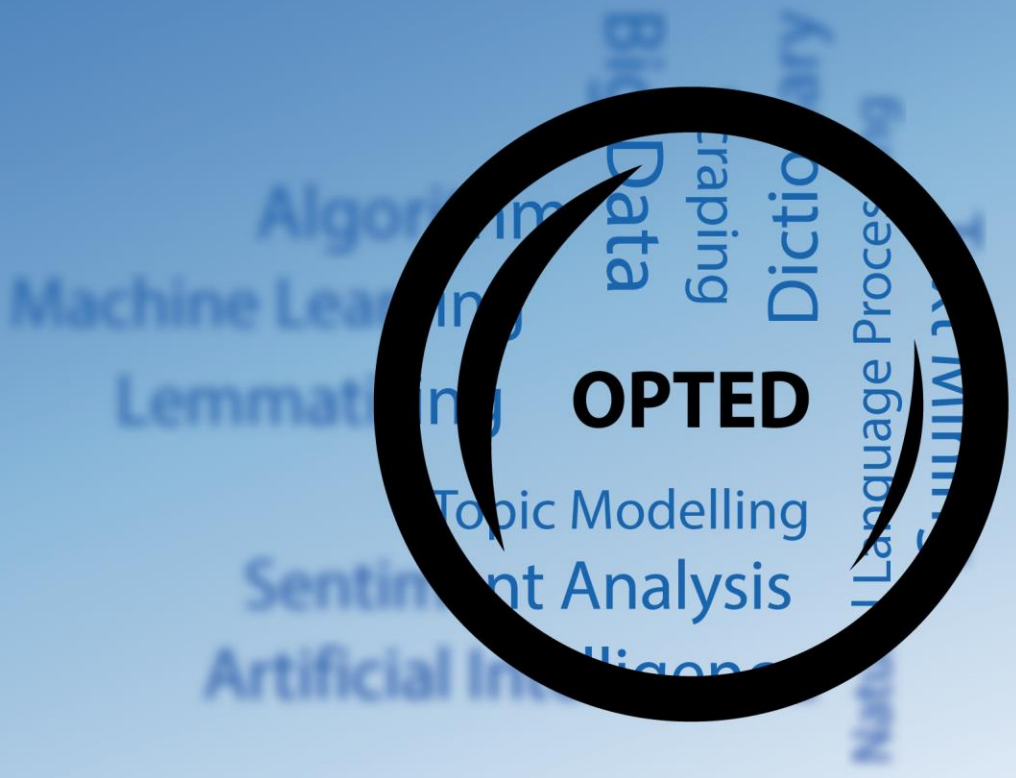
OPTED

Deliverable D5.2

Jan Schwalbach, Christian Rauh, Alexander Dahlheimer, Sven-Oliver Proksch, Miklós Sebők



Disclaimer

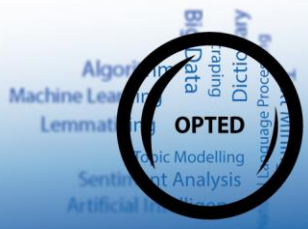
This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

PU

Type

R



OPTED
Observatory for Political Texts in European Democracies:
A European research infrastructure

Dataset ParlLawSpeech
Deliverable D5.2

OPTED WP5 team
Jan Schwalbach[1/2], Alexander Dahlheimer[1], Christian Rauh[3], Sven-Oliver Proksch[1], Miklós Sebők[4]

[1] University of Cologne
[2] GESIS Leibniz Institute for the Social Sciences
[3] WZB Berlin Social Science Center
[4] Institute for Political Science, Centre for Social Sciences, Budapest

Due date: September 2022

Table of Contents

Purpose

The project **OPTED: Observatory for Political Texts in European Democracies** (Horizon 2020 Grant agreement 951832) outlines a European Research Infrastructure facilitating the large-scale computational analysis of political texts in Europe. **Work Package 5** focuses on texts produced in the decision-making processes of national and supranational parliaments.

