Digital Parliamentary Data in Action (DiPaDA 2024) – Introduction

Daniel Brodén[1,†], Mats Fridlund[1,†], Matti La Mela[2,†] and Albert Wendsjö[3,†]

[1]*Department of Literature, History of Ideas, and Religion, University of Gothenburg*
[2]*Department of ALM, Uppsala University*
[3]*Department of Political Science, University of Gothenburg*

Abstract

The workshop Digital Parliamentary Data in Action 2024 (DiPaDa 2024) took place in Reykjavik, Iceland, on 28 May, co-located with The 8th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2024). The workshop, along with its predecessor organised in Uppsala in 2022, supports the advancement of research using parliamentary datasets, which present both opportunities and challenges for interdisciplinary research and infrastructure development especially in the digital humanities and social sciences.

Keywords

parliamentary data, digital humanities, parliamentary studies

1. Introduction

The workshop Digital Parliamentary Data in Action 2024 (DiPaDa 2024) took place in Reykjavik, Iceland, on 28 May, co-located with The 8th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB, 2024). This was the second edition of the workshop, following its inaugural event in Uppsala, Sweden, in 2022 (La Mela, Norén, and Hyvönen 2022). The workshop brought together scholars primarily from the humanities, social sciences and data sciences to showcase ongoing work on curating and utilising digital parliamentary data in Digital Humanities research, resources, and applications.

The 2024 DiPaDa workshop, along with its predecessor, supports the advancement of research using parliamentary datasets, which present both opportunities and challenges for interdisciplinary research and infrastructure development. As highlighted by its title, Parliamentary Data in Action, the workshop focuses on the active use of data as a means to further Humanities and Social Sciences (HSS) knowledge about parliaments, rather than solely emphasising the technical feasibility of developing digital corpora, tools and infrastructures. In the 2020s, parliamentary datasets are increasingly being utilised by researchers to explore and analyse politics,

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, Reykjavik, Iceland, May 28, 2024.
[†] These authors contributed equally.
daniel.broden@lir.gu.se (D. Brodén); mats.fridlund@gu.se (M. Fridlund); matti.lamela@abm.uu.se (M. La Mela); albert.wendsjo@gu.se (A. Wendsjö)
0000-0002-5914-1516 (D. Brodén); 0000-0002-5759-0027 (M. Fridlund); 0000-0003-0340-9269 (M. La Mela); 0000-0002-7627-5372 (A. Wendsjö)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

policy-making and parliamentary culture, as well as to investigate broader societal and cultural dynamics through parliamentary texts as lenses. However, current research often adopts more interdisciplinary and context-sensitive approaches (see Guldi 2024), which have underscored the limitations of existing datasets by highlighting the need for improvements in areas such as OCR-quality, metadata and annotation to better support nuanced analyses.

At the same time, the drive toward standardisation of parliamentary data infrastructures, such as the Parla-CLARIN[1] initiative, has increased the requirements for structured datasets and cross-national compatibility, raising important questions about accessing data that often spans multiple historical layers. One prominent example is ParlaMint, a parliamentary corpus that currently contains uniformly annotated parliamentary debates from over 29 countries (Erjavec et al. 2024). In Sweden, the SWERIK project is digitising Swedish parliamentary debates (1867-2022) into linked open data. Notably, it has introduced a system of versioning that tracks the evolution of text recognition and annotation quality across different parts of the dataset (Yrjänäinen et al. 2024), improving transparency and usability for researchers. Moreover, accessibility has also been a key focus in historical parliamentary data projects. For instance, the Finnish ParliamentSampo project utilises linked open data and semantic web technologies to both publish and enrich parliamentary debates data. It also provides an interactive user interface that facilitates searching and filtering within the debates (Hyvönen et al. 2025). In this context, the two DiPaDa workshops have aimed to promote dialogue between scholars who curate parliamentary datasets and those who use them to answer their research questions.

Based on the presentations at this years' workshop, we would like to draw out a few more general tendencies in current research on parliamentary data.

2. Tendencies at this year's workshop

2.1. Research-driven studies

Firstly, while the first workshop primarily centered on datasets and their research potentials, the papers presented at DiPaDa 2024 placed stronger emphasis on findings from research-driven studies. This tendency seems to highlight a growing focus on concrete research outcomes rather than solely on investigating the the affordances of the underlying infrastructures. Several papers at the DiPaDA 2024 workshop examined concepts and thematic frameworks within political discourse as reflected in the parliamentary debates. These papers explored a range of topics, including democracy, artificial intelligence (AI), the marketisation of education and access rights to nature.

We also see that the papers demonstrated, on the one hand, an increased awareness of the specific contexts of national parliaments and their respective languages, but on the other hand, there is a discernible trend towards extending studies beyond individual national parliaments or languages. This may result from the wider availability of parliamentary corpora and adoption of shared standards, along with the development of multilingual research tools. For instance, Žagar and Moats (2025, in this issue) analyse how politicians discuss 'AI' in the UK and Slovenia, while Skubic (2024, workshop presentation) studied gender representation and power relations

[1] https://github.com/clarin-eric/parla-clarin

in three European parliaments. In a similar vein, La Mela (2024, workshop presentation) studied the conceptual history of 'allemansrätten', a Nordic right of public access to nature, through the use of Finnish and Swedish parliamentary debates.

In addition, there are studies that compared debates not only across different parliaments, but also over time. For example, Turunen and Ihalainen (2025, in this issue) also examine how the use of democracy varies across parliaments and how its meaning has developed across time. Similarly, Brodén et al. (2025, in this issue) analyse the development of word compounds over time to identify wider shifts in discourse related to 'terrorism' and 'markets'.

2.2. Application of novel and advanced methods

Secondly, and partly related to the above, we can observe that the workshop papers employed a wide range of methods in their analyses of parliamentary debates. Examples of methods include topic modeling, word embeddings, word co-occurance networks and Named Entity Recognition (NER). Some papers also combine these computational methods with close reading, adopting a mixed method design to deepen their investigations.

Several of the papers presented various approaches to studying language and language use in the parliaments. Ristilä (2025, in this issue) applies computational methods to explore how different foreign languages are used in the Finnish parliament, complementing the analysis with close readings to gain deeper insights into the results. Stefánsdóttir and Ingason (2025, in this issue) explore the effects of transcription on the recorded style of the speeches of MPs. Olsson et al. (2025, in this issue) showcase the use of word vectors to study semantic shifts and the changing word contexts for the concept of terrorism.

In addition, there are presentations where several NLP tools were used in combination. For example, Bleier and Zeilinger (2025, in this issue) apply and compare three different topic modeling methods with the aim of creating a subject index for the Holy Roman Empire's Diet Records. Wendsjö (2024, workshop presentation) employed NER to identify places mentioned in the Swedish Parliament and combined this with topic modeling to examine how different locations were framed over time.

2.3. Large Language Models

Thirdly, while only a few presentations at the DiPaDa 2024 workshop explicitly focused on what is commonly referred to as AI, a significant topic of discussion was the application and potential of Large Language Models (LLMs) and other AI associated methods in research on parliamentary datasets. While ongoing advancements in NLP are likely to continue providing new opportunities for scholarly work on parliamentary debates, it is important to critically reflect on both the possibilities and the challenges these methodologies currently present. In particular, we highlight two concerns that have been central to recent discussions on the use of LLMs in research.

Recent research has shown that LLMs like ChatGPT open new avenues for analysing large-scale corpora more easily than before. For example, GPTs have been found to outperform crowdworkers, across a large series of text annotation tasks (Gilardi, Alizadeh, and Kubli 2023). However, using proprietary or closed-source LLMs, such as ChatGPT, present significant

challenges. Researchers have limited insights into how these models are trained, what data they are trained on and, perhaps most critically, what specific version is being used during analysis (Palmer, Smith, and Spirling 2024). This has important implications for research, as it can significantly influence the conclusions drawn from studies and affect the transparency, and also the replicability of results (Barrie, Palmer, and Spirling 2024). Some of these concerns are addressed in the workshop proceedings. For instance, Karlsson et al. (2025, in this issue) found that the choice of particular sampling strategies can have a substantial impact on the conclusions derived from the analysis of parliamentary debates.

While it may still be some time before the scholarly community reaches consensus on 'best practices' for using these methodologies, several current suggestions have emerged: i) use open-source LLMs that can be run locally to enhance reproducibility; ii) consider how measurement errors might affect downstream conclusions and, if possible, adjust for these errors; and iii) demonstrate the robustness of conclusions across different models, prompts and pre-processing decisions. Although these suggestions are hardly definite, we believe it is crucial for researchers to critically reflect on the research design implications when incorporating LLMs into their studies. Key questions to consider include: Is the use of an LLM necessary? What gains does it offer over more traditional approaches? Can the study be easily replicated? How might the chosen method influence the downstream analysis and thus the conclusions? While these questions are not new to academic research, nor unique to the use of LLMs, they can easily be overlooked amid the current hype surrounding these technologies.

3. Contributions and peer review

The workshop accepted both long (20 min) and short (10 min) presentations and the proceedings consequently include long articles (8–16 pages) as well as short (4–8 pages) articles. Both the paper proposals for the workshop and the post-proceedings underwent a peer-review process, with two anonymous reviewers with expertise in digital research using historical parliamentary data. Additionally, the workshop had an international program committee that approved the workshop program and the papers for the post-proceedings. The DiPaDA 2024 workshop received twelve paper proposals, of which eleven were accepted for presentation. Of these, eight papers (seven long papers and one short paper) were submitted for review in the post-proceedings and were accepted for publication.

Organisers and Program Committee The DiPaDA 2024 workshop was organised by Daniel Brodén, Mats Fridlund, Matti La Mela, and Albert Wendsjö. The Program Committee included the following scholars:

- Kaspar Beelen, The Alan Turing Institute
- Daniel Brodén, GRIDH, University of Gothenburg
- Kimmo Elo, University of Turku
- Tomaž Erjavec, Jozef Stefan Institute
- Darja Fišer, University of Ljubljana
- Mats Fridlund, GRIDH, University of Gothenburg
- Jo Guldi, Emory University

- Eero Hyvönen, Aalto University and University of Helsinki (HELDIG)
- Pasi Ihalainen, University of Jyväskylä
- Matti La Mela, Uppsala University
- Måns Magnusson, Uppsala University
- Bruno Martins, University of Lisbon
- Costanza Navarretta, University of Copenhagen
- Gustaf Nelhans, SSLIS, University of Borås
- Fredrik Mohammadi Norén, Malmö University
- Leif-Jöran Olsson, Språkbanken Text, University of Gothenburg
- Jouni Tuominen, Aalto University and University of Helsinki
- Albert Wendsjö, University of Gothenburg
- Magnus P. Ängsal, University of Gothenburg
- Patrik Öhberg, The SOM Institute, University of Gothenburg

Acknowledgments

We would like to sincerely thank everyone who supported the workshop, especially the programme committee, the anonymous peer-reviewers, the session chairs, all participating authors and the DHNB 2024 conference organisers. Furthermore, the DiPaDa 2024 workshop was organised with support from the two Swedish national research infrastructures Huminfra (funded by VR, contract no. 2021-00176) and Swe-Clarin (funded by VR, contract no. 2017–00626). We also extend our gratitude to the DiPaDA 2022 organisers, for laying the groundwork for this year's event.

