ParlEE plenary speeches V4 data set

Annotated full-text of 18 million sentence-level plenary speeches of eight European legislative chambers

Christine Sylvester, University of Strathclyde
sylvesterca@gmail.com
https://christineasylvester.weebly.com/

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

February 12, 2024

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; Khokhlova, Aleksandra; Yordanova, Nikoleta; Greene, Zachary, 2024, "ParlEE plenary speeches V4 data set: Annotated full-text of 18 million sentence-level plenary speeches of eight European legislative chambers", https://doi.org/10.7910/DVN/TLKVWY, Harvard Dataverse.

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 Estonia, Malta and portions of Bulgaria, Denmark, Greece, Latvia, and Lithuania. We would like to thank Saad Obaid for his assistance implementing the classification approach in the new cases. We also wish to thank a collection of research assistants including Carl Gammeltoft, Thor Renner, Elli-Maria Luud, Kerty Sirman, Emily Wallerstrom, Harry Ioannou, Eva Keruze, Trine Zagare, Manvydas Džiaugys, Matthias Portelli, Sasha Caruana Pedersen, Antoni Gawel, Antonina Baczyńska, Patrycja Kowalczyk, Aleksandar Nikolov and Lyuba Nicheva 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, Dritsa and Louridas (2018) Greek Parliament Proceedings 1989-2019, Darǵis et al. (2018) LinkedSaeima and Ogrodniczuk and Nitoń (2020) Polish Parliamentary Corpus 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 partner Anastasia Ershova for her 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 V4 contains the full-text speeches from eight legislative chambers for Bulgaria, Denmark, Estonia, Greece, Latvia, Lithuania, Malta and Poland covering 2009-2019, extending the breadth of countries included in the ParlEE Plenary Speech dataset to a total of 28 parliaments. Like ParlEE Plenary Speeches V1, V2, and V3 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, V2 and V3, which comprised 20 European legislative bodies and provides further variation on a number of dimensions of interest to the study of European politics. These additional eight parliaments (Bulgaria, Denmark, Estonia, Greece, Latvia, Lithuania, Malta and Poland) vary by geographic locale, domestic institutional structures, and their positions within broader European Union (EU) governance. 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.

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 Bulgarian 'Narodno Sabranie', the Danish 'Folketing', the Estonian 'Riigikogu', the Greek 'Elliniko Kinovoulio', the Latvian 'Saeima', the Lithuanian 'Seimas', the Maltese 'House of Representatives', and the Polish 'Sejm'.

We began by identifying those countries from our selection where existing raw speech data had already been compiled and published. The Rauh et al. (2020) ParlSpeech data set provided an invaluable collection of raw speech data for a large portion of Denmark.[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 Bulgarian, Latvia, Lithuania, and Poland.[2] The Dritsa and Louridas (2018) Greek Parliament Proceedings 1989-2019 data set provided a substantial portion of raw speeches for much of our time window of interest for Greece.[3] The Darģis et al. (2018) LinkedSaeima data set provided a substantial portion of raw speeches for much of our time window of interest for Latvia.[4] The Ogrodniczuk and Nitoń (2020) Polish Parliamentary Corpus data set provided the remainder of data for Poland.[5] While these existing data sets contained a substantial portion of plenary session speeches for our period of interest (2009-2019), we extended Bulgaria, Denmark, Greece, and Lithuania by scraping official parliamentary sources where the existing data did not reach our full time window of interest. For the remaining parliaments' (Estonia and Malta), we scraped plenary debates from the official parliamentary repositories using customized scripts.[6]

[1] For Denmark, the ParlSpeech V2 provided data from 2009 through 2018.
[2] For Bulgaria, ParlaMint 2.1 provided data from 28 October 2014 through the end of 2019. For Latvia, ParlaMint 2.1 provided data for 2018 through 2019. For Lithuania, ParlaMint 2.1 provided data from 16 November 2012 through the end of 2019. For Poland, ParlaMint 2.1 provided data from 12 November 2015 through the end of 2019.
[3] For Greece, the Greek Parliament Proceedings 1989-2019 data set provided speeches from 2009 through the end of 2018.
[4] For Latvia, the LinkedSaeima data set provided speeches from 2009 through the end of 2017.
[5] For Poland, the Polish Parliamentary Corpus provided the missing data to complete Poland from 2009 through 11 November 2015.
[6] For Bulgaria, we scraped speeches from January 2009 through 27 October 2014 from https://www.parliament.bg/bg/plenaryst. For Denmark, we scraped speeches for 2019 from https://www.ft.dk/da/dokumenter/dokumentlister/referater?session=20181. For Estonia, we scraped speeches from January 2009 through December 2019 from https://stenogrammid.riigikogu.ee/et?rangeFrom=01.01.2009&rangeTo=31.12.2020&singleDate=&phrase=&type=ALL

