Comparative European Legislative Research in the Age of Large-Scale Computational Text Analysis – A Review Article

Miklós Sebők, Centre for Social Sciences, Budapest
Sven-Oliver Proksch, University of Cologne
Christian Rauh, WZB Berlin Social Science Center
Péter Visnovitz, Centre for Social Sciences, Budapest
Gergő Balázs, Centre for Social Sciences, Budapest
Jan Schwalbach, GESIS – Leibniz Institute for the Social Sciences

Corresponding author: Miklós Sebők, sebok.miklos@tk.hu, Centre for Social Sciences, Budapest, Hungary

ABSTRACT

Advances in data accessibility and analytical methods opened new frontiers for comparative studies of European legislative activities. However, these advances still need to be fully harnessed by legislative scholars for multiple reasons. We provide an overview of extant research agendas to identify these reasons and explore the opportunities for tapping the potential of Big Data and quantitative text analysis. We present significant data collection efforts, such as ParlSpeech, the Comparative Agendas Project and CLARIN, and highlight their respective value for, primarily, large-N comparative research focusing on EU member states and the Union itself. Our review highlights the most consequential gaps in the literature and shortcomings of available data and analysis. These include the lack of extensive historical and geographical coverage, missing harmonisation and cross-linking between separate efforts, no unified speech and document (bill, law) databases, and the unavailability of good quality full-text variables.

Keywords: legislative studies, comparative politics, European politics, quantitative text analysis

Introduction

In this article we provide an overview of extant research agendas and explore the opportunities for tapping the potential of Big Data and quantitative text analysis in comparative legislative research. We present significant data collection efforts, such as ParlSpeech, the Comparative Agendas Project and CLARIN, and highlight their respective value for, primarily, large-N comparative research focusing on EU member states and the Union itself. Our review highlights the most consequential gaps in the literature and shortcomings of available data and analysis. These include the lack of extensive historical and geographical coverage, missing harmonisation and cross-linking between separate efforts, no unified speech and document (bill, law) databases, and the unavailability of good quality full-text variables.

Parliaments are vital venues for political representation and policymaking in democratic states. In parliaments, elected representatives publicly communicate with each other and their voters (Back et al., 2021; Proksch and Slapin, 2014). This allows them to position themselves towards broader societal cleavage lines or specific initiatives (Borghetto and Chaqués-Bonafont, 2019), with legislative debates constituting 'the formal end-games of a long political process' (Laver, 2021). Besides informing the broader political discourse by highlighting political issues and stances in speeches, members of parliament (MPs) also propose or amend bills, thereby fixing political preferences in binding laws and regulations. Thus, the content of parliamentary debate and its legislative documents offer invaluable information on the political agendas and conflict lines structuring collective decision-making in democracies.

Analysing and comparing the content of legislative speeches and documents have thus been a long-standing focus of empirical political science. However, our ability to extract systematic

information from large legislative text corpora at scale has changed over recent decades by using algorithmic approaches that treat (qualitative) text as (quantitative) data (Goplerud, 2021; Grimmer and Stewart, 2013). This review article outlines the tremendous potential and practical limits of exploiting these tools to enhance our understanding of the functioning of representative democracy. Our assessment aims to review the extant literature and spot gaps to explore how the field could and should proceed in leveraging available Big Data sources and newly developed innovative methods. These gaps (and the respective takeaways) include a lack of historical and geographical coverage, missing harmonisation and cross-linking between separate efforts and the unavailability of good-quality full-text variables.

We proceed in two steps. First, we showcase successful applications of text-as-data methods to critical questions of legislative politics. Scholars have effectively used text data to measure essential concepts such as issue attention, ideological positioning and polarisation, rhetorical strategies, or legislative influence. This demonstrates the promise that text-as-data methods hold for understanding the inner workings of democracies. However, readily available, machine-readable text corpora of speeches, bills, and laws across a broad set of countries present a bottleneck for innovative comparative research.

We, thus, secondly, report the findings from an encompassing stock-tacking exercise on the availability of parliamentary text corpora across democracies in the European Union (EU) and beyond. We provide readers with an aggregate overview and a searchable database. We also discuss a few of the broadest data collection efforts thus far (the ParlSpeech, the Comparative Agendas Project, and the CLARIN infrastructure). These projects highlight the fact that the availability of parliamentary text data has massively improved.

But we also show that geographical and temporal biases persist, that access to legislative documents is less developed when compared to speeches, that our ability to link meaningfully different types of text produced in the legislative process is still suboptimal, and that datasets with good quality, machine-readable full-text data are still the exception, not the rule. Many of these observations also apply to data collection efforts outside of Europe (and CAP, indeed, is a global project).

By showcasing the potential of modern text-as-data methods and outlining which corresponding data sets are already available to develop, test, and implement new tools and research questions, we hope to encourage political scientists to push legislative studies further (in a similar manner how the study of elections developed through joint infrastructure building over the last decades) . We also conclude that beyond advancing tools and methods, comparative legislative studies as a field would profit from more joint investments in data collection and data harmonisation.

Text-as-data approaches in comparative legislative studies

Comparative legislative research comprises a variety of research streams on different aspects of parliamentary institutions, actors, and processes. The diversity of relevant research questions on parliamentary behaviour and outputs is mirrored in equally diverse methodological approaches to empirical inquiry. Interviews and surveys, for instance, have been successfully exploited to understand the motives of legislative actors (Bailer, 2014). Experimental methods are used to study the policy preferences and voting practices of parliamentarians and their responsiveness to public preferences (Druckman et al., 2014). And the analysis of voting patterns, specifically roll-call votes, has revealed much about the empirical structuring of democratic decision-making (Carroll and Poole, 2014).

While the appropriateness of methodological choices is and will remain a function of the research question, it is important to highlight that much of what parliaments do is represented in some form of text – often incredibly large amounts of text. Between 1988 and 2019, for example, MPs in the UK House of Commons gave almost two million speeches totalling almost 370 million words (Rauh et al., 2020). To take another example, between 1994 and 2016, the European Commission proposed 2,228 binding directives and regulations containing almost 4.7 million words (Rauh, 2021). And these words matter: they encapsulate the political priorities, stances, and choices that move and structure democratic societies (Slapin and Proksch, 2014).

That is why this review article focuses on text-as-data methods: extracting systematic information from such large text corpora holds, in our view, the biggest potential for the future of empirical comparative legislative studies. Historically, scholars interested in analysing and comparing the content of legislative texts had to sift through printed volumes of plenary protocols or adopted bills

before manually extracting systematic comparable information or by hired coders. The high costs of such approaches often led to selective analyses that were limited in scope and comparative insights across countries and time.

The past two decades, however, saw significant methodological advances for overcoming these obstacles, changing our thinking about the prospects of comparative legislative studies. Legislative texts and speeches can now be analysed with the help of a wide array of computer-assisted and (partially) automated methods. Tools in this vein retrieve references, geographical information, sentiment, or emotions in texts with flexible dictionary searches (Osnabrügge et al., 2021; Sebők et al., 2017; Proksch et al., 2019b; Rauh, 2018; Rauh and De Wilde, 2018) or algorithms based on minimal human input (Watanabe, 2021; Widmann and Wich, 2022).

Various algorithms offer means to classify, cluster, or scale texts either in a supervised manner, where the machine learns from text annotated by humans, or in an unsupervised way, where the machine learns patterns from the data itself (Quinn et al., 2010; Rheault and Cochrane, 2020; Slapin and Proksch, 2008). Text reuse algorithms offer a means to study the evolution of documents or the diffusion of ideas across texts (Cross and Hermansson, 2017; Gava et al., 2021; Wilkerson et al., 2015). The rapidly evolving field of Natural Language Processing (NLP) constantly pushes the methodological toolkit of content parsing and classification (Jurafsky and Martin, 2000; Sebők and Kacsuk, 2021; Sebők et al., 2022). Furthermore, the advent of reliable machine translation technologies offers promising avenues for enabling comparative work across different languages (Lucas et al., 2015; Lind et al., 2019; De Vries et al., 2018; Proksch et al., 2019a).

Combining this rapid development of automated text analysis methods with the importance of text in parliamentary processes offers path-breaking opportunities for empirical inquiry: researchers are no longer constrained by the need to read and interpret every text unit. This is not to say that human interpretation has become dispensable. On the contrary, supervised algorithms do not work without high-quality human annotations of exemplary texts while unsupervised algorithms must be validated against the 'gold standard' of human interpretation. But once good training data exists and algorithms have been appropriately validated for the concept of interest, the information processing capacities of automated text analysis algorithms are infinite in principle, limited only by the availability of machine-readable text data on parliamentary activity.

To demonstrate this potential further, the remainder of this section briefly introduces extant recent applications of text-as-data algorithms to core questions that have moved scholars of legislative studies of different stripes: Which political issues are covered by speeches and laws? How and along which lines do MPs position themselves? What rhetorical strategies do they use for mobilisation or persuasion? And how does the parliamentary process influence the substance of laws? This overview is necessarily incomplete, and we focus on exemplary comparative research projects mostly in European parliaments while presenting the most significant American applications. These examples are meant to showcase what these new tools have to offer for understanding the inner workings of modern democracies.

What issues do MPs focus on?

The question of what issues figure in the political process has traditionally been the domain of agenda-setting studies. They rest on the view that significant political power lies in limiting public consideration of specific issues (Bachrach and Baratz, 1962) and evolved around 'punctuated

equilibrium theory'. This hypothesis expects the general stability of the political agenda, sometimes interrupted by periods of high attention and dramatic policy shifts (Baumgartner et al., 2006). This dynamic has been examined from the perspective of various actors (politicians, presidents, government and opposition parties, lobby groups etc.) with the help of a wide variety of documents, including legislative speeches and documents (Borghetto and Belchior, 2020; Green-Pedersen, 2023; Vliegenthart et al., 2013).

In the past decade, substantial efforts were invested in creating extensive datasets under the joint Comparative Agendas Project (CAP) coding scheme to test further and develop this theory. As of the last available data, these projects also include text data on 14 legislatures.[i] These are, however, mostly produced by massive human coding efforts, thus creating high costs and limiting the breadth of the research scope.

