ParlEE plenary speeches data set

Annotated full-text of 21.6 million sentence-level plenary speeches of eight EU states

Christine Sylvester, University of Strathclyde
sylvesterca@gmail.com

Zachary Greene, University of Strathclyde
zacgreene@gmail.com
http://www.zacgreene.com/

Benedikt Ebing, University of Mannheim

Sept 9, 2022

Citations:

Substantial time and effort was involved in the collection and preparation of this dataset and as such, we would be most grateful if those using the data in parts, in full or quoting this note could please cite the resources as:

Sylvester, Christine; Greene, Zachary; Ebing, Benedikt, 2022, "ParlEE plenary speeches data set: Annotated full-text of 21.6 million sentence-level plenary speeches of eight EU states", https://doi.org/10.7910/DVN/ZY3RV7, Harvard Dataverse, V1

Acknowledgments:

We are grateful to our ever-diligent research assistants Minh Duc Bui and Marlena Kuc for their help with the crawling of raw text from Belgium, France, Hungary, Ireland and Portugal as well as portions from Austria, Germany, and the UK. We wish to thank Inês Carneiro, Clémence Deswert, Raphael Dugmanits, Lindsey Flagler, Bence Hamrak, Dániel Kovarek, Inês Renda, Thomas Schober, Joana Segonds, and Lauren Toner for their assistance in the validation of our two tiers of classification. This effort was greatly aided in its early stages by the excellent raw speech data provided by Rauh et al. in their ParlSpeech V1 data set. We especially wish to thank our project partners at the University of Würzburg, previously University of Mannheim, Goran Glavaš and Fabian Schmidt for their help problem solving when complications and roadblocks arose. Finally, this effort was assisted by generous funding support from the Norface Governance “Democratic governance in a turbulent age” program as part of the EUINACTION project.

Description:

ParlEE Plenary Speeches contains the full-text speeches from eight legislative chambers for Austria, Belgium, France, Germany, Hungary, Ireland, Portugal and the United Kingdom, covering 2009-2019. This data set provides the plenary speeches split to the sentence-level annotated with date, speaker, party, EU vs. Domestic politics classification, and relevant policy area (using the Comparative Agendas Project coding scheme).

Data Collection:

The initial countries included in this data set provide variation on a number of dimensions of interest to the study of European politics. These eight countries (Austria, Belgium, France, Germany, Hungary, Ireland, Portugal, and the UK) vary by geographic locale, domestic institutional structures, and their positions within broader European Union (EU) governance. Take for example electoral system – these eight represent a mix of first-past-the-post (UK), second ballot majority runoff (FR), open and closed list proportional representation (AT, BE, PT), mixed member (DE, HU), and single transferable vote (IE). These eight represent a mix of donor or recipient states for EU resources. They provide a varied mix of countries for those interested in political behaviour, party politics, and/or representation. Subsequent countries will be added in due course.

For each of these countries we selected the body most focused on domestic governance and legislation. This final list comprised the following eight parliaments: the Austrian 'Nationalrat', the Belgian 'Chambre des Représentants', the French 'Assemblée Nationale', the German 'Bundestag', the Hungarian 'Országgyűlés', the Portuguese 'Assembleia da República', and the United Kingdom 'House of Commons'.

We began by identifying those countries from our selection where existing raw speech data had already been compiled and published. The Rauh et al. (2017, 2020) ParlSpeech data sets provided an invaluable collection of raw speech data for several of our countries of interest, specifically portions from Austria, Germany, and the UK. While these existing data sets contained the bulk of plenary session speeches for our period of interest (2009-2019), we extended each of these three by scraping from official parliamentary sources where the existing data sources did not reach to the end of our time window of interest in 2019. For the remaining five parliament's we scraped plenary debates from the official parliamentary repositories using customized scripts.[1] Each parliamentary repository required a unique approach to scraping speeches. While some such as Ireland provide an API to ease the collection of debate and member data not all provide easy to access documentation.

[1] For Austria, we scraped speeches from 14 December 2018 through the end of 2019 from https://www.parlament.gv.at/PAKT/PLENAR/#. For Belgium, the official parliamentary record provides dual translations of all speeches in both Flemish and French. To simplify the data collection effort we elected to scrape the French translations of the speech text. We scraped speeches from January 2009 through December 2019 from https://www.dekamer.be/kvvcr/showpage.cfm?section=/cricra&language=nl&cfm=dcricra.cfm?type=plen&cricra=cra&count=all&legislat=51. For France, we scraped speeches from January 2009 through December 2019 from http://www.assemblee-nationale.fr/15/debats/. For Germany, we scraped speeches from 15 December 2018 through the end of 2019 from https://pdok.bundestag.de/index.php?start=pp. For Hungary, we scraped speeches from January 2009 through December 2019 from https://www.parlament.hu/orszaggyulesi-naplo-elozo-ciklusbeli-adatai. For Ireland, we scraped speeches from January 2009 through December 2019 from https://www.oireachtas.ie/en/debates/find/. For Portugal, we scraped speeches from January 2009 through December 2019 from https://www.parlamento.pt/ActividadeParlamentar/Paginas/IntervencoesPlenario.aspx. For the United Kingdom, we scraped speeches from 18 December 2019 through the end of 2019 from https://hansard.parliament.uk/commons.