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 Bulgaria, no API or open data portal was provided that included the relevant speech and member information. Session data was available in HTML with weak structures, speakers and their meta-data were not included. The stenographic protocols provided no information about the parties of politicians. For Denmark, no API or open data portal was provided that included the relevant speech and member information. Session data was available in HTML with clear structures using different HTML-classes and HTML–attributes that distinguished explanatory text from actual speeches. The stenographic protocols were not available on the English version of the parliamentary website. For Estonia, an open data API with partial English documentation was provided. No documentation of the data model or how entities linked to each other was provided. Stenographic protocols were available in HTML and while the speakers and their meta-data was provided in weak structure, no information on political parties was included in the stenographic protocols. For Greece, no API or open data portal was provided that included the relevant speech and member information. Data was provided in .doc format as plain text. For Lithuania, an open data API that links to the stenographic protocols was provided though no English documentation was provided. Stenographic protocols were available in HTML though they do not use custom attributes or classes to structure the data, and no special indication of speakers and their meta-data was provided. For Malta, no API or open data portal was provided that included the relevant speech and member information. Though stenographic protocols were shared this information was provided in two-column .doc format and did not indicate information on political parties.

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.[7] 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.

. For Greece, we scraped speeches from January 2019 through December 2019 from https://www.hellenicparliament.gr/en/Praktika/Synedriaseis-Olomeleias. For Lithuania, we scraped speeches for 2009 through 15 November 2012 from http://www3.lrs.lt/pls/inter/w5_sale.kad_ses. For Malta, we scraped speeches from 2009 through 2019 from https://parlament.mt/en/10th-leg/plenary-session/?type=committeedocuments.

[7] 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
Bulgaria (BG)	ParlEE_BG_plenary_speeches	817MB	2,306,384	19	АБВ, Атака, ББЦ, БНД, БНС, БСП, ВОЛЯ, ГЕРБ, ДПС, ДСБ, КБ, НДСВ, ОДС, ОП, ПБГ, ПФ, РБ, РЗС, СК	1,145
Denmark (DK)	ParlEE_DK_plenary_speeches	1.3GB	2,642,184	17	ALT, DF, EL, IA, JF, KD, KF, LA, NB, NQ, RV, S, SF, SIU, SP, T, V	469
Estonia (EE)	ParlEE_EE_plenary_speeches	425MB	1,527,141	8	Eesti Keskerakond, Eesti Konservatiivne Rahvaerakond, Eesti Reformierakond, Eesti Vabaerakond, Eestimaa Rahvaliit, Erakond Eestimaa Rohelised, Erakond Isamaa ja Res Publica Liit, Sotsiaaldemokraatlik Erakond	478
Greece (GR)	ParlEE_GR_plenary_speeches	1.3GB	3,886,000	13	DIMAR, DISY, EK, EL, KKE, LAE, LAOS, MERA, New Democracy, PASOK, SYRIZA, To Potami, XA	914
Latvia (LV)	ParlEE_LV_plenary_speeches	163MB	922,298	17	Attīstībai/Par!, Jaunā konservatīvā partija, Kam pieder valsts?, Latvijas Ceļš, Latvijas Krievu savienība, Latvijas Reģionu apvienība, Latvijas Sociālistiskā partija, Latvijas Zaļā partija, Nacionālās apvienības Visu Latvijai!– Tēvzemei un Brīvībai/LNNK, Par Labu Latviju, Pilsoniskā Savienība, Sociāldemokrātiskā partija Saskaņa, Tautas partija, Vienotība, Visu Latvijai!, Zatlera Reformu partija, Zaļo un Zemnieku savienība	345
Lithuania (LT)	ParlEE_LT_plenary_speeches	530MB	2,052,504	12	Darbo partija, Drasos Kelias, Krikščioniu partija, Liberalu ir centro sajunga, Liberalu sajudis, Lietuvos lenku rinkimu akcija, Lietuvos sarašas, Lietuvos socialdemokratu partija, Lietuvos valstieciu ir žaliuju sajunga, Tautos pažangos partija, Tevynes sajunga – Lietuvos krikščionys demokratai, Tvarka ir teisingumas	893
Malta (MT)	ParlEE_MT_plenary_speeches	390MB	1,486,965	3	Partit Demokratiku, Partit Laburista, Partit Nazzjonalista	132
Poland (PL)	ParlEE_PL_plenary_speeches	642MB	3,158,343	35	German Minority, Inicjatywa Polska, KORWIN, Konfederacja, Kongres Nowej Prawicy, Kukiz'15, Lewica, Nowoczesna, PSL Piast, Partia Demokratyczna, Platforma Obywatelska,	1487