Therefore, scholars have started to develop supervised text classification algorithms that learn from extant human-coded data along the CAP scheme to classify a possibly large amount of virgin texts into these categories. Using, among others, bills from the U.S. congressional record, Hillard et al. (2008) demonstrated that a supervised classifier provides accurate estimates of issue proportions across the corpus, achieves similar levels of reliability as human coding, and may reduce the bill classification costs by 80% or more. Hansen et al. (2019) used human-coded agenda items of individual parliamentary sessions and trained an automated classifier based solely on the words in the tile of agenda items. This replicated human codes with an accuracy of around 96%. More recent studies successfully deployed machine learning algorithms to achieve human-level precision scores for the CAP classification task (Sebők and Kacsuk, 2021; Sebők et al., 2022).

Beyond the CAP-coding scheme, scholars have used unsupervised topic models – a class of algorithms optimising the distribution of words over a pre-specified number of topics (Blei, 2012) – to generate insights into parliamentary agendas. Greene and Cross (2017), for example, study the issue agenda of the European Parliament across more than 200,000 speeches by more than 1,700 MEPs between 1999 and 2014. They show substantial variation in the issue agenda over time and how it responds to external events such as the Euro Crisis. Quinn et al. (2010) trained a topic model on more than 118,00 speeches in the U.S. Senate from 1997 to 2004, revealing meaningful and interrelated speech topics that can be validated along known topics of roll call votes or specific events on the parliamentary calendar.

Such topic distributions are also essential for arguments in the issue salience theory of partisan competition, which posits that some parties can 'own' issues which they try to emphasise in the legislative arena (Geese, 2020). Text-as-data methods are increasingly used in this subfield: Osnabrügge et al. (2021), for example, trained an algorithm on labelled topics from party manifestos then to classify almost 300,000 parliamentary speeches in New Zealand. Across 44 pre-defined topics, the classifier approached the interpretation of trained human coders (though validity still varied significantly by topic). The authors then used machine-generated data to show that an electoral reform increases the prevalence of topics related to party competencies and that topics vary heavily by the gender of MPs.

Partisan attention to specific topics can also be studied by unsupervised algorithms. Finseraas et al. (2021) used a structural topic model to identify MP attention to climate change with data on responses to an oil price shock. They found that MPs whose constituencies faced high political costs of climate policies tried to avoid environmental topics, while less affected MPs talked more about investments in green energy in response to the abrupt movements of oil prices.

For research questions focussing on specific issues, sometimes even computationally much less demanding text-as-data tools can generate hitherto unavailable insights. Contributing to the vibrant debate about the potential 'Europeanisation' of national parliaments, for example, Rauh and De Wilde (2018) developed and validated flexible references to the polity, politics and policies of the EU. Applying these dictionaries to more than 2.5 million plenary speeches from four European legislative bodies over more than 20 years showed that national parliamentary emphasis on EU affairs had grown together with the legislative empowerment of the EU. But they also highlighted that parliamentary salience of EU affairs is primarily driven by governing parties, decreases in election time and is negatively associated with citizens' Euroscepticism – leading to the conclusion of a substantial opposition deficit in parliamentary debates about EU affairs. Overall, the study of Europeanization via the analysis of legislative speech is one of the leading areas of applying large-N research designs within legislative studies (De Ruiter and Schalk, 2017; Hoerner, 2019; Kinski, 2021; Lehmann, 2023; Navarro and Brouard, 2014; Winzen et al., 2018).

How do MPs position themselves in legislative conflict?

The question of how political representatives position themselves regarding broader societal conflict lines or with a view to specific issues is driving many theories of electoral partisan competition in democracies. In this domain, parliamentary speeches hold a distinct advantage over other relevant information sources such as expert or citizen surveys and party manifestos: they are distributed more evenly over time and are available throughout the electoral cycle. Moreover, depending on institutional context, they are less moderated by party leadership and party consensus and may thus offer insights into within-party differences.

Unsurprisingly, much of the development of political science text-as-data methods has thus focussed on scaling efforts targeted to place speeches on latent dimensions of political conflict. An early pathbreaking tool was the Wordscores algorithm proposed by Laver et al. (2003). This algorithm relies on relative word frequencies of very few reference texts with known positions on the latent scale (e.g. manifestos of the most 'leftist' and the most 'right-wing' party) to use these as predictors in a regression model to predict the position of 'virgin' texts on that scale. The original paper showed that this approach reliably recovers pro- and anti-government stances in the Irish parliamentary debates. Aiming at reliable measures of the ideological and policy positions of MPs as expressed in their speeches, Slapin and Proksch (2008) then proposed the unsupervised 'Wordfish' algorithm, which – based on the assumption of an overarching unidimensional policy space – automatically optimises document discrimination along differences in relative word frequencies.

Lauderdale and Herzog (2016) extended this model to account for multiple dimensions of political conflict while controlling for agenda-specific patterns of partisan discrimination. The latter point highlights that agenda control needs to be incorporated when applying scaling models to legislative debates: naive estimations of ideological positions in parliaments risk failing as government-opposition dynamics will pre-determine much of the rhetorical conflicts.

This point is, for instance, demonstrated by a dictionary-based sentiment analysis showing that opposition parties consistently use more negative and government parties more positive language in bill-specific debates even though the opposition parties may come from opposite ends of the ideological spectrum (Proksch et al., 2019b). Therefore, scaling results are very good at estimating conflicts but struggle with recovering broader ideological placements of parties. Political systems

also matter in this respect: this effect is observably different in institutions with less pronounced government-opposition dynamics, such as in the European Parliament (Proksch and Slapin, 2010).

In sum, caution is warranted in interpreting and validating automatically retrieved scales from legislative speeches. Yet still, extant work shows that meaningful actor positions can be retrieved relative to the agenda individual debates cover. Along this line, Bernauer and Bräuninger (2009) use the Wordscores algorithm to study within-party preferences heterogeneity in speeches of the German Bundestag, recovering meaningful links to intra-party faction memberships. Finally, Goet (2019) uses over 200 years of UK House of Commons data to argue for supervised scaling methods over unsupervised approaches.

How do legislators speak and reason?

Researchers also use supervised and unsupervised machine learning methods to learn more about the actors, such as members of parliament, who produce massive amounts of text and their rhetorical strategies. Textual characteristics can be identified and linked to speakers, helping researchers to draw conclusions about legislators' personalities, styles, preferred tones, the type of language they rely on and the level of conflict their interactions entail. Dictionary-based sentiment analysis helps capture government-opposition conflict in bill debates (Proksch et al., 2019b), showing that sentiment analysis helps delineate the differences between governing and opposition parties. Sentiment dictionaries have been translated or are available specifically for other languages. For example, a dictionary-based analysis is suitable to track moral reasoning in texts (Kraft, 2018) or analyse legislative cycles' effect on party behaviour in parliaments (Schwalbach, 2022).

Legislators' language use can also be measured by the complexity of their sentences: Lin and Osnabrügge (2018) demonstrate that German parliamentarians tend to make their speeches less complicated when their constituents are relatively poor and less educated. Similar methods have been applied to political communication at the European Union level (Rauh et al., 2020). Other possibilities include the analysis of populist rhetoric or other rhetorical tools used by politicians (Decadri and Boussalis, 2020).

How much does the legislature influence law-making?

Most text-as-data applications in comparative legislative studies have focused on speeches. However, parliaments produce many other types of texts that are rarely analysed but contain much information about the political process behind the highly visible plenary debates (Remschel and Kroeber, 2022). The strength of parliaments' influence on the output of the legislative process has been successfully analysed by automatically comparing legislative proposals or bills (in many cases originated by the executive) and the adopted forms (which went through the legislative process). Applying text reuse algorithms to large legislative corpora has, for example, worked for analysing EU-level parliamentary decision-making. Cross and Hermansson (2017) looked at the changes between proposals and the final legislative outcome using a 'minimum edit distance' algorithm. This algorithm assesses how much text must be inserted, removed, or transposed to transfer the draft document into the adopted output text. One of their key findings was that legislative amendments are influenced by the different formal rules structuring inter-institutional negotiation between the European Parliament and the Council of Ministers.

Rauh (2021) used a similar algorithm to compare proposals from the European Commission to the full texts of the laws adopted by the Council and the European Parliament. He finds that the

Commission's agenda-setting power varies across its different departments and is constrained by rules prescribing a more active involvement of the parliament.

Gava et al. (2021) applied a similar research design to study around 1,700 bills in the Swiss parliament, finding substantial variation in the degree to which they are changed in parliament. Such text reuse approaches may also inform the analysis of preference attainment, the degree to which parliaments adopt interest group proposals into law. While this brief review hardly does full justice to the field of legislative studies as a whole or to all text-as-data developments over the last years, these snapshots related to four thematic areas demonstrate how much insight can be gained at comparatively low costs from approaching parliamentary text data computationally.

Legislative text data sources for large-scale comparative studies

Primary and secondary sources of national legislative corpora

The overview of extant research shows the potential for empirical legislative studies that lie in large-scale text analysis. However, the review also showed that most research designs are based on single-country studies or small-n comparative analysis at best. In our view, the primary reason for the lack of more Large-N investigations in legislative studies is the limited availability of machine-readable representations of textual output from parliaments. This bottleneck is partly due to the fact that parliamentary text data are usually provided by nation-state or supranational archives (this is true of both legislative documents and transcripts of speeches). Their repositories are designed to help political, legal and media stakeholders find specific text items. In contrast, researchers willing to embark on computational text analysis need access to comprehensive corpora in machine-readable and, ideally, standardised formats across country cases.

