply contains an index of the name of the parliament in the specific file.

In the case of the Belgian Parliament only, the variable **Language** is included to distinguish between speeches delivered in French and speeches delivered in Flemish. Given the bilingual nature of the Belgian legislature, the language of the speakers may change within a debate and this variable could be useful to researchers[2]. To do that, we use Google's Compact Language Detector 3 wrapped in the *cld3* package in R, which guesses the language of each string based on what the majority of the words are. The possible values are "fr" for French, "nl" for Dutch/Flemish and "NA" in the cases where the algorithm could not reliably determine the language (i.e. when the speaker used both languages equally or when the speech was too short etc.)

## Corpus Description

Table 2 summarizes the available corpora per country. We hope this data is useful. For any discrepancies or mistakes in the data, please contact the authors.

[2]The same is not the case in the Irish Dáil Éireann where although Gaelic is the first official language and English is recognised as the second official language, only some MPs are fluent in the language and use it frequently and most only introduce their speeches in Irish. The application of the same algorithm to determine the language, as in the Belgian data, returns around 90% of the speeches in English.

| Parliament | File Size (MB) | Time Frame | Legislatures/Periods | Total Speeches | Av. Speeches per Session | Av. Terms per Speech (w/out Chair) | Unique Terms/Tokens in Corpus* | Unique Speakers |
|---|---|---|---|---|---|---|---|---|
| BE Chambre | 72 | 1999 - 2019 | 50, 51, 52, 53, 54 | 138136 | 111 | 257 (281) | 240163 | 876 |
| DE Bundestag | 168 | 1998 - 2017 | 14, 15, 16, 17, 18 | 242841 | 208 | 325 (529) | 577132 | 1904 |
| ES Congreso | 100 | 2000 - 2019 | 7, 8, 9, 10, 11, 12 | 199610 | 157 | 257 (568) | 203506 | 1973 |
| GR Vouli | 370 | 2000 - 2019 | 10, 11, 12, 13, 14, 15, 16, 17 | 883790 | 260 | 155 (201) | 543569 | 3422 |
| IE Dáil | 144 | 1997 - 2019 | 28, 29, 30, 31, 32 | 469639 | 216 | 161 (193) | 129588 | 426 |
| PT Parlamento | 109 | 1999 - 2019 | 8, 9, 10, 11, 12, 13 | 607719 | 773 | 84 (94) | 320496 | 3853 |

*** Transforming to lowercase and removing numbers and punctuation** .

Table 2: Corpus Description.

# References


[1] Paulo Almeida, Manuel Marques-Pita, and Joana Gonçalves-Sá. "PTPARL-D: Annotated Corpus of 44 years of Portuguese Parliament debates". In: *arXiv preprint arXiv:2004.12502* (2020).

[2] Elisabetta De Giorgi and AL Dias. "Portuguese Observatory on Parliamentary Dynamics Database (POPaD): Information on Legislative Process, Scrutiny Activity and Speeches in the Portuguese Parliament.(Development version)[Datafile and Codebook]". In: *Development version)[Datafile and Codebook]* (2019).

[3] Christian Rauh, Pieter De Wilde, and Jan Schwalbach. "The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states". In: (2017).

[4] Christian Rauh and Jan Schwalbach. "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies". In: (2020).