References

Barrie, Christopher, Alexis Palmer, and Arthur Spirling. 2024. "Replication for Language Models Problems, Principles, and Best Practice for Political Science."

Erjavec, Tomaž, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, et al. 2024. "ParlaMint II: advancing comparable parliamentary corpora across Europe." *Language Resources and Evaluation,* 1–32.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.

Guldi, Jo. 2024. "The Revolution in Text Mining for Historical Analysis is Here." *The American Historical Review* 129 (2): 519–543.

Hyvönen, Eero, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2025. "Publishing and using parliamentary Linked Data on the Semantic Web: ParliamentSampo system for Parliament of Finland." *Semantic Web* 16 (1): SW–243683.

La Mela, Matti, Fredrik Norén, and Eero Hyvönen. 2022. "Digital Parliamentary Data in Action (DiPaDA 2022): Introduction." In *Digital Parliamentary Data in Action (DiPaDA 2022) workshop,* 1–8. https://ceur-ws.org/Vol-3133/paper00.pdf.

Palmer, Alexis, Noah A Smith, and Arthur Spirling. 2024. "Using proprietary language models in academic research requires explicit justification." *Nature Computational Science* 4 (1): 2–3.

Yrjänäinen, Väinö Aleksi, Fredrik Mohammadi Norén, Robert Borges, Johan Jarlbrink, Lotta Åberg Brorsson, Anders P Olsson, Pelle Snickars, and Måns Magnusson. 2024. "The Swedish parliament corpus 1867–2022." In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024),* 16100–16112.

Finding Patterns Across Multiple Time-Series Datasets: 'Democracy' in the British, Finnish and Swedish Parliaments and Press in the Twentieth Century

Risto Turunen[a], Pasi Ihalainen[a]

[a] *University of Jyväskylä*

Abstract

This study investigates the contextual variation of nouns and adjectives associated with 'democracy' in twentieth-century Britain, Sweden and Finland. By quantitatively analysing parliamentary debates and conservative and liberal press sources from three countries, we explore the interplay between political and media discourses on democracy. Our methodology involves visualising the use of democratic terms over time and applying correlation analysis to detect patterns between multiple datasets. We use the Pearson correlation coefficient to quantify similarities between word frequency time series, identifying strong and weak relationships. Our findings reveal the strongest correlations between similar political terms within the same dataset and between identical terms in different national datasets from the same country. Notably, there is a statistically strong correlation between media and parliamentary discourses, partly explained by entanglements between journalism and political parties. Although transnational correlations are weaker than intra-national ones, they still suggest common trends across nations. Close reading highlights key periods where democratic discourse peaked, corresponding to global events like the challenge of totalitarianism in the 1930s, the rise of the left in the 1940s, the emergence of social movements and the extension of democracy after 1968, and the dissolution of the Eastern bloc in the early 1990s. This study introduces time-series methods to quantitative and comparative conceptual history. We demonstrate empirically a strong linkage between political discourses in parliament and the press, showing that parliamentary speech has often been in sync with broader societal debates – at least in the case of democracy and the quality press.

Keywords

Democracy, text mining, time-series analysis, newspapers, parliamentary debates

1. Introduction

Studies on parliamentary discourse may be criticised for concentrating on the speeches of isolated political elites. To test this claim, in this article we analyse the macro-level contextual variation of nouns and adjectives related to 'democracy' in major British, Finnish and Swedish forums of political discourse in the twentieth century: the national parliaments and leading newspapers and periodicals. We do so by comparing parliamentary plenary debate data (from *Hansard* including both houses of the British parliament, Eduskunta and Riksdag) to press data (Britain: *Guardian* and *Times*; Finland: *Helsingin Sanomat* and *Suomen Kuvalehti*; Sweden: *Dagens Nyheter* and *Svenska Dagbladet*). By including both parliamentary debates and papers that were generally recognised to have either liberal or conservative leanings, our study offers a fresh perspective on the relation between discourses on democracy articulated by politicians and journalists reporting on and contributing to them. We acknowledge that these two groups were closely connected; especially in the early twentieth century,

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, Reykjavik, Iceland, May 28, 2024.
EMAIL: risto.j.turunen@jyu.fi (A. 1); pasi.ihalainen@jyu.fi (A. 2)
ORCID: 0000-0002-8898-1274 (A. 1); 0000-0002-5468-4829 (A. 2)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

they often consisted of the same people. We also recognise that a purely quantitative approach to mentions of democracy does not measure the degree to which a paper spoke for or against it. Particularly in the interwar period, critics of democracy were often most active in writing opinion pieces in editorials, columns and letters to the editor. Quantification thus reflects the intensity of debate on democracy rather than its evaluation. The ideal would have been to include papers with a wider range of political affiliations but suitable comparative data from each country were not yet available.

Our focus on 'democracy' here is derived from our ongoing project, writing long-term comparative conceptual history of representative democracy (Political Representation: Tensions between Parliament and the People from the Age of Revolutions to the 21st Century | University of Jyväskylä). The concept of democracy was at the heart of politics in much of the twentieth century, starting with calls to extend suffrage and electoral reforms that preceded or followed the First World War (in Finland in 1906 under Russian sovereignty, in Britain and Sweden in 1918), and parliaments have constituted the core of the representative form of democracy. Debate on the state of democracy was accelerated by concerns about a crisis of parliamentarism and democracy under the threat of authoritarian regimes in the 1920s and 1930s. What was becoming generally called 'parliamentary democracy' survived in these countries during the Second World War despite both internal and external threats. After a period of procedural and controlled but ideologically divided democracy in the 1950s (Müller, 2011; Corduwener, 2016; Conway, 2020), the concept of democracy started to undergo major shifts: its evaluation became near-universally positive, and its scope and content extended from constitutional questions to many more areas of life, especially after the student protests of 1968 (Kurunmäki, Nevers, and te Velde, 2018; Scholl, 2024).

We have previously used word embeddings to locate words whose contextual similarity to 'democratic' has decreased or increased over time, combining them with bigram analysis by manually classifying word pairs including 'democratic'. In addition to demonstrating extension of the concept quantitatively, we detected a conceptual expansion of democracy regarding its increased abstractness and proceduralisation towards the end of the twentieth century (Bonin et al., under review). As it is crucial to understand whether such changes were driven by elected representatives in the parliament, journalists in the press, both in interaction, or through other forums such as theoretical debates in the social sciences, we are here comparing discursive trends in the first two of these forums.

Our methodological choice stems from the observation that quantitative time-series analysis has not been applied much in writing computer-assisted conceptual history or digital history in general. Though interest in text mining, especially applied to parliamentary debates, has been rising (Blaxill and Beelen, 2016 on regression analysis on the British parliament; Ihalainen et al., 2022 on the conceptual history of the British parliament; Elo and Karimäki, 2021 and Hyvärinen et al., 2023 on conceptual analysis of the Finnish parliament; Brodén et al., 2023 on the conceptual history of the Swedish parliament; Turunen, 2024 on text network analysis in the Finnish parliament), the similarities between multiple time series have been identified based on manual examination of visualisations, with only a few exceptions thus far (Wevers et al., 2020). Although time series of words or topics are often generated automatically in digital humanities (Lansdall-Welfare et al., 2017; Snickars, 2022), quantitative time-series methods used in economic and demographic history (Hudson and Ishizu 2017, 129–162) are rare in historical text mining. While the human brain is good at spotting patterns in visualisations (Cairo, 2016), applying time-series methods might help historians to distinguish between real and imagined patterns. Besides our general interest in a more data-driven approach to time-series data, we are specifically interested in methods that automatically find connections across multiple datasets of different character. One common criticism of computational analysis is that it focuses on one dataset at a time, thus making interpretations more one-dimensional than traditional research designs in which multiple source groups are woven together into a rich narrative (Arguing with Digital History – Working Group, 2017). Experiments such as ours may help develop a digital humanities methodology that has both breadth, through large-scale datasets, and depth, by connecting multiple datasets. We want to contribute to recent methodological discussions in digital humanities that aim to bridge the traditional divide between quantitative and qualitative scholarship (Roller, 2023).

In what follows, we begin by presenting our selected datasets and providing a critical assessment of our methodological approach to correlation analysis, addressing both its benefits and limitations. Then, we calculate correlation coefficients for the relative frequencies of the vocabulary of 'democracy' across datasets and offer a detailed interpretation of the findings, supported by visual

representations of the data. To contextualise these quantitative results, we conduct a comparative close reading of national parliamentary debates on democracy, examining both interwar and postwar periods. The analysis culminates in five empirically grounded theses. To conclude, we explore future directions for finding patterns across multiple time-series datasets.

2. Data

Our main datasets include parliamentary plenary debates from Britain, Finland and Sweden between 1920 and 2001, accessed through the People and Parliament interface (for the exemplary Dutch corpus see People & Parliament uu.nl) that we have constructed in collaboration with Utrecht University Research Software Lab. The interface provides parallel text-mining tools for national datasets from nine North-West European countries since the nineteenth century (Ihalainen et al., 2022). Its British data has been enriched by Jo Guldi and Steph Buongiorno, Southern Methodist University, Dallas, the Finnish data by the Semantic Computing Research Group, Aalto University, and the Swedish data by Fredrik Norén's team, Umeå University. Parliamentary debates encompass several political ideologies simultaneously, reflecting ideologically diversified discourses. Undoubtedly this data may be biased due to certain groups being overrepresented, directing agenda-setting or talking about democracy in great numbers and in ways that differ from mainstream understandings.

In addition, we selected influential quality newspapers with a broadly known ideological inclination from each country, collecting the data from several providers: for the *The Times* and *The* [*Manchester*] *Guardian* / *Observer* from the I-Analyzer interface of Utrecht University, *Suomen Kuvalehti* and *Helsingin Sanomat* from the interfaces of the National Library of Finland and Merkki (Media Museum and Archives), and *Svenska Dagbladet* and *Dagens Nyheter* from the National Library of Sweden's interface. These were all published in the capital except *The Guardian* and can broadly be categorised into conservative and liberal/leftist strands even if they had not all declared themselves organs of a specific party (Rydén 2001, 153; Sandlund 2001, 317, 329; Vares and Siltala, 2016, 248–249, 187–195, 234–235, 255–258, 437–438; Jensen-Eriksen et al., 2019, 48–55; Chapman, 2005, 116–117; McNair, 2011, 52–54). To better understand the potential influence of ideological biases on references to democracy, we selected newspapers that differ in political leanings. This approach helps identify whether frequency trends in discussions about democracy are linked to a newspaper's political bias.