One of the work package's objectives is to highlight that data linking different parliamentary texts in machine-readable formats promise more systematic insights into the functioning of parliamentary democracies in Europe: we want to be able to exploit computational text analysis to learn about systematic patterns of how decisions are made from initial bill proposals, over publicly visible debates of elected representatives on the plenary floor, up until the finally adopted and collectively binding law. Linking different types of texts that parliamentary democracies produce, in short, promise new systematic insights on how collectively binding decisions are made and shaped.

Thus far, however, computational text analysis has been mostly successfully applied to study parliamentary debates only. It has been much more rarely exploited to study the policies on which parliaments decide in and through such debates, though. This is neither a problem of lacking analytical tools, nor of the principled availability of the required texts. Rather, the most significant hurdle is accessing these texts in readily machine-readable formats that would also allow to substantially link the different types of texts that parliaments produce.

Therefore, the work package team built **ParlLawSpeech**, a prototype of a systematic and encompassing text data collection semi-automatically collected from publicly available archives which combines ***machine-readable full-text vectors of legislative bills, speeches, and finally adopted legislation linked through a common identifier*** across several European countries.

In this deliverable we outline ***the collection, structure, and availability of the resulting ParlLawSpeech files***. In a parallel deliverable (D5.5), we illustrate the analytic potential and encourage future users along ***three exemplary application tutorials*** that put these data to use with transparent code in the free and open-source R environment.

Data - what *ParlLawSpeech* offers

Coverage

Two considerations drove the choice of countries/parliaments in the initial ParlLawSpeech prototypes. On the one hand, the initial data collection should be usable for diverse comparative research projects by including smaller and larger EU member states from different regions of the Union (including the EU itself).

On the other hand, we relied on our prior inventory of parliamentary text sources in Europe and our earlier collections of parliamentary speeches (ParlSpeech) to identify those official archives from which we could collect the respective text data within the time and resource constraints of the project. In a very labor-intensive process, all data have been collected by own web scraping scripts customized to each parliamentary archive and/or legal database (for which APIs are available in very few cases only).

The table below summarizes the countries and time periods we can eventually include in the ParlLawSpeech prototype.

Country	Code	Parliament	Time.period
Austria	AT	Nationalrat	1996-2019
Croatia	HR	Hrvatski sabor	2003-2020
Czech Republic	CZ	Poslanecká snemovna Parlamentu Ceské republiky	2002-2021
Germany	DE	Bundestag	2009-2021
Spain	ES	Congreso de los Diputados	1996-2019
EU	EU	European Parliament	1999-2019

Data collection and data structure

For each of the above-mentioned countries (including the EU), ParlLawSpeech provides three separate data files: one for bills, one for laws, and one for plenary speeches. Compared to extant political science corpora, ParlLawSpeech offers *two key innovations*. First, the bill and law corpora are - to the best of our knowledge - the most encompassing full-text vectors of legal documents handled by parliaments. Second, a common identifier that links across these three sets of documents opens up new analytical opportunities of linked text data analysis (illustrated, for example, in the tutorials provided in the parallel deliverable D5.5). The *ParlLawSpeech data files* are structured along the following columns:

- **Corpus_bills_[country].RDS**:
  Data on legislative bills tabled in the respective parliament
  - ***bill_ID***: Identification number of the bill document, following the conventions of the respective parliament
  - ***bill_title***: Bill title as provided in the original parliamentary archive
  - ***initiator***: Sponsor of the respective bill as provided in the original parliamentary documentation
  - ***bill_text***: Full text of the bill
  - ***initiation_date***: Day on which the bill was tabled (**TO DO**: harmonize to YYY-MM-DD throughout!)
  - ***speech_procedure_ID***: A unique ParlLawSpeech identifier linking bills, speeches, and law

- **Corpus_laws_[country].RDS**:
  Data on laws finally adopted by the respective parliament
  - ***law_ID***: Identifier of the law text as used in the respective parliament or the official legislation database of the country
  - ***title_law***: Title of the law as provided in the original archive
  - ***law_text***: Full text of the law
  - ***adoption_date***: Day on which the law was published
  - ***speech_procedure_ID***: A unique ParlLawSpeech identifier linking bills, speeches, and law (here *NA* for speeches that were not directly related to a specific bill/law procedure)

