scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Finnish parliamentary speeches dataset

Salla Simola[1] & Jeremias Nieminen[2,3] ✉ Janne Tukiainen[4]

In this Data Descriptor, we introduce a dataset on Finnish parliamentary speeches and the background characteristics of politicians who speak in the parliament. The database includes all parliamentary speeches in Finland from the beginning of the existence of the Finnish parliament (years 1907-2018). These data have been used in previous studies to analyze questions related to political polarization, effects of media on parliamentary speeches, and political representation. The long time period makes it possible to study how historical changes are reflected in parliamentary speeches.

Background & Summary

Parliamentary speech data can be utilized to study various political phenomena, including partisanship and political polarization[1–3], effects of media on politicians' behavior[4], and political representation[5,6]. Unlike voting data, which is constrained by party discipline, parliamentary speeches can be used to better infer politicians' policy positions. This is important in Finland, as the party discipline is traditionally quite high in Finland[7]. Although party discipline may also impact parliamentary speeches, the parties' control of speeches is likely to be smaller than that of voting, and previous literature has indeed found party discipline to be smaller compared to voting records at least in Switzerland[8].

In this data article, we describe a dataset covering all parliamentary speeches in Finland (years 1907-2018). The data described here has been previously utilized in three separate papers by the authors of this article. First, we have used the data to estimate speech differences between various group in the Parliament utilizing methods[1] developed for estimating group differences in text data. In one study, we estimate the evolution of the differences between left-wing and right-wing parties during the last century in Finland[3]. We find that the differences peaked in the 1970s and were driven by the differing stances of left wing and right wing parties towards the Soviet Union. Although we observe an increasing trend in polarization during the latest decades, the observed levels of polarization are still very modest historically, which highlights the advantage of the long time frame available in the data described in this data article.

There are also other previous applications of our data[4,6]. In one article[4], we use a subset of this data, and complement it with calendar date information, which is not available nor easily collectable for the whole dataset described in this paper. In that paper, we estimate the effects of the presence of TV cameras in the Parliament on political polarization and other outcomes describing politicians' behavior. We study this question in a difference-in-differences setting where we utilize the introduction of TV cameras in 1988 for the first Thursdays of every month (while question hours on other Thursdays were not televised until 2008). The main results from the paper are that the introduction of TV cameras in the Parliament increased the speech differences between government and opposition MPs, while it did not affect the speech differences between left-wing and right-wing parties. In another study[6], in turn, we investigate speech differences between various intra-party groups.

Similar types of datasets have also been used in other countries like Norway[2]. Although there are also other projects that have digitized Finnish parliamentary speeches. See especially the ParliamentSampo project (https://seco.cs.aalto.fi/projects/semparl/en/). The advantage of our data is that our data is also preprocessed and stemmed, and thus "ready for analysis". Our data is also in an accessible and compact format for analyses that want to utilize parliamentary speeches from many different years as opposed to limited searches targeted at specific speeches or speakers, where other sources such as the ParliamentSampo are more useful. In our data, we have both raw and modified OCR output of all parliamentary speeches in Finland, linked to politicians' characteristics. Our data collection efforts began before any other large-scale parliamentary speech datasets from Finland had been introduced.

[1]Wolt, Stockholm, Sweden. [2]Department of Economics, Turku School of Economics, University of Turku, Turku, Finland. [3]Labour Institute for Economic Research (Labore), Helsinki, Finland. [4]Department of Economics, Turku School of Economics, University of Turku, Rehtorinpellonkatu 3, 20500, Turku, Finland. ✉e-mail: jeremias.nieminen@utu.fi

The Section 2 of this paper describes the original data sources used to construct our datasets. Section 3 describes our datasets on parliamentary speeches. This includes, e.g., descriptions of variables available in our datasets. Section 4 outlines possibilities for potential usage of these data.

Methods

Data sources and preprocessing methods. We use openly available pdf records and images of parliamentary minutes as our main source for constructing the dataset. Data for 2001-2018 (https://www.eduskunta.fi/FI/taysistunto/Sivut/Taysistuntojen-poytakirjat.aspx) and 1907-2000 (https://avoindata.eduskunta.fi/#/fi/digitoidut/) are available from the Parliament of Finland. Some of the years have multiple sittings in a single pdf file while other years have separate pdf files for each sitting during the year. The process for transforming OCR output into structured dara files involves multiple steps tailored to different time periods. First, OCR-generated text pages from 1907 to 2015 are combined into single text files for each PDF document. This step consolidates pages in sequential order and inserts a unique tag indicating the beginning of each page. For files from 2016 to 2018, text extraction is conducted using a tool called pdf2text, while similarly tagging the beginning of each page.