Table 1: Datasets.

DATASET	YEARS	VOLUME
Hansard	1920-2001	1,102,089,681 words
Riksdag	1920-2001	259,165,079 words
Eduskunta	1920-2001	117,228,145 words
The Times	1920-2001	3,996,728,581 words
The Guardian / Observer	1920-2001	3,493,400,706 words
Svenska Dagbladet	1920-2001	893,880 pages
Dagens Nyheter	1920-2001	976,866 pages
Suomen Kuvalehti	1920-2001	245,603 pages
Helsingin Sanomat	1920-1989	706,575 pages

Table 1 summarises the datasets used. The extent of the datasets varies significantly: the dataset on the two houses of the British parliament contains many more words than the unicameral or less talkative others, suggesting that there may be more variation in word frequencies over time in the Finnish and Swedish parliaments as variation tends to be larger in smaller samples. Additionally, there was no access to directly comparable frequency data for the Finnish liberal newspaper (*Helsingin Sanomat*) from 1990 to 2001. Consequently, comparisons between the Finnish parliament and media data are limited to 1920–1989, whereas in Britain and Sweden, the datasets are complete for the whole period

1920–2001. As a conservative Finnish paper we are using *Suomen Kuvalehti*, which came out weekly. The search terms were 'democracy' and 'democratic' for the UK; 'demokrati', 'demokratin', 'demokratins', and 'demokratisk*' for Sweden; and 'demokratia*' and 'demokraatti*' for Finland.

The quality of text recognition also differs somewhat across the parliamentary datasets, though all the used datasets have been edited and are constantly being enriched. Our assessments, which are based on sampling 100 parliamentary speeches from each year and then measuring the quality of text recognition with the LAS tool (Mäkelä, 2016), suggests that the Hansard dataset edited by Jo Guldi's research group has the best quality (98.2–99.4% per word), while the edition of the Riksdag dataset we used delivers relatively the weakest quality (84.7–90.0% per word) which would be higher (around 91.3% on average) with the recently updated SweRik dataset.

3. Methodology

Quantitative analysis based on the Pearson correlation coefficient (PCC) is a standard statistical method. It measures how strongly two variables are linearly related, i.e., how much one variable (e.g., height) changes in units of its standard deviation when the other variable (e.g., weight) changes by one standard deviation. In time-series similarity analysis, the PCC measures whether word frequencies increase and decrease simultaneously. The PCC values can range from -1 to 1: a value of 1 indicates a perfect positive correlation where every increase in word frequency in dataset A is matched by a simultaneous increase in dataset B. Conversely, a value of -1 indicates a perfect negative correlation, where every increase in word frequency in dataset A corresponds to a simultaneous decrease in dataset B. The closer the PCC values are to 0, the weaker the relationship between the two variables (Derrick and Thomas, 2004, 194–196). The PCC measures the strength of linear relationships; thus, it might not be best for assessing time-series similarity when the relationship between variables is non-linear. The efficacy of the PCC in measuring non-linear relationships remains a topic of discussion (van den Heuvel and Zhan, 2022).

The main strengths of the PCC are its mathematical simplicity and easy interpretability: the mean of the time series is calculated, and for each data point – such as the annual frequencies in the following case study – the deviation from the mean is determined. This constitutes all the data preparation required before comparing the similarity of two time series. Another benefit is that the PCC is relatively insensitive to noise, meaning it can tolerate the yearly variations typical of historical datasets quite well (Kianimajd et al., 2017). However, one limitation is its sensitivity to extreme outliers (such as when an individual yearly frequency is 100 times larger than any other), which can significantly distort the mean that forms the basis of time-series similarity analysis. It is always advisable to plot time-series data before conducting the correlation analysis to identify potential outliers, as we have done. Alternatively, one can employ descriptive statistics, such as identifying the maximum yearly values in the dataset, or more advanced statistical tests like the Z-Score, which measures how many standard deviations yearly values are from the mean, to spot outliers. To mitigate the problem of outliers, one could use larger sliding windows, basing the word frequency values on decades, not years. This approach should help smooth out extreme outliers and prevent radical distortion of the mean.

PCC allows for the comparison of time-series datasets of different scales – a common situation in historical research – as the method is not based on completely isolated frequencies of temporal data points but on their standard deviations from the mean. In addition to handling diverse types of datasets, the PCC is adaptable to different measurements used for understanding change over time. For example, we use relative word frequency for all parliaments and British newspapers, i.e., the frequency of the search term divided by the frequency of every word in a given calendar year. For the Swedish and Finnish press, however, for interface-specific reasons, we use relative page frequency, i.e., the number of pages on which the search term is mentioned divided by the total number of pages that year. To align the parliamentary time series with the press data in Figures 4 and 5, we multiplied the relative word frequencies by 100 for consistent visualisation. As the measurements differ for the time being, we cannot say, for example, that 'democracy' was discussed more in the Finnish parliament than in the Finnish media. However, we can determine whether changes in discussion of 'democracy' over time are synchronous across datasets, which is our main interest here.

To our knowledge, the PCC has not been used in historical research previously to find similarities across multiple text datasets. Based on our following case study, however, we consider it a useful addition to the toolkit for digital historians. There are several ways to measure the similarity between time series, each of which highlights different kinds of similarity. Thus, the first challenge for humanities scholars is to precisely define which kind of similarity is relevant for the research question. For example, Spearman correlation does not consider the absolute sizes of individual peaks and troughs in historical time series but is based on the ranks of changes over time. Spearman correlation might be preferred to the PCC if the relationship between variables is expected to be non-linear, though some studies suggest that the PCC can detect certain types of non-linear relationships (van den Heuvel and Zhan, 2022). If one is interested in the similarity of time series that do not change simultaneously but with lagged effects, then Dynamic Time Warping might be a good option (Berndt and Clifford, 1994).

We focus on the PCC values because our goal is to rank the time series that are most similar, rather than to calculate thresholds for statistically significant similarities. Nevertheless, we have counted statistical significance. Our method calculates the correlation between pairs of time-series data and adjusts for autocorrelation by considering how each data point relates to the one immediately before it, known as lag-1 autocorrelation. By accounting for this immediate similarity, the method reduces the original sample size to more accurately reflect the amount of truly independent information in the data. Since autocorrelation implies that consecutive data points are related, the effective sample size is smaller than the actual number of observations. This adjusted sample size is then used to compute the statistical significance (p-value) of the correlation. However, this approach assumes that only the immediate past influences each data point and that the overall data is stable over time. If the time series have more complex dependencies or trends, this method cannot fully capture them.

4. Results

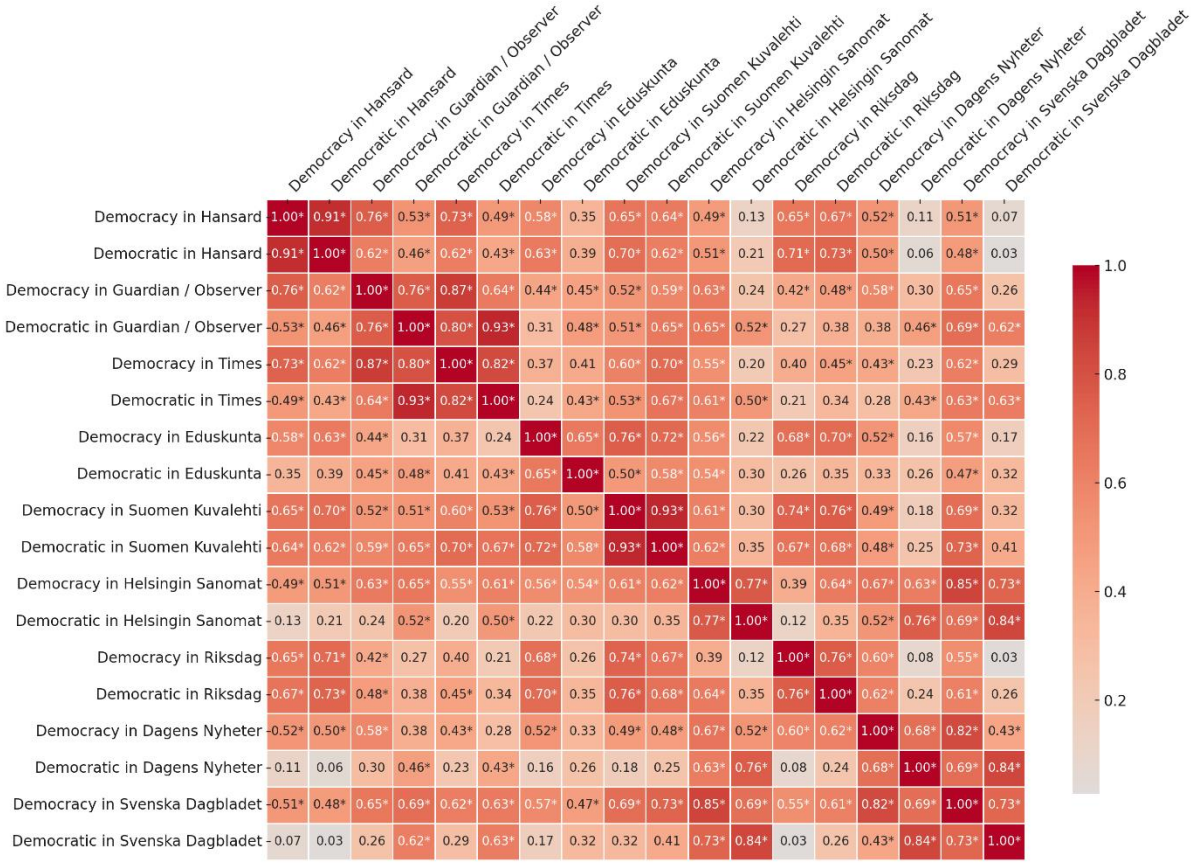


Figure 1: Correlation heatmap for annual relative frequencies of 'democracy' and 'democratic', Britain, Finland and Sweden, 1920–2001. The symbol * indicates significance at the 0.0001 level ($p < 0.0001$).