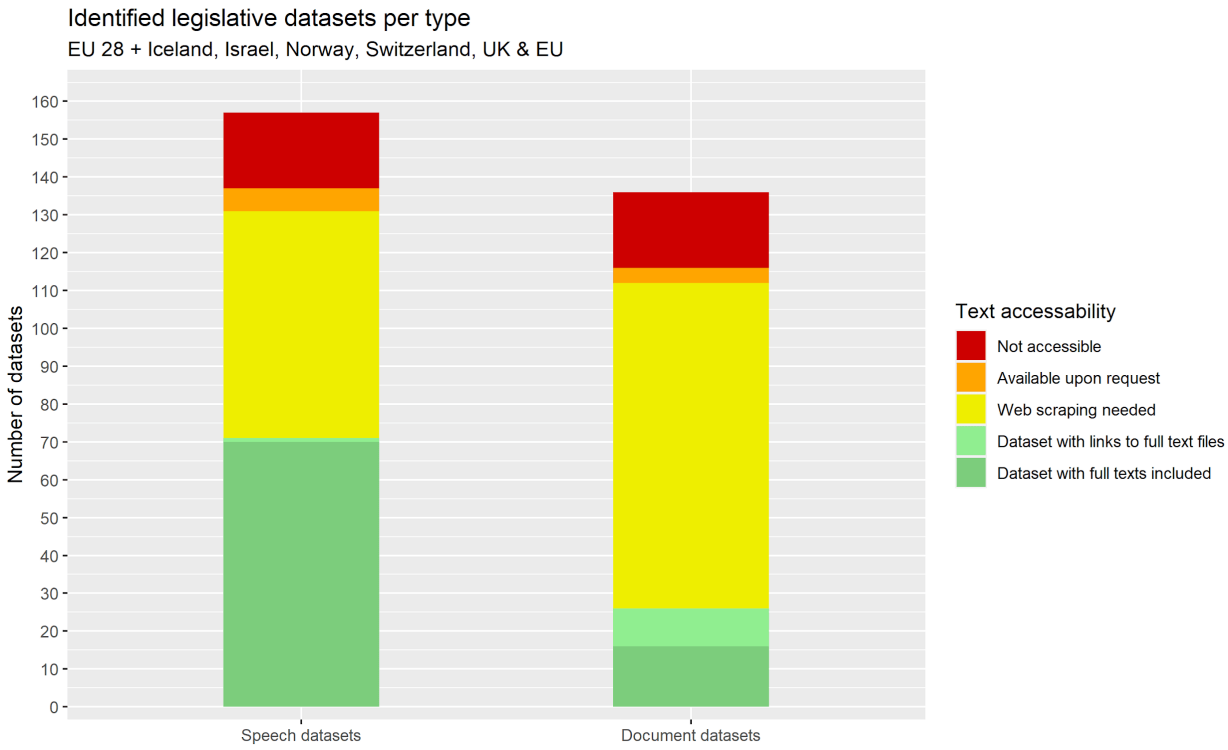
Such collections exist, even if they have a limited geographical scope or representation of metadata along with the text of legislative speeches and documents. Moreover, the validation of the data available is often beyond the means of individual researchers or small projects (which may come to the fore when it comes to, inter alia, the precision of speech transcripts). Therefore, they should serve as a stepping-stone towards creating a comprehensive parliamentary speech and document database for Large-N comparative legislative studies employing a text-as-data methodology. In what follows, we offer such an inventory based on a thorough review conducted within the framework of OPTED, an EU-funded Horizon 2020 project aimed at surveying accessible research infrastructures to analyse political texts.

We collected text corpora and accompanying metadata in the 27 EU member states, plus Iceland, Israel, Norway, Switzerland, and the UK, as well as for various supranational European legislative institutions (case selection was informed by the geographical coverage of the OPTED project). Our data inventory includes at least one data source per country, often the official data source associated with the respective legislative body. Recognising that many existing databases belong to projects with specific research agendas (such as CAP or ParlSpeech), we also provide an overview of those with the most country/sub-projects in the following sections.

Our data collection efforts yielded 293 distinct legislative text data sources.[ii] This presents the most comprehensive overview of available legislative text data in Europe (. Figure 1 shows the number of identified data sources distinguishing parliamentary *speeches* from the other text *documents* parliaments produce (bills, laws etc.). Since the productive application of text analysis algorithms depends on full-text access (although title-based shortcuts are also widely used), we also indicated the difficulty of obtaining full-text data. Full texts are available for almost half of the speech data sources, but access is more difficult for legislative documents. Furthermore, a

sizeable share of data sources offers full texts in sufficient structures to apply additional web scraping methods for building parliamentary text corpora (as in the case of datasets with links to full text).

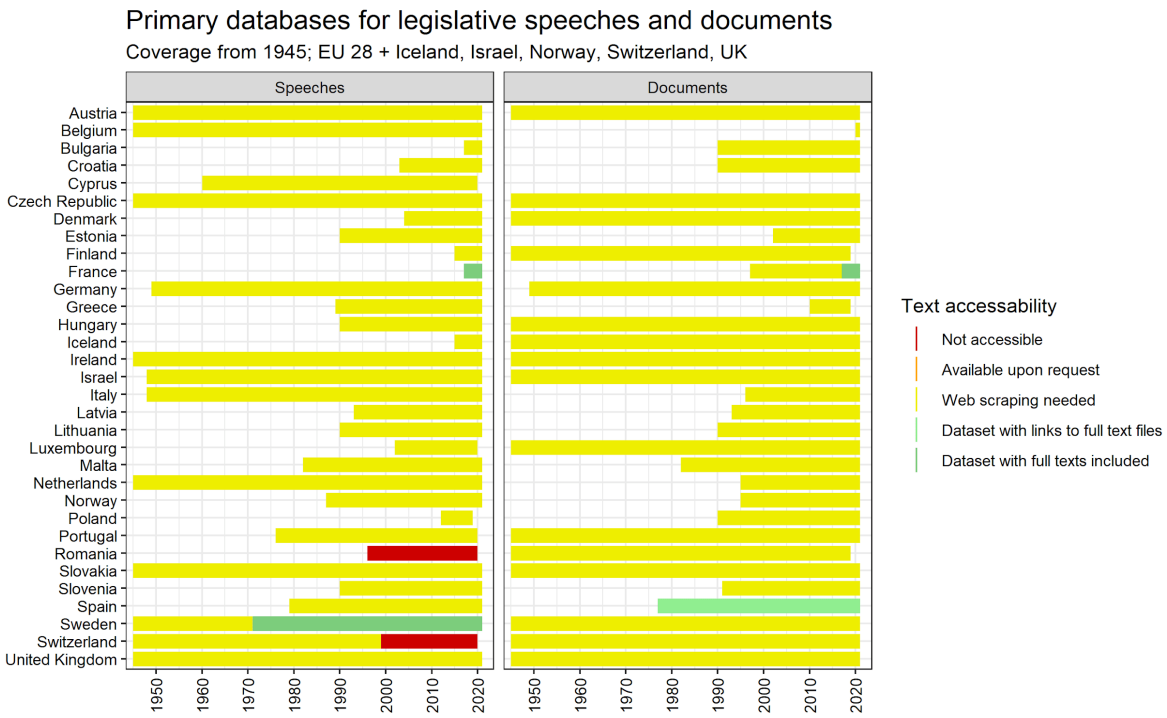
Figure 1. Legislative datasets per type.



Our review covered both primary and secondary data sources. *Primary databases* are collections made publicly accessible in a searchable format by the institution producing the legislative documents (the respective parliaments in most cases). Such primary sources were available for all countries under investigation. Still, none of these provided data that would be readily importable into the programming environments needed for modern text analysis (such as python or R, for example).

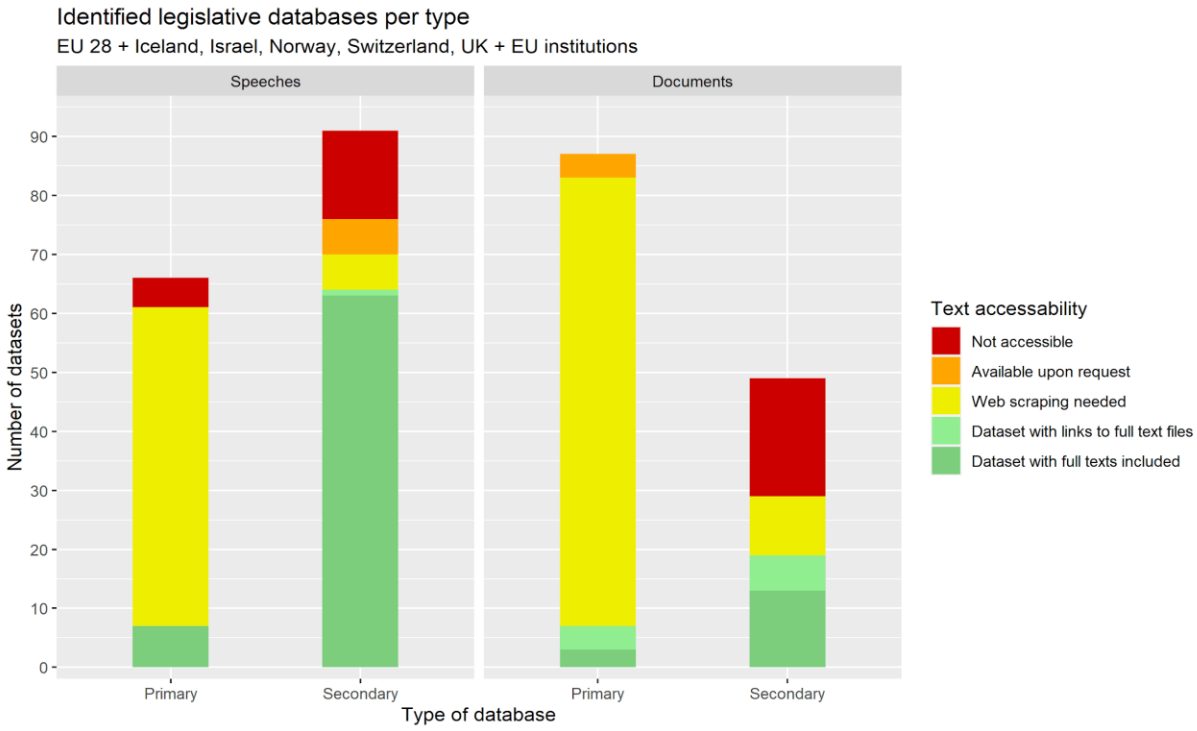
Moreover, the covered timespan of primary sources varies heavily. Figure 2 shows the years since 1945 (before which most countries have no data) covered by the primary legislative databases. Both speech and document data were available for the past two decades for most European countries. However, they mostly require web scraping to create processable files (as opposed to some secondary databases, see below). Therefore, these national primary databases are potential resources but require additional work before using them in meaningful text-as-data analyses. A key takeaway, therefore, is that creating awareness among the managers of official archives of the promises that modern computational tools offer is critical for offering better data access for scientists and the public. This view was reinforced by a thematic workshop which brought together scholars, practitioners and archivists (Kiss and Sebők, 2022).