Next, tagged versions of these consolidated files are created to aid in speech extraction. Creating "tagged" files means that we tag sections of text using regular expressions such that it is then easier to separate the speeches from all other text. Special characters, numbers-only rows, and hyphenated line breaks are removed. Each parliamentary minute consists of different parts such as speaker lists and the discussion part. Because of this, key sections of the text, such as the start and end of discussions, subtitles, session breaks, and session ends, are marked using regular expressions.

```
starttags = \
 [r"((Keskustelu|Yleiskeskustelu)[\s]*(:|;|\.|\n))",
  r"((Keskustelu|Yleiskeskustelu)[\s]*[^.:;\n]{0,60}[\s]*(jatkuu|jatkui)(:|;|\.|\n))",
  r"((Keskustelua|Yleiskeskustelua)[\s]*(jatkettiin|jatketaan)(:|;|\.|\n))",
  r"((\nKeskustelu|\nYleiskeskustelu)[\s]*(jatkuu|jatkui)[^.:;\n]{1,60}[\s]*?(:|;|\.|\n))"
 ]
```

Tags for speakers are also created. For example, speaker tags are created in the following way using regex:

```
speakertags = [
  r"([A-ZÅÄÖ][\w\-.]* [\w\-.]* [\w\-]*[\s]*/[\w\-]*[ \-\(]*[\w ]*[\)]*:)",
  r"([A-ZÅÄÖ][\w\-]*inisteri [A-ZÅÄÖ][\w\-.]*[\s]*[\w\-]*[\s]*[\w\-]*[\s]*[\(]*[\s]*[\w]*[\s]*[\)]*:)",
  r"([A-ZÅÄÖ][\w\-]* ja [\w]*inisteri [A-ZÅÄÖ][\w\-.]*[\s]*[\w\-]*[\s]*[\w\-]*[\s]*[\(]*[\s]*[\w]*[\s]*[\)]*:)",
  r"\n[ ]*(Puhuja:)",
  r"NEWPAGE_\d+[ ]*(Puhuja:)",
  r"(Puhuja \(korokkeelta\):)"
]
```

As the OCR output contains some errors, we have a separate step where the speaker's name is matched to the closest match among the members of parliament for the given term. Additionally, we have manually taken into account a large number of name changes.

In the third step, text files are split into individual speeches (using the tags created earlier) and saved in CSV format. The process identifies and separates discussion sections and question hours. Speech sections are organized with details like section start pages, and each speaker's speeches are separated using speaker-specific tags. Parenthetical content is removed and page breaks are replaced with whitespace characters.

The process is the following: 1.) Split into speech sections based on DISCUSSION_STARTS - DISCUSSION_ENDS tags. If other implicit speech end tags found (SESSION_BREAK, SESSION_ENDS, SUBTITLE), split on them. Split into question hours based on QUESTION_HOUR - SESSION_ENDS tags. 2.) Find speechsection startpage number. 3.) Print section to file, 4. Remove parenthetical additions from text and save to file with filename and speechsection startpage nr, 5. Split to speeches based on SPLIT SPEAKER/SPEAKER CHAIR tags, 6. Replace pagebreak with XXX and other whitespace characters with " ".

The final step involves cleaning and preprocessing each speech. Initial lines without speaker tags, date rows, page numbers, and procedural text are removed. Speaker names are isolated from the rest of the text. Further preprocessing includes language detection, symbol expansion, conversion to lowercase, removal of stopwords, and stemming to normalize words. This preprocessing steps make the data suitable for analysis and modeling. The final data has both the *raw* and *cleaned* speeches. Including the raw speeches makes it possible for users to conduct their own preprocessing steps if they wish to do so.

The share of speeches successfully retrieved by the automated script is well above 90 percent for most decades in an audit of a randomly selected subset of transcripts. This manual audit was conducted by a research assistant at the early stages of the project (in October 2019). The script has been updated and improved since, and thus, the number may be interpreted as a lower bound for the coverage of the data.

