Miklós Sebők,* Csaba Molnár** & Anna Takács***

Levelling up quantitative legislative studies on Central-Eastern Europe: Introducing the ParlText CEE Database of Speeches, Bills, and Laws

Intersections. EEJSP
10(4): 106–125.
https://doi.org/10.17356/ieejsp.v10i4.1327
https://intersections.tk.hu

[*] [sebok.miklos@tk.hun-ren.hu] (HUN-REN Centre for Social Sciences, Budapest)
[**] [molnar.csaba@tk.hun-ren.hu] (HUN-REN Centre for Social Sciences, Budapest)
[***] [takacs.anna@tk.hun-ren.hu] (HUN-REN Centre for Social Sciences, Budapest)

Abstract

The availability of ready-made textual corpora for research is crucial for social scientists, especially in the current era of rapid advancements in natural language processing (NLP) and artificial intelligence (AI) methods. Despite various useful contributions that address issues of accessibility and standardisation when it comes to such corpora, in many cases, they have limitations related to scope, geographical coverage, and time frame. This concern is particularly significant in the context of political research on Central-Eastern Europe (CEE), for which such deployment-ready databases are few and far between. In this research note, we bridge part of this gap by making available a new database: ParlText CEE. The database, prepared under the auspices of the V-Shift Momentum project at the HUN-REN Centre for Social Sciences, covers almost 1.9 million text vectors and metadata for parliamentary speeches, bills, and laws for Czechia, Hungary, Poland, and Slovakia for the period from 1990–1991 to 2022–2024. The datasets encompass relevant dates, texts, titles, and, in the case of the speech corpora, parliamentary agendas, speaker names, and parties. All data are also linked based on unique identifiers following the ParlLawSpeech standard. This paper introduces the specifics of the 1.0 release of ParlText CEE and contemplates its possible use cases.

Keywords: Central-Eastern Europe; legislative studies; legislative database; parliamentary speeches; bills and laws

1 Introduction

The rapid emergence of text-as-data approaches and natural language processing methods (NLP) in the 2010s opened up vast new opportunities for political research in general and quantitative legislative studies in particular (Grimmer & Stewart, 2013; Brady, 2019; Slapin & Porksch, 2014). The most significant requirement of conducting quantitative text analysis is finding relevant and directly usable corpora associated with adequate metadata (such as the socio-economic background of members of parliament (MPs) who make speeches, see Grossman & Pedahzur, 2020, p. 254). Quantitative legislative studies can utilise such databases of political debates (Bächtiger, 2014) and legal documents (Martin & Vanberg, 2014, p. 439), including bills (draft laws) and adopted laws. Using these data with NLP methods can help reveal the still hidden patterns and characteristics of political behaviour and governance and extend – still prevalent – single-country research designs to various jurisdictions and languages in a comparative manner.

There are many precursors in the form of projects aimed at creating structured datasets on legislatures for the Central-Eastern European (CEE) region. Among others, the ParlSpeech (Rauh & Schwalbach, 2020) and ParlLawSpeech (Proksch et al., 2024) datasets, the Comparative Agendas Project (Baumgartner et al., 2019), and the ParlEE database (Sylvester et al., 2024) are notable examples of such contributions. In all cases, they cover at least one of the so-called Visegrád countries (Czechia, Hungary, Poland, and Slovakia, which we use as a synonym for CEE). These datasets mainly contain information on legislative speeches, but the Comparative Agendas project also collected information, for example, on legislative documents (bills and laws). Except for the Comparative Agendas Project, they mainly focus on recent decades.

Despite various useful contributions and existing datasets that address issues of accessibility and standardisation, they tend to have several limitations in terms of the scope of metadata, geographical coverage, and timeframe. This concern is particularly significant in the context of political research on CEE, for which such deployment-ready databases are few and far between. Although the region's countries are generally regarded as a mostly homogenous group, they are, in practice, different from each other in several critical respects (Wolchik & Curry, 2018). Differences arise if we focus on, e.g. their social diversity, party systems, institutional settings, or the relevant actor types of policy-making. Databases are crucial tools that permit access to valid inferences for such comparative research questions.

Besides the limited scope and disjointed nature of legislative datasets on CEE, the other main problem of quantitative researchers is data accessibility. Although legislative archives are publicly available for CEE countries, at least for the period starting with the democratic transition of 1990, they are often difficult for data scientists to navigate. APIs are sometimes available,[1] but they are not amenable to text analysis, web scraping, and data cleaning also faces challenges. This is partly understandable: traditionally, these archives were designed to serve the needs of legislative staff, political actors, citizens, or journalists (Joshi & Rosenfield, 2013); thus, they are less amenable to systematic data collection, even in a Western European context (Kiss & Sebők, 2022). However, empirical researchers have different data needs from other stakeholders: they search for comprehensive data (e.g., the full population of speeches in a given period) in a standardised, structured, machine-readable format. Fortunately, most legislatures offer full texts and metadata for speeches and legal documents, but in many cases, the onus is still on the researcher to process them into a structured format.

In this research note, we bridge part of this gap regarding the CEE region by presenting a new database: ParlText CEE. The database was prepared under the auspices of

[1] The Czech (https://www.psp.cz/sqw/isp.sqw) and Hungarian (https://www.parlament.hu/alkalmazasok) APIs are available after registration, while the Polish one (https://api.sejm.gov.pl/sejm.html) is available without further registration.

the V-Shift Momentum project at the HUN-REN Centre for Social Sciences with the financial support of the Hungarian National Laboratory for Artificial Intelligence.[2] It includes data on the unicameral legislatures of Hungary and Slovakia and the lower chambers of the bicameral Czech and Polish legislatures (the Chamber of Deputies and the Sejm). The main advantages of the ParlText CEE dataset compared to other databases are its wider time frame, larger metadata collection, and a relational database structure for its distinct subcorpora covering all plenary activities in terms of speeches, bills, and laws. The 1.0 version of the dataset currently covers almost 1.9 million text vectors and metadata for the subcorpora of the legislative processes of all four CEE countries. The time frame in all cases covers the democratic period, starting in the early 1990s until the 2020s. Metadata include relevant dates (such as the initiation or adoption of bills), texts, titles, and, in the case of the speech corpora, parliamentary agendas, speaker names, and parties. An additional contribution of the project is that all data are linked based on a unique identifier following the ParlLawSpeech standard. This allows for connecting each bill to its adopted final version as law and all the plenary debates that took place in connection to them.

The ParlText CEE database was built on an open-science framework. All data is published in public repositories, providing access based on the CC BY-NC license (Attribution-NonCommercial 4.0 International), constituting its only official version. This paper introduces the specifics of the 1.0 release of ParlText CEE and contemplates its possible use cases. After briefly introducing some precursor datasets in the field of legislative studies, we present the structure of the ParlText CEE database through a description of the main variables in the database's codebook. Next, we detail the data linkage methods, followed by an overview of some descriptive statistics of the database. In the Conclusion, we suggest potential use cases for the database in political science and beyond.