Figure 2. Primary databases for legislative speeches and documents.



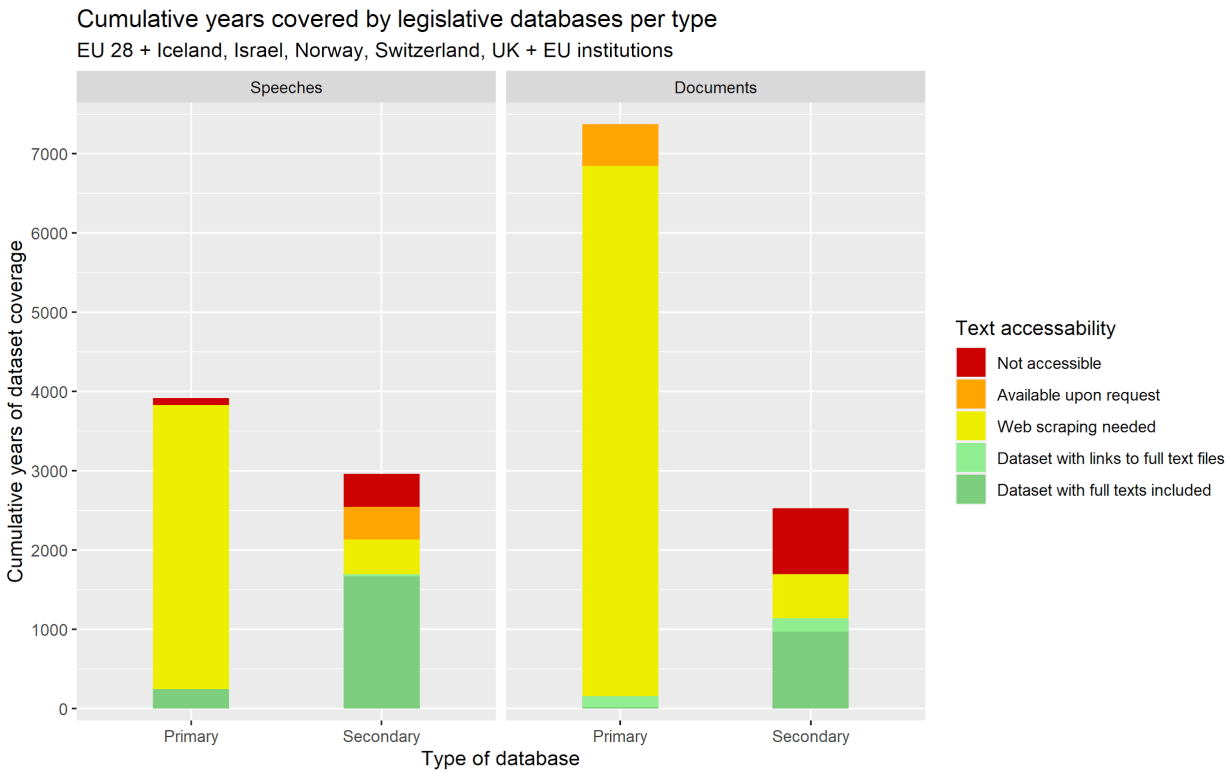
As opposed to primary databases, *secondary datasets* have been collected directly for scientific purposes and are therefore processed and structured to some extent. Their usefulness for automated text analysis still differs according to the purposes they were originally collected for. Secondary datasets are also unevenly distributed across time but notably space, and there is a substantial disparity between datasets for speeches and legislative documents. Figure 3 shows that in terms of *legislative speech* collections, secondary datasets are far more numerous than primary databases (66 vs 91), while in terms of *legislative documents*, primary sources outnumber secondary datasets (87 vs 49). These indicate the academic focus: scholars seem to have dedicated significantly more attention to collecting and analysing parliamentary speeches than legal documents and laws.

Figure 3. Identified legislative databases by type.



These trends are also illustrated by the number of years covered in secondary datasets. They cover less than half as many years as primary sources would offer. Scholarly attention prefers to focus on legislative speeches in this case as well: Figure 4 illustrates that for every year of primary database coverage, there is a significantly more considerable amount of secondary dataset coverage for speeches than for documents.

Figure 4. Cumulative years covered by legislative databases per type.



Moreover, most (country-specific) secondary datasets come from a handful of international projects. Most legislative corpora (67% of the total) belong to one of three projects: ParlSpeech (8 datasets, 6%), CAP (48 datasets, 34 %) and CLARIN (34 databases, 27 %). Only the remaining 46 datasets (33%) result from individual database building. Figure 5 shows the share of the three main projects within secondary datasets, while Figure 6 confirms that they are also dominant in terms of the number of years covered.

Figure 5. Share of top projects in legislative datasets.

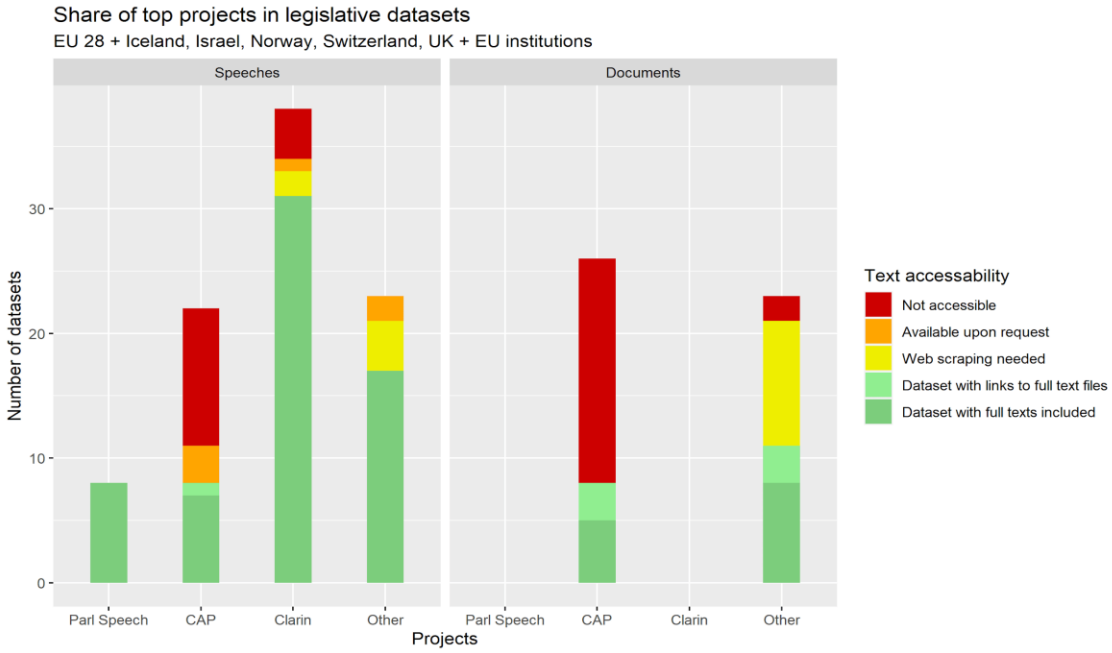
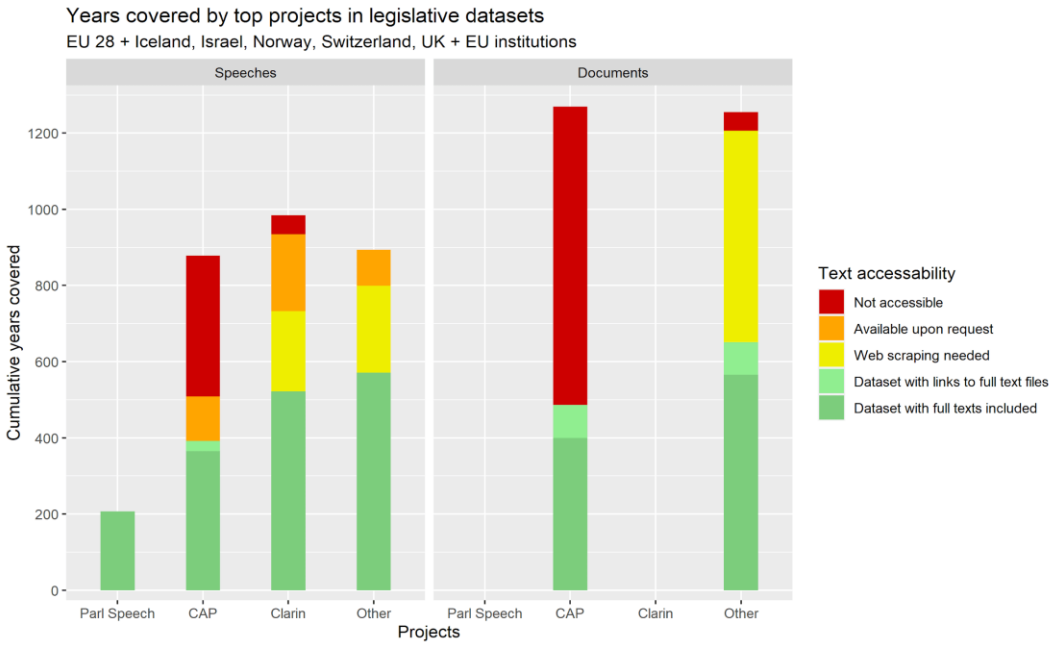


Figure 6. Years covered by top projects in legislative datasets.



