The *ParlSpeech* V2 data set
Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies

CHRISTIAN RAUH
WZB Berlin Social Science Center
christian.rauh@wzb.eu
www.christian-rauh.eu

JAN SCHWALBACH
University of Cologne
schwalbach@wiso.uni-koeln.de

June 13 2020

For using any of the described data in parts or in full or for quoting this note, please cite the resource as:

Rauh, Christian and Jan Schwalbach (2020) 'The ParlSpeech V2 data set: Full-text corpora of 6.31 million parliamentary speeches in the key legislative chambers of nine representative democracies, *Harvard Dataverse*, V1, https://doi.org/10.7910/DVN/L4OAKN.

Acknowledgements

We are particularly indebted to Pieter De Wilde who was strongly involved in the original *ParlSpeech* data collection (Rauh *et al.* 2017) and supported it also through the participation in the Erasmus Academic PADEMIA network, funded by the European Union Lifelong Learning Programme. Furthermore, we are extremely grateful for the diligent research assistance of Jakob Angeli, Kris Best, Niklas Mäkinen, and Pavel Šatra who supported the work presented here with their local knowledge and adapting some of the scraping routines, and with assembling missing meta-data from external sources. All remaining errors are, of course, our own.

1. Introduction

Politics happens persistently in and through some form of text. Modern political science thus invests heavily in methods that can extract systematic data from large text corpora (for overviews see e.g. Cardie and Wilkerson 2008; Grimmer and Stewart 2013; Monroe and Schrodt 2008). Paralleling these methodological advances, computational power has increased and manageable software implementations of state-of-the-art text analysis algorithms proliferate (see especially Benoit *et al.* 2018).[1] The comparative analysis of political text corpora is furthermore facilitated by the increasing quality and availability of automated translation tools (e.g. De Vries *et al.* 2018; Lucas *et al.* 2015; Proksch *et al.* 2018). In result, exploiting the wealth of empirical information contained in large political text collections has become much more tangible for applied researchers.

However, realizing this potential strongly hinges on data availability. While more and more political text is available online in principle, bringing the various, often only rather loosely structured sources into a machine-readable format that is readily amenable to automated analysis still presents a major hurdle. Collecting and cleaning relevant political texts from different sources, and also annotating them consistently with relevant meta-data can quickly consume the time and resources that are available in a typical applied project.

Against this background we aim to contribute to the public availability of annotated full-text vectors that offer relevant empirical information for a broad range of political science research questions (for other efforts in this vein, see Baturo *et al.* 2017; Gijs Schumacher *et al.* 2016; Merz *et al.* 2016). With this note we thus release full-text vectors and meta-data of more than 6.3 million parliamentary speeches held in the key legislative chambers of Austria, the Czech Republic, Germany, Denmark, the Netherlands, New Zealand, Spain, Sweden, and the United Kingdom, covering periods between 21 to 32 years up until recently.

This update extends the original ParlSpeech data collection (Rauh *et al.* 2017) which originated in a project studying the partisan salience of European Union affairs in domestic plenary debates (Rauh 2015; Rauh and De Wilde 2018) In comparison to the first version, additional speeches were collected in the context of Schwalbach's dissertation project analysing the strategic behaviour of parties in national parliaments and we also wanted to cover the Brexit period in the House of Commons as well the period after the entry of AfD to the German Bundestag. These particular research interests initially drove the selection of countries and time periods covered. But given the amount of speeches, the rather broad geographical and temporal coverage, as well as the political importance of parliamentary debates in representative democracies, we think that the corpora are

[1] See also https://cran.r-project.org/web/views/NaturalLanguageProcessing.html (last accessed: 10.03.2020)

also highly useful for other researchers interested in questions of representation, partisan politics or political text analysis more generally.

The remainder of this note thus aims at allowing interested researchers a quick assessment of the provided resources. Section 2 briefly describes the scraping and splitting routines by which the speeches were collected and assembled into the corpora. Section 3 provides a guide to the individual corpora, lists the contained variables and briefly discusses the vectors of speech texts.