We have visualised our results in a correlation heatmap (Figure 1): the higher the value, the stronger the correlation between the yearly relative frequencies of terms across datasets. At the top left corner in Figure 1, one can see that 'democracy in Hansard' correlates perfectly with itself (PCC = 1.00) because it is the same time series. Based on the matrix, the correlation is strongest between similar political terms in the same dataset, e.g., the relative frequency of the noun 'democracy' and attribute 'democratic' over time in a national parliament (in *Hansard* 0.91, Riksdag 0.76, and Eduskunta 0.65). This finding is entirely expected – it makes sense that two words related to democracy rise and fall together when discussed in the same forum. Methodologically, it suggests that PCC values capture something essential about the similarity of time series.

There is some interesting variation between 'democracy' and 'democratic' through time (Figure 2). Such variation may have data-specific reasons such as the emergence and disappearance of parties that include 'democracy' or 'democratic' in their name. These parties include *Suomen Kansan Demokraattinen Liitto* (the Finnish People's Democratic Union, an umbrella for the Finnish far left and communists), which entered the Finnish parliament in 1945. It often spoke about democracy but of a specific Soviet kind and withered away in the early 1990s, resulting in fewer references to 'democratic' in the data. Similarly, *Ny Demokrati* (New Democracy, a right-wing populist party) entered the Swedish parliament in the early 1990s, contributing to a peak in discussions about "democracy" at that time.

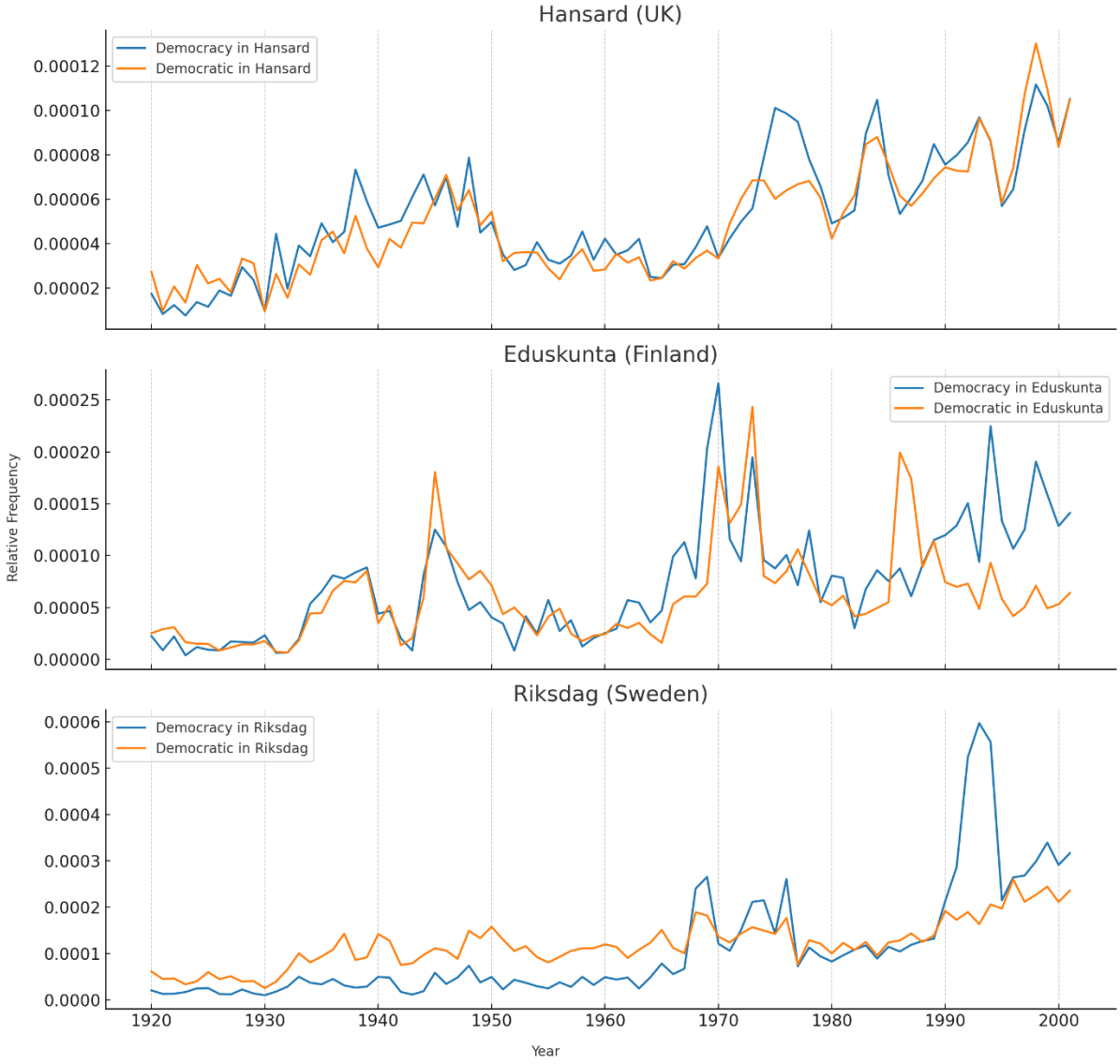


Figure 2: Relative frequency of 'democracy' and 'democratic' in the national parliaments, 1920–2001.

Another strong set of correlations is evident when the same political term appears in similar datasets from the same country, e.g., the frequency of 'democracy' in more liberal and more conservative papers in Britain (0.87), Finland (0.61) and Sweden (0.82). This pattern suggests that ideological differences have had a rather modest impact on the volume of newspaper discourse about democracy since the 1920s, though there is some variation in intensity over time. Our perhaps most important finding is the statistically strong correlation between media and parliamentary discourses, with values ranging from 0.55 to 0.76 for the term 'democracy'. This observation challenges any assumption that parliamentary speech would be somehow elitist and distinct from broader societal debates – at least in the examined leading newspapers with party connections. Regarding discussions on democracy, it appears that parliamentary and media discourses are well aligned in all three countries. Next, we focus on the simultaneous rise of democratic discourses in various national contexts, as trends present in the data are more straightforward to interpret through close reading than those that are absent. Nevertheless, trends showing a decline in references to democracy could be equally significant and might be interpreted using external sources, including prior research and other datasets.

4.1. Interwar and wartime discourses on democracy

While there were national specificities in discourse on democracy, selective close reading of the debates confirms shared trends between the three countries in most decades of the twentieth century. The first period of synchronicity in discourses on democracy concerns the 1920s, 1930s and 1940s. In the British parliament, discourse on 'democracy' and things 'democratic' was still rather limited in relative terms in the early 1920s, in the immediate aftermath of the introduction of universal male and limited female suffrage (1918), yet was intensified to some extent preceding the introduction of equal women's suffrage (1928). In the 1930s, there was a considerable rise in the frequency of both terms, reflecting reactions to rising authoritarian regimes on the Continent. Representative government was renamed 'parliamentary democracy' in several West European countries at this time due to growing concern about the future of democracy in Europe, combining two previously often-contested concepts into a name for a *national* political system (Kurunmäki and Ihalainen, 2024). After the German assimilation of Austria in 1938, for example, the Conservative leader Winston Churchill spoke to the House of Commons in favour of a military alliance between Britain and France against Germany as "[a]ll this time the vast degeneration of the forces of Parliamentary democracy will be proceeding throughout Europe" (Hansard Corpus (HC), 24 March 1938). It was a major speech act that a leading Conservative redefined the British political system as parliamentary democracy.

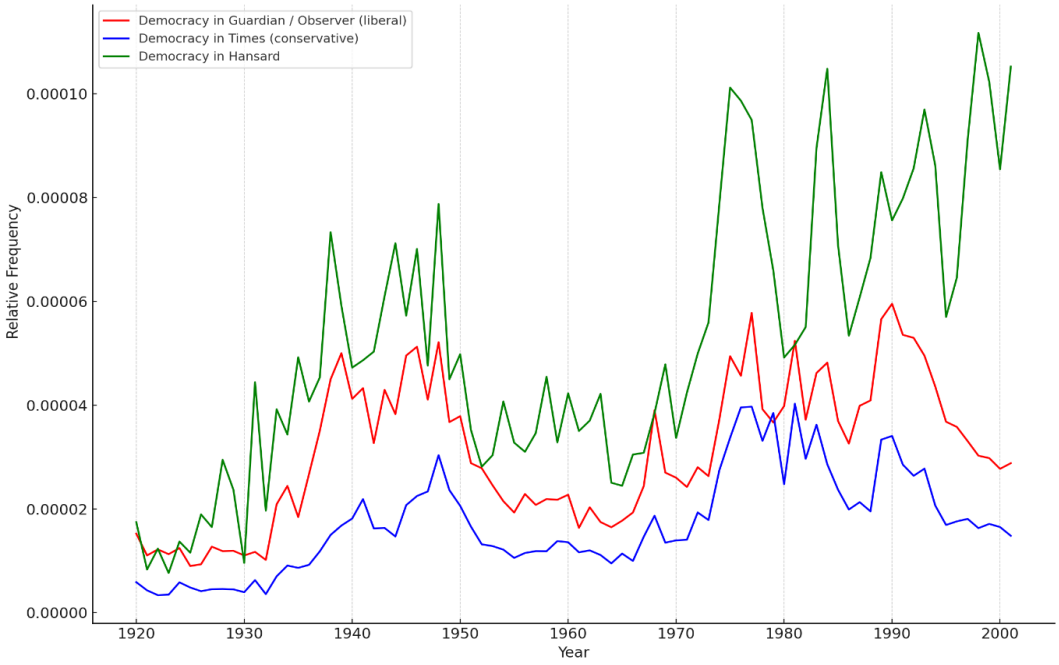


Figure 3: Relative frequency of 'democracy' in the British parliament and leading papers, 1920–2001.

Figure 3 shows parallel trends in two leading British newspapers in which discourse on 'democracy' was activated in 1933 after Hitler came to power, and rose steadily for the rest of the decade. In the case of *The Times*, however, old reservations about mass democracy among Conservatives may have been reflected in considerably lower frequencies than in the liberal *Manchester Guardian*, for which 'democracy' was becoming a key political concept between 1933 and 1939. The latter may also have observed Continental developments with more concern, eagerly printing related news reports.

All three forums saw further peaks in references to democracy during and immediately after the Second World War, as 'democracy' became a key element in discourse on the defence of the Western democratic order, first against the Nazis and then against the extension of the Soviet power in the east of Europe (which was being felt very concretely in Finland). There was a peak in *The Times* in 1941 as the United States joined the battle against Nazi Germany and again in both British papers in 1948, at the early stages of the Cold War. In the postwar situation, not only the Allied victory in 1945 but also Labour's rise to power in Britain that autumn and deteriorating international relations intensified debates on democracy in parliament and a left-leaning paper. The British parties continued to disagree on the extent of 'economic democracy', for instance, but agreed on defending Western democracy against so-called people's democracies (Saunders, 2019, 192, 194).