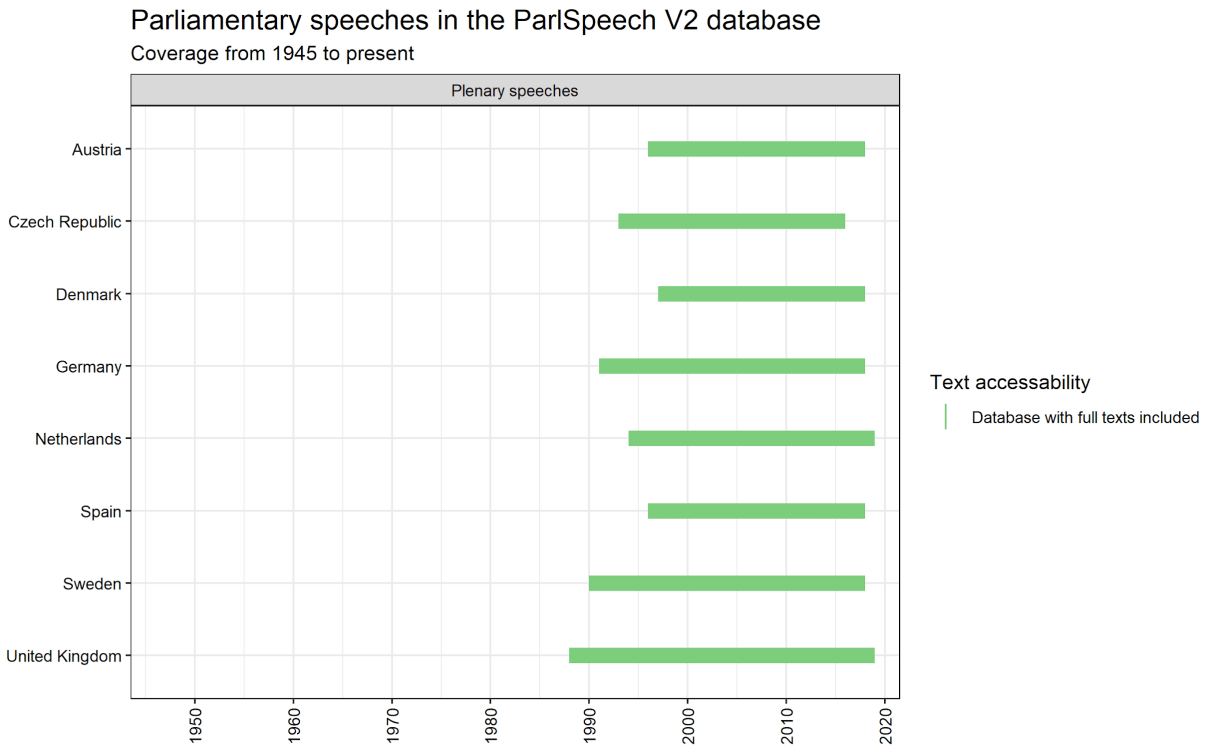
In conclusion, access to text data is still a significant hurdle when it comes to applying the

modern tools of computational text analysis across space, time, and different types of legislative texts. To outline more clearly what is available already, we now dig deeper into the three broadest datasets and their features for text analysis approaches. In turn, we present the ParlSpeech, CAP and CLARIN corpora and then make general conclusions about the overall coverage of all datasets and legislative text availability across Europe.

The ParlSpeech corpora

ParlSpeech is one of the most accessible and complete parliamentary speech corpora. It provides unique access to annotated full texts of parliamentary debates across national parliaments. The extended ParlSpeech V2 includes 6.31 million parliamentary speeches from eight European countries and New Zealand (Rauh and Schwalbach, 2020). ParlSpeech V2 consists of separate datasets for each national parliament (in bicameral systems, speeches of the chamber with more legislative competencies are covered). The time frames differ, but all parliamentary datasets cover at least the 1997 to 2016 period. Figure 7 shows the covered timespan of the legislative speech data for the European countries in the database.

Figure 7. Parliamentary speeches in the ParlSpeech V2 database.



The eight corpora have 11 identical variables in a column structure stored in .rds files (accessible in the R environment but easily transferable to other formats). Speech variables include (1) the date, (2) the agenda item addressed (partial), and (3) the number of tokens (i.e. the length of the speech). Speaker variables cover name and party, including the party ID linking to Döring and Regel's Party Facts database (Döring and Regel, 2019) and whether the speaker is acting as a chairperson. The ParlSpeech corpora represent an ideal format and structure for text as data approaches. Its essential limitation lies in its geographic and temporal scope. Furthermore, ParlSpeech includes only plenary legislative speeches but no other legislative texts like bills.

The Comparative Agendas Project legislative corpora

The CAP corpora aim to track the political agenda of democracies, and it is the second most extensive collection of legislative speeches and texts in Europe (Baumgartner et al., 2019). Several data sources were used for its creation, including speeches delivered in national parliaments. Thirteen European countries have legislative datasets collected with CAP methodology, containing various document types from different time ranges.[iii] As the CAP project focuses on political agendas, the collected legislative corpora are limited to specific types. Most datasets cover parliamentary questions, considered more indicative of the legislature's political agenda than plenary speeches in general. CAP databases collected bills submitted to, laws enacted by national parliaments, and some countries adopted national budgets. Figures 8 and 9 present an overview of the coverage of CAP databases.[iv]

Figure 8. Parliamentary speeches in the CAP project.

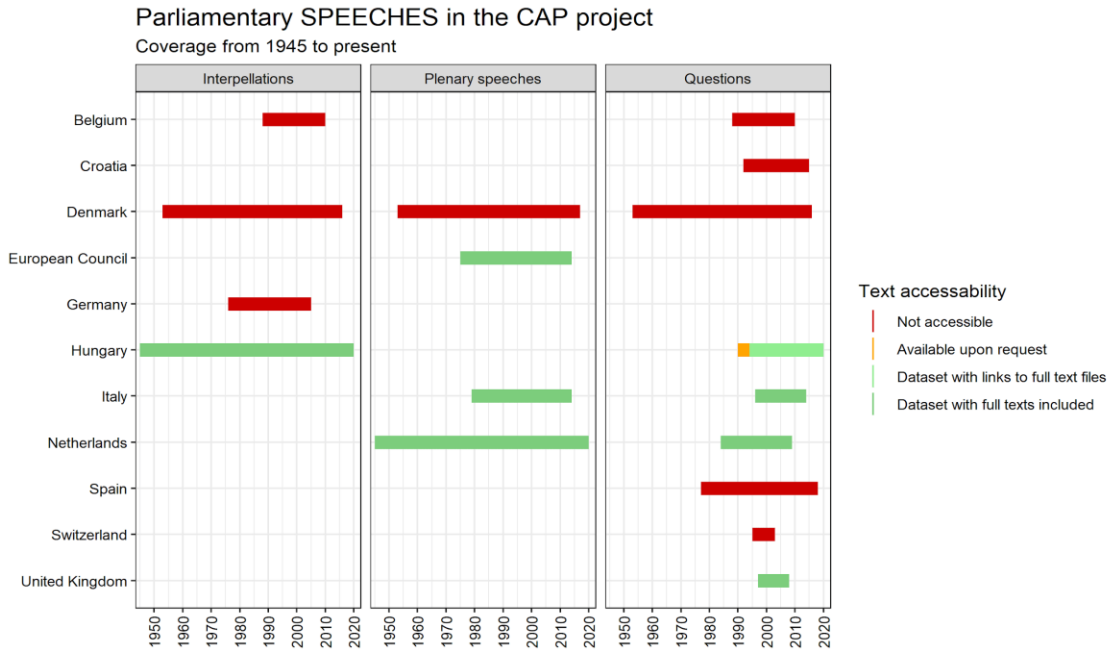
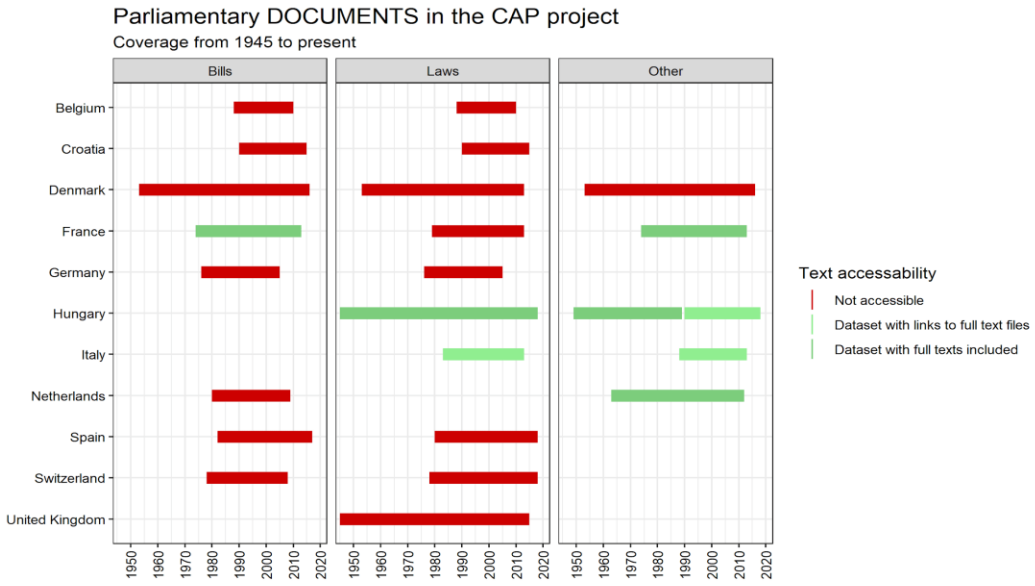


Figure 9. Parliamentary documents in the CAP project.