2 Precursors of ParlText CEE

Although the CEE region often lags behind its Western European counterparts in terms of the availability of ready-made textual corpora, there are some important precursors that a project aimed at collecting and publishing data for the countries of the Visegrad Four can build on. Such databases that offer machine-readable corpora for the CEE region include CLARIN, CAP, ParlEE, ParlSpeech, and the ParLawSpeech project (for a more detailed overview, see Sebők et al., 2025). As Table 1 presents, at least one of the ParlText CEE target countries is included in these datasets.

[2] The ParlText CEE database has no connection to the British Parliament's or the Australian legislature's Teletext service of the same name (ParlText). For more info on these services, see Parliament of Australia (1991). Department of Parliamentary Reporting Staff – Report for – 1990–91, https://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;query=Id%3A%22publications%2Ftabledpapers%2FHPP032016008744%22;src1=sm1, and Select Committee on Broadcasting Minutes of Evidence, 1998. and Select Committee on Broadcasting Minutes of Evidence (1998). ANNEX 1. THE PARLIAMENTARY CHANNEL https://publications.parliament.uk/pa/cm199798/cmselect/cmbroad/984/8071503.htm

Table 1 Precursor projects of ParlText CEE

Dataset name	N of polities	Coverage of CEE countries	Domains	Maximum timespan
CLARIN	29	Czechia, Hungary, Poland	Speech	2015–2022
CAP	27	Hungary	Speech, bill, law, media etc.	1000–2023
ParlEE	28	Czechia, Hungary, Poland, Slovakia	Speech	2009–2019
ParlSpeech	9	Czechia	Speech	1987–2019
ParlLawSpeech	8	Czechia, Hungary	Speech, bill, law	1993–2022

The Common Language Resources and Technology Infrastructure (CLARIN) project, one of the EU's so-called ESFRI roadmap of major research infrastructures, developed several datasets on different European countries. The ParlMint dataset contains the annotated corpora of 29 European countries and autonomous region's parliamentary debates, which–at the time of writing–makes it the dataset with the most comprehensive coverage of legislative debates in Europe in a unified structure (Erjavec et al., 2023a; Erjavec et al., 2023b; Kuzman, 2023). Its first wave included Bulgaria, Croatia, Czechia, Hungary, Latvia, Lithuania, Poland, and Slovenia. The second one added Bosnia and Herzegovina, Estonia, Romania, Serbia, and Ukraine from the CEE region, mainly for 2015–2022.

The datasets are available in XML format. The corpora were pre-processed (tokenisation and lemmatisation) and linguistically annotated, and several types of metadata were also included (e.g., gender of the speaker). The ParlaMint corpora[3] are also divided into periodical subcorpora, such as speeches made during the COVID-19 pandemic. However, it does not contain information on legal documents (bills and laws) and, therefore, is not directly applicable to the joint analysis of all legislative procedures. Furthermore, the processed datasets follow the TEI XML technical standard, which is more common in the digital humanities and is not directly compatible with the workflows in the programming languages most prevalent in the social sciences (such as R and Python) due to its unique data structure. The corpora were developed based on web-scraped data (Mikušek, 2024).

The second significant collection of legislative data is associated with the Comparative Agendas Project (CAP).[4] This international collaboration of several dozen country projects investigates several arenas of the policy agenda (Baumgartner et al., 2019). Besides countries, the datasets also contain information on both supranational entities (e.g. the European Union) and substate-level regions (e.g. the State of Florida) for different periods (but mainly for the decades around the 2000s). The CAP project's datasets contain information not only on legislative speeches (although in most cases only on a selected,

[3] https://www.clarin.eu/parlamint

[4] https://www.comparativeagendas.net/datasets_codebooks

important component of them, e.g. parliamentary questions or State of the Union speeches) but also on legal documents such as laws and bills. For the purpose of a unified legislative textual database for CEE, it has two drawbacks. First, as CAP researchers focus on classifying documents by their policy content, including the full corpus is not a requirement, and this is often missing. Second, the CEE region is scarcely covered, with datasets available only from Croatia, Hungary, and Poland.

Third, the V3 version of the ParlEE project[5] encompasses the parliamentary speeches of 28 polities covering the timespan between 2009 and 2019 (Sylvester et al., 2024). These speeches were broken down into sentences and annotated with date, speaker, party, references to the EU governance, and policy topics (based on the Comparative Agendas Project classification method). Although all target countries of ParlText CEE are included in ParlEE, the dataset does not contain information on bills and laws. Moreover, its time frame is shorter than what may feasibly be covered with publicly available data.

The fourth important precursor is the ParlSpeech project.[6] It also comprises an extended corpus of legislative speeches, including over six million parliamentary speeches from nine countries (Rauh & Schwalbach, 2020). In addition to the speeches' texts, it includes essential metadata, such as date, agenda item title, and party names based on Döring and Regel's Party Facts database (Döring & Regel, 2019). Here, the CEE region is only represented by Czechia, and similarly to the abovementioned databases (with the exception of CAP), the dataset is speech-only.

The ParLawSpeech project[7] extends the corpora of ParlSpeech into new text domains: bills and laws. It covers data from eight legislative bodies, including the European Parliament. Variables in these datasets focus on relevant metadata (such as dates, speakers/initiators, agenda titles, and party affiliations in the case of speeches, where available) alongside the full-text vectors for all bills, laws, and speeches. The covered time frame differs by country, but at least 11 years of data are available for all legislatures, mostly covering the 2010s. The novelty of the dataset lies in linking the three domains of texts: researchers can find the bill discussed by a given parliamentary speech and the respective law text after the bill's adoption. The dataset includes two countries from the CEE region, Czechia and Hungary, and, similarly to its predecessor project, ParlSpeech, it covers a more limited time frame than what is publicly available (Proksch et al., 2024).

Finally, we mention an additional database that was less of a precursor than a complementary corpus. The Vitrin Démocratique database (Tremblay-Antoine et al., 2024) contains valuable information on European Parliament debates in both the speech's original language and its English translation for the period 2014–2023. Similarly to the EP corpus of ParlLawSpeech, it brings more information to the table on politicians representing the CEE region and allows for a multi-level analysis of MP behaviour.

[5] https://pureportal.strath.ac.uk/en/datasets/parlee-plenary-speeches-v3-data-set-annotated-full-text-of-10-mil
[6] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/L4OAKN
[7] parllawspeech.org

3 Database structure

The ParlText CEE database builds on its predecessors by building on available data for Czechia and Hungary, extending their time frame and metadata collection via data linkage, and presenting entirely new data collections for Poland and Slovakia[8]. The ParlText CEE database comprises three UTF-8 encoded corpora in a .rds format for each country: bills, laws, and parliamentary speeches. The respective texts and metadata were web-scraped directly from the legislatures' archives to obtain newly added data points. Tables 2 to 4 present the variables from these three collections.