2. Data collection

Having selected relevant countries along our original research goals, data collection focussed on the domestic parliamentary chamber with most competences in legislation and government formation. In result, the *ParlSpeech* data set covers the Austrian *Nationalrat*, the Czech *Poslanecká sněmovna Parlamentu*, the Danish *Folketing*, the Dutch *Tweede Kammer,* the German *Bundestag*, the New Zealand *House of Representatives*, the Spanish *Congreso de los Disputados*, the Swedish *Riksdag*, as well as the UK *House of Commons.*[2]

For each of these parliaments we identified the most encompassing online database that provides digital full-text access to the plenary debates in the respective chamber. These online sources were then scraped with scripts customized to the structures and formats of the respective database. Afterwards, a second set of scripts – again customized to the specific formats supplied – automatically removes boilerplate from the downloaded material, splits it into individual speeches, and annotates the resulting observations with a time stamp, the full speaker's name, his or her party membership and/or specific parliamentary roles, a running speech number within day, and partially information on the agenda item under which a given speech was held (more information on the variables in section 3 below). Table 1 lists the entry points to the databases we used.

[2] Note that the Finish *Eduskanta* is not part of the updated *ParlSpeech* data set. Upon careful inspection we discovered OCR errors in the pdf source files as well as related errors in the routines distinguishing individual speeches therein. Without a Finish native-speaker, however, we are unfortunately not in the position to correct these errors reliably. Researcher interested in the Finish debates may still refer to ParlSpeech V1 (Rauh *et al.* 2017) and we are happy to share our extant data cleaning routines.

Parliament	URL of entry point	Scrape date
AT: *Nationalrat*	https://www.parlament.gv.at/PAKT/STPROT/	26.11.2019
CZ: *PSP*	http://www.psp.cz/eknih/index.htm	11.05.2016
DE: *Bundestag*	http://dipbt.bundestag.de/doc/btp	25.05.2014
	https://pdok.bundestag.de/index.php?start=pp	22.01.2019
DK: *Folketing*	https://www.ft.dk/da/dokumenter/dokumenter-foer-2004_05	24.09.2017
	https://www.ft.dk/da/dokumenter/dokumentlister/referater	27.11.2019
ES: *Congreso*	http://www.congreso.es/portal/page/portal/Congreso/Congreso/Publicaciones	05.11.2019
SE: *Riksdag*	https://www.riksdagen.se/sv/dokument-lagar/?doktyp=prot	24.08.2019
NL: *Tweede Kamer*	https://zoek.officielebekendmakingen.nl/uitgebreidzoeken/parlementair	22.08.2019
NZ: *HoR*	https://www.parliament.nz/en/pb/hansard-debates/historical-hansard/	18.05.2018
	https://www.parliament.nz/en/pb/hansard-debates/rhr/	14.08.2019
UK: *HoC*	http://www.parliament.uk/business/publications/hansard/commons/by-date	22.01.2016
	https://hansard.parliament.uk/search/Debates	06.01.2020

Table 1: Scraped online speech archives

Speeches in the Austrian *Nationalrat* were collected from the archive on the parliament's website. The debates are stored in a user-friendly structured pdf format until 1996. Earlier protocols are also available until 1920. However, these are scanned originals, prone to OCR errors.

For the Czech *Poslanecká sněmovna Parlamentu* we scraped html versions of the plenary protocols from the documentation website of the Chamber of Deputies. Before splitting and cleaning them, the raw html files were externally converted to UTF-8 encoding. Speaker and party annotation rested on external lists generated by scraping the individual sub pages of Czech MPs.

We scraped the plenary protocols of the German *Bundestag* originally as plain txt or asc files from a parliamentary ftp server and then used a set of regular expressions as well as external MP lists to split the stenographic debate protocols along formatting regularities observed in the textual structure. The information for the additional speeches in ParlSpeech V2 (date > 2014-09), however, were harvested from the pdf documents that are available on the pdok website of the Bundestag. Researchers wishing to extend the data further should note that the *Bundestag* offers xml access in the meantime.