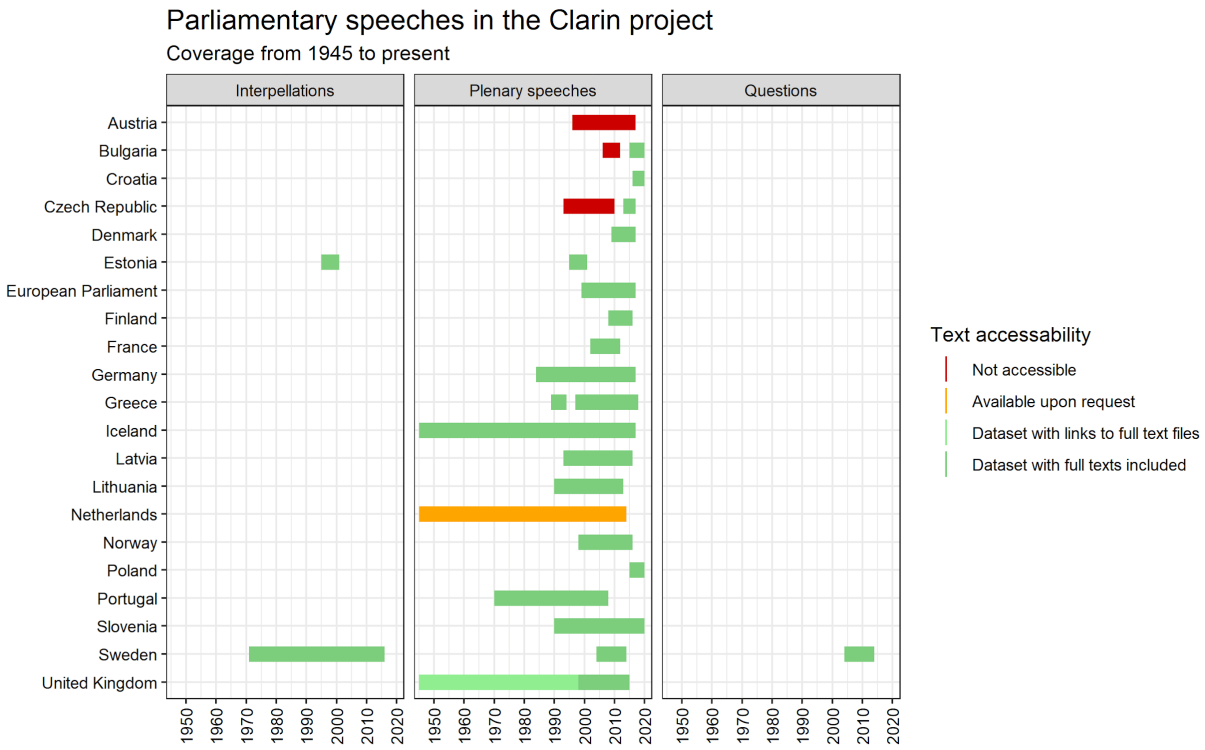
CAP datasets cover more countries than ParlSpeech. However, the number of countries is still low and unevenly distributed (for instance, it features only one Eastern European country). Moreover,

most datasets are unsuitable for textual analysis, as they do not include the full text of speeches and documents, as most projects focused on policy issues rather than making the full text available. Most databases comprise limited metadata, including the topic, the date, and the speaker's name.

The CLARIN parliamentary corpora

The CLARIN project offers the most comprehensive coverage of parliamentary speeches in Europe (de Jong et al., 2022). CLARIN datasets were prepared with a multidisciplinary research agenda in mind. It provided access to 26 parliamentary speech databases through the national repositories of the network. Furthermore, it was funding the *ParlaMint: Towards Comparable Parliamentary Corpora* project at the time of writing (Erjavec et al., 2022). It aimed to produce additional standardised and linguistically processed parliamentary datasets focused on COVID-19 issues and reference ones for comparison. Altogether the CLARIN project includes 34 publicly available speech datasets, mostly plenary speeches. Figure 10 presents the temporal coverage of the CLARIN datasets across countries.

Figure 10. Parliamentary speeches in the CLARIN project.

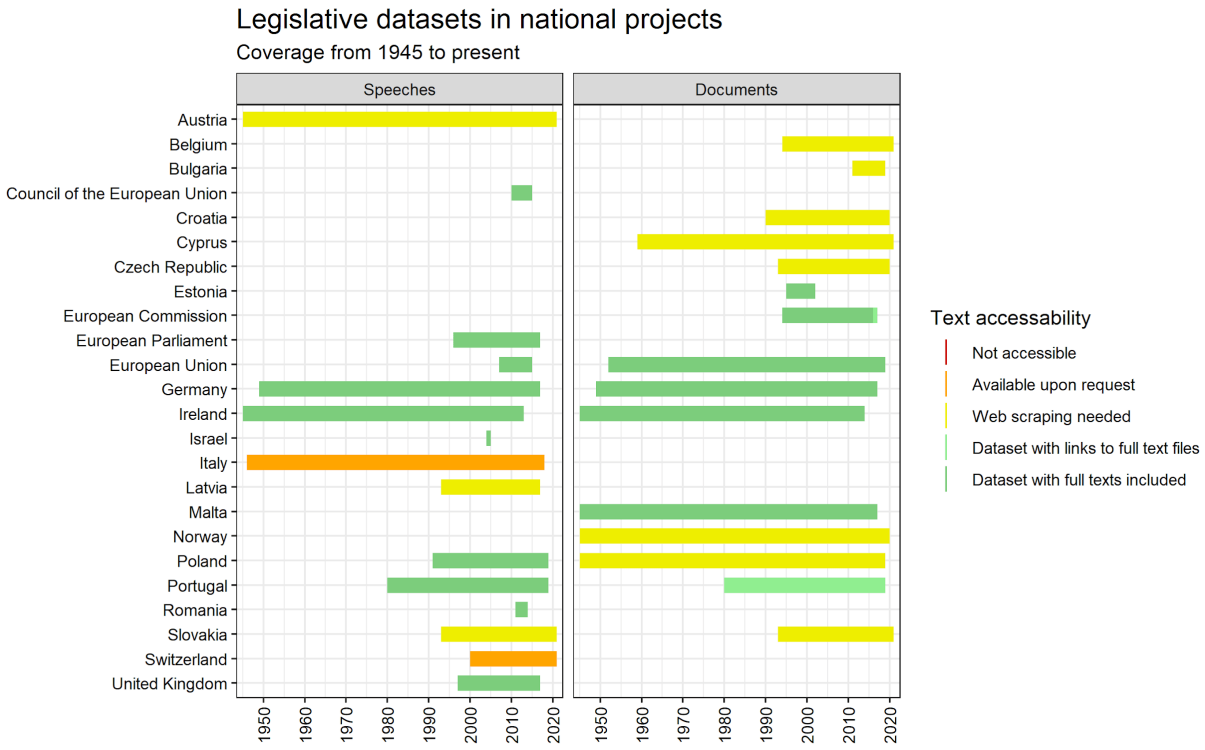


Most CLARIN datasets are suitable for text-as-data analysis, even if the data formats and labelling techniques are somewhat different from those primarily used in the social sciences. The speech corpora are highly processed and richly tagged, using TEI standards for encoding. Most parliamentary corpora are tokenised, lemmatised, and many are tagged for part-of-speech components (such as nouns and verbs). The corpora are primarily available in XML format. The fundamental limitations are similar to that of ParlSpeech: geographic and time limits; moreover, it only covers the input side of the legislative process, as it fails to include collections of legal texts like laws.

A European-level comparison

Secondary datasets outside the abovementioned three large projects are fragmented in coverage and data accessibility. Out of 20 European countries with national projects that produced datasets, only 13 have datasets available as single-file, directly importable databases. Most national speech datasets focus on plenary speeches, while most legislative document datasets present laws. Figure 11 summarises legislative speech and document corpora from national collections as a whole.

Figure 11. Legislative datasets in national projects.

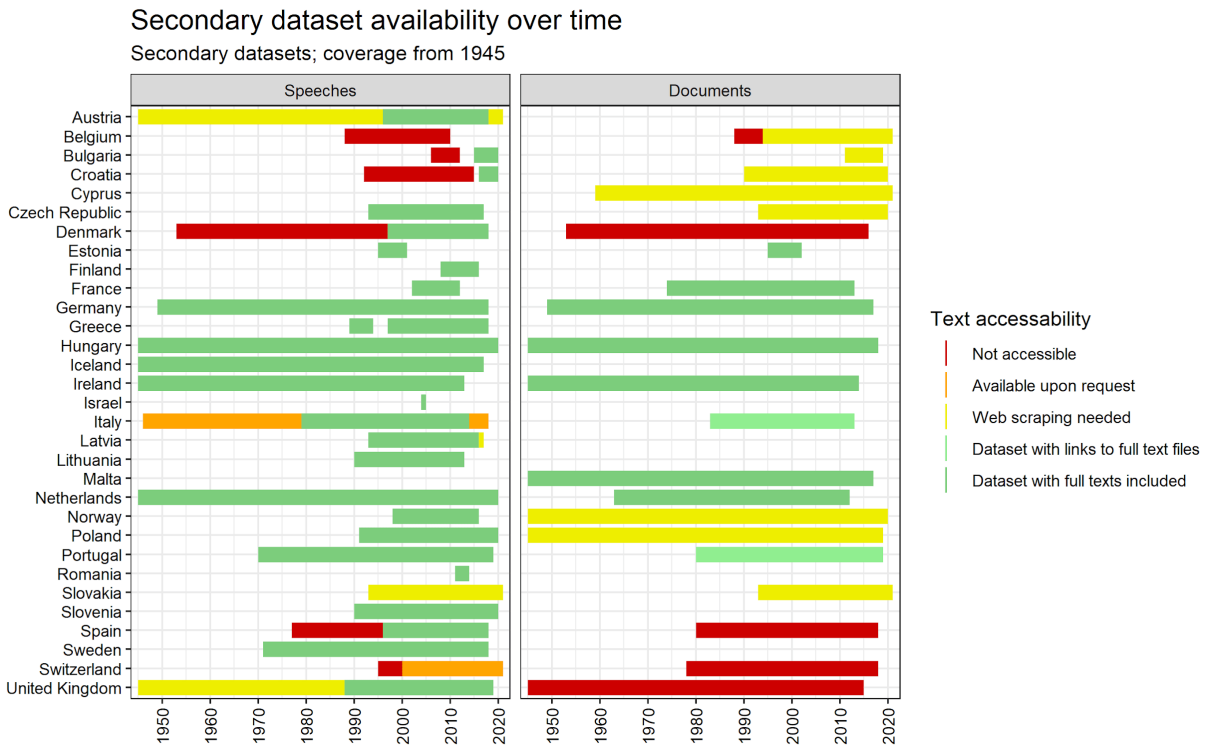


Datasets about European legislative speeches and documents are fragmented and incoherent. Many researchers attempted to build their own textual datasets (often driven by question-specific selection). As a result, there is a proliferation of textual corpora about parliamentary speeches and

laws. However, the available datasets are far from offering comprehensive coverage for European legislative processes, and there is a danger that different quality standards are present in the data.