For Austria, an open data portal was available, but the actual content like stenographic protocols was not directly provided and instead they provide XMLs that link to the stenographic protocols. The stenographic protocols were not provided in the English-language interface and where provided they were available in HTML with limited HTML-attributes provided. For Belgium, while an open data API was offered retrieval of stenographic protocols was not possible and no documentation for the API was available. In this case, we resorted to accessing PDFs formatted with two-columns to reflect the two languages that all debates are simultaneously translated into. Limiting our efforts to a single column within the PDFs left us with a substantial cleaning effort to identify and efficiently remove large portions of boilerplate text. For France, the parliamentary interface allows for bulk downloads for the most recent legislative sessions in XML-format. Data of earlier sessions was only available in HTML and the format of those HTMLs changed over time (from 2009-2019). The HTML structure of documents for earlier sessions was generally helpful (making use of classes to distinguish different types of texts and providing ids to link politicians to meta-data). Politicians and their roles were weakly-structured in the session HTMLs with no information about the political parties of politicians listed in the session protocols. Documentation of the XML-schema was provided in French only, with the no stenographic protocols shared on the English version of the interface. For Germany, stenographic protocols were provided in XML-format and allowed for bulk downloading, but while documentation of the schema was provided it was only accessible in German and through the German-language interface. For Hungary, no API or open data portal was provided that included the relevant speech and member information. The parliamentary website blocked access to data crawlers and provided no English version of the stenographic protocols. Data was weakly structured HTML-format, functionally a list of paragraphs, and did not include HTML-attributes to indicate useful meta-data like speaker names, parliamentary roles, or political parties. For Ireland, while they did offer an API for debates or members of parliament, and debates were available in xml-format, no xml-schema was provided. For Portugal, no API or open data portal was provided that included the relevant speech and member information and though Portuguese stenographic protocols were shared this information was in either plain text or as a PDF. For the United Kingdom, an API was provided and given the breath of existing data sources we were only forced to scrape a few days of speeches.

Once collected and compiled into a single dataset per country we transformed the speeches to the sentence-level incorporating additional variables on sentence ordering as needed to ensure statement or speech continuity was clear. Finally, we set about the task of cleaning and correcting the raw speech text and additional corresponding variables to remove errors or inconsistences that are somewhat common when accessing raw speeches across many years from official parliamentary sources. This cleaning process involved removing boilerplate information, page numbers, correcting corrupted special characters, and finally ensuring consistent naming conventions for each party and speaker.[2] The final corpora are saved by parliament with UTF-8 encoding as a comma-separated values (csv) file. Table 1 provides a general overview for the makeup of each parliament's corpora.

[2] See the variable description below for detail on the specific steps taken for each applicable variable as a part of this cleaning process.

Table 1. Overview of ParlEE Plenary Speeches corpora

Parliament	File Name	File Size	Observations	Total Parties	Party Names	Total Speakers
Austria (AT)	ParlEE_AT_plenary_speeches	1.4 GB	1,526,362	10	BZÖ, FPÖ, Grüne, JETZT, LIF, NEOS, other, ÖVP, SPÖ, STRONACH	536
Belgium (BE)	ParlEE_BE_plenary_speeches	307 MB	471,107	14	CD&V, cdH, Ecolo-Groen, FDF, LDD, MLD, MR, N-VA, Open Vld, other, PS, PVDA-PTB, sp.a, Vlaams Belang	428
France (FR)	ParlEE_FR_plenary_speeches	1.8 GB	4,672,476	14	ECOLO/EELV/PE/LV, FDG, FI, G.s/PS, GE, LAREM, LC/NC, MODEM, MR, other, PCF, PRG/RRDP, LR/UMP/R-UMP, UDI/UDI-A-I/UDI-AGIR/UDI-I	4185
Germany (DE)	ParlEE_DE_plenary_speeches	1.1 GB	2,866,035	7	Afd, CDU/CSU[3], FDP, Gruene, other, PDS/LINKE, SPD	2014
Hungary (HU)	ParlEE_HU_plenary_speeches	378 MB	1,769,307	12	DK, Együtt, Fidesz, független, Jobbik, KDNP, LMP, MDF, MSZP, other, Párbeszéd, SZDSZ	810
Ireland (IE)	ParlEE_IE_plenary_speeches	1.2 GB	3,988,025	13	Anti-Austerity Alliance, Fianna Fáil, Fine Gael, Green Party, Labour Party, other, People Before Profit Alliance, Progressive Democrats, Sinn Féin, Social Democrats, Socialist Party, Solidarity, Workers and Unemployed Action	342
Portugal (PT)	ParlEE_PT_plenary_speeches	281 MB	1,278,005	11	BE, CDS-PP, CH, IL, L, other, PAN, PCP, PEV, PS, PSD	888
United Kingdom (UK)	ParlEE_UK_plenary_speeches	1.5 GB	6,767,026	12	APNI, Conservative Party, DUP, GPEW, Labour Party, LibDem, other, Plaid Cymru, SDLP, SNP, UKIP, UUP	1172