Institutional setting. The Finnish parliamentary system is distinctive in its very candidate-centered system that sets it apart from many other democracies. The system is described by many authors[9,10]. The Finnish parliament, known as the *Eduskunta*, operates under an open-list electoral system. Finland's system has been described

Column Name	Description	Example Value
year	parliamentary year	1907
file	the pdf source file for the speech	PTK_1907_II
speaker	speaker name as extracted from the text	Ed. Kirves, Ed. Palmön
corr_name	suggested spelling correction for speaker name; empty if correct	, Palmén
speaker_id	speaker id in mps-ministers.csv	753, 1431
speech_section_ startpage	Start page of the speech in the pdf counterpart	120
language	Speech language as detected by langdetect library	fi
speech_raw	Speech without symbols and with line breaks replaced by "XXX"	SPEAKER Ed. Kirves: Minä olisin hyvällä syyllä toivonut, ettäXXXkaikki ne asiakirjat, jotka koskevat elinkeinolain uudistuksia,XXXtoisin sanoen myöskin entinen
speech_clean	Lowercased and stemmed speech without stopwords, line breaks, symbols and punctuation	hyvä syyl toivonu kaik asiakirj koskev elinkeinol uudistuks tois sanoe myösk entin

Table 1. Description of Parliamentary Speeches Dataset.

as being among the most candidate-centered systems in the world, as MPs in Finland have considerable freedom to express their views[9]. This flexibility extends to parliamentary debates, where MPs are largely free to speak without party restrictions.

The Finnish parliament underwent dramatic changes starting in 1907, when Finland adopted a unicameral legislative model. This shift also gave the right to vote for almost 90% of the adult population, a significant leap from the restricted electorate in the 1905 elections[11,12]. Despite the broad representation, the new parliament's powers were initially limited as any legislative proposals required the approval of the Russian Emperor. During World War I, this dynamic tightened further, with no parliamentary meetings held in 1915 or 1916 due to increased Russian control[11].

The question hours introduced in 1960s-designed to make plenary discussions more dynamic-impose strict time limits on speeches. Similarly, debates, added to the plenary repertoire in 2012, are also subject to time restrictions. In these contexts, the Speaker of Parliament has discretion over speaker selection[10]. Otherwise there are no limits on speech length[3].

Plenary sessions have grown in prominence over time[9]. Especially opposition MPs, party leaders, and smaller parties often take advantage of plenary sessions[9]. Today, these sessions are less about shaping legislative content and more about reaching voters and gaining media attention[13]. Indeed, the media plays a crucial role in amplifying the reach of parliamentary speeches. Finland's first live radio broadcast of a plenary session occurred in 1926, but regular radio coverage started much later[3]. Television followed, with the first televised session in 1960, and consistent broadcasts began in the 1980s. By the 2010s, plenary sessions continued to attract significant viewership, with televised debates often reaching hundreds of thousands of Finns[3]. For a deeper exploration of Finland's political institutions there are sources that delve into that[14,15].

Data Records

Dataset on parliamentary speeches. The dataset is available at Figshare (reference number: 28028732), with this section being the primary source of information on the availability and content of the data being described. Please find the privacy statement and impact assessment for the data described and research projects using this data (https://sites.utu.fi/intrapol/data/, in Finnish). Table 1 describes the dataset on parliamentary speeches. The dataset[16] is deposited in Figshare. The data is for years 1907-2018, except for years 1915 and 1916 when the Parliament did not gather. The column year identifies the parliamentary year in which each speech was delivered. The file column provides references to the PDF source files containing the speeches, serving as archival identifiers for locating the original records. The speaker column lists the names of individuals delivering the speeches, as extracted from the text. The column corr_name offers suggested corrections for speaker names if any discrepancies or errors are detected in the original text. The speaker_id column links speakers to unique identifiers such that they can be linked to speaker characteristics data.

The speech_section_startpage column specifies the starting page of each speech within the corresponding PDF document, aiding in pinpointing the exact location of the text within the original pdf files. The language column identifies the language of each speech (Finnish or Swedish) using a language detection library (langdetect).

For textual analysis, the dataset includes both raw and processed versions of the speech content. The column speech_raw includes OCR output where line breaks are replaced by a placeholder "XXX" and symbols have been removed. The speech_clean column provides a preprocessed version of the speech, where text has been lowercased, stemmed, and stripped of stopwords, punctuation, and line breaks, offering a format readily usable for text analysis.

Dataset on speaker characteristics. Table 2 describes the dataset on speaker characteristics. This dataset offers valuable metadata that can be used to examine the demographic dimensions of politics, such as the descriptive representation of different groups in politics.