Table 2 Description of the core variables in the speech corpora[9]

Variable	**Type**	**Description**
speech_ID	string	The unique identifier of the speech (ISO_S_YYYYMMDD_N)
link	string	The link to the given speech
agenda	string	Agenda item title under which the speech was given (9998 if there is none, as in the case of pre/post-agenda speeches)
electoral_cycle	string	Electoral cycle during which the speech was given
speechnumber	integer	The rank order of the speech within the given session
speaker	string	The name of the speaker
chair	logical	Dummy variable indicating whether the speaker is the parliamentary chair
date	date	The date of the speech (YYYY-MM-DD)
speech_text	string	The text of the speech
bill_ID	string	The id of the bill(s) discussed (ISO_B_YYMMDD_N); if multiple bills are related to the speech, their IDs are separated by a comma; if the speech is not connected to any bills, it is marked with 9998

Data linkage was implemented with the help of three different ID types: bill_ID, law_ID and speech_ID. All three serve as the unique identifiers of the observations of their respective datasets. The bill_ID serves as a common link across the corpora (we return to the details of the linking process below). Bill_IDs are based on the conventions of the respective parliament or the official legislative database of the country. The same applies to

[8] We would like to express our gratitude to Jakub Szabó for his contributions related to the dataset on Slovak legislative speeches.

[9] The Slovakian speeches corpus in its 1.0 version does not contain the following variables: *link*, *agenda*, *speechnumber*, and *bill_ID*

law_IDs. In contrast, speech_IDs were generated after data collection. There are instances where the given speeches were not related to a specific bill (or law), annotated as '9998'. Missing values were marked by '9999' in the datasets.

The first corpus consists of the full-text vectors of parliamentary speeches and their respective metadata, similar to the ParlEE and ParlSpeech datasets (Table 2). A unique speech_ID was created, comprising four segments separated by an underscore. This consists of the countries' ISO 3166 codes, the letter S (as an abbreviation for speech), the date of the speech, and a marker of the chronological positions of the speech within the day. As it is essential for both validation and archiving purposes, the URL of the original link to legislative websites/APIs for the given speech is also provided. The agenda items are also listed, allowing for combining texts and filtering for specific debates (agenda titles also served as the basis for linking speeches to bills and, via the bills, laws). The name of the chair is included to track legislative activity (they contextualise agenda items) and allows for excluding procedural information. The dates of the speeches (YYYY-MM-DD) and the electoral cycles, which were calculated based on the dates, are also provided. The latter are essential for connecting speeches to political variables such as legislative majorities and government periods.

Table 3 Description of the core variables in the bills corpora

Variable	**Type**	**Description**
bill_ID	string	A unique ParlText CEE identifier to the bill, also used for linking across corpora (ISO_B_YYMMDD_N)
bill_link	string	Link to the bill's parliamentary data sheet
electoral_cycle	string	The electoral cycle during which the bill was introduced
bill_title	string	The title of the given bill
date_introduced	date	The date of the bill's introduction (YYYY-MM-DD)
number_document	integer	The record number under which the bill was introduced
bill_text	string	The text of the bill

The second dataset, presented in Table 3, is a collection of bills introduced in parliament comprising titles, cleaned bill texts, and metadata. Similarly to the corpora of speeches, the bills' respective links, their date of introduction, and the electoral cycles are also provided. The original record numbers (number_document) were kept for traceability, as they serve as the official identification for the bills. Bill_IDs were created based on the respective legislative systems. They encapsulate the countries' ISO 3166 codes, the letter B for the bill, the proposal's date, and the original document's number.

The corpus of adopted laws contains their titles and cleaned full-length texts (Table 4). The respective links on which the laws can be accessed are also included, as well as their date, year, and electoral cycle of publication. Law IDs were constructed in a similar manner to bills and speeches. They contain the countries' ISO 3166 codes, the letter L for law, the year of publication in YYYY format, and the record number under which they were published, all separated by an underscore.

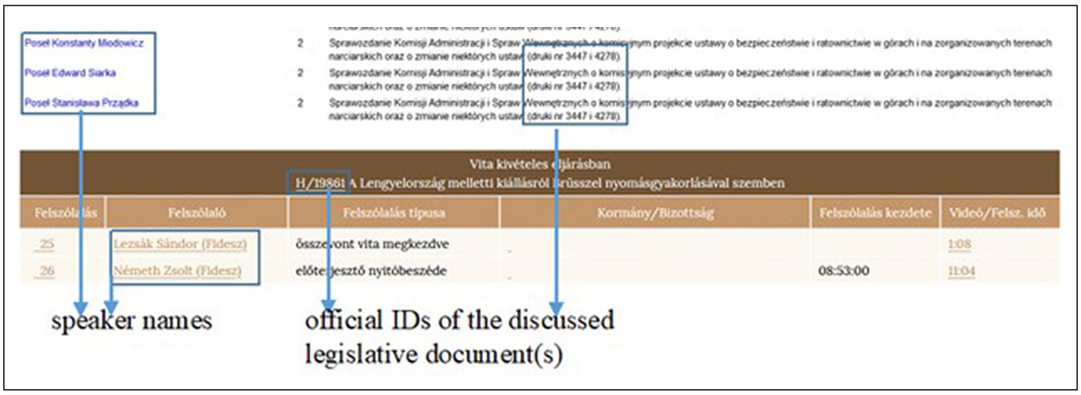
Table 4 Description of the core variables in the laws corpora

Variable	**Type**	**Description**
law_ID	string	The unique identifier of the law text (ISO_L_YYYY_N)
law_link	string	Link to the law text
law_text	string	Full text of the law
electoral_cycle	string	The electoral cycle during which the law was introduced
year_published	integer	The year of the law's publication
number_published	integer	Record number under which the law was published
date_published	date	The date of the law's publication
law_title	string	The title of the law
bill_ID	string	The unique identifier of the bill whose accepted version the law is (ISO_B_YYMMDD_N)

4 The linkage structure of ParlText CEE

The structure of ParlText CEE allows for the creation of linkages with other database formats. We followed the standards defined by the creators of ParlLawSpeech (Proksch et al., 2024). Here, we only present the basics of this approach. The three corpora, the legislative speeches, bills, and laws are linked on the country-dataset level. The logic of the linkage is rooted in information on the legislative procedure. In a generic process, after bills are introduced, a decision is made (in many cases by the speaker/president of the legislative body) on whether they can proceed first to committees and then to the plenary. If they are put on the plenary legislative agenda, they are discussed among MPs in legislative speeches. If a required majority of members of parliament (MPs) support a bill and they adopt it, the bill becomes law. In short, some speeches discuss bills (as prospective but not necessarily adopted laws). Some speeches are unrelated to bills and laws: they can, inter alia, be procedural in nature, connected to resolutions of the given house of parliament, or pre/post-agenda political debates.

The linkage of different corpus types can open up new avenues for research seldom leveraged in legislative studies. However, the methodological process leading there is riddled with challenges. While the Polish and Hungarian legislature's official website contained structured information on the relationship between speeches and the bills, the same did not hold for the Czech and Slovakian parliament. Figure 1 shows two examples of the Polish and Hungarian parliaments' websites. They represent well-structured tables containing hyperlinks to the texts of parliamentary speeches under the speakers' names. In the third column of the Polish database, the agenda titles are listed and contain the official IDs of the discussed legislative documents (in this example, 3447 and 4278). In the Hungarian parliament's database, the header contains the same information on the agenda item title and the official ID(s) of the discussed documents (in this example, H/19861).