In ParlSpeech V1, the Spanish data were scraped from the website of the *Congreso* where all protocols are available as pdf files since 1977. However we realised OCR errors in the earlier records and resort to txt files provided on the *Congreso* website for the updated data presented here. These are of much better quality, but are only available backwards until 1996.

The Swedish plenary protocols were accessed via the website of the *Sveriges Riksdag*. Besides various other documents the website provides all parliamentary debates since 1971 in different file formats,

such as txt or html. Due to the clear and user-friendly structure of these files, cleaning and splitting them into individual speeches was comparatively straightforward.

Parliamentary records from the *Tweede Kamer*, at least after 1994, were accessed as already very well-structured xml files provided through the central online access point of the Dutch government.[3] Parsing and meta-data collection relied on the provided xml tags.

Speeches in the New Zealand *House of Representatives* were scraped from the official website as well as the linked historical Hansard. All debates until 2003 are stored in pdf format that is well transformable into txt files. For all earlier speeches, however, it must be noted that the individual days of debate are combined in so-called volumes. These volumes include all speeches of about one month and were first separated into individual days before being processed further.

Finally, the UK *House of Commons* debates were accessed via the digital Commons Hansard which conveniently offers plenary protocols as comparatively well-structured html documents which facilitates the extraction of speech texts and relevant metadata. However, since the original ParlSpeech data collection, the website has been completely remodelled. For the update presented here we have re-scraped all data from this new source but realized that a number of speeches (including a.o. full debate days in the period October 2004 to May 2006) are missing in the revamped online archive. Having verified that these speeches are truly missing along the records offered by www.theyworkforyou.com (last accessed: 28.02.2020), we filled the identified gaps with the corresponding speech from the ParlSpeech V1 release (Rauh *et al.* 2017).

All scraping and processing routines were implemented in the R environment (R Core Team 2019),[4] resorting mainly to the base text processing capabilities[5] as well as the *stringr* (Wickham 2015), *rvest* (Wickham 2016), *RCurl* (Temple Lang and R Core Team 2016), and *XML* (Temple Lang and CRAN Core Team 2015) packages.

With regard to data quality, we took several steps to ensure that the provided data frames match the information provided in the respective parliamentary archives. Before being scaled up, the scraping and splitting routines were developed and tested on individual elements distributed across the whole coverage of the respective parliamentary archive. Also all further steps involved an intense back-and-forth between the online resources and the raw texts on the one hand and the

[3] With a country specialist we pre-selected only those items that contain political debates on the floor, excluding, e.g., practical announcements by the President of Parliament ('Mededelingen'), order of the agenda and planning of activities ('Regeling van werkzaamheden'), celebrations and commemorations ('Beëdiging(en) van', 'Afscheid', 'Herdenking') or closure ceremonies ('Sluiting').
[4] Munzert et al. (2015) offer a didactically very useful introduction to interacting with common web technology from within R.
[5] https://en.wikibooks.org/wiki/R_Programming/Text_Processing (last accessed: 11.03.2020)

resulting data frames on the other. Finally, a range of random sample debate comparisons per country verified that raw debate protocols and text vectors are identical with regard to speech content and speaker characteristics in the protocols each parliament provides.

3. Data description

3.1 Corpora and speech availability

The resulting speech and meta-data vectors are stored as one corpus per parliamentary chamber. We provide them as .rds files (save version 3, `R` >= 3.5.0) for three reasons. First, this is the format we work with in our own substantial applications. Second, given the large size of the corpora, saving them as serialized files with gzip compression seems reasonable. Third, the format provides no significant barriers as R is free and open-source software that has multiple options to locally export the data into the preferred format.[6]

Table 2 below gives an overview of the corpora contained in the *ParlSpeech* V2. It should initially be noted that periods for which full text is available varies in the scraped data sources. Thus, start and end point of the covered period differs over the nine parliaments. Rather than harmonising the investigation period we opted for publishing as many speeches as could be fetched along the above described procedures.[7] Within these temporal constraints, however, the provided text vectors represent *a census of plenary debate in the given parliament*. Of course, we cannot completely rule out technical errors in the original database sources (except for removing clear duplicates). Yet, during numerous random sample checks and visual analyses of the temporal speech distributions (Figure 1 below) we did not find any unusual breaks except for typical low periods during summer, Christmas and government formation.