In the Finnish parliament, debate on 'democracy' was rather modest in the 1920s, after a major ideological confrontation on the constitutional form of democracy in the Civil War of 1918 (Ihalainen, 2017) and the compromise of the republican constitution in 1919, but was intensified as a reaction to the rise of rightist extremism – both at home and internationally – in the 1930s (Figure 4). Multiple parties simultaneously increased their discourses surrounding democracy in the 1930s, although their evaluations of it were often contradictory. In 1935, for example, democracy was defended by left-liberals such as Urho Toivola, a progressivist MP who had made a diplomatic career in London, Paris and the Finnish delegation to the League of Nations, while democracy as a political system continued to be criticised by the far right. According to Toivola, the collapse of parliamentary democracy in a number of European countries over the preceding years was not as much due to the weaknesses of the system as to the fact that it had not been properly implemented, for instance with the survival of the monarchy within democratic constitutions (Eduskunta (EK), Urho Toivola, 22 March 1935). For an MP from the autocratic Patriotic People's Movement (IKL), by contrast, the prevalent democratic system appeared as no more than an application of "Marxist social democracy" (EK, Kaarlo Kares, 26 February 1935), which reflects transnational far-right patterns of thought of the era, reproducing associations between democracy and socialism.

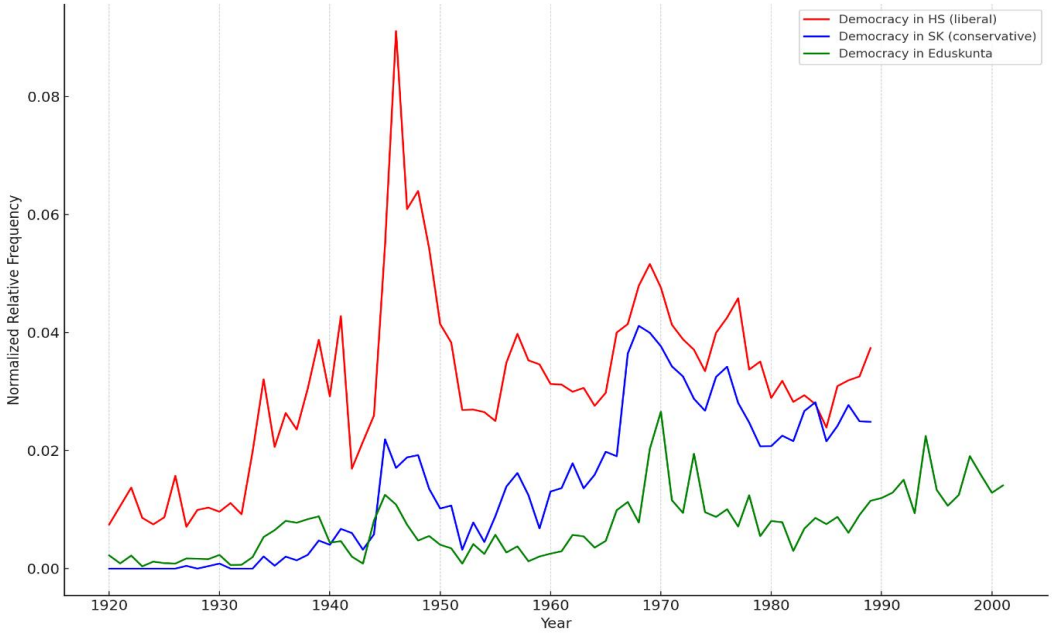


Figure 4: Relative frequency of 'democracy' in the Finnish parliament and leading papers, 1920–2001.

During the Second World War the Finnish parliament remained rather quiet about democracy despite the Winter War being framed as defence of Western democracy, partly due to alliance with Nazi Germany in 1941–44. By 1945, with the re-entry of communists to parliament, there was plenty of discourse on democracy, as communists often emphasised their pro-Soviet understanding of 'socialist' or 'communist' democracy, contrasting it with the 'bourgeois', 'Western' or 'capitalist' democracy that they rejected but which was defended with rising intensity by the other parties.

Although uses of the term 'democracy' stayed relatively stable in the Swedish parliament after the introduction of universal suffrage, much like in Britain and Finland, democracy talk expanded considerably in the 1930s as a reaction to the rise of Nazism in Germany and the consequent construction of a national democratic tradition aimed at uniting the Swedish parties that had previously disagreed on the desirability of mass democracy and parliamentarism (Kurunmäki, Ihalainen and Peksujeff, 2025). After a declining trend in the 1920s and early 1930s, there was a peak in references to democracy in the press in 1933 and again as Europe came closer to war (Figure 5). A dominant idiom in the Riksdag was not so much 'democracy' as such but rather 'democratic states', favoured mainly by communists, while for the social democrats and the centre parties, 'democratic society' was becoming a normative concept used to justify a variety of policies (Riksdag (RD), Erik von Heland, Centre Party, 14 April 1937) aiming at building a "people's home", or a Nordic model.

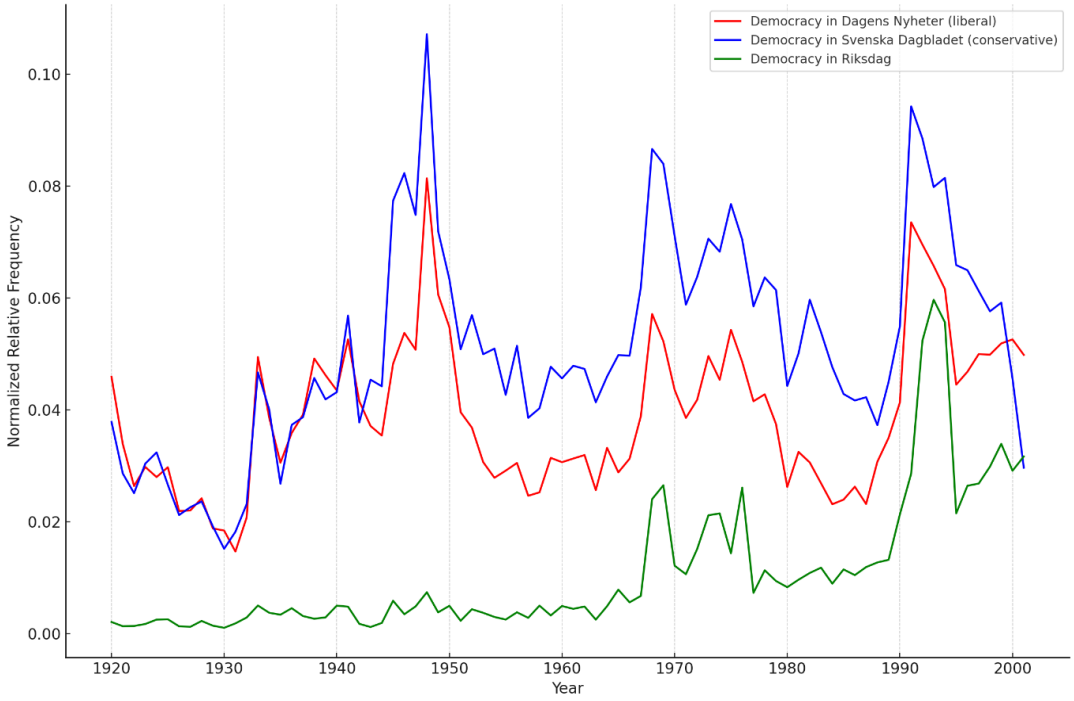


Figure 5: Relative frequency of democracy in the Swedish parliament and leading papers, 1920–2001.

The trends in discussing democracy seem highly synchronous between the leading Swedish conservative and liberal newspapers. Both non-socialist parties had had doubts about mass democracy but started to write more about democracy as a reaction to Hitler's rise to power and the consequent Swedish reconciliation about democracy (Kurunmäki and Ihalainen, 2024). The Second World War and its immediate aftermath with the victory of Western or Soviet-type 'people's' democracy over right-wing dictatorships reinforced this trend, making moderate conservatives more active advocates of Western democracy as opposed to Soviet power, as in all Western European countries. Capitalist and communist varieties of democracy competed in domestic politics, in Europe and globally (Corduwener, 2016). Concern about the future of parliamentary democracy at the time of the communist takeover in Czechoslovakia (1948), for instance, is visible as a peak in Swedish press discourses.

4.2. Postwar discourses on democracy

A second period of synchronicity between the three countries and their parallel forums of debate concerns the 1950s, 1960s and the 1970s. In all these countries, there was a general downhill trend in discourse on democracy in the 1950s and early 1960s, which quantitatively corroborates previous research on controlled democracy that emphasised voting in elections over active citizen participation in postwar Europe (Müller, 2011; Conway, 2020; see also Allen and Mirwaldt, 2010, on references to democracy in party manifestos) and the consequent limited discourse on the nature and state of democracy as the established political system.

Democracy started to be discussed more intensively everywhere by the 1970s, as historians of democracy have pointed out (Conway, 2020; Corduwener, 2023; Müller, 2011; Gilcher-Holtey, 2018). Once the shortcomings of citizens' participation in political decision-making had been exposed by some intellectuals and the transnational student movement around 1968, discussion on the future of democracy was reactivated. In the British parliament, there was just a modest peak at the end of the 1960s during a Labour majority, while the peak is much more evident in both newspapers, though *The Guardian* (change of name in 1959) published more enthusiastically about democracy than the more conservative *The Times*.

In the 1970s, the British parliament was involved in the extension of the semantic domain of 'democratic' (Bonin et al., under review). While the parliamentary debate extended and 'democracy' found its way to more and more policy sectors, becoming diversified in its meanings, the basic genre of parliamentary debate changed little over time. In the meantime, to compete with television, newspapers were increasingly covering a wider variety of topics other than politics. Instead of parliamentary debates they rather wrote about celebrities, published human interest stories or moved towards investigative journalism that might challenge politicians (Salminen, 1991, 253–254; Jonsson, 2002, 137, 205; Temple, 2008, 169, 186). Such changes in European media structures may have contributed to a relative decline in intensity of discourses on democracy in the 1970s and 1980s in the press, if not in parliament.

The overwhelmingly dominant parliamentary idiom in Britain was 'industrial democracy' as advocated by the Labour government in 1974–79. This idiom was in line with traditional socialist calls to extend democracy to the 'economic'. As Secretary of State for Trade Peter Shore put it, "[t]he Government are committed to carrying through as soon as possible a programme for the radical extension of industrial democracy in both the private and public sectors" (HC, 5 August 1975). Margaret Thatcher won a Conservative majority in the elections of 1979 and 1983, followed by a further peak in parliamentary debate on democracy, the dominant topic being her local government bill, which the Labour opposition described as "a naked attack on local democracy and local freedom and rights" (HC, Jack Cunningham, 28 March 1984). These peaks in democracy talk are visible in both newspapers.