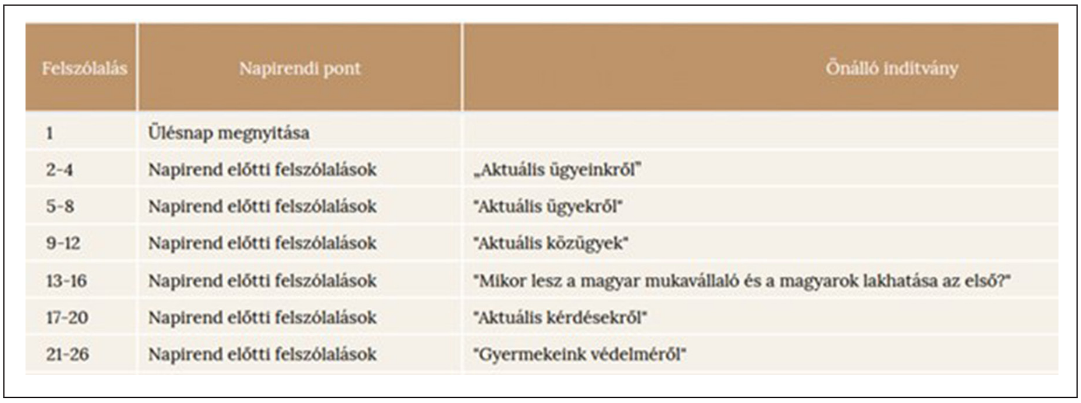
Poseł Konstanty Miodowicz	2	Sprawozdanie Komisji Administracji i Spraw Wewnętrznych o komisyjnym projekcie ustawy o bezpieczeństwie i ratownictwie w górach i na zorganizowanych terenach narciarskich oraz o zmianie niektórych ustaw (druki nr 3447 i 4278)
Poseł Edward Siarka	2	Sprawozdanie Komisji Administracji i Spraw Wewnętrznych o komisyjnym projekcie ustawy o bezpieczeństwie i ratownictwie w górach i na zorganizowanych terenach narciarskich oraz o zmianie niektórych ustaw (druki nr 3447 i 4278)
Poseł Stanisława Prządka	2	Sprawozdanie Komisji Administracji i Spraw Wewnętrznych o komisyjnym projekcie ustawy o bezpieczeństwie i ratownictwie w górach i na zorganizowanych terenach narciarskich oraz o zmianie niektórych ustaw (druki nr 3447 i 4278)

Vita kivételes eljárásban
H/19861 A Lengyelország melletti kiállásról Brüsszel nyomásgyakorlásával szemben

Felszólalás	Felszólaló	Felszólalás típusa	Kormány/Bizottság	Felszólalás kezdete	Videó/Felsz. idő
25	Lezsák Sándor (Fidesz)	összevont vita megkezdve	-		1:08
26	Németh Zsolt (Fidesz)	előterjesztő nyitóbeszéde	-	08:53:00	11:04

Figure 1 Examples of linking speech and bill texts from the Polish and Hungarian legislatures' websites

It was more difficult to extract the same information from the official websites of the Czech and Slovak parliament. Although the Czech legislature developed a similar structure for most electoral cycles, in some cases, extra effort was necessary. For these electoral cycles, the chair of the assembly sessions regularly presented the new agenda item title in bold on the official website, as shown in Figure 2. In these cases, until the next bold agenda item name, all speeches were connected to the given agenda item. The underlined bold speaker names tagged the speeches' beginning. In most cases, the agenda item title also contained the official IDs of the discussed legal documents. An easy way to validate this method for linking data is to check chair interventions, as they regularly enlist agenda items at the beginning of the session in the same order as they are discussed later.



Felszólalás	Napirendi pont	Önálló indítvány
1	Ülésnap megnyitása	
2-4	Napirend előtti felszólalások	„Aktuális ügyeinkről"
5-8	Napirend előtti felszólalások	"Aktuális ügyekről"
9-12	Napirend előtti felszólalások	"Aktuális közügyek"
13-16	Napirend előtti felszólalások	"Mikor lesz a magyar mukavállaló és a magyarok lakhatása az első?"
17-20	Napirend előtti felszólalások	"Aktuális kérdésekről"
21-26	Napirend előtti felszólalások	"Gyermekeink védelméről"

Figure 2 Examples of the extraction of agenda item names (Czech and Hungarian)

Another problem with linking data occurs when an agenda item is devoted to multiple legal documents. In such cases, we connected all documents mentioned in the agenda item name to the speeches listed under the agenda item. As bills are the central element of the linkage procedure (since all laws originated in bills, but not vice versa), bill IDs are included for each observation of the legislative speech dataset and the dataset of laws. A separate code was assigned for speeches unrelated to bills and laws (see above).

We illustrate the linking process by using an example from the Czech corpora. The bill_ID 'CZE_B_190626_535' was introduced under the title 'o bezpečnosti práce' ('On occupational safety') on 26 June 2019. The legislative debate started almost two years later, on 23 March 2021. It was a relatively short debate, as in addition to the speaker of the Chamber of Deputies presiding over the debate, only eight speeches were held connected to the law, and these speeches were presented by only five speakers. After the first debate, the bill was adopted on 23 April 2021 after three speeches presented by two speakers. The law was approved by both chambers on 9 June 2021 under the title 'o bezpečnosti práce v souvislosti s provozem vyhrazených technických zařízení a o změně souvisejících zákonů' (in English 'On occupational safety in connection with the operation of reserved technical equipment and on changes to related laws') as law 250/2021. This example shows how procedural metadata can be leveraged in a potential debate-focused analysis of a single bill.

5 Descriptive statistics

Table 5 summarises the three corpora (laws, bills, and speeches) by country. The time frame slightly differs between the countries, but they are all designed to encompass at least 30 years. Most corpora contain data from the post-transition period at the earliest (early 1990s) until the latest available data points (between 2022 and 2024). The selection of relevant parliamentary websites/URLs was based on expert decisions. The final sources included scraped files from legal databases, national parliamentary websites, and databases from previous data collections prepared by our research team (cap.tk.hu). The values presented for the Polish datasets are subject to change upon additional validation in progress at the time of submission.

A validation check on the corpora partly relied on an R script developed by the ParlLawSpeech team. The multi-step process encompassed both automated and manual checks. First and foremost, general completeness was tested, searching for duplicates and missing values in the texts and metadata columns. Further inspection was needed in the case of duplicates to check for, among other things, identically worded speeches. Verification of the number of observations against the source websites was also done after the scraping process, and further cross-validation using alternative sources, such as the above-presented ParlaMint database (when available). The expected quality of text content was ensured by extracting random samples and manually spotting error patterns (such as headings, footers, unnecessary breaks in texts, or even incorrect encoding). We also conducted a uniqueness check of the respective links using random manual evaluation.