- **Corpus_speeches_[country].RDS**:
  Data on plenary speeches in the respective parliament (in chronological order across and within sessions)
  - ***speech_ID***: Unique ID for speeches per session within corpus that reflects the chronological order
  - ***date***: Day on which the speech was given (YYY-MM-DD)
  - ***agenda***: Agenda item under which the speech was given (following the conventions of the respective parliament)
  - ***text***: Full text of the respective speech
  - ***speaker***: Name of the person having given the plenary speech
  - ***party***: Party and/or partisan faction of speaker
  - ***chair***: Logical variable indicating whether the speaker is the organisation chair of the respective session
  - ***speech_procedure_ID***: A unique ParlLawSpeech identifier linking bills, speeches, and law

We collected all of these data by setting up own web scraping scripts (relying primarily on programming tools from the R environment such as the rvest package) that had to be individually customized to each parliamentary archive and/or legal database. In the absence of dedicated APIs or common identifiers in the parliamentary archives, we matched bills, speeches and laws on the basis of careful qualitative analysis of the parliamentary and/or legislative archives prior to web scraping. Specifically, we first inspected the bill archive of each parliament to find the locally unique ways of referring to a bill - typically a local numbering system encoding legislative periods, dates, and running numbers of bills therein. In cases where this pattern is repeated annually or per legislative term, the identifier had to be extended with a unique temporal tag. We then matched these identifiers in the respective agenda documents for plenary debates and added the bill identifier to all speeches given under the matching agenda item. Analogously, we inspect the respective national archive of laws to learn how bills are referred to in either the laws' texts or preferably the meta data provided by the archive. Once the pattern was established, we matched laws and bills with a respective string-matching procedure on the outputs of the web scraper.

If not otherwise indicated, all variables are provided as UTF-8 encoded strings. The lists above show the minimal data structure that is available across all countries and periods. Where readily available in the source archives, individual ParlLawSpeech files include additional meta information, for example on speaker roles in the speeches data set, or on document types, committees involved, and voting results in the bill data sets (in the case of the EU also including Celex IDs and legal bases). Note that in the case of the EU, the bill data set refers to legislative proposals from the European Commission.

We are furthermore currently in exchange with OPTED Work Package 8 to include Wikidata IDs for speakers as well as PartyFacts IDs for partisan factions so as to increase data linkage potential further.

Note that the columns with meta information as well as the full texts of bills and laws follow the conventions and document structures that the respective parliamentary archive provides. Staying close to the original conventions has two advantages. First, it allows users to filter the ParlLawSpeech data along external knowledge about the respective parliament. For example, researchers may want to isolate only debates related to specific document numbers or document titles that have been identified through the respective parliamentary website or other qualitative research. Second, sticking to the original document and data formats provides researchers with maximum freedom regarding text cleaning and pre-processing choices. For example, researchers may decide whether or not the recitals and justifications often provided with a legal document should enter their text analyses.

The downside of largely sticking to the original formats is, however, that comparative analysis especially across countries might require additional text cleaning (for an example, see Tutorial 2 in Deliverable D5.3) or (dis-)aggregation steps.

One exemplary issue in that regard is *speech segmentation*: In deciding of what counts as an individual speech, the ParlLawSpeech data strictly follow the conventions of the stenographic protocol in the respective parliament - each debate contribution that is either delimited by a clear HTML/XML tag or is introduced by speaker name and function in text files counts as an individual speech. In some parliaments, however, interventions from the audience appear as individual speeches, while in others they are not recorded in the minutes at all. In yet other cases, interventions appear as inserts in the main speech (e.g. in the German Bundestag). We provide a raw and a cleaned version of the speech text in the latter instance. In any case, we advise researchers interested in comparisons across parliaments to be aware of speech segmentation questions and the potential need to aggregate, disaggregate, or filter certain speeches in the light of their specific research question.