[3] While the cmp_party variable combines the CDU and CSU as one party, these two can be easily separately if needed using their individual names listed in the party variable.

Classification:

One of our primary interests in the development of this data set was to provide a tool for the investigation of shifts in the substantive or topical components of parliamentary speech by individual members, parties, governments, countries and temporally. We next set about a two-tiered process of classification for these sentence-level speeches. One aim of our research efforts was to explore variation in speech over different levels of governance. Our first classification step divided text into one of two policy dimensions – focusing on distinguish references to EU governance from other, more domestic focused speech. We trained our classifier on existing EU topical coded speech data from the Comparative Agendas Project (CAP) (Baumgartner et al. 2019). The result being a variable for each sentence where '1' references if the text discusses EU governance and '0' if it did not. We recognize this approach may be seen as a substantial threshold for classification; however, believe it was important to ensure those classified as pertaining to EU governance are highly likely to be accurately predicted. The variety of topics then falling into the mirror category of pertaining to domestic governance are those instances that discuss the myriad of issues and internal dimensions the domestic political institutions are responsible for.

A second aim of our research efforts was to explore shifts within specific policy areas. We use the CAP major categories coding scheme to identify 21 distinct policy areas (Baumgartner et al. 2019). To this initial 21 categories we had one additional pertaining specifically to Fisheries. We trained this classifier (*xlm-roberta-base* from the *transformers* package in Wolf et al. 2020) on existing annotation data from CAP.[4] We conducted a validation of these two tiers of classification using native language policy experts, who were asked to conduct a manual validation of a randomly drawn subsample of sentences from all policy areas and across both EU/domestic governance.

This is the first release of these classifications which will be improved upon in later iterations of this data set. Table 2 provides the micro-F1 scores for both classifiers by country. Table 3 provides the micro-F1 scores for all countries by CAP policy area.

Table 2. Micro F1 for EU and CAP classifiers by country

Country	**EU F1 score**	**CAP F1 score**
Austria	69.2	46.2
Belgium	40.8	47.0
France	41.4	42.4
Germany	73.9	47.9
Hungary	79.2	63.8
Ireland	73.3	56.6
Portugal	52.6	61.1
UK	85.7	63.9

[4] The aggregated existing CAP codes data contained a collection of text from Questions to Parliament, Formal Legislative Documents, and Plenary Speeches. This training data also comprised text in five languages: Danish, Dutch, English, Hungarian, and Spanish. We split the sample 80:20 into training and validation. As the distribution of topics discussed by parliament naturally varies by parliament and time we sought to help account for the imbalance of policy area labels by training up *xlm-roberta-base* until convergence of macro-F1 on a manual validation set.

Table 3. Micro F1 for all countries by CAP major category

Major Category	**F1 score**
Macroeconomics (1)	53.7
Civil Rights (2)	37.0
Health (3)	71.8
Agriculture (4)	49.1
Labour (5)	42.4
Education (6)	62.2
Environment (7)	50.4
Energy (8)	74.6
Immigration (9)	48.2
Transportation (10)	69.0
Law & Crime (12)	47.7
Social Welfare (13)	36.6
Housing (14)	55.2
Domestic Commerce (15)	48.9
Defence (16)	51.7
Technology (17)	45.8
Foreign Trade (18)	22.3
International Affairs (19)	42.5
Government Operations (20)	59.7
Public Lands (21)	34.6
Culture (23)	38.8
Fisheries	33.3

Variables:

All eight data sets provide a similar structure of raw text and unique identifiers. They contain 12 standard variables with the structure of each detailed below. Further, in some instances, additional variables are incorporated as a result of unique items provided by official parliamentary sources in the automated text scraping process.