Table 5 Summary of the laws, bills, and speeches corpora by country

Country	Data	Count	Time frame	Source
Czechia	*Laws*	3,214	1990–2023	wolterskluwer.cz
	Bills	5,284	1990–2023	nrsr.sk, psp.cz
	Speeches	*574,548*	1990–2023	psp.cz
Hungary	*Laws*	4,303	1994–2022	cap.tk.hu
	Bills	7,498	1994–2022	cap.tk.hu
	Speeches	*487,877*	1994–2022	cap.tk.hu
Poland	*Laws**	*5,716*	1991–2023	sejm.gov.pl
	*Bills**	*9,488*	1991–2023	sejm.gov.pl
	*Speeches***	*261,802*	2011–2023	sejm.gov.pl
Slovakia	*Laws*	4,260	1990–2023	slov-lex.sk
	Bills	*NA*	NA	nrsr.sk
	Speeches	*423,952*	1994–2023	psp.cz, nrsr.sk

Notes: * Estimated counts pending final validation; ** Estimated count for the 1991–2023 period is about 716,000.

The dataset contains all speeches, bills and laws until the end of the last closed electoral cycle in the case of Hungary, Poland and Slovakia. Because the current electoral cycle of Czechia is near to its end, we decided to scrape data until the end of 2023.

The most important results of these processes for the fully finished datasets are summarised below (Table 6). We present the number of textual, link and ID duplicates in the three corpora for each country and indicate if linking through the datasets was possible. It is important to note that due to the structure of the Czech and Slovak parliament's website, the same link is associated with multiple speeches, as they are listed on the same page. The high number of speech duplicates is usually caused by identically worded texts (such as greetings) and can be cross-referenced by checking unique links. However, further cleaning and deeper investigation may be needed in the case of speeches, as link- and ID duplicates can signal scraping errors.

A set of figures was also generated for the Hungarian corpora to check the dataset's quality and explore potential outliers and data errors. For instance, visualising distributions can help detect potential anomalies after data collection. Figure 3 depicts the number of Hungarian bills introduced between 1994 and 2021. As seen on the plot, there are no systematic errors in the corpus (missing values or unusual trends due to scraping mistakes, for example). The number of bills is reasonably balanced throughout the period, except for some outliers.

Table 6 Summary of the validation check on the corpora

		Czechia (2013–2023)	**Hungary (1994–2022)**
Text duplicates (NAs)	*Laws*	0 (0)	0 (0)
	Bills	33 (33)	31 (2)
	Speeches	*2,581* (0)	67,080 (2)
bill_ID duplicates (9998s)	*Laws*	0 (0)	0 (0)
	Bills	0 (0)	0 (0)
	Speeches	*191,527* (60,956)	35,232 (0)
Link duplicates (NAs)	*Laws*	0 (0)	– (–)
	Bills	0 (0)	0 (0)
	Speeches	*166,983* (0)	35,232 (1)
Linking		Validated	Validated

Note: Table 6 provides an overview of the results of the validation processes for the Czech and Hungarian corpora. We present the number of duplicates in texts, bill_IDs, and links, marking the number of NAs, as well as signalling if linking through the datasets is possible. The results of the two law corpora overall indicate that the files are clean, and the preprocessing and linking through bill_IDs were done correctly. Identical bill texts in the Czech case indicate missing data. Although negligible, this required further investigation in the Hungarian corpus. In the case of speeches, the large quantity of bill_ID duplicates indicated that multiple speeches were made on the same bill, while 9998s marked speeches unrelated to an agenda point. Link duplicates in the Czech speech corpus arose due to the structure of the source website and did not signal a real issue.

A similar observation can be made regarding the fluctuation of the number of laws published throughout the years (Figure 4). Although there are periods with higher frequencies, the number of published laws fluctuates within a reasonable range. Notably, the extreme values in the two datasets seem to align well, pointing toward potential institutional factors (such as the number of sessions).

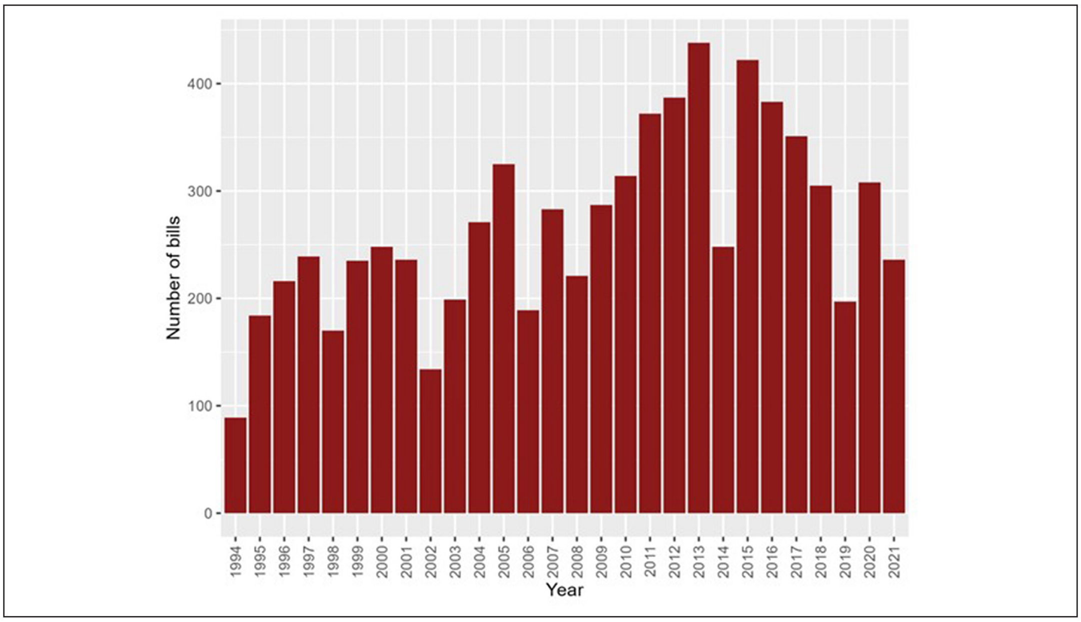


Figure 3 Distribution of Hungarian bills between 1994 and 2021

Note: Figure 3 shows the distribution of bills introduced in Hungary between 1994 and 2021. Since 2022 was an election year, we decided to exclude these observations from our figures (the electoral cycle ended in May).