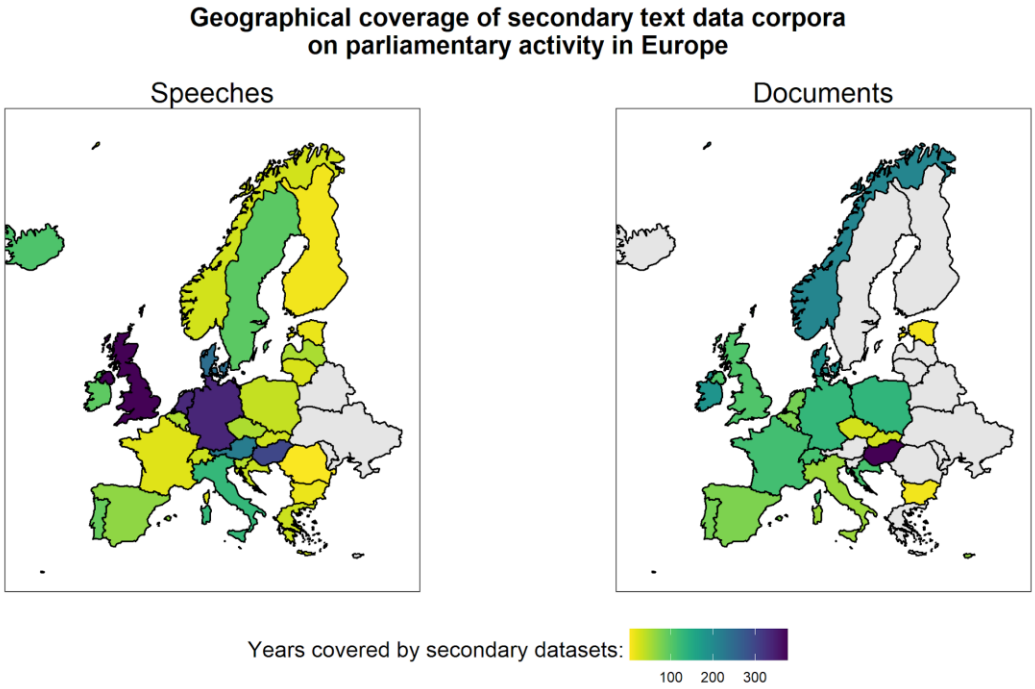
Data about parliamentary speeches and laws is available from the last two decades in several European countries, although genuine, large-N comparative work is still difficult because the datasets are often not comparable. Transnational projects like CAP or CLARIN made real comparative work possible, but even those remain severely limited geographically and temporally. Figure 12 shows the overall temporal coverage of existing secondary corpora in terms of time range and across text types.

Figure 12. Secondary dataset availability over time.



Finally, Figure 13 shows the years covered by secondary speech and document corpora in European countries, strongly suggesting substantial geographic differences in data availability.

Figure 13. *Geographical coverage of secondary text data corpora on parliamentary activity in Europe.*



In conclusion, systematic efforts are needed to overcome the limitations and fragmentation of the existing data infrastructure. Primary databases offer the potential to augment the existing and fragmented secondary datasets both on legislative speeches and legislative documents. However, the potential of parliamentary speech data can only fully be tapped if systematic linkages can be made to legislation (and the text in which it is captured). Parliaments offer access to these related documents but often use a different database or do not precisely link speeches to the debates.

Realising the potential of analysis depends on high-quality representations of the produced text data. Scholars require machine-readable, annotated data frames of systematically linked parliamentary speeches and legislative texts with comparable structures across various countries and long timespans. Parliamentary websites are usually geared towards a non-academic audience

and are designed to identify individual documents rather than offering long-term comparisons across countries, rarely providing sufficiently annotated full-text corpora.

Conclusion

Parliamentary speeches and legislative documents are becoming increasingly available for automated textual analysis. This offers the possibility of analysing previously unimaginable quantities of data. In political science, these techniques are applied for, among other things, comparative agenda studies, ideological scaling of speakers, measuring the law-making power of the legislature and the complexity of speeches.

The need to collect and annotate the necessary text data creates high transaction costs and a significant barrier to entry. Thus, a European infrastructure on political text analysis should facilitate a broader engagement in parliamentary text data collections and their substantial analysis. A coherent database that is well-structured for social research is still needed. While some researchers focus exclusively on legislative proceedings, bills and laws, others examine debates in different EU institutions. Researchers sometimes provide a combination of various speeches, but not necessarily linked to law-making (see the EUSpeech project – Schumacher et al. (2020)). Despite multiple attempts, such as the LinkedEP dataset (Van Aggelen et al., 2017), a coherent speech and legislative database is missing.

National legislative databases are inapt for text-as-data analysis, owing to the need for clear, annotated, and pre-processed text. Various corpora-building projects have been initiated, such as the CLARIN Project, the Comparative Agendas Project (CAP) and the ParlSpeech database. Nonetheless, scholarly projects' timespan, coverage and accessibility exhibit significant

differences. Even international initiatives suffer from certain shortcomings, especially regarding countries from Eastern and Southern Europe.

In sum, legislative text data offers substantial untapped potential for researchers interested in democratic politics and policymaking, but systematic infrastructure building remains a pivotal step towards realising this potential. In some cases, research-oriented infrastructure building even lags behind civic or media coverage (see, for instance, the data visualisations of German parliamentary speeches by Zeit Online)[v]. Such scientific infrastructure building could benefit not just legislative studies but social research at large: text analysis as a method and machine-readable state-of-the-art text data sources can be deployed in a number of research agendas beyond those focusing on parliaments. Furthermore, the blueprints generated for interlinked parliamentary sources can also serve as templates for communities working on similar (and via government-initiated bills evidently connected) fields, such as executive politics and the production of secondary legislation by the regulatory state.

Funding

The research project was supported by the OPTED H2020 project of the European Union (grant nr. 951832); by the Hungarian Ministry of Innovation and Technology NRDI Office and the European Union, in the framework of the RRF-2.3.1-21-2022-00004 Artificial Intelligence National Laboratory project; and by the V-SHIFT Lendület research group of the Hungarian Academy of Sciences.

ORCID iDs

Miklós Sebők, Centre for Social Sciences, Budapest, https://orcid.org/0000-0003-0595-2951

Sven-Oliver Proksch, University of Cologne, https://orcid.org/0000-0002-6130-6498

Christian Rauh, WZB Berlin Social Science Center, https://orcid.org/0000-0001-9357-9506

Péter Visnovitz, Centre for Social Sciences, Budapest, (no ORCID)

Gergő Balázs, Centre for Social Sciences, Budapest, https://orcid.org/0000-0002-1355-7524

Jan Schwalbach, GESIS – Leibniz Institute for the Social Sciences, https://orcid.org/0000-0002-6990-8098

Notes

[i] See: https://www.comparativeagendas.net/datasets_codebooks

[ii] The data collection is made public on the following link: https://opted.eu/results/inventories/. A description of the results is available here: https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D5.1.pdf. The number of identified datasets in our figures stand for disaggregated datasets for specific data types in specific countries. During the review, we broke down larger databases into separate datasets if a database itself held corpora belonging to multiple countries or institutions, or if it included data about different types of legislative speeches and texts. For example, a database from a specific country that included plenary speeches, interpellations and written questions was broken down to three separate datasets, each accounting for one type of text. We acknowledge the possibility of additional sources being available for any given country case.

[iii] The European countries covered by CAP include Belgium, Croatia, Denmark, France, Germany, Hungary, Italy, the Netherlands, Portugal, Spain, Switzerland, and the UK.

[iv] The databases included in our review only partially cover those presented Baumgartner et al. (2019) as part of the CAP project. The differences are due to datasets temporarily inaccessible on their respective project websites. Datasets that we could not download or open were omitted from our review.

[v] https://www.zeit.de/politik/deutschland/2019-09/bundestag-jubilaeum-70-jahre-parlament-reden-woerter-sprache-wandel#s=weltraum

References

Bachrach, Peter and Morton S Baratz (1962) Two Faces of Power. *American political science review* 56(4): 947-952.

Back, Hanna, Marc Debus and Jorge M Fernandes (2021) *The Politics of Legislative Debates.* Oxford University Press.

Bailer, Stefanie (2014) Interviews and Surveys in Legislative Research. *The Oxford handbook of legislative studies*. Oxford University Press, 167-193.

Baumgartner, Frank R, Christian Breunig and Emiliano Grossman (2019) *Comparative Policy Agendas: Theory, Tools, Data.* Oxford University Press.

Baumgartner, Frank R, Christoffer Green-Pedersen and Bryan D Jones (2006) Comparative studies of policy agendas. *Journal of European public policy* 13(7): 959-974.

Bernauer, Julian and Thomas Bräuninger (2009) Intra-Party Preference Heterogeneity and Faction Membership in the 15th German Bundestag: A Computational Text Analysis of Parliamentary Speeches. *German Politics* 18(3): 385-402.

Blei, David M (2012) Probabilistic topic models. *Communications of the ACM* 55(4): 77-84.

Borghetto, Enrico and Ana Maria Belchior (2020) Party Manifestos, Opposition and Media as Determinants of the Cabinet Agenda. *Political Studies* 68(1): 37-53.

Borghetto, Enrico and Laura Chaqués-Bonafont (2019) Parliamentary Questions. *Comparative Policy Agendas: Theory, Tools, Data*. First edition ed.: Oxford University Press, 282-299.