Peaks after 1968 were more obvious in Helsinki than in Westminster and considerable in the Finnish press, too, the conservative periodical opening to a wider variety of societal issues, approaching the range of the liberal newspaper. Even if interest in democracy was increasingly synchronic, ideological differences continued. A peak in parliamentary discourse can be explained with election victories of the left (1966), the 1968 generation and a major populist party entering parliament (1970) and the general radicalisation of societal debate between constitutionalists and communists. In 1970, Western representative democracy was defended by conservatives such as Tuure Junnila (EK, 10 March 1970), while a pro-Soviet communist leader called for extending the cooperation agreement with the Soviet Union while defining his political group as defenders of "peace and democracy" (EK, Taisto Sinisalo, 15 September 1970). As for the press, references to democracy tended to decline after 1968.

In Sweden, the immediate postwar period saw plenty of press debate on the future of democracy, supported by the political scientist Herbert Tingsten criticising communism as the editor-in-chief of *Dagens Nyheter* (Engblom, 2002, 27). In parliament, the debate on democracy intensified later, in and after 1968, not only motivated by a transnational student challenge to representative democracy but possibly also by understanding of social democratic Sweden as an alternative model for future democracy. Preparations for a parliamentary reform (1971) activated constitutional debate. In 1969, democracy debates in the Swedish parliament were overwhelmingly about local democracy and to a lesser extent about the position of the monarch in a modern parliamentary democracy. The new

constitution in 1975 is a peak in press discourse on democracy. As soon as a modernised constitutional monarchy had been established, references to democracy became rarer in the years that followed.

We end up with some less distinct synchronicity in the 1990s. Enthusiasm about the fall of the Eastern bloc around 1990 – visible in considerable peaks in media discourse on democracy – evaporated quickly from the British parliament and press, with decreasing interest in questions of democracy. In the late 1990s discourse on democracy diversified dramatically as a Labour majority under Tony Blair's leadership set off to discuss the renewal of the British political system. Yet in both examined newspapers references to democracy declined in the 1990s, which raises the question whether, despite all its attempts at spin, New Labour was able to get its novel discourses on democracy through in the media. This decline may be an indication of growing disenchantment with politics in general and democracy in particular, but is rather explained by the changing relationship between politicians and journalists – developing towards mutual hostility (Temple, 2008, 170–171). The rising parliamentary trend, at the same time, is partly explained by the plans of the New Labour to reform representative democracy by increasing citizen participation (Allen and Mirwaldt, 2010; Bonin et al., under review).

Enthusiasm about the fall of the Eastern bloc appeared in each country but did not last, in Finland partly due to the impact the fall of the Soviet Union had on the economy and on ideological bankruptcy for the far left. A major peak in democracy debate was seen in 1994 as Finland was joining the European Union. An opponent might insist that the special features of 'Nordic democracy' had been disregarded in membership negotiations (EK, Erkki Pulliainen, The Greens, 6 September 1994), while a supporter viewed Finland as "a country increasingly moving towards European democracy" (EK, Margareta Pietikäinen, Swedish People's Party, 2 December 1994).

In Sweden, there was a renewed rise in debates on democracy both in the press and parliament in the early 1990s. The newspapers reacted more strongly to the fall of the Soviet Union in 1991. In parliament, there was a spike in 1993 and 1994, following not so much from the decision to apply for EU membership but rather from the rise of a new neoliberal, right-wing populist party called New Democracy to parliament for the election period 1991–94. Swedish debates on Europe ended up reflecting on the consequences for representative democracy of a referendum on EU membership in 1994 but also on vindications of a particularly Swedish type of democracy (see Ihalainen, 2013). In the late 1990s, references to democracy in the conservative paper drop well below that of the liberal paper, which may be related to disappointments with the triumph of Western democracy that did not materialise, but conclusions on causality call for more extensive contextualising close reading.

Overall, quantitative data from the three countries illustrates the well-known fact that there was a rise in the use of 'democracy' and 'democratic' in parliaments and the press over time, while major downhills are observable. Based on the statistical analysis, transnational correlations of political terms were not as strong as intra-national correlations, yet they were clear in the PCC values; e.g., for the frequency of 'democracy' they varied from 0.58 to 0.68 between the three parliaments. The strongest correlation was between the Finnish and Swedish parliaments, reflecting centuries of shared and entangled political cultures, whereas the relation between Finnish and British parliaments was the weakest, these political cultures being far apart. The transnational trends are nevertheless clearer than anticipated, and give credence to the theses that:

1. Several parliamentary democracies reacted strongly to Hitler's rise to power in Germany to conceptualise their national tradition and political system increasingly through democracy.
2. The post-Second World War victories of the left in national elections and the rising threat of the Soviet Union leading to the Cold War activated democracy talk everywhere.
3. The dip in democracy talk in the controlled democracies of the 1950s and early 1960s was a transnational phenomenon that can be illustrated quantitatively.
4. 'Democratic' was redefined from around 1970, in the aftermath of social movements calling for alternatives to representative democracy in and after 1968, with the domain of democracy being extended and expanded.
5. Spikes in references to democracy with the fall of the Eastern bloc in the early 1990s were followed by diversified discourses on democracy, though with a declining trend that supports the thesis that both disenchantment with democracy and distance between politicians and the media were increasing, the discrepancy between parliamentary and press discourses being most striking in Britain in the late 1990s.

5. Conclusion

The Pearson correlation coefficient (PCC) provides meaningful insights for analysing similarities across multiple time-series datasets. The method is mathematically straightforward, computationally efficient and simple to interpret. Yet, correlation does not imply causation: increases or decreases in word frequencies across datasets might occur simultaneously for unrelated reasons, hence our initial interpretations of potential contextual explanations remain somewhat speculative.

To validate findings derived from the PCC, further investigation is necessary, through detailed contextualising qualitative analysis or alternative text-mining methods, such as quantifying the linguistic contexts surrounding the search terms. Substantial historical background information is needed to interpret the correlations observed across datasets, for example, between media and parliamentary discussions or between different nations. Questions to investigate include the involvement of the same actors in both the press and parliaments and the impact of changing media structures on the relationship between newspapers and politicians (including potential influence on parliamentary speaking), all questions of causality that only historical interpretation can answer.

Future research should seek more details on the similarities across time series. For example, dividing our timeline of 1920–2001 into smaller segments could help identify more exact periods of historically strong or weak correlations. Our preliminary manual investigation suggests that transnational correlations across parliaments were weaker at the beginning of the twentieth century than at the end, when transnational interaction had generally increased. This hypothesis could be empirically tested using the PCC. To add depth to our correlation analysis, resources allowing, we could introduce additional granularity by incorporating metadata such as party affiliation, gender, age or government position, which is partly available for the three parliaments.

Bringing in newspapers with a wider ideological spectrum would be useful. Although our initial analysis of newspapers across political orientations did not detect strong ideological trends using two terms associated with democracy, it is possible that exploring political words specific to certain groups, such as varieties of socialists, might reveal different correlation patterns based on metadata across datasets. By segmenting newspaper content into genres such as news stories, editorials, columns and letters to the editor, for instance, we could determine whether the trends across newspapers of different political orientations are driven by news content, or if focusing solely on opinion pieces would produce different results. Moreover, we could test methods developed to filter out irrelevant mentions of democracy, such as references to specific political parties (Tinits, 2023).

The most noteworthy empirical result here is the high correlation between parliamentary and press debates in most historical periods, which suggests that parliaments were neither out of touch with extra-parliamentary debates nor merely reactive to media. Parliaments and newspapers influenced each other, parties having an interest in the content of the press (most distinctly in Sweden) and MPs sometimes serving as editors-in-chief themselves (Rydén, 2001, 157), citing newspapers frequently and taking the initiative by moulding their discourse to influence the public through press reports. The relationship between politicians and the press changed radically over our research period and has continued to evolve after it: while reporters were traditionally subservient to MPs, representative politicians have increasingly turned into clients of journalists. Newspapers have reported less and less about parliaments and MPs compete for their attention, while the media has become more critical of politicians, particularly in Britain since the 1990s. The relative influence of the media on politics has increased (Temple, 2008, 71). Our observation that parliamentary and press discourses on democracy in Britain were no longer in sync in the 1990s may find an explanation in this changed relationship.

A further empirical result was that trends in intensity of debates on democracy proved be relatively transnational between the three countries, with peaks in the late 1930s, late 1940s, late 1960s and early 1970s and in the 1990s. This development reflects entanglements arising from increased transnational interaction, not only European integration but also shared longer traditions of parliamentary culture and the quality press in Britain and the Nordic countries (Salminen, 1991, 251).

A more extensive range of political terms and datasets would be necessary to construct larger arguments about common patterns between parliament, the media and theoretical debates (which we did not touch on here due to unavailability of data), nationally or transnationally. For example, while

democracy undeniably undergoes transnational shifts in the twentieth century, more research is needed to show how its transnationality compares with other key terms of political thinking (such as isms, welfare discourses, or rising environmentalism). "Compared to what" is at the heart of quantitative thinking (Tufte, 2001) in both natural sciences and humanities, and correlation analysis can serve as a great tool to distinguish local patterns from global patterns.

Acknowledgements

This research has been funded by the Research Council of Finland Professor grant numbers 336709 and 345111 as well as by the University of Jyväskylä Rector's strategic funding.

References

Allen, Nicholas, and Katja Mirwaldt. 2010. "Democracy-speak: party manifestos and democratic values in Britain, France and Germany". *West European Politics* 33(4): 870–893.

Arguing with Digital History – Working Group. 2017. "Digital history and argument: White paper". Roy Rosenzweig Center for History and New Media. https://rrchnm.org/argument-white-paper/.

Berndt, Donald, and James Clifford. 1994. "Using dynamic time warping to find patterns in time series". KDD Workshop.

Blaxill, Luke, and Kaspar Beelen. 2016. "A feminized language of democracy? The representation of women at Westminster since 1945". *Twentieth Century British History* 27(1): 412–449.

Bonin, Hugo, Pasi Ihalainen, Berit Janssen, Jani Marjanen, and Risto Turunen. Under review. "Extension or expansion? Quantitative conceptual analysis of parliamentary uses of 'democratic' in Sweden and Britain, 1950s-1990s".

Brodén, Daniel et al. 2023. "The diachrony of the new political terrorism: Neologisms as discursive framing in Swedish parliamentary data 1971–2018". The 7th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2023).

Cairo, Albert. 2016. *The Truthful Art: Data, Charts, and Maps for Communication*. San Francisco: New Riders.

Chapman, Jane. 2005. *Comparative Media History*. Cambridge: Polity.