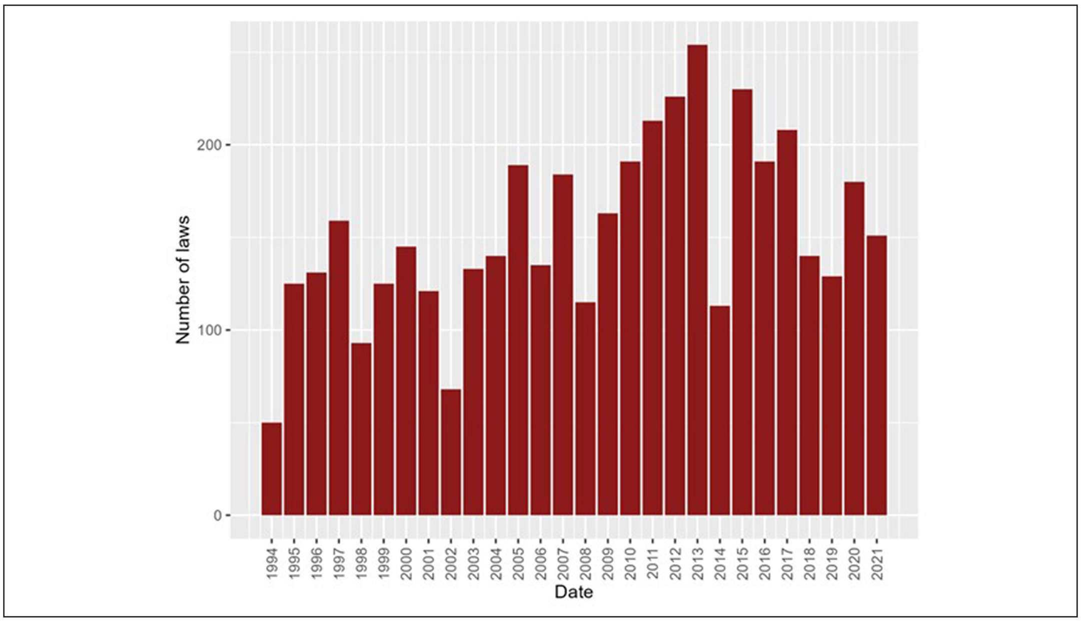


Figure 4 Distribution of Hungarian laws between 1994 and 2021

Note: Figure 4 shows the distribution of laws published in Hungary between 1994 and 2021. Observations from 2022 are excluded, as it was an election year.

A similar pattern emerges when examining the frequency of speeches delivered by MPs over the same period (Figure 5). Again, the data is consistently distributed with occasional outlier spikes, indicating a relatively stable pattern of speech activity. Moreover, comparing these patterns with those observed in the bills that were introduced and laws published, it is evident that while there may be correlations, each dataset possesses its own unique characteristics and dynamics.

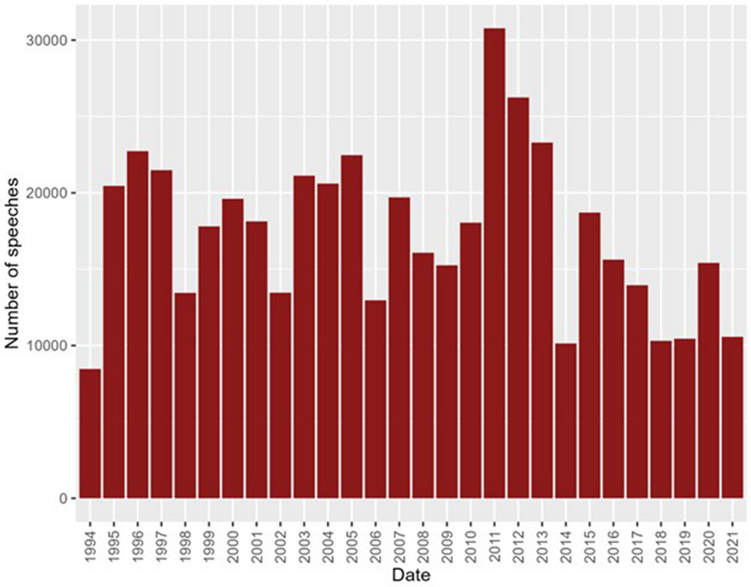


Figure 5 Distribution of Hungarian speeches between 1994 and 2021

Note: Figure 5 shows the distribution of speeches held in the Hungarian parliament between 1994 and 2021. The histogram indicates a relatively stable pattern of speech activity. Observations from 2022 are excluded, as it was an election year.

Another valuable approach to identifying anomalies involves examining the length of bills and laws linked to each other, measured in terms of their character count, as depicted in Figure 6. While a close and nearly linear correlation is evident, there are prominent outliers in the length of bills.

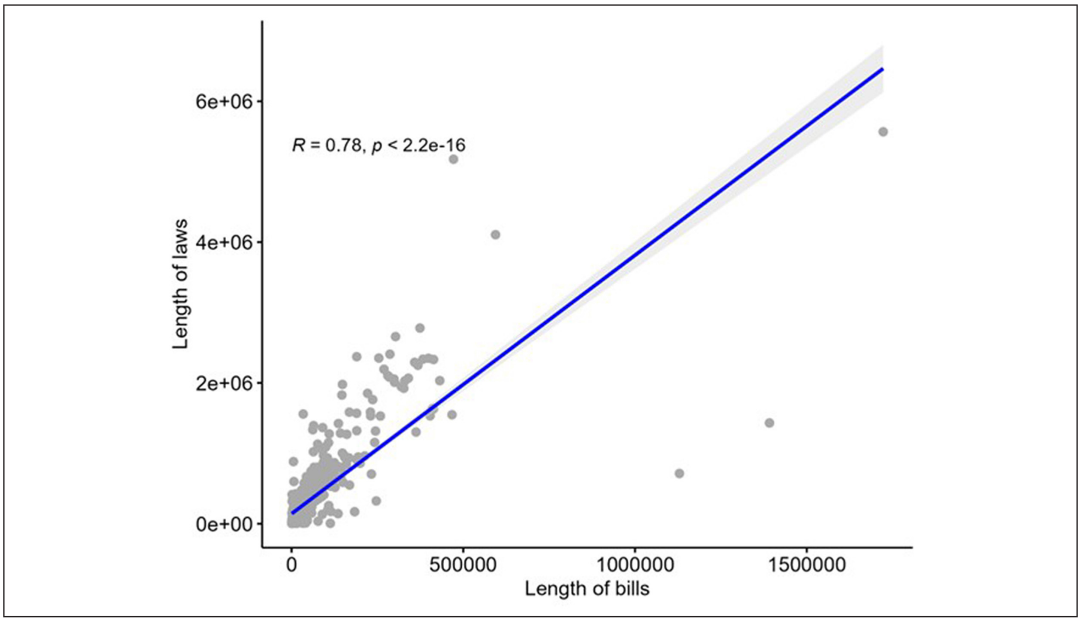


Figure 6 Correlation between the length of laws and bills in Czechia between 2013 and 2023

Note: Figure 6 depicts the correlation between the length of the *bills* measured in characters (x-axis) and the length of the *laws* measured in characters (y-axis). Based on the linear approximation, a strong correlation is assumed between the two.

6 Conclusion

The potential use cases of ParlText CEE are associated with a wide range of research questions concerning legislative processes, trends, and patterns of legislative activity. One such research area may be polarisation (e.g., measuring polarisation by the diversity of the legislative speeches of different political groups on the same bills). Researchers of the politics of parliamentary debate (see Back et al., 2022) and legislative studies more generally (see Benoît & Rozenberg, 2020) can utilise the new database as an input for discourse analysis, investigation of policy frames, or party issue ownership.

Figure 7 illustrates one such application: the frequency of speeches in the Hungarian parliament during each plenary meeting, categorised by political parties (from 1994 to 2022). By examining the patterns and fluctuations depicted in the figure, researchers can gain insight into the varying degrees of legislative activity of political parties over time and directly juxtapose this with their other legislative actions (the proposal of bills, for instance) as well as their issue or policy topic attention .