The column *year* specifies the year of the parliamentary term, while *term* indicates whether it is a full parliamentary term or only spring, fall, or first term. The *minister_term* column records the ministerial term for

Column Name	Description	Example Value
year	Year of the parliamentary term	1930
term	Parliamentary session type	second, full
minister_term	Term of the minister if applicable, often empty if not relevant	(empty)
speaker_id	Unique ID for the speaker	0
full_name	Full name of the speaker, including all first and last names	Mikkel Emil Mikko Eemeli Aakula
last_name	Speaker's last name	Aakula
first_names	Speaker's given names	Mikkel Emil Mikko Eemeli
party	Political party (parliamentary group) affiliation	Maalaisliiton eduskuntaryhmä
dates	Date range of the speaker's term in the parliamentary group	21.10.1930 - 31.08.1936
titles	Any titles or honorifics, if applicable	(empty)
female	Indicator if the speaker is female (1 for female, 0 for male)	0
birthplace	Speaker's place of birth	Kokemäki
electoral_districts	Electoral districts associated with the speaker, including date ranges	Turun läänin pohjoinen vaalipiiri 21.10.1930 - 31.08.1936
first_district	The primary or initial electoral district	Turun
education	Speaker's education, open field	kansakoulu, maanviljelyskoulu 1903
profession	Speaker's profession, open field	maanviljelijä, kunnallisneuvos
birthday	Speaker's birthday	12.06.1879
birthyear	Speaker's birth year	1879.0
startyear	Start year for MP's current term in the parliamentary group	1930.0
endyear	End year for MP's term in the parliamentary group	1936.0
minister	Indicator if the speaker held a ministerial position (1 for yes, 0 for no)	0
coalition	Coalition groups the speaker was affiliated with during the term	Maalaisliiton eduskuntaryhmä, Kansallisen kokoomuksen eduskuntaryhmä
govparty	Indicator if the speaker's party was part of the governing party (1 for yes, 0 for no)	1
govseats	Number of seats held by the government coalition in that session	134
pmparty	Indicator if the speaker's party was the prime minister's party (1 for yes, 0 for no)	0
partyname_el_data	party name from election data attached to the main data	Maalaisliitto
voteshare_el_data	party vote share from last election (from election data)	12,60
seats_el_data	number of seats in last election (from election data)	20
seatshare_el_data	party seat share in last election (from election data)	0.2
tenure_pg	number of years served consecutively in the parliamentary group	6
experience	number of years served in the parliament (in any group, at any time before)	6

Table 2. Description of Columns in Parliamentary Speaker Dataset.

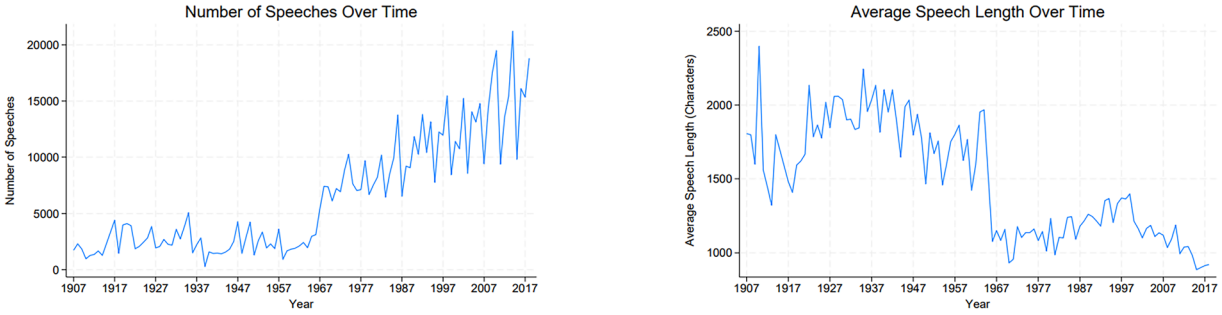
speakers who held ministerial positions, though this is often empty for non-ministers. Each speaker is assigned a unique identifier in the *speaker_id* column. This can be used to link these data on speaker characteristics to the parliamentary speech dataset.

The *full_name* column records the complete name of the speaker, including all first and last names, while *last_name* and *first_names* separate these components for ease of analysis. The *party* column identifies the speaker's political party affiliation, and *dates* captures the duration of the speaker's parliamentary service in the current parliamentary group. Titles or honorifics, if applicable, are noted in the *titles* column.

