ParlEE plenary speeches V2 data set

Annotated full-text of 15.1 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/

Feb 10, 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; Ershova, Anastasia; Khokhlova, Aleksandra; Yordanova, Nikoleta; Greene, Zachary, 2023, "ParlEE plenary speeches V2 data set: Annotated full-text of 15.1 million sentence-level plenary speeches of six EU legislative chambers", https://doi.org/10.7910/DVN/VOPK0E, 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 Romania and Slovakia as well as portions from Finland. We wish to thank Jakub Andacký, Ville Tapani Haapanen, Simon Humer, Juuso Järviniemi, Ivana Malčić, Mara Matcu, Meri Mentula, Martin Procházka, Borna Turalija, and Simón Wiegers for their 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. ParlSpeech V2 and Erjavec et al. ParlaMint 2.1 data sets. We especially wish to thank our project partner at the University of Würzburg, previously University of Mannheim, Goran Glavaš for his 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 V2 contains the full-text speeches from six legislative chambers for Croatia, European Parliament (EP), Finland, Netherlands, Romania, and Slovakia, covering 2009-2019, extending the breath of countries included in the ParlEE Plenary Speech data set to a total of 14 parliaments. Like ParlEE Plenary Speeches V1, 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, which comprised eight European legislative bodies and provides further variation on a number of dimensions of interest to the study of European politics. These additional six parliaments (Croatia, European Parliament (EP), Finland, Netherlands, Romania, and Slovakia) 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 Croatian 'Sabor', the EP, the Finnish 'Eduskunta', the Dutch 'Tweede Kamer', the Romanian 'Camera Deputaților', and the Slovakian 'Národná Rada'.

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 Finland and the Netherlands. 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 Croatia as well as the Netherlands. We further expanded the speech data for Croatia by using Filip Rodik's GitHub repository of Sabor sessions[1]. The Van Aggelen et al. (2016) Talk of Europe data set provided a portion of the relevant EP speeches. While these existing data sets contained a substantial portion of plenary session speeches for our period of interest (2009-2019), we extended both Finland and the EP by scraping official parliamentary sources where the existing data did not reach our full time window of interest. For the remaining parliaments' (Romania and Slovakia), we scraped plenary debates from the official parliamentary repositories using customized scripts.[2] At the time of collection, neither of the official parliamentary repositories provided an API to ease the collection of debate and member data. Each parliamentary repository required a unique approach to scraping speeches.

For Finland, no API or open data portal was provided that included the relevant speech and member information. The parliamentary website provided no English version of the stenographic protocols

[1] This repository contained parliamentary sessions 5 through 9 from https://github.com/rodik/Sabor.
[2] For Finland, we scraped speeches from December 2015 through December 2019 from https://www.eduskunta.fi/FI/search/Sivut/Vaskiresults.aspx. For Romania, we scraped speeches from January 2009 through December 2019 from https://www.cdep.ro/pls/steno/steno.home?idl=2. For Slovakia, we scraped speeches from January 2009 through December 2019 from https://www.nrsr.sk/dl/?lang=en. For the EP, we scraped speeches from July 2017 through December 2019 from https://www.europarl.europa.eu/portal/en.

and all data was provided as PDFs. Further, the parliamentary website used JavaScript for the search to find documents complicating the crawling process. For Romania, no API or open data portal was provided that included the relevant speech and member information. Data was weakly structured HTML-format, without specific attributes or classes to structure their HTML. The stenographic protocols provided no information about political party affiliation for politicians. For Slovakia, no API or open data portal was provided that included the relevant speech and member information. The stenographic protocols were stored in a digital library separate from their respective documents and required crawling/downloading of the respective documents separately. Documents were provided in .docx format as plain text. The stenographic protocols provided no information about political party affiliation for politicians. For the EP, an open data portal was provided that included the relevant speech and member information. Member reference data, such as national party affiliation was further verified with additional MEP identification sources from the official parliamentary repository.

Once collected and compiled into a single dataset per legislative body 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.[3] 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.