					Polska Jest Najwazniejsza, Polska Plus, Polska Razem, Polskie Stronnictwo Ludowe, Porozumienie, Prawica Rzeczypospolitej, Prawo i Sprawiedliwosc, RLN, Razem, Ruch Ludowo-Narodowy, Ruch Narodowy, Ruch Palikota, Socjaldemokrajca Polska, Sojusz Lewicy Demokratycznej, Solidarna Polska, Stronnictwo Demokratyczne, Teraz!, Twój Ruch, Unia Europejskich Demokratów, Unia Polityki Realnej, Unia Pracy, Wiosna, Wolni i Solidarni, Zieloni	

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 distinguishing references to EU governance from other, more domestic-focused speech. We trained a classifier on existing EU topical coded speech data from the Comparative Agendas Project (CAP) (Baumgartner et al. 2019). We describe the specifics of both classifiers below. The classification results in a variable at the sentence level 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 (we constructed both classifiers using *xlm-roberta-base* from the *transformers* package in Wolf et al. 2020) on existing annotation data from CAP.[8] 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**
Bulgaria	50.0	42.2
Denmark	63.4	49.6
Estonia	64.5	48.2
Greece	67.8	61.6
Latvia	32.0	44.8
Lithuania	44.4	53.6
Malta	66.0	55.4
Poland	78.6	49.6

[8] 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[9]

Major Category	**F1 score**
Macroeconomics (1)	47.5
Civil Rights (2)	31.4
Health (3)	60.2
Agriculture (4)	53.6
Labour (5)	35.5
Education (6)	62.0
Environment (7)	47.3
Energy (8)	66.6
Immigration (9)	24.3
Transportation (10)	54.2
Law & Crime (12)	37.6
Social Welfare (13)	41.0
Housing (14)	31.4
Domestic Commerce (15)	43.2
Defence (16)	48.1
Technology (17)	48.5
Foreign Trade (18)	41.1
International Affairs (19)	46.8
Government Operations (20)	61.9
Public Lands (21)	35.5
Culture (23)	52.0

Variables:

All eight 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 Estonia, Lithuania, Malta and Poland also provide a paragraph-level identifier. For these 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.*

[9] This table reflects the by policy area average for Bulgaria, Denmark, Greece, Estonia, Latvia, Lithuania, Malta and Poland.

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 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 recorded 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 Bulgaria, Denmark, Estonia, Lithuania and Malta. The parliamentary agenda provides a succinct title or summary for the primary planned topic of the speeches contained within.

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.

Darģis, Roberts, Ilze Auziņa, Uldis Bojārs, Pēteris Paikens and Artūrs Znotiņš (2018) Annotation of the Corpus of the Saeima with Multilingual Standards. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Dritsa, Konstantina, and Louridas, Panagiotis. (2018) Greek Parliament Proceedings 1989-2019 (2.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.2587904.

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.

Ogrodniczuk, Maciej and Bartłomiej Nitoń (2020) New developments in the Polish Parliamentary Corpus. In Darja Fišer, Maria Eskevich, and Franciska de Jong, editors, Proceedings of the Second ParlaCLARIN Workshop, pages 1–4, Marseille, France. European Language Resources Association (ELRA).

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, Ŕ´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.