The dataset includes a binary indicator in the *female* column, specifying whether the speaker is female. Geographic data is provided in the *birthplace* column, which records the speaker's place of birth, and *electoral_districts*, which details the electoral districts represented by the speaker, including associated date ranges. The *first_district* column highlights the initial electoral district for each speaker.

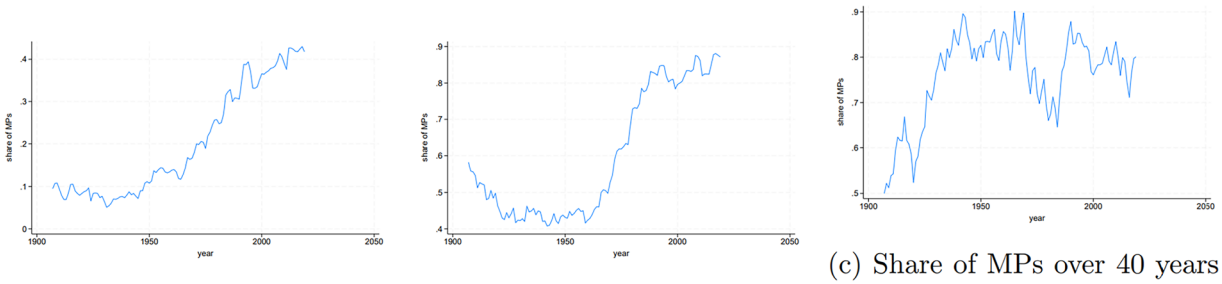
Political roles and affiliations are captured through several columns. The *minister* column indicates whether the speaker held a ministerial position, while *coalition* lists coalition groups the speaker was affiliated with during the parliamentary term. The *govparty* column specifies whether the speaker's party was part of the governing coalition, and *govseats* provides the number of seats held by the government coalition in the session. The *pmparty* column identifies whether the speaker's party was the prime minister's party during the term.

The dataset includes also variables matched from election results data. These variables are *partyname_el_data*, *voteshare_el_data*, *seats_el_data*, *seatshare_el_data*. The election results data are recorded on party level while parties in our data are parliamentary groups. As sometimes parliamentary groups split or merge, sometimes the parliamentary group MP belongs to is not the same they were members of when elected. Thus, we cannot match all speaker-year observations with the latest election results. We can match around 90 % of speaker-year observations with past party-level election results.



(a) Number of speeches (b) Length of speeches

Fig. 1 Number of speeches and speech length.



(a) Female share of MPs (b) Highly educated MPs share (c) Share of MPs over 40 years old

Fig. 2 Characteristics of MPs.

The main reason for this discrepancy is that the politicians' parliamentary group is one that did not participate in the elections. In addition, there is a reason related to minor issues in the original data source. Namely, for parliamentary groups that have changed their names at some point, the original MP characteristics data retrieved from the Parliament of Finland sometimes contains only the new version of the party name for those MPs that have been in the Parliament during both the new and old party name eras. The most notable example would be the Centre Party (Suomen Keskustan eduskuntaryhmä) which used to be called Agrarian League (Maalaisliitto). MPs who served in Parliament during both eras have been classified as belonging to the Centre Party, even during the period when the party was called the Agrarian League. As we have done in our other papers[3,4], when running analyses, it is possible to simply consider various versions of the same party (e.g., Centre Party and Agrarian League) to be the same party in order to overcome this issue.

Finally, we have also calculated variables describing politicians' experience. The first of these, *tenure_pg* indicates tenure in the parliamentary group, and *experience* indicates how many years the politician has been in the Parliament in any parliamentary group at any time before the particular year.

It would also be possible to link our data to ParlGov and PartyFacts codes. However, implementing this would involve several steps that are better left to the researchers' discretion. Firstly, our data is at the speaker-year level, so there would be multiple alternatives regarding, e.g., how to link cabinet-level codes to our data. For instance, in years with multiple cabinets, one might choose to link the last cabinet of the year or perhaps instead the one that served the longest during the year. Additionally, mapping parties from the ParlGov data to our dataset would require researchers to make decisions, such as whether to treat parties and their predecessors as the same entity when linking to the ParlGov codes.

Technical Validation

We plot the trends in the number of speeches and the average length of speeches over time in Fig. 1. In 1967, the government question hour was introduced. As shown in Panel a of Fig. 1, there is jump at 1967 in the number of speeches and the number of speeches per year has been increasing since then. At the same time in 1967, we also observe that the average speech length went down substantially (Panel b), also consistent with the introduction of the question hour. Thus, the data is consistent with this big known change made to the parliamentary system.