[3] 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**
Croatia (HR)	ParlEE_HR_plenary_speeches	755 MB	2,596,857	33	BDSH, BM 365-BM 365 I NZ, BUZ, Change, DC, Demokrati, GLAS, HDS/HRAST/HSP AS/IDS-PGS-RI, HDSSB, HDZ, HGS, HNS, HRID, HSD, HSLS, HSU-SNAGA-NZ i NP/SDP/SNAGA, HSP, HSS, HSU, ID, IDS, Laburisti, MOST, Naprijed Hrvatska!, NLM, NLSP, Novi vai, OraH, PH, Reformisti, SDAH, SDSS, Živi zid	764
Finland (FI)	ParlEE_FI_plenary_speeches	429 MB	2,150,109	15	KD, Kesk, Kok, KP, liik, M11, PS, RKP, SDP, SIN, tl, Uv, Vas, VIHR, Vr	456
Netherlands (NL)	ParlEE_NL_plenary_speeches	798 MB	4,990,179	15	50PLUS, CDA, CU, D66, DENK, FvD, GL, PvdA, PvdD, PVV, SGP, SP, Van Haga, vKA:GKVK, VVD	504
Romania (RO)	ParlEE_RO_plenary_speeches	243 MB	1,122,625	13	ALDE, Grupul parlamentar progresist, Minoritati, PC, PD-L, PMP, PNL, PP-DD, PRO România, PSD, UDMR, UNPR, USR	1,330
Slovakia (SK)	ParlEE_SK_plenary_speeches	407 MB	2,040,571	12	HZDZ, KDH, LSNS, MOST – HÍD, OLaNO, SaS, SDKÚ – DS, SIET, SMER – SD, SMK, SNS, SR	463
European Parliament (EP)	ParlEE_EP_plenary_speeches	831 MB	2,270,429	*Data set compiles and presents both membership in EP political groups as well as national party membership*		2,057

Classification:

One of our primary interests in the development of the ParlEE data sets was to provide a tool for the investigation of shifts in the substantive or topical components of parliamentary speech by individual members, parties, factions, 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 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.[5]

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 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
Croatia	78.9	61.6
Finland	40.8	54.7
Netherlands	80.0	47.8
Romania	71.0	65.6
Slovakia	85.7	64.9

[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.

[5] This validation process was completed for all five national legislative bodies.

Table 3. Micro F1 for all countries by CAP major category[6]

Major Category	**F1 score**
Macroeconomics (1)	48.8
Civil Rights (2)	45.1
Health (3)	68.6
Agriculture (4)	65.2
Labour (5)	36.8
Education (6)	71.9
Environment (7)	52.6
Energy (8)	74.0
Immigration (9)	23.8
Transportation (10)	68.3
Law & Crime (12)	51.5
Social Welfare (13)	41.4
Housing (14)	48.5
Domestic Commerce (15)	42.4
Defence (16)	58.3
Technology (17)	44.0
Foreign Trade (18)	59.4
International Affairs (19)	53.1
Government Operations (20)	66.8
Public Lands (21)	37.9
Culture (23)	40.0
Fisheries	55.6

Variables:

All six data sets provide a similar structure of raw text and unique identifiers. All six contain 10 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. Further, the EP data set provided additional national-level comparative variables given its unique position as a collection of multiple nations' representatives.

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 source for Slovakia provided a paragraph-level identifier. For Slovakia, the *speechnumber* variable is comprised of paragraphs – the *parapgraphnumber* variable – which in turn is comprised of sentences listed as the *sentencenumber* variable. For the national legislative bodies (Croatia, Finland, Netherlands, Romania, and Slovakia), 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.*

[6] This table reflects the by policy area average for Croatia, Finland, Romania, Netherlands and Slovakia.

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 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 the EP and Croatia. The parliamentary agenda provides a succinct title or summary for the primary planned topic of the speeches contained within.

For the EP, we sought to incorporate additional variables to both explore intra-EP variation by country as well as build a bridge to ease comparisons between EP speeches and national parliamentary debates. The EP data set contains an additional six variables. The variable *country* denotes the member state that elected the speaking MEP. The variable *language* identifies the specific language being spoke for each sentence. The variable *mode* was provided by the official parliamentary source to separate spoken versus written statements. Within the EP data set, the *party* variable denotes the EP parliamentary group to which the MEP is a member at the time of the statement. However, the variable *national_party* identifies the MEPs political affiliation within their domestic political party system. Given the variety of political parties within the EP, especially the prevalence of smaller parties not always politically active within domestic parliaments, we incorporated additional party-level variables with the inclusion of *parlgov_id* and *em_party*. The

former denotes the identification number given to the applicable MEP's national party affiliation by Döring et al. in the ParlGov 2022 data set. The variable *em_party* denotes the identification number given to the applicable MEP's national party affiliation by Schmitt et al. in the Euromanifesto Study data set.

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.

Döring, Holger; Huber, Constantin; Manow, Philip, 2022, "ParlGov 2022 Release", https://doi.org/10.7910/DVN/UKILBE, 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.

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.

Schmitt, Hermann, Braun, Daniela, Popa, Sebastian A., Mikhaylov, Slava, & Dwinger, Felix (2016). European Parliament Election Study 2014, Euromanifesto Study. GESIS Data Archive, Cologne. ZA5162 Data file Version 1.0.0, https://doi.org/10.4232/1.5162.

Hollink, Dr Laura, Astrid Van Aggelen, Martijn Beunders, Martijn Kleppe, Max Kemman (2015). "Talk of Europe – The debates of the European Parliament as Linked Open Data." DANS. http://dx.doi.org/10.17026/dans-2xg-umq8.

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.