ParlEE plenary speeches V3 data set

Annotated full-text of 10 million sentence-level plenary speeches of six EU legislative chambers

Christine Sylvester, University of Strathclyde
sylvesterca@gmail.com

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

May 5, 2023

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, 2023, "ParlEE plenary speeches V3 data set: Annotated full-text of 10 million sentence-level plenary speeches of six EU legislative chambers", https://doi.org/10.7910/DVN/IETNNB, Harvard Dataverse, V1.

Acknowledgments:

We are grateful to our ever-diligent research assistants Benedikt Ebing, Minh Duc Bui and Marlena Kuc for their help with the crawling of raw text from Cyprus, Italy and portions of Sweden. We would like to thank Saad Obaid for his assistance implementing the classification approach in these new cases. We also wish to thank a collection of research assistants including Alberto de Leon, Alicia Alexandersson, Elisa Frelin, Emily Wallerstrom, Erika Bejerano Moreno, Francesco Bromo, Harry Ioannou, Jakub Andacký, Klara Vrabl, Martin Procházka, Nives Hüll, Paolo Gambacciani, and Raul Villegas Santana for their help with data cleaning and/or assistance in the validation of our two tiers of classification. This effort was aided in its early stages by the excellent raw speech data provided by the Rauh et al. (2020) ParlSpeech V2, Erjavec et al. (2022) ParlaMint 2.1, Pančur et al. (2020) siParl 2.0 data sets. We especially wish to thank our project partner at the University of Würzburg Goran Glavaš for his help problem solving when complications and roadblocks arose. We wish to express our deep gratitude and thanks to our project partners, Nikoleta Yordanova and Anastasia Ershova for their help identifying language expert research assistants for the data validation work along with support and encouragement throughout this extension of ParlEE. 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 V3 contains the full-text speeches from six legislative chambers for Cyprus, Czechia, Italy, Slovenia, Spain, and Sweden, covering 2009-2019, extending the breadth of countries included in the ParlEE Plenary Speech dataset to a total of 20 parliaments. Like ParlEE Plenary Speeches V1 and V2, 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 countries included in this data set are an expansion of the initial ParlEE Plenary V1 and V2, which comprised 14 European legislative bodies and provides further variation on a number of dimensions of interest to the study of European politics. These additional six parliaments (Cyprus, Czechia, Italy, Slovenia, Spain, and Sweden) vary by geographic locale, domestic institutional structures, and their positions within broader European Union (EU) governance. These six 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 six parliaments: the Cypriot 'House of Representatives', the Czechian 'Poslanecká sněmovna', the Italian 'Camera dei deputati', the Slovenian 'Državni zbor', the Spanish 'Congreso de los Diputados', and the Swedish 'Riksdag'.

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 from Czechia, Spain, as well as a large portion of Sweden.[1] The Erjavec et al. (2022) ParlaMint 2.1 data sets provided an additional useful collection of raw speeches for a portion of our time window of interest for Czechia, Slovenia, and Spain.[2] We further expanded the speech data for Slovenia by using Pančur et al. (2020) siParl 2.0 data set.[3] While these existing data sets contained a substantial portion of plenary session speeches for our period of interest (2009-2019), we extended Sweden by scraping official parliamentary sources where the existing data did not reach our full time window of interest. For the remaining parliaments' (Cyprus and Italy), we scraped plenary debates from the official parliamentary repositories using customized scripts.[4]

Each parliamentary repository required a unique approach to scraping speeches. While some national parliaments provide an API to ease the collection of debate and member data not all provided easy to access documentation. For Cyprus, no API or open data portal was provided that included the relevant speech and member information. Multiple parliamentary sessions were concatenated in PDF documents and stenographic protocols were concatenated over a larger period of time. The stenographic protocols contained no information about political parties. Extracting the relevant speech and speaker names require

[1] For Czechia, the ParlSpeech V2 provided data from 2009 through 3 June 2016. For Spain, the ParlSpeech V2 provided data from 2009 through 20 December 2018. For Sweden, the ParlSpeech V2 provided data from 2009 through 21 December 2018.

[2] For Czechia, ParlaMint 2.1 provided data for the remainder of 2016 through the end of 2019. For Slovenia, ParlaMint 2.1 provided data for August 2014 through the end of 2019. For Spain, ParlaMint 2.1 provided data for 2019.

[3] The siParl 2.0 data set provided the remainder of data for Slovenia from 2009 through August 2014.

[4] For Cyprus, we scraped speeches from January 2009 through December 2019 from https://www.parliament.cy. For Italy, we scraped speeches from January 2009 through December 2019 from https://www.camera.it/leg18/221. For Sweden, we scraped speeches for 2019 from https://www.riksdagen.se/sv/dokument-lagar/?doktyp=prot.

the use of Optical Character Recognition to extract text from the PDFs as the encoding was broken. For Italy, while an open data API with a SPARQL endpoint was offered, documentation, small-scale examples, description of their data model, and the stenographic protocols schema was only accessible in Italian. For Sweden, an open data API with extensive documentation in Swedish was provided. We required only a small number of stenographic protocols and used the HTML version. Stenographic protocols were available as HTML or PDF. The HTMLs were a list of paragraphs without further HTML-attributes and speakers, roles and parties were not specifically indicated in the HTML. The stenographic protocols were not available on the English version of the parliamentary website.