Regarding the dataset on politicians' characteristics, Fig. 2 confirms what is already known to be true, namely, that both the share of female MPs (Panel a) and the share of "highly educated" MPs (Panel b) have substantially increased during the last century, thus validating the data. Panel c shows that the majority of MPs have always been older than 40 years old.

Usage Notes

These data have been used in previous studies to analyze questions related to political polarization[3], effects of media on parliamentary speeches[4], and political representation[6]. The long time period makes it possible to study how historical changes are reflected in parliamentary speeches.

Code availability

Stata code to produce the figures in this article is included in the repository for this dataset.

Received: 20 December 2024; Accepted: 23 April 2025;
Published online: 23 June 2025

References

1. Gentzkow, M., Shapiro, J. M. & Taddy, M. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica* **87**, 1307–1340, https://doi.org/10.3982/ecta16566 (2019).
2. Fiva, J. H, Nedregård, O., & Øien, H. Group identities and parliamentary debates. Working paper available at https://www.jon.fiva.no/docs/FivaNedregardOien.pdf (2024).
3. Simola, S, Nieminen, J., & Tukiainen, J. A Century of Partisanship in Finnish Political Speech. *Aboa Centre for Economics Discussion Paper No. 160* https://ace-economics.fi/kuvat/dp160.pdf (2023).
4. Nieminen, J, Simola, S., & Tukiainen, J. Effects of increased transparency on political divides and MP behavior: Evidence from televised question hours in the Finnish Parliament. *Legis. Stud. Q.* https://doi.org/10.1111/lsq.12439 (2023).
5. Lippmann, Q. Gender and lawmaking in times of quotas. *J. Public Econ*. **207** https://doi.org/10.1016/j.jpubeco.2022.104610 (2022).
6. Nieminen, J, Simola, S., & Tukiainen, J. Evolution of within-party group differences: Evidence from 110 years of parliamentary speech. *Aboa Centre for Economics Discussion Paper No. 161*, https://ace-economics.fi/kuvat/dp161.pdf (2023).
7. Pajala, A. Government vs opposition voting in the Finnish parliament Eduskunta since World War II. *Eur. J. Gov. Econ.* **2**, 41–58 (2013).
8. Schwarz, D., Traber, D. & Benoit, K. Estimating intra-party preferences: Comparing speeches to votes. *Polit. Sci. Res. Methods* **5**, 379–396 (2017).
9. Poyet, C. & Raunio, T. Reconsidering the electoral connection of speeches: The impact of electoral vulnerability on legislative speechmaking in a preferential voting system. *Legis. Stud. Q.* **46**, 1087–1112, https://doi.org/10.1111/lsq.12314 (2021).
10. Poyet, C., & Raunio, T. Finland: Legislative speechmaking in a changing parliament. In Bäck, H., Debuc, M., & Fernandezn, J. M. (eds.) *The Politics of Legislative Debates*, 329–350 (Oxford Univ. Press, 2021).
11. Paloheimo, H. Eduskuntavaalit 1907-2003. In Ollila, A. & Paloheimo, H. (eds.) *Kansanedustajan työ ja arki. Suomen eduskunta***100**, 173-369 (2007).
12. Jyränki, J. & Nousiainen, A. *Eduskunnan muuttuva asema*. (Edita, 2006).
13. Pekonen, K. *Puhe eduskunnassa*. (Vastapaino, 2011).
14. Karvonen, L. *Parties, Governments and Voters in Finland: Politics under Fundamental Societal Transformation*. (ECPR Press, 2014).
15. Karvonen, L, Paloheimo, H. & Raunio, T. *The Changing Balance of Political Power in Finland*. (Santérus Academic Press, 2016).
16. Simola, S., Nieminen, J. & Tukiainen, J. Finnish parliamentary speeches dataset. *figshare* https://doi.org/10.6084/m9.figshare.28028732 (2025).

Acknowledgements

We thank Ada Grönlund for collecting election data variables added to our main data. This research is funded by the European Union (ERC, INTRAPOL, 101045239). Views and opinions expressed are only those of the authors, however, and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

Salla Simola originally started the project and cleaned the data. Jeremias Nieminen checked, augmented and improved some of the codes. All authors contributed equally to the drafting and editing of this Data Descriptor manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.N.