Interested researchers should furthermore note the cross-national variation in the number of available speeches. Technically an individual speech (row) in our corpora refers to a continuous bit of text that was spoken on the plenary floor and that was clearly assigned to one and only one individual speaker as marked in the respective data source. The resulting variation in the number of such speeches thus has three possible sources. First and most importantly, it reflects mainly different parliamentary traditions and operations. While the ideal-typical 'talking parliament' in the

[6] The `R` software environment can be acquired via https://www.r-project.org/ (last accessed: March 11 2020). A guide for importing/exporting data into and from R can be found at https://cran.r-project.org/doc/manuals/r-release/R-data.html (last accessed: March 11 2020).

[7] For the Tweede Kamer and the House of Commons, e.g., older records exist (see for example Eggers and Spirling 2014; Marx and Schuth 2010), but are mostly supplied in different databases and/or different formats. Setting up corresponding scraping and splitting scripts, however, exceeded the resources and our own research purposes.

UK handles most of its business through many (often rather interactive) speeches on the plenary floor, most of the policy work in 'working parliaments' such as the German Bundestag happens in committees while the plenary floor is reserved to fewer, prepared, but also longer speeches. Second, variation in the number of speeches is most likely driven by differing number of partisan factions with speaking rights in the respective parliaments (on the politics of speech-making see e.g. Bäck *et al.* forthcoming; Proksch and Slapin 2015). Third, some nuisances of protocolling have to be noted. While the House of Commons protocols treat interventions and interruptions from the floor as separate speech acts, similar actions are hidden within the text of the main speaker in the German Bundestag (and were removed in the original data collection by Rauh 2015). Depending on the specific research questions and theory, aggregations or filtering of individual speeches may be advisable.

3.2 Variables

All nine corpora have an identical column structure with 11 variables. The variable **`date`** stores the day of speech delivery in a character vector (YYYY-MM-DD). For researchers wishing to study interaction patterns, the variable **`speechnumber`** stores the running order of speeches within the respective day. Along these two variables, the original temporal order of speeches can be restored within each corpus.

The column **`speaker`** holds a character vector with the full name of the person having given the respective speech as provided in the official protocol or other parliamentary sources.[8] Researchers wishing to exploit this variable to match individual level data should consider that speaker names might slightly diverge from external lists, e.g. with regard to nicknames, middle initials or titles.

The variable **`party`** contains a character vector storing the partisan faction[9] of the speaker as given in the plenary protocol or matched by external, time-stamped MP lists. We have already harmonized slight temporal differences in the protocols' references to the same party.[10] Researchers wishing to match party-level data along this character variable should be also aware of possible discrepancies between party abbreviations along parliamentary conventions and those of external

[8] The UK House of Commons debates posed a particular challenge in this regard as they often only give the surname of a speaker (especially if she or he has held a speech on the same day already) which are often not unique. We have tackled this issue already along two steps. First, based on external MP lists we set up an algorithm that looks backwards and fills in the full name of speakers appearing twice or more often within a session. Second, we automatically corrected short-form names if the respective surname was unique among the MPs in office at the day of the respective speech. For (only) 902 case we could not uniquely identify the speaker and store solely the given surname (in consequence also leaving the party identifiers empty in these cases).

[9] In the German case, CDU and CSU form one faction. Researchers wishing to distinguish these two parties should recode the party variable along external MP lists. Note that PartyFacts provides a unique id (211) for this faction, matching the Manifesto data.