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.[5] 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 legislative body's corpora.

[5] 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
Cyprus (CY)	ParlEE_CY_plenary_speeches	256 MB	766,459	12	AKEL, Citizens Alliance, DIKO, DISY, EDEK, ELAM, Environmentalists, EVROKO, Latino Religious Group Representatives, Representatives of Armenian Religious Group, Representatives of Maronite Religious Group, Solidarity	151
Czechia (CZ)	ParlEE_CZ_plenary_speeches	299 MB	1,528,806	12	ANO 2011, CSSD, KDU-CSL,KSCM, ODS, Piráti, SPD, STAN, SZ, TOP 09, Úsvit, VV	670
Italy (IT)	ParlEE_IT_plenary_speeches	1.2 GB	2,463,502	28	AdC, ALA, AP, Api, AzP, CD, CI-SC, CP, FDI, FI, FLI, IDV, IV, LEGA-LN, LEU, M5S, MDP, MISTO, MRN, NCD, Noi Sud, PD, PDL, PID, PLI, PT, SEL-SI, UDC	1658
Slovenia (SI)	ParlEE_SI_plenary_speeches	299 MB	1,408,450	14	DeSUS, DL, IMNS, LDS, Levica, NSi, PS, SD, SDS, SLS, SMC, SNS, ZaAB, Zares	443
Spain (ES)	ParlEE_ES_plenary_speeches	272 MB	1,160,340	36	Amaiur, Anova, Barcelona en Comu, BNG, CatComu, CCa-PNC-NC, CDC, CiU, Compromís, Compromís-Podemos-EUPV, Cs, CUP, DL, EAJ-PNV, EH Bildu, En Comu Podem, Equo, ERC, EUiA, FAC, Geroa Bai, ICV, IU, Izquierda Plural, JxCat-Junts, MÉS, Nafarroa Bai, PDeCAT, PP, PRC, PSOE, UDC, UP, UPN, UPyD, Vox	1042
Sweden (SE)	ParlEE_SE_plenary_speeches	873 MB	3,116,238	11	AfS, C, FP, KD, L, M, MP, S, SD, SPI, V	908

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, and countries over time. 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.[6] We conducted a validation of these two tiers of classification using multiple 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 for these legislative bodies 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 average 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**
Cyprus	73.7	51.8
Czechia	71.4	53.4
Italy	42.9	68.0
Slovenia	83.3	51.6
Spain	47.1	58.0
Sweden	85.7	55.0

[6] 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[7]

Major Category	**F1 score**
Macroeconomics (1)	52.9
Civil Rights (2)	32.4
Health (3)	73.0
Agriculture (4)	64.0
Labour (5)	43.0
Education (6)	56.6
Environment (7)	67.8
Energy (8)	65.5
Immigration (9)	49.2
Transportation (10)	61.4
Law & Crime (12)	39.1
Social Welfare (13)	38.9
Housing (14)	38.2
Domestic Commerce (15)	51.6
Defence (16)	40.5
Technology (17)	54.8
Foreign Trade (18)	66.7
International Affairs (19)	44.6
Government Operations (20)	67.1
Public Lands (21)	46.0
Culture (23)	68.6

Variables:

All six data sets provide a similar structure of raw text and unique identifiers. They contain 13 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 Italy also provides a paragraph-level identifier. For this country, 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. 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

[7] This table reflects the by policy area average for Cyprus, Czechia, Italy, Slovenia, Spain and Sweden.

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 parliamentary procedural roles that ensure the smooth functioning of a parliament, like the Speaker, where it might be valuable to researchers to differentiate these individuals.

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*. Additionally, the Party Facts – Version 2021 (Bederke et al. 2021) party identification numbers have been incorporated where applicable as the variable *partyfacts*.

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 where available and is provided for Czechia, Italy, Spain, and Sweden. 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 Czechia, Spain, and Sweden. The official parliamentary sources for Cyprus, Czechia, Slovenia, and Sweden, 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.

References:

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

Bederke, Paul; Holger Döring, Sven Regel (2021): "Party Facts – Version 2021", https://doi.org/10.7910/DVN/GM8LWQ, Harvard Dataverse, V1.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, et al. (2022): “The ParlaMint corpora of parliamentary proceedings”, Lang Resources & Evaluation, https://doi.org/10.1007/s10579-021-09574-0.

Pančur, Andrej, Tomaž Erjavec, Milan Ojsteršek, M. Šorn, and Neja Blaj Hribar (2020): “Slovenian parliamentary corpus (1990–2018) siParl 2.0.” Retrieved from http://hdl.handle.net/11356/1300.

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 and 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.