Conway, Martin. 2020. *Western Europe's Democratic Age: 1945–1968*. Princeton: Princeton University Press.

Corduwener, Pepijn. 2016. "Democracy as a contested concept in post-war Western Europe: A comparative study of political debates in France, West Germany, and Italy". *The Historical Journal* 59(1): 197–220.

Corduwener, Pepijn. 2023. *The Rise and Fall of the People's Parties: A History of Democracy in Western Europe since 1918*. Oxford: Oxford University Press.

Derrick, Timothy, and Joshua Thomas. 2004. "Time series analysis: The cross-correlation function". In *Innovative Analyses of Human Movement*, edited by Nicholas Stergiou, 189–205. Human Kinetics Publishers.

Elo, Kimmo, and Jenni Karimäki. 2021. "Luonnonsuojelusta ilmastopolitiikkaan: Ympäristöpoliittisen käsitteistön muutos parlamenttipuheessa 1960–2020". *Politiikka* 63(4): 373–402.

Engblom, Lars-Åke. 2002. "Tidningar dör men pressen lever (1945–1958)". In *Den svenska pressens historia efter 1945*, vol. 4, 20–133. Stockholm: Ekerlids förlag.

Gilcher-Holtey, Ingrid. 2018. “Political participation and democratization in the 1960s: The concept of participatory democracy and its repercussions.” In *Democracy in Modern Europe: A Conceptual History*, edited by Jussi Kurunmäki, Jeppe Nevers, and Henk te Velde. Oxford: Berghahn Books.

Hudson, Pat, and Mina Ishizu. 2017. *History by Numbers. An Introduction to Quantitative Approaches*. 2nd edition. London: Bloomsbury Academic.

Hyvärinen, Matti et al. 2023. “‘Democracy’ and ‘people’s power’ in the Finnish Parliament – the struggle between representative, participatory and direct Democracy’. *Redescriptions: Political Thought, Conceptual History and Feminist Theory* 26(2): 117–40.

Ihalainen, Pasi. 2013. “From a despised French word to a dominant concept: The evolution of ‘politics’ in Swedish and Finnish parliamentary debates”. In *Writing Political History Today*, edited by Willibald Steinmetz, Ingrid Holtey, and Heinz-Gerhard Haupt, 57–97. Frankfurt & New York: Campus.

Ihalainen, Pasi. 2017. *The Springs of Democracy: National and Transnational Debates on Constitutional Reform in the British, German, Swedish and Finnish Parliaments, 1917–1919*. Helsinki: The Finnish Society for Literature.

Ihalainen, Pasi et al. 2022. “Building and testing a comparative interface on Northwest European historical parliamentary debates: Relative term frequency analysis of British representative democracy”. In *Digital Parliamentary Data in Action, CEUR Workshop Proceedings*, 3133:52–68. http://ceur-ws.org/Vol-3133/paper04.pdf.

Jensen-Eriksen, Niklas, Aleksi Mainio, and Reetta Hänninen. 2019. *Suomen suurin: Helsingin Sanomat 1889–2019*. Helsinki: Siltala.

Jonnson, Sverker. 2002. ”Tv förändrar världen (1958–1975)”. In *Den svenska pressens historia efter 1945*, vol. 4, 134–247. Stockholm: Ekerlids förlag.

Kianimajd, Adell et al. 2017. “Comparison of different methods of measuring similarity in physiologic time series”. IFAC-PapersOnLine 50(1), 11005–11010, https://doi.org/10.1016/j.ifacol.2017.08.2479

Kurunmäki, Jussi, Jeppe Nevers and Henk te Velde (Eds.). 2018. *Democracy in Modern Europe: A Conceptual history*. New York: Berghahn.

Kurunmäki, Jussi, and Pasi Ihalainen. 2024. “Hur demokrati blev nationell identitet i Sverige – en begreppshistorisk analys”. *Historisk Tidskrift för Finland* 109(1): 1–33.

Kurunmäki, Jussi, Pasi Ihalainen, and Kai Peksujeff. 2025. “Begreppet parlamentarism i svensk debatt och historiepolitik fram till andra världskriget”, *Scandia* 91(2).

Lansdall-Welfare, Thomas. 2017. “Content analysis of 150 years of British periodicals”. *Proceedings of the National Academy of Sciences* 114(4): E457–E465.

McNair, Brian. 2011. *An Introduction to Political Communication*. Fifth Edition. London: Routledge.

Müller, Jan-Werner. 2011. *Contesting Democracy: Political Ideas in Twentieth-century Europe*. New Haven, CT: Yale University Press.

Mäkelä, Eetu. 2016. “LAS: An integrated language analysis tool for multiple languages”. *Journal of Open Source Software* 1(6), 35. http://dx.doi.org/10.21105/joss.00035.

Roller, Ramona. 2023. “Theory-driven statistics for the digital humanities: presenting pitfalls and a practical guide by the example of the Reformation.” *Journal of Cultural Analytics*, 7(4).

Rydén, Per. 2001. “Guldåldern (1919–1936)”. In *Den svenska pressens historia 1897–1945*, vol. 3. 142–265. Stockholm: Ekerlids förlag.

Salminen, Esko. 1991. “Sitoutumattomuuden ja laajenevan informaation aika 1950–1980”. In *Sanomalehdistö sodan murroksesta 1960-luvulle*: *Suomen lehdistön historia*, vol. 3, edited by Päiviö Tommila et al., 141–306. Kuopio: Kustannuskiila.

Sandlund, Elisabeth. 2001. “Beredskap och repression (1936–1945)”. In *Den svenska pressens historia 1897–1945*, vol. 3., 266–381. Stockholm: Ekerlids förlag.

Saunders, Robert. 2019. “Doubtful democrats: Democracy in Britain since 1800”. *Journal of Modern European History* 17(2): 184–195.

Scholl, Stefan. 2024. “Demokratie“. In *Das 20. Jahrhundert in Grundbegriffen: Lexikon zur historischen Semantik in Deutschland*, edited by Ernst Müller, Barbara Picht, and Falko Schmieder. Basel/Berlin: Schwabe. https://doi.org/10.31267/Grundbegriffe_62216520

Snickars, Pelle. 2022. “Modeling media history: On topic models of Swedish media politics 1945–1989”. *Media History* 28(3), 403–424. DOI: 10.1080/13688804.2022.2079484

Temple, Michael. 2008. *The British Press*. Maidenhead: Open University Press.

Tinits, Peeter. 2023. “Finding environmental discourse in historical newspapers: a topic model workflow for query disambiguation”. *DHNB2023 Conference Proceedings*, https://journals.uio.no/dhnbpub/issue/view/875

Tufte, Edward. 2001. *The Visual Display of Quantitative Information*. Cheshire: Graphics Press.

Turunen, Risto. 2024. “Ideologies as conceptual networks: towards a data-intensive approach”. *Journal of Political Ideologies*, 1–23. https://doi.org/10.1080/13569317.2024.2366812

van den Heuvel, Edwin, and Zhan Zhuozhao. 2022. “Myths about linear and monotonic associations: Pearson’s r, Spearman’s ρ, and Kendall’s τ”. *The American Statistician* 76:1, 44–52, DOI: 10.1080/00031305.2021.2004922

Vares, Vares, and Sakari Siltala. 2016. *Sanan ja kuvan vuosisata. Suomen Kuvalehti 1916–2016*.

Wevers, Melvin, Jianbo Gao, and Kristoffer Nielbo. 2020. “Tracking the consumption junction: Temporal dependencies between articles and advertisements in Dutch newspapers”. *Digital Humanities Quarterly* 14(2).

"*Türkensteuer, Anschlagmodus, Gemeiner Pfennig*", what else? Digital Scholarly Editing on the Basis of Topic Modelling Applied to the Holy Roman Empire's Imperial Diet Records of 1576

Roman Bleier[a] and Florian Zeilinger[a,b,c]

[a] *Universität Graz, Universitätspl. 3, Graz, 8010, Austria*
[b] *Commission for Modern History of Austria (KNGÖ), c/o Institut für Geschichtswissenschaften und europäische Ethnologie, Universität Innsbruck, Innrain 52, Innsbruck, 6020, Austria*
[c] *Historical Commission at the Bavarian Academy of Sciences and Humanities, Alfons-Goppel-Str. 11, München, 80539, Germany*

Abstract

A key challenge for the long-term editorial project *Deutsche Reichstagsakten* is the lack of a cross-edition subject index, alongside the complexity of the existing specialised indices. These factors hinder the searchability of texts and their comparability with sources from other European proto-parliamentary estates assemblies. While these indices are largely consistent with contemporaneous language, they are shaped by editorial selection and primarily designed from a print-focused perspective. In this article, we test and compare three Topic Modelling methods frequently used in Digital Humanities, particularly for modern parliamentary records. These methods facilitate the automatic identification of language-based word clusters, which can be labelled with new subject terms to create updated indices and critically evaluate on editorial practices. Using the minutes from the Imperial Diet Records of 1576, we applied and compared Gensim and Mallet, both based on Latent Dirichlet Allocation, as well as the transformer-based BERTopic. All three methods produced strong results for approximately 100 topics. These include not only content- and function-related topics—traditionally prioritised in editorial work—but also topics related to the functioning and procedural aspects of the Imperial Diet. Such findings highlight and support recent research efforts into the functioning and symbolic dimensions of these historical assemblies.

Keywords[1]

Digital Scholarly Edition, History, Holy Roman Empire, Imperial Diet, Early Modern Period, Topic Modelling, Latent Dirichlet Allocation, Gensim, Mallet, BERTopic

1. Introduction

In this article, we present findings from a project focused on the records of the Holy Roman Empire of the German Nation's Imperial Diets (*Reichstage*, RTT, singular: RT).[2] This project is part of a larger investigation into the Imperial Diet Records (*Reichstagsakten*, RTA), the funding for which is still pending: Our aim is to construct a corpus of minutes from several, ideally all, existing 16th-century RTA editions, and to create a cross-edition research database of deliberation topics and participants. For this concise report we conducted tests to compare the application of three tools commonly employed

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, May 28, 2024, Rejkjavik, Iceland
EMAIL: roman.bleier@uni-graz.at (A. 1); florian-zeilinger@live.at (A. 2)
ORCID: 0000-0003-4674-1042 (A. 1); 0009-0009-1666-0701 (A. 2)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

[2] "Geschichtswissenschaftliches Edieren in der digitalen Transformation: Eine Annäherung am Beispiel einer Themengeschichte der Reichstage der zweiten Hälfte des 16. Jahrhunderts". As digitalisation progresses, for many long term editorial projects the question of continuation, future editorial concepts and subsequent use cases in times of digital transformation arises.

for Topic Modelling (TM). Additionally, we discuss the computer-generated topics and reflect on the existing editorial indexing and indices of the RTA editions.