[10] As, for example, with regard to the German 'PDS/LINKE' which has also figured as "Die Linke", "LINKE" or "PDS", respectively.

lists[11]. However, we facilitate linkage with party-level data also along the **party.facts.id** variable. For most parties in our corpora, it points to the numeric identifier for the respective party in Döring and Regel's (2019) Party Facts database which in turn offers straightforward linkages e.g. to the Chappel Hill Expert surveys (Bakker *et al.* 2015), the Manifesto Data (Volkens *et al.* 2014), or the ParlGov database (Döring and Manow 2018).

The variable **chair** is a Boolean vector indicating whether the speech was given by the contemporaneously acting chairperson of the respective parliament or his/her step-ins (as, for example, the *Speaker* of the House of Commons, the *Talman* in the Swedish Riksdag, or the *Bundestagspräsident* in Germany). These speeches virtually always serve for debate organisation only. Note that the protocolling conventions in some of the parliaments do not provide party information and sometimes also no speaker-name information of the respective chair.

In many applications, speech length might be a useful filter or normalization benchmark. Thus, the variable **terms** stores the overall length of the speech text calculated as the number of continuous symbol chains separated by whitespaces.

The variable **agenda** indicates the name of the agenda item under which an individual speech was held, again as provided in the respective parliamentary archive. This variable was not part of the original ParlSpeech data collection but is geared to assist scholars wishing to select specific (types of) debates (see, for example, Proksch *et al.* 2018) or to examine debate-specific word frequency distributions (see, for example, Lauderdale and Herzog 2016). Column 8 in Table 2 indicates the availability of the agenda information. For technical reasons mostly lying in the structure of the respective online databases, we can unfortunately not provide this information for all speeches in the corpora.[12] For instances in which researchers require additional information on the agenda, our best possible advice is to match speeches along the date and speaker order to external sources on the parliamentary agenda.

Finally, the variable **text** stores the raw text of the speech as given in the original plenary protocol. We have only removed boilerplate such as headers, footers, page numbers, and formatting markers as well as motions, bill texts, or order papers that were not read out aloud on the plenary floor. Storing the full text of speeches in partially rather long strings results in large files, but is driven by

[11] For the Spanish case, the variable indicates the parliamentary faction. For many large parties, this is identical to the respective party. However, the parliamentary factions can also consist of alliances of up to three parties. In this case no party.facts.id is available.

[12] Note that the House of Commons has some minor gaps in this regard, due to missing speeches in the new online archive (see section 2 above). For all speeches sourced from ParlSpeech V1, agenda information in the House of Commons corpus V2 is unfortunately not available. We hope that the Hansard Archives may fill these gaps again and thus refer the interested user to the daily HoC agendas accessible via https://hansard.parliament.uk/commons/ (last accessed: March 11 2020).

the desire to allow other researchers as much methodological choice as possible. On the one hand, text pre-processing can have drastic effects on the conclusions drawn from bags-of-word models (e.g. Denny and Spirling 2018; Greene *et al.* 2016). On the other hand, advanced natural language processing tools frequently resort to the grammatical structure of the original text (e.g. Van Atteveldt *et al.* 2008). Thus, providing 'raw' texts provides the most flexible way forward.[13] From this point of departure `R` packages such as *spacy*R (Benoit and Matsuo 2020), *tm* (Feinerer and Hornik 2019), *quanteda* (Benoit *et al.* 2018), or *ko*R*pus* (Michalke 2016) provide handy wrappers to easily create the text and term representations needed for different text analysis algorithms.

4. Summary

Table 2 summarizes the available corpora while Figure 1 plots the temporal distribution of available full-text speeches per parliamentary chamber and month. We sincerely hope that these data stimulate further text-based research on questions of representation, party politics and political language more generally. If these resources are useful for you, please be so kind to cite the source and do let us know about your work!

[13] While an explicit declaration of the encoding did not work in all instances, the text vectors all represent sequences of valid UTF-8 characters as per stringi::stri_enc_isutf8(corpus$text).