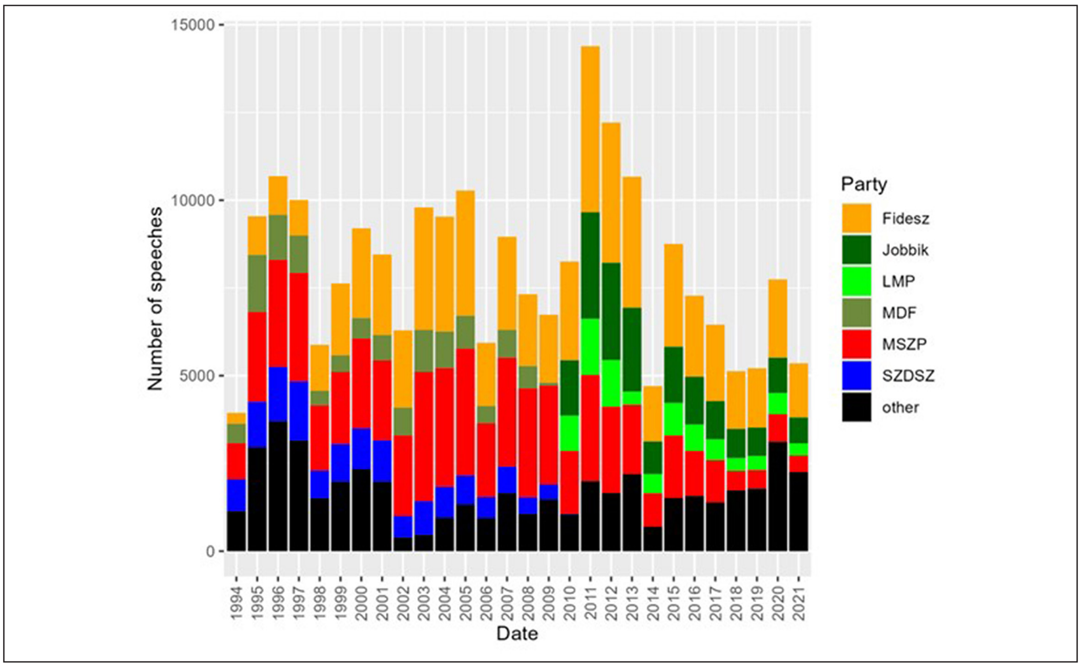


Figure 7 Number of speeches made on each session day by parties (Hungary, 1994–2021)

Note: Figure 7 is a bar plot of the number of speeches held by larger parties (see legend) in the Hungarian parliament between 1994 and 2021. Speeches held by the chair of the respective sessions were excluded, as they regularly do not contain relevant information besides the agenda item titles and names of the next speaker. Observations from 2022 are excluded since it was an election year.

Given the relative scarcity of legislative data for the Central-Eastern European region, ParlText CEE may fill a gap in our understanding of legislative politics by allowing for expanding research designs that mainly focus on the U.S. or Western Europe. It may also serve as a stepping stone for generating new ideas and understanding the unique features of legislative politics in the region. It contributes to pre-existing data collections by applying a wider time frame, larger metadata collection, and a relational database structure for its distinct subcorpora, covering all plenary activities regarding speeches, bills, and laws.

An additional contribution of the project is that all data are linked based on unique identifiers following the ParlLawSpeech standard, allowing each bill to be connected to its adopted final version as law and all the plenary debates that took place in connection with it. Finally, the procedures developed for the 1.0 release of ParlText CEE can be readily replicated with additional countries in the region and beyond with the help of the detailed description of procedures and open-access repository of data and scripts.

In conclusion, we briefly present three potential use cases for the new database that would allow for the extension of branches of the literature to the CEE region. The analysis of gender issues related to legislative politics is a mainstay in the relevant West European and North American literature. Ash et al. (2024) investigated nonverbal reactions during legislative debates. The authors employed a latent Dirichlet allocation on a corpus of German state parliamentary speeches to quantify gender congruency by analysing the relative usage of topics associated with each gender. Their findings suggest a potential bias: although female MPs receive more reactions altogether (both positive and negative) than men, 'women's topics' usually generate less interaction, especially when associated with male MPs.

This topic would make a good subject for comparative research with the inclusion of the CEE region due to its smaller proportion of female MPs. Figure 8 depicts the share of speeches made by male and female speakers in the Hungarian Parliament, highlighting the gender disparity in legislative representation within the region. This visualisation supplements current research by providing empirical data on the gender composition of parliamentary speakers. Such data is essential for understanding the dynamics of parliamentary debates. By examining this ratio, we can explore how the smaller proportion of female speakers may influence legislative processes and interactions, thereby contributing to a more comprehensive analysis of gender issues in legislative studies.

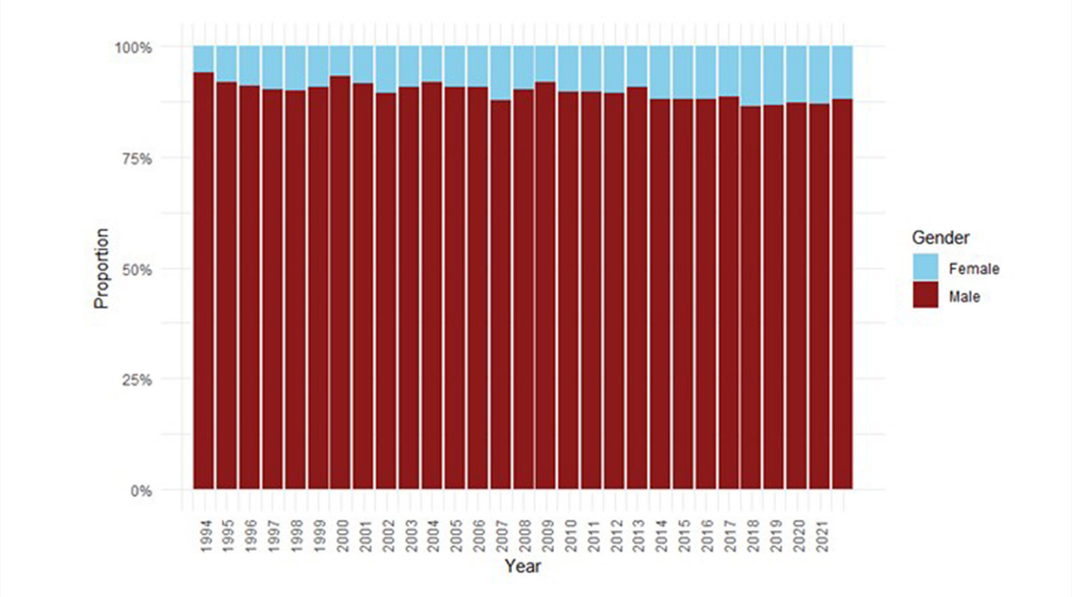


Figure 8 Proportion of speeches made by male and female speakers in the Hungarian Parliament between 1994 and 2022

Note: Figure 8 shows the proportion of speeches made by female and male speakers (including persons who were not MPs in the given electoral cycle) in the Hungarian Parliament between 1994 and 2022. Speeches delivered by the chairs of the respective sessions were excluded to avoid distorting the proportions.