A unique identifier for each sentence, the variable *instance_id* is provided to ease the temporal ordering of speeches. The variable *date* notes the day a speech was given and is structured as DD/MM/YYYY. Within each *date* the variable *speechnumber* provides an identification number to group sentences made at one time by one speaker. Within each *speechnumber* the variable *sentencenumber* provides an id for each sentence of a speech. Collectively, the variables *date*, *speechnumber* and *sentencenumber* aid those researchers interested in specific debates. In addition to these two speech and sentence ordering variables, the structuring of raw text in the official parliamentary sources for Belgium, France, Hungary, Ireland, and Portugal provides a paragraph-level identifier. For these five countries, the *speechnumber* variable is comprised of paragraphs – the *parapgraphnumber* variable – which in turn is comprised of sentences listed as the *sentencenumber* variable. Two variables have been included to assist scholars wishing to explore cross-national variation. The country name abbreviation is variable *iso3country,* while the name of the specific legislative body is variable *parliament.*

The variable *speaker* lists the name of the person who made the statement according to official parliamentary sources[5]. Detailed attention has been paid to this variable to eliminate or reduce the frequency of errors and inconsistency in speaker names. While official parliamentary sources commonly use nicknames, variations in spelling, or honorary titles, this data set aims to reduce these inconsistencies by providing a single standardized naming convention for each individual speaker. One exception being ministerial titles where it might be valuable to researchers to differentiate when a speaker is speaking as 'the Minister' versus as an individual member.

The variable *party* states the partisan alignment of each speaker. Temporal variation in party naming has been synchronised where the official parliamentary sources varied in their naming conventions. To further aid scholars seeking to explore party-level factors, the Manifesto Project (Volkens et al. 2021) party identification numbers have been incorporated where applicable as the variable *cmp_party*.

The *text* variable provides the raw original speech as it was given in parliament at the time. This variable splits speeches to the sentence-level and is encoded in UTF-8. As noted in the data collection process steps, this text incorporated as few changes as possible. Data cleaning steps involved correcting special characters that had corrupted along with the removal of page numbers and boilerplate text such as headers or footers. Disaggregated raw text lends itself well to meet the needs of researchers with a variety of methodological interests. This sentence-level data can be aggregated up with relative ease to explore variation over a myriad of measures such as member, party, country, and across time or issue to name just a few. Further, this fine-grained textual structure lends itself particularly well to natural language processing tools that exploit the grammatical structure of raw text.

Finally, the collection and compilation of these datasets involved a two-step classification process to provide greater information on the topics discussed within these legislative bodies. Both classification steps were completed using data from the *text* variable. The first classification, variable *EU*, identifies if a text discusses the European Union – its institutions, competencies, or authority – and is structured as a simple dichotomous variable. The second classification, variable *policyarea,* identifies the specific policy topic of a text. We used the CAP major categories coding scheme as a framework for identifying possible policy topics (Baumgartner et al. 2019). This micro-level topic identification can aid scholarly work seeking to explore issue shifts by member, parties, Governments, countries, or temporally.

The variable *agenda* notes the topic title given to a selection of speeches by official parliamentary sources and is provided for Austria, Belgium, France, Germany, Ireland, and the UK. The parliamentary agenda provides a succinct title or summary for the primary planned topic of the speeches contained within.

The variable *chair* is a dummy used to denote when, according to official parliamentary sources, the speaker was acting as the chairperson or their stand-in. Official parliamentary sources only provided this chair information for Austria, Germany, and the UK. For Portugal, official parliamentary sources provide a *speakerrole* variable that lists the formal ministerial title of the speaker (if they had one). These variables are especially useful to identify speeches focused almost solely on procedural or organisational matters, rather than say substantive policy or position statements.

[5] For Ireland, the parliamentary record of speakers provides a hyperlink to individual member pages. These pages provide details on the parliamentarians' official positions, length of tenure and other relevant biographical information.

References:

Baumgartner, Frank R., Christian Breunig, and Emiliano Grossman (eds), Comparative Policy Agendas: Theory, Tools, Data (Oxford, 2019; online edn, Oxford Academic, 21 Mar. 2019), https://doi.org/10.1093/oso/9780198835332.001.0001.

Rauh, Christian, Pieter De Wilde, Jan Schwalbach, (2017): "The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states", https://doi.org/10.7910/DVN/E4RSP9, Harvard Dataverse, V1.

Rauh, Christian; Jan Schwalbach, (2020): "The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies", https://doi.org/10.7910/DVN/L4OAKN, Harvard Dataverse, V1.

Volkens, Andrea, Tobais Burst, Werner Krause, Pola Lehmann, Theres Matthieß, Sven Regel, Bernhard Weßels, Lisa Zehnter, (2021): The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2021a. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). https://doi.org/10.25522/manifesto.mpds.2021a.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R´emi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush, (2020): Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics pp. 38–45, https://www.aclweb.org/anthology/2020.emnlp-demos.6.