Parliament	File size (MB)	Period	Speeches	Av. speeches per month	Av. terms per speech (w/out chair)	Unique terms in speeches	Unique speakers	Unique partisan factions	Agenda item available?
AT ***Nationalrat***	545	1996-01-15 2018-12-13	199,481	792	305 (579)	508,508	875	BZÖ, FPÖ, Grüne, JETZT, Jetzt – Liste PILZ, LIF, NEOS, ÖVP, PILZ, SPÖ, STRONACH	throughout
CZ ***Poslanecká sněmovna Parlamentu***	387	1993-01-01 2016-06-03	329,135	1,286	157 (240)	290,576	1,096	ANO, CMSS, CMUS, CSSD, HSD-SMS, HSDMS, KDS, KDU-CSL, KSCM, LB, LSNS, LSU, Nez.-SZ, Nezarazení, ODA, ODS, ONH, SPR-RSC, SZ, TOP09, US, US-DEU, Úsvit, VV	n/a
DE ***Bundestag***	857	1991-03-12 2018-12-14	379,545	1,249	291 (460)	770,524	2,374	AfD, CDU/CSU, FDP, GRUENE, PDS/LINKE, SPD	From 2009-10-27
DK ***Folketing***	762	1997-10-07 2018-12-20	772,18	3,558	143 (231)	400,837	766	ALT, CD, DF, EL, FF, FP, FRI, IA, JF, KD, KF, KRF, LA, LH, NQ, NY, RV, S, SF, SIU, SP, T, UFG, UP, V	throughout
ES ***Congreso***	509	1996-03-27 2018-12-20	262,276	1,037	254 (579)	171,489	1,504	GC-CiU, GC-DL, GCC, GCC-NC, GCs, GCUP-EC-EM, GER, GER-ERC, GER-IU-ICV, GIP, GIU, GIU-ICV, GMX, GPP, GPSOE, GUPyD, GV-PNV, GV EAJ-PNV	throughout
NL ***Tweede Kamer***	1,027	1994-12-20 2019-07-04	1,143,366	4,188	142 (162)	450,418	1,209	50PLUS, CDA, CU, D66, DENK, FvD, GL, GPV, LPF, other, PvdA, PvdD, PVV, RPF, SGP, SP, VVD	n/a
NZ ***House of Represenatives***	1,002	1987-09-16 2019-07-24	925,766	2,772	149 (165)	359,806	475	ACT, Alliance, Christian Democrat Party, Christian Heritage Party, Conservative Party, Future, Green, Independent, Labour, Liberal, Mana, Mana Wahine, Maori, Mauri Pacific, National, NewLabour, NZ First, Progressive, Te Tawharau, United Future, United NZ	throughout
SE ***Riksdagen***	905	1990-10-02 2018-12-21	365,56	1,274	348 (350)	633,657	1,894	C, FP, KD, L, M, MP, NYD, S, SD, V	throughout
UK ***House of Commons***	2,451	1988-11-22 2019-12-17	1,956,223	5,892	186 (194)	898,762	2,175	APNI, Birkenhead Social Justice, Change UK, Con, DUP, GPEW, Lab, LibDem, PlaidCymru, Referendum, Respect, SDLP, SDP, SNP, The Independents, UKIP, UKUP, UPUP, UUP	throughout (with gaps)