Overview of the available data files

The data have been initially compiled as .rds files for programming use in the free and open-source R environment. Users working in other environments can easily export them from R to any other format, using either base R's export functions or add-on packages such as haven or feather, for example.

The table below summarizes the individual data files provided in ParlLawSpeech, indicating the file size (MB), the number of variables offered, and the number of documents (bills, laws, speeches) therein.

In ***total***, the current state of the ParlLawSpeech prototype comprises ***1.98 GB of data***, including full texts of ***23,093 bills***, ***15,459 laws***, and ***1,367,130 plenary speeches***.

Country	File	Size	Variables	Observations
Austria	Corpus_bills_austria.RDS	52.65	10	5989
Austria	Corpus_laws_austria.RDS	46.08	7	3059
Austria	Corpus_speeches_austria.RDS	279.97	12	205110
Croatia	Corpus_bills_croatia.RDS	122.12	12	3690
Croatia	Corpus_laws_croatia.RDS	32.57	2	2974
Croatia	Corpus_speeches_croatia.RDS	162.01	9	405273
Czech Republic	Corpus_bills_CZ.RDS	96.81	5	1846
Czech Republic	Corpus_laws_CZ.RDS	6.32	4	714
Czech Republic	Corpus_Speeches_CZ.RDS	140.64	7	391629
EP	Corpus_bills_EP.RDS	101.66	9	6924
EP	Corpus_laws_EP.RDS	65.91	8	6654
EP	Corpus_speeches_EP.RDS	212.95	9	300172
Germany	Corpus_bills_germany.RDS	83.30	10	1764
Germany	Corpus_laws_germany.RDS	15.12	6	1173
Germany	Corpus_speeches_germany.RDS	244.39	13	188232
Spain	Corpus_bills_spain.RDS	27.47	10	2880
Spain	Corpus_laws_spain.RDS	21.47	5	885
Spain	Corpus_speeches_spain.RDS	267.21	11	268343

Data access

The ParlLawSpeech data are meant to be public (see also the outlook below) but we first plan to carefully validate them with the scientific case study we conduct for the upcoming deliverable D 5.3. Once this is done, we will offer a data dump with a permanent Digital Object Identifier under a CC0 1.0 Universal (CC0 1.0) Public Domain Dedication via the Harvard Dataverse. More importantly, the ParlLawSpeech data will also be integrated the final overarching protype of *OPTED data infrastructure* (due in September 2023). For Deliverable D 5.6, we are also working on online apps for more targeted data extraction and visualization (see also the outlook below).

To prove that the data have been actually collected for now, the datafiles listed above are currently provided for review purposes via a dedicated **Dropbox**. These files can also be used to reproduce the tutorials we provide in the parallel deliverable D5.5.

Outlook

Expanding on the data collection provided in this deliverable, the OPTED WP5 team in conjunction with other OPTED work packages has finished or currently works on the following themes:

- Training tutorials (parallel deliverable D5.5): we illustrate the analytic potential and encourage future users along ***three exemplary application tutorials*** that put the ParlLawSpeech data to use with transparent code in the free and open-source R environment. In addition, we are currently in exchange with work Packages 3 and 9 about an integrated online presentation of all tutorials produced by the OPTED project.

- Scientific case study (deliverable D5.3): We are preparing a scholarly article that puts the potential of linked parliamentary text data to use in a comparative manner across the covered European democracies. Specifically, we plan to measure the semantic similarity between speeches of partisan members of parliament and the bill proposals to then study whether and which debates are reflected in the ultimately adopted law.

- Public access tools (deliverable D5.6): Particularly with a non-scholarly audience in mind (a.o., data journalists and interested citizens), we want to offer tools that allow the extraction of descriptive visualisations or more targeted text corpora directly from a public access website. Initial prototypes are available here and here. Currently we are in exchange with Work Packages 7 and 9 on how to best integrate such tools into a common infrastructure.

- Further data linkage: We are in exchange with work package 8 on how to improve linkages of ParlLawSpeech to other data sets further. One specific aim is to enhance meta-information on parliamentary speakers with links to biographical or party databases through common identifiers.