A second theme of interest is environmental and climate policy. Investigating the impact of economic shocks on parliamentary discourse offers valuable insights into policy priorities. Finseraas et al. (2021) employed a difference-in-differences approach with a structural topic model to analyse speeches in oil-dependent Norway during the 2014–2015 oil price shock. They anticipated a decline in environmental discussions in oil-producing regions. However, their findings suggest a continued emphasis on the 'green shift' topic, potentially reflecting a strategic pivot towards green investment. While CEE countries differ in their energy dependence, a similar research design could be illuminating, revealing how, e.g., the dependence on Russian energy shapes parliamentary discourse on sustainability.

A third topic of interest is related to the transformation of European party systems. Schwalbach (2023) examined party interaction in parliamentary debates across Denmark, Germany, Netherlands, and Sweden following the entry of populist radical-right parties (PRRPs). Utilising correspondence analysis and dictionaries, the study focused on daily parliamentary interactions. While his findings suggest limited overall realignment of the traditional government-opposition structure, debates on immigration revealed a significant polarising effect by PRRPs. Given the growing prominence of PRRPs in the CEE region, a similar investigation into the V4 parliaments would be valuable.

Acknowledgements

The research project was supported by the European Union within the framework of the RRF-2.3.1-21-2022-00004 Artificial Intelligence National Laboratory Program, the V-Shift Momentum Project of the Hungarian Academy of Sciences, and the European Union's Horizon 2020 research & innovation programme under Grant Agreement no. 951832. We appreciate the contributions and/or research assistance of Zoé Baumgartner, Levente Csóka, Márton Fadgyas, Tamás Fehér, Kristóf Hadadi, Ákos Holányi, Norbert Huszti, Rebeka Kiss, Viktor Kovács, Richárd Lehoczki, and Norbert Nagy.

Data availability statement

The data collected under the ParlText project and the repository made for the replication of the tables and figures in this article are available at parltext.org.

References

Ash, E., Krümmel, J. & Slapin, J. B. (2024). Gender and reactions to speeches in German parliamentary debates. *American Journal of Political Science,* online first. https://doi.org/10.1111/ajps.12867

Back, H., Debus, M. & Fernandes, J. M. (Eds.) (2022). *The Politics of Legislative Debates.* Oxford University Press. https://doi.org/10.1093/oso/9780198849063.001.0001

Baumgartner, F. R., Breunig, C. & Grossman, E. (2019). The Comparative Agendas Project: Intellectual Roots and Current Developments. In F. R. Baumgartner, C. Breunig & E. Grossman (Eds.) *Comparative Policy Agendas. Theory, Tools, Data* (pp. 3–16). Oxford University Press. https://doi.org/10.1093/oso/9780198835332.003.0001

Bächtiger, A. (2014). Debate and Deliberation in Legislatures. In S. Martin, T. Saalfeld & K. W. Strøm (Eds.), *The Oxford Handbook of Legislative Studies* (pp. 146–167). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199653010.013.0008

Benoît, C. & Rozenberg, O. (Eds.) (2020). *Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures.* Edward Elgar Publishing. https://doi.org/10.4337/9781789906516

Brady, H. E. (2019). The Challenge of Big Data and Data Science. *Annual Review of Political Science, 22*(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

Döring, H. & Regel, S. (2019). Party Facts: A database of political parties worldwide. *Party Politics, 25*(2), 97–109. https://doi.org/10.1177/1354068818820671

Erjavec, T. et al. (2023a). Multilingual comparable corpora of parliamentary debates ParlaMint 4.0. http://hdl.handle.net/11356/1859

Erjavec, T. et al. (2023b) Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0. http://hdl.handle.net/11356/1860

Finseraas, H., Høyland, B. & Søyland, M. G. (2021). Climate politics in hard times: How local economic shocks influence MPs attention to climate change. *European Journal of Political Research*, *60*(3), 738–747. https://doi.org/10.1111/1475-6765.12415

Grimmer, J. & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Grossman, J. & Pedahzur, A. (2020). Political Science and Big Data: Structured Data, Unstructured Data, and How to Use Them. *Political Science Quarterly, 135*(2), 225–257. https://doi.org/10.1002/polq.13032

Joshi, D. & Rosenfield, E. (2013). MP Transparency, Communication Links and Social Media: A Comparative Assessment of 184 Parliamentary Websites. *Journal of Legislative Studies, 19*(4), 526–545. https://doi.org/10.1080/13572334.2013.811940

Jurafsky, D. & Martin, J. H. (2024). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (Third Edition Draft)* [Unpublished Manuscript]. https://web.stanford.edu/~jurafsky/slp3

Kiss, R. & Sebők, M. (2022). Creating an enhanced infrastructure of parliamentary archives for better democratic transparency and legislative research: Report on the OPTED forum in the European Parliament (Brussels, Belgium, 15 June 2022). *International Journal of Parliamentary Studies, 2*(2), 278–284. https://doi.org/10.1163/26668912-bja10053

Kuzman, T. et al. (2023). Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 4.0. http://hdl.handle.net/11356/1864

Martin, L. W. & Vanberg, G. (2014). Legislative Institutions and Coalition Government. In S. Martin, T. Saalfeld & K. W. Strøm (Eds.), *The Oxford Handbook of Legislative Studies* (pp. 437–455). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199653010.013.0014

Mikušek, O . (2024). One Year of Continuous and Automatic Data Gathering from Parliaments of European Union Member States. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024* (pp. 149–153). ELRA; ICCL. https://aclanthology.org/2024.parlaclarin-1.22/

Proksch, S.-O., Rauh, C., Sebők, M., Schwalbach, J. & Hetzer, L. (2024). *The ParlLawSpeech Dataset.* [Unpublished manuscript].

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

Schwalbach, J. (2023). Talking to the populist radical right: A comparative analysis of parliamentary debates. *Legislative Studies Quarterly*, *48*(2), 371–397. https://doi.org/10.1111/lsq.12397

Sebők, M., Proksch, S.-O., Rauh, C., Visnovitz, P., Balázs, G. & Schwalbach, J. (2025). Comparative European legislative research in the age of large-scale computational text analysis: A review article. *International Political Science Review, 46*(1), 18–39. https://doi.org/10.1177/01925121231199904

Slapin, J. B. & Proksch, S.-O. (2014). Words as Data: Content Analysis in Legislative Studies. In S. Martin, T. Saalfeld, K. W. Strøm (Eds.), *The Oxford Handbook of Legislative Studies* (pp. 127–145). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199653010.013.0033

Sylvester, C., Khokhlova, A., Yordanova, N. & Greene, Z. (2024). ParlEE plenary speeches V4 data set: Annotated full-text of 18 million sentence-level plenary speeches of eight European legislative chambers. *Harvard Dataverse,* V1. https://doi.org/10.7910/DVN/TLKVWY

Tremblay-Antoine, C., Jacob, S., Dufresne, Y., Poncet, P., & Dinan, S. (2024). An open window into politics: A structured database of plenary sessions of the European Parliament. *European Union Politics, 25*(3), 605–622. https://doi.org/10.1177/14651165241239637

Wolchik, S. L. & Curry, J. L. (2018). Democracy, the Market, and the Return to Europe: From Communism to the European Union and NATO. In S. L. Wolchik & J. L. Curry (Eds.), *Central and East European Politics. From Communism to Democracy,* 4th ed. (pp. 3–30). Rowman & Littlefield