Table 2: Overview of the *ParlSpeech* V2 corpora

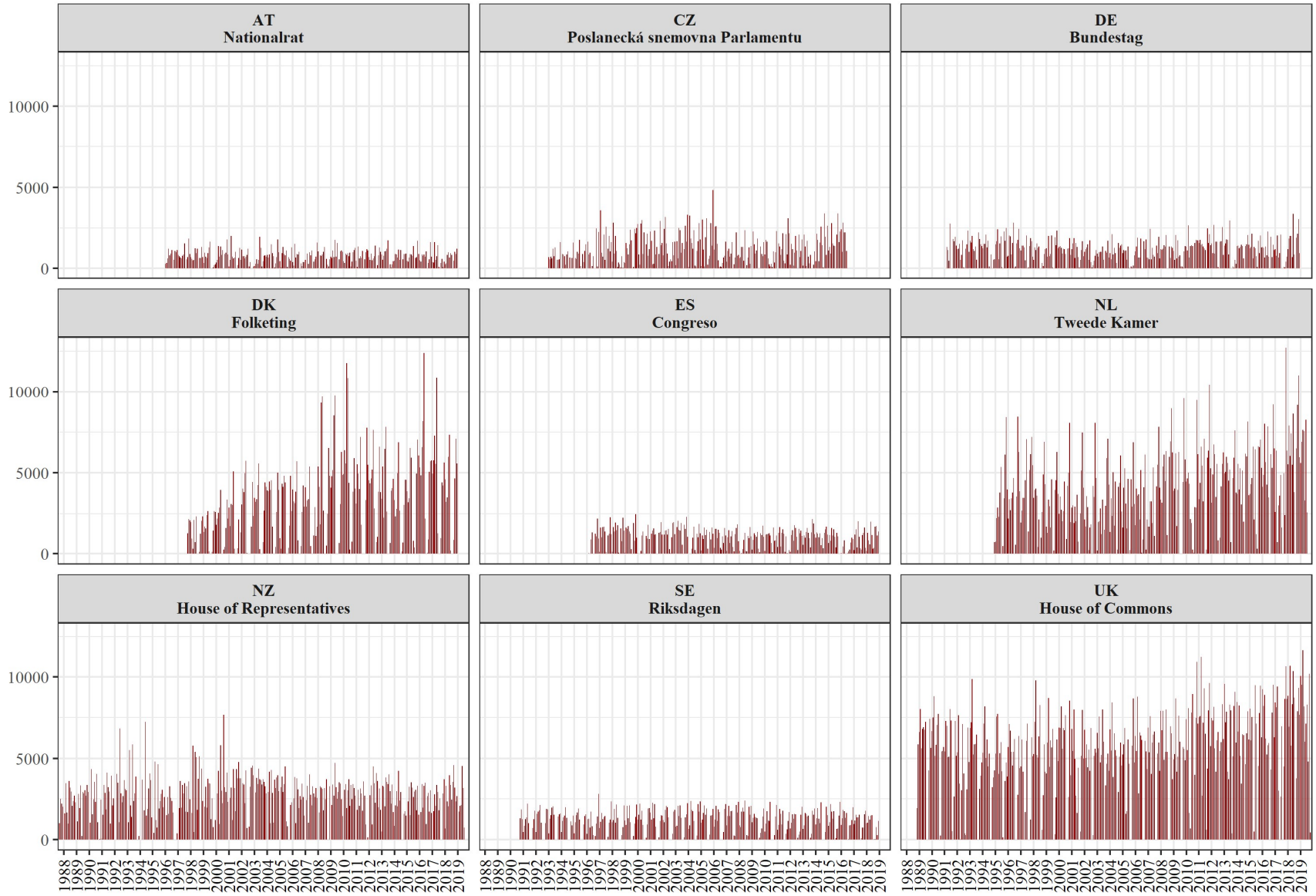

Figure 1: Temporal distribution of available speeches per parliamentary chamber

4. References

Bakker, R. et al. (2015) 'Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010', *Party Politics* 21(1): 143–152.

Baturo, A., Dasandi, N. and Mikhaylov, S. J. (2017) 'Understanding state preferences with text as data: Introducing the UN General Debate corpus', *Research & Politics* 4(2): 2053168017712821.

Benoit, K. et al. (2018) 'quanteda: An R package for the quantitative analysis of textual data', *Journal of Open Source Software* 3(30): 774.

Benoit, K. and Matsuo, A. (2020) *spacyr: Wrapper to the 'spaCy' 'NLP' Library. R package version 1.2.1*, 2020.

Cardie, C. and Wilkerson, J. (2008) 'Text Annotation for Political Science Research', *Journal of Information Technology & Politics* 5(1): 1–6.

De Vries, E., Schoonvelde, M. and Schumacher, G. (2018) 'No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications', *Political Analysis* 26(4): 417–430.