Carroll, Royce and Keith Poole (2014) Roll call analysis and the study of legislatures. *The Oxford handbook of legislative studies*. 103-125.

Cross, James P and Henrik Hermansson (2017) Legislative amendments and informal politics in the European Union: A text reuse approach. *European Union Politics* 18(4): 581-602.

de Jong, Franciska, Dieter Van Uytvanck, Francesca Frontini, et al. (2022) Language matters. The European research infrastructure CLARIN, today and tomorrow. *CLARIN. The infrastructure for language resources*. de Gruyter, 31-57.

De Ruiter, Rik and Jelmer Schalk (2017) Explaining cross-national policy diffusion in national parliaments: A longitudinal case study of plenary debates in the Dutch parliament. *Acta Politica* 52: 133-155.

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

Decadri, Silvia and Constantine Boussalis (2020) Populism, party membership, and language complexity in the Italian chamber of deputies. *Journal of Elections, Public Opinion and Parties* 30(4): 484-503.

Döring, Holger and Sven Regel (2019) Party Facts: A database of political parties worldwide. *Party politics* 25(2): 97-109.

Druckman, James N, Thomas J Leeper and Kevin J Mullinix (2014) The experimental study of legislative behaviour. *The Oxford handbook of legislative studies*. Oxford University Press Oxford, 194-212.

Erjavec, Tomaž, Maciej Ogrodniczuk, Petya Osenova, et al. (2022) The ParlaMint corpora of parliamentary proceedings. *Language resources and evaluation*. 1-34.

Finseraas, Henning, Bjørn Høyland and Martin G Søyland (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.

Gava, Roy, Julien M Jaquet and Pascal Sciarini (2021) Legislating or rubber-stamping? Assessing parliament's influence on law-making with text reuse. *European Journal of Political Research* 60(1): 175-198.

Geese, Lucas (2020) Immigration-related Speechmaking in a Party-constrained Parliament: Evidence from the 'Refugee Crisis' of the 18th German Bundestag (2013–2017). *German Politics* 29(2): 201-222.

Goet, Niels D (2019) Measuring Polarization with Text Analysis: Evidence from the UK House of Commons, 1811–2015. *Political Analysis* 27(4): 518-539.

Goplerud, Max (2021) Methods for Analyzing Parliamentary Debates. *The Politics of Legislative Debates*. Oxford University Press, 72-90.

Green-Pedersen, Christoffer (2023) Why all this attention on issue competition? *The Routledge Handbook of Political Parties* 1.

Greene, Derek and James P Cross (2017) Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis* 25(1): 77-94.

Grimmer, Justin and Brandon M Stewart (2013) Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3): 267-297.

Hansen, Dorte Haltrup, Costanza Navarretta, Lene Offersgaard, et al. (2019) Towards the Automatic Classification of Speech Subjects in the Danish Parliament Corpus. CEUR Workshop Proceedings, 166-174.

Hillard, Dustin, Stephen Purpura and John Wilkerson (2008) Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics* 4(4): 31-46.

Hoerner, Julian M (2019) Adding Fuel to the Flames? Politicisation of EU Policy Evaluation in National Parliaments. *Politische vierteljahresschrift* 60(4): 805-821.

Jurafsky, Daniel and James H Martin (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, N.J: Prentice Hall.

Kinski, Lucy (2021) *European representation in EU national parliaments.* Springer Nature.

Kiss, Rebeka and Miklós Sebők (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.

Kraft, Patrick W (2018) Measuring Morality in Political Attitude Expression. *The Journal of Politics* 80(3): 1028-1033.

Lauderdale, Benjamin E and Alexander Herzog (2016) Measuring Political Positions from Legislative Speech. *Political Analysis* 24(3): 374-394.

Laver, Michael (2021) Analyzing the Politics of Legislative Debate. *The Politics of Legislative Debates*. Oxford University Press, 21-33.

Laver, Michael, Kenneth Benoit and John Garry (2003) Extracting Policy Positions from Political Texts Using Words as Data. *American political science review* 97(2): 311-331.

Lehmann, Felix (2023) Talking about Europe? Explaining the salience of the European Union in the plenaries of 17 national parliaments during 2006–2019. *European Union Politics* 24(2): 370-389.

Lin, Nick and Moritz Osnabrügge (2018) Making comprehensible speeches when your constituents need it. *Research & Politics* 5(3): 1-8.

Lind, Fabienne, Jakob-Moritz Eberl, Tobias Heidenreich, et al. (2019) Computational Communication Science| When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction. *International Journal of Communication* 13: 21.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, et al. (2015) Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis* 23(2): 254-277.

Navarro, Julien and Sylvain Brouard (2014) Who Cares About the EU? French MPs and the Europeanisation of Parliamentary Questions. *The Journal of Legislative Studies* 20(1): 93-108.

Osnabrügge, Moritz, Sara B Hobolt and Toni Rodon (2021) Playing to the Gallery: Emotive Rhetoric in Parliaments. *American political science review* 115(3): 885-899.

Proksch, Sven-Oliver and Jonathan B Slapin (2010) Position Taking in European Parliament Speeches. *British Journal of Political Science* 40(3): 587-611.

Proksch, Sven-Oliver and Jonathan B Slapin (2014) *The Politics of Parliamentary Debate: Parties, Rebels and Representation.* Cambridge University Press.

Proksch, Sven-Oliver, Christopher Wratil and Jens Wäckerle (2019a) Testing the Validity of Automatic Speech Recognition for Political Text Analysis. *Political Analysis* 27(3): 339-359.

Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, et al. (2019b) Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches: Multilingual Sentiment Analysis. *Legislative Studies Quarterly* 44(1): 97-131.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, et al. (2010) How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science* 54(1): 209-228.

Rauh, Christian (2018) Validating a sentiment dictionary for German political language—a workbench note. *Journal of Information Technology & Politics* 15(4): 319-343.

Rauh, Christian (2021) One agenda-setter or many? The varying success of policy initiatives by individual Directorates-General of the European Commission 1994–2016. *European Union Politics* 22(1): 3-24.

Rauh, Christian, Bart Joachim Bes and Martijn Schoonvelde (2020) Undermining, defusing or defending European integration? Assessing public communication of European executives in times of EU politicisation. *European Journal of Political Research* 59(2): 397-423.

Rauh, Christian and Pieter De Wilde (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, Christian and Jan Schwalbach (2020) The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. V1 ed.: Harvard Dataverse.

Remschel, Tobias and Corinna Kroeber (2022) Every Single Word: A New Data Set Including All Parliamentary Materials Published in Germany. *Government and Opposition* 57(2): 276-295.

Rheault, Ludovic and Christopher Cochrane (2020) Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora. *Political Analysis* 28(1): 112-133.
Schumacher, Gijs, Nicolai Berk, Christian Pipal, et al. (2020) EUSpeech V2. DOI: 10.17605/OSF.IO/C79HA.
Schwalbach, Jan (2022) Going in circles? The influence of the electoral cycle on the party behaviour in parliament. *European Political Science Review* 14(1): 36-55.
Sebők, Miklós and Zoltán Kacsuk (2021) The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach. *Political Analysis* 29(2): 236-249.
Sebők, Miklos, Zoltán Kacsuk and Ákos Máté (2022) The (real) need for a human touch: testing a human–machine hybrid topic classification workflow on a New York Times corpus. *Quality & Quantity* 56(5): 3621-3643.
Sebők, Miklós, Csaba Molnár and Bálint György Kubik (2017) Exercising control and gathering information: the functions of interpellations in Hungary (1990–2014). *The Journal of Legislative Studies* 23(4): 465-483.
Slapin, Jonathan B and Sven-Oliver Proksch (2014) Words as Data: Content Analysis in Legislative Studies. *The Oxford handbook of legislative studies*. First Edition ed. New York, NY: Oxford University Press, 126-144.
Slapin, Jonathan B and Sven-Oliver Proksch (2008) A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science* 52(3): 705-722.
Van Aggelen, Astrid, Laura Hollink, Max Kemman, et al. (2017) The debates of the European Parliament as Linked Open Data. *Semantic Web* 8(2): 271-281.
Vliegenthart, Rens, Stefaan Walgrave and Brandon Zicha (2013) How preferences, information and institutions interactively drive agenda-setting: Questions in the B elgian parliament, 1993–2000. *European Journal of Political Research* 52(3): 390-418.
Watanabe, Kohei (2021) Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures* 15(2): 81-102.
Widmann, Tobias and Maximilian Wich (2022) Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text. *Political Analysis*. DOI: 10.1017/pan.2022.15. 1-16.
Wilkerson, John, David Smith and Nicholas Stramp (2015) Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach. *American Journal of Political Science* 59(4): 943-956.
Winzen, Thomas, Rik De Ruiter and Jofre Rocabert (2018) Is parliamentary attention to the EU strongest when it is needed the most? National parliaments and the selective debate of EU policies. *European Union Politics* 19(3): 481-501.

Miklós Sebők is a Senior Research Fellow at the Centre for Social Sciences in Budapest and the co-leader of Work Package 5 in the OPTED H2020 project tasked with the creation of a new Europe-wide database of legislative speeches and documents.

Sven-Oliver Proksch is co-leader of work package 5 in the OPTED project and Professor of Political Science and Chair for European and Multilevel Politics at the Cologne Center for Comparative Politics at the University of Cologne.

Christian Rauh is a senior researcher in the Global Governance unit of the WZB Berlin Social Science Center. In the OPTED project he primarily supports the provision of parliamentary text data with a particular emphasis of enabling targeted data extraction via online tools.

Péter Visnovitz was a Junior Research Fellow of the poltextLAB at the Centre for Social Sciences in Budapest and a research assistant in the OPTED project.

Gergő Balázs was a research assistant in the OPTED project at the Centre for Social Sciences in Budapest.

Jan Schwalbach was a doctoral researcher at the Cologne Center for Comparative Politics at the University of Cologne and is currently a postdoctoral researcher at GESIS. In the OPTED project he primarily supports the creation of the new *ParlLawSpeech* data set.