Denny, M. J. and Spirling, A. (2018) 'Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It', *Political Analysis* 26(2): 168–189.

Döring, H. and Manow, P. (2018) 'Parliaments and governments database (ParlGov): Information on parties, elections and cabinets in modern democracies.', *Web Page* Accessed: May 31 2018(http://www.parlgov.org/).

Döring, H. and Regel, S. (2019) 'Party Facts: A database of political parties worldwide', *Party Politics* 25(2): 97–109.

Eggers, A. C. and Spirling, A. (2014) 'Ministerial Responsiveness in Westminster Systems: Institutional Choices and House of Commons Debate, 1832–1915', *American Journal of Political Science* 58(4): 873–887.

Feinerer, I. and Hornik, K. (2019) *tm: Text Mining Package. R package version 0.7-7*, 2019, available at https://cran.r-project.org/web/packages/tm/tm.pdf.

Gijs Schumacher, Martijn Schoonvelde, Denise Traber, Tanushree Dahiya and Erik de Vries (2016) 'EUSpeech: a New Dataset of EU Elite Speeches', *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text.*

Greene, Z., Ceron, A., Schumacher, G. and Fazekas, Z. (2016) 'The Nuts and Bolts of Automated Text Analysis. Comparing Different Document Pre-processing Techniques in Four Countries', *Open Science Framework* November 1(osf.io/ghxj8).

Grimmer, J. and Stewart, B. (2013) 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts', *Political Analysis* 21(3): 267–297.

Lauderdale, B. E. and Herzog, A. (2016) 'Measuring Political Positions from Legislative Speech', *Political Analysis* 24(3): 374–394.

Lucas, C. et al. (2015) 'Computer-Assisted Text Analysis for Comparative Politics', *Political Analysis* 23(2): 254–277.

Marx, M. and Schuth, A. (2010) 'DutchParl. The Parliamentary Documents in Dutch', in *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. 17 May 2010, Valletta, Malta: European Language Resources Association (ELRA), available at http://www.lrec-conf.org/proceedings/lrec2010/summaries/263.html.

Merz, N., Regel, S. and Lewandowski, J. (2016) 'The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis', *Research & Politics* 3(2): 2053168016643346.

Michalke, M. (2016) *koRpus: An R Package for Text Analysis (Version 0.06-4)*, 2016.

Monroe, B. and Schrodt, P. (2008) 'Introduction to the Special Issue: The Statistical Analysis of Political Text', *Political Analysis* 16(4): 351–355.

Munzert, S., Rubba, C., Meißner, P. and Nyhuis, D. (2015) *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, Wiley, available at http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/111883481X.

Proksch, S.-O., Lowe, W., Wäckerle, J. and Soroka, S. (2018) 'Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches', *Legislative Studies Quarterly* (Online first).

R Core Team (2019) *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.

Rauh, C. (2015) 'Communicating supranational governance? The salience of EU affairs in the German Bundestag, 1991-2013', *European Union Politics* 16(1): 116–138.

Rauh, C. and De Wilde, P. (2018) 'The opposition deficit in EU accountability: Evidence from over 20 years of plenary debate in four member states', *European Journal of Political Research* 57(1): 194–216.

Rauh, C., De Wilde, P. and Schwalbach, J. (2017) 'The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states', *Harvard Dataverse* V1, available at https://doi.org/10.7910/DVN/E4RSP9.

Temple Lang, D. and CRAN Core Team (2015) *XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98-1.3.*, 2015.

Temple Lang, D. and R Core Team (2016) *RCurl: General Network (HTTP/FTP/...) Client Interface for R. R package version 1.95-4.8.*, 2016.

Van Atteveldt, W., Kleinnijenhuis, J. and Ruigrok, N. (2008) 'Parsing, Semantic Networks, and Political Authority Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles', *Political Analysis* 16(4): 428–446.

Volkens, A. et al. (2014) *The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR)*, Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB).

Wickham, H. (2015) *stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.0.0.*, 2015.

Wickham, H. (2016) *rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3.2.*, 2016.