2. The State of Editing, Indexing and Researching Imperial Diets

The periodic RTT, convened at various locations and at irregular intervals spanning several months or years until 1662, were among the most important institutions of the Holy Roman Empire. Similar to other European estate assemblies, it functioned as a precursor to the modern parliament (Hébert, 2014; Kewes et al., 2022; Blockmans, 2024).[3] Based on Leopold von Ranke's concept formulated during his study of the Reformation, the RTA have been edited since the 19th century by the Historical Commission at the Bavarian Academy of Sciences and Humanities (e.g., Neuhaus, 2006, 43)[4] and they form the basis of all RT research. The matters deliberated upon during the RTT are fundamental themes in the constitutional history of the Holy Roman Empire (Moraw, 1980, 165). They not only reflect the issues deemed significant by the imperial power elites but also influenced political dynamics (Kampmann, 2023, 593) and societal discourse (Ristilä/Elo, 2023, 1) in a culturally distinct manner (Stollberg-Rilinger, 2005, 10–12, 116). Hence, most editions have indexed these themes using corresponding subject terms. These terms serve to identify and label the topics of deliberation and negotiation documented in the records. From a linguistic perspective, these topics comprise words within a particular semantic field that co-occur frequently. The identification thus relies largely on recurring phrases or patterns. Some of the topics were already labeled by contemporaries and were interpreted as such by editors, who employ specific terms as heuristic tools to index relevant textual passages

However, unlike modern parliamentary minutes indices (articles in: La Mela/Norén/Hyvönen, 2022), there is currently a lack of comprehensive cross-edition indexing for RTT's topics, resulting in a lack of homogeneity.[5] Moreover, the complex subject indices based on editorial interpretation and topic selection, are still rooted in a print-oriented perspective (e.g., indices linking to pages or longer text passages) and with their hermetic editorial terminology that, while largely but not entirely in line with the contemporaneous language, makes it difficult to reflect on them (Gotthard, 2002, 467) and hinders the recognition of the international dimension of estates assemblies (see above). For instance, the "*Gemeiner Pfennig*" was a specific direct imperial tax on wealth and income, serving as an alternative to the "*Reichsmatrikel*" concerning the imperial estates which prevailed after 1544 (Lanzinner, 2012/2023, 139f.). In the context of the 16th century, it is occasionally subsumed under "*Reichssteuer, Türkensteuer*" (Leeb, 2013, Register), at other times under "*Türkenhilfe, Steuermodus*" referring to the perceived Ottoman threat and defence and war against the Ottomans (Leeb et al., 2023, Sachregister).

Building upon Gabriele Haug-Moritz's research (2021), which examines the RT as a proto-organisational institution with its own partially formalised processes, logics and myths, one can argue that such institutions tend to develop a core of favored topics that are characterised by their capacity for consensus and easy justifiability. However, they also face the risk of becoming entrenched in their own "topic history" (Kieserling, 1994, 175). Additionally, depending on their trajectory, editorial enterprises may uphold certain forms of arbitrary indexing and keywording for a long time. TM allows for the identification of underlying themes or topics within extensive text collections, such as the RT editions. However, topics generated with this method align with the historical texts and diverge from the existing editorial indexing methods. Thus, this approach enables the study of editorial practices and sources in novel ways.

[3] On the criticism of labeling the RT as a parliament: Hartmann 2017, 19–21. The opposite position is, for instance, represented by: Haug-Moritz 2021; the University of Oxford's platform: https://intellectualhistory.web.ox.ac.uk/recovering-europes-parliamentary-culture-1500-1700.

[4] By 2024, there are currently 43 edited volumes by the Historical Commission on RTT and Imperial Assemblies from this period. Still missing RTA (for the 16th century): 1500, 1518, 1530, 1597/98 (and for the "long 16th century" also), 1603, 1608, 1613. Almost every edition contains minutes or differently named texts corresponding in function and structure, except those on 1513, 1521, and 1522. The first detailed minutes are available with the one from 1544 (Gotthard, 2002, 466).

[5] At a conference in Graz in 2022 (Brantner/Rammer 2022) researchers discussed the need for a controlled vocabulary. One of these terms could be the category "communication topic".

3. Our Sources: Imperial Diet Minutes

At RTT the rulers themselves and/or their authorised envoys, as personal representatives relatively bound by instructions (Zeilinger, 2023), should deliberate and make decisions on matters including: internal security, the judicial system, taxes, defence against external danger, monetary policy and more. Even envoys from foreign rulers from Portugal to the Ottoman Empire were present.

The RTT's procedures became increasingly formalised in the course of the 16th century (Hartmann, 2017, 11f., 275–336; Haug-Moritz, 2023, §§80–84), leading to the emergence of more varied text types. Once announced by the king or emperor of the House of Austria, and with the participants gathered at the meeting place, the RT commenced with the reading of the imperial proposition. The main topics outlined in this manner were then separately deliberated in three curiae before reaching an agreement. By means of negotiation files, views were exchanged with the monarch until a collective decision was reached which became part of the Imperial Recess (*Reichsabschied*) that established new imperial laws. The scribes of the individual delegations documented the discussions for the respective imperial estates, resulting in multiple, often slightly differing, minutes for the same meetings, often with varying levels of detail and signifcance (cf. different minutes on the same days: Leeb et al., 2023).

Minutes have a unique place among the various RTA text types, as they document the daily deliberations in specific curiae and other assemblies throughout an RT in a structured form and they are always part of an RT edition. Similar to their modern successors, they document "votes" (albeit paraphrases of speeches) and the speakers (mostly the rulers and/or the group of councillors representing the imperial estate, rarely individuals) during discussions. However, minutes are not direct representations of historical reality; instead, they are constructs with their own logic, pragmatics and semantics (Bedos-Rezak, 2011, 11), as well as referenceability and truth effects (Plener/Werber/Wolf, 2023, V, VII). They reflect cultural-specific protocol practices and institutional self-presentation more than actual debates. When comparing RTA with modern parliamentary data, there are fundamental biases on three levels: (1) deliberations were not faithfully recorded verbatim, as they were documented by contemporaries with specific backgrounds and political agendas using culturally specific language, (2) in the RTA editions different editors selected minutes (e.g., only some of the existing minutes of the Council of Princes) and edited them based on guidelines which changed over time, (3) the existing editorial indices can be excessive and selective, constituting an integral part of a longstanding editorial tradition. However, they are also a crucial tool for enhancing the accessibility of the content within the RT editions.

To contextualise the results of our study, we compared the subject index of the digital 1576 edition (Leeb et al., 2023) with the retro-digitised 1556/57 edition (Leeb, 2013). The subject terms in the 1556/57 edition adhere to the editorial guidelines established in 1988 but were slightly modified for the 1576 edition. In the 1576 edition, these terms comprise more abstract analytical terms and therefore less specific ones (approx. 150 subject terms) without complex nested general indices of people, places and topics concurrently (e.g., "Augsburg"). Furthermore, terms relating to the "functioning" of the procedure (Haug-Moritz, 2020) were subordinate to the "functional" deliberation content in the 1556/57 edition, whereas they were separately recorded for the 1576 edition.

4. Our Data and Preprocessing

The project discussed in this article primarily used the minutes of the RTA edition of 1576, as both authors were members of the respective editorial team. The digital edition of the RTA of 1576 was developed and published in GAMS, the digital long-term repository of the University of Graz (Bleier/Zeilinger/Vogeler, 2022), and the minutes' transcriptions are accessible as TEI through the GAMS API. Our corpus encompasses the texts of 16 minutes, with seven minutes fully transcribed. For any gaps in the full transcriptions[6] partial transcriptions of other minutes are provided. Each transcription consists of a series of day entries of one or more paragraphs, resulting in a corpus of 472 day entries ranging from 3 to 5,000 tokens (words). The entire corpus contains approx. 330,000 tokens.

[6] Caused by the membership of the recording imperial estates in one of several committees meeting in parallel or by the non-documentation of the unofficial religious negotiations.

In the TEI files editorial subject terms are assigned to each of the minutes' day entries, facilitating the extraction of individual days' text with unique IDs, dates, and subject terms. An essential preprocessing step involved extracting relevant text from the TEI and removing data that is not part of the original text, such as editorial commentaries.

When processing day entries, Mallet and Gensim[7] performed well, as LDA generally handles longer texts effectively. However, for BERTopic, it was necessary to create smaller entities,[8] leading to the decision to use paragraphs due to their manageable length and unique IDs assigned in the TEI. Paragraphs with very little text (below 5 tokens) were removed as they typically only record the date (e.g., a holiday) and do not add much content. Following initial tests, large paragraphs were split two or three times based on punctuation, resulting in approx. 10,000 individual paragraphs or texts, with the majority consisting of 30 to 60 tokens each. There is certainly potential for improvement regarding the chunk size used, and we plan to explore the results using individual sentences in the future.

The minutes are written in Early New High German (*Frühneuhochdeutsch*), a language characterised by a lack of orthography and correspondingly by great linguistic variance, even within the same document (e.g., the town Regensburg appears as *Regenspurg*, *Regenßpurg*, *Regenspurgkh*, and even Latin *Ratisbona*, *Ratispona*…). There are few electronic corpora[9] available for historical languages, and our TM appears to be the first for this particular language. NLP-tools trained on modern New High German often struggle with such texts, making preprocessing a challenging task. Consequently, it was necessary (1) to compile a specific list of stopwords through the iterative generation of topics, eliminating the words that are irrelevant to deliberation topics (starting with: *aber, abermals, abermalß, aim, ain, aine, ainem, ainen, ainer, aines, aineß, all, allain, allaine, alle, allein, alleine, allem, allen, aller, alles, alleß, alls, allso, allß, allßo, als, also, alß, alßo, am, an, ane, …*). (2) All words were converted to lower-case. Other preprocessing steps, such as (3) removing punctuation, were only necessary for the first two tools. Stemming and lemmatisation were not performed due to the lack of suitable tools. In retrospect, it is evident that additional normalisation would be beneficial in future TM and would considerably improve the results of all three TM tools.

5. Topic Modelling

The Latent Dirichlet Allocation (LDA), an enhanced Bayesian generative probabilistic model for extracting topics from large text collections (Blei et al., 2003; id., 2012), was until recently the most frequently used method for TM in Digital Humanities (DH) (Althage, 2022, 256–258, criticism: 265f.), and thus, we selected it for our project. We used the implementations provided by Mallet, a Java-based package for TM, and the Python library Gensim (Althage, 2022, 261f.). Additionally, we experimented with the transformer-based library BERTopic which is currently preferred over LDA as it also considers the semantic relations (contexts) within sentences through word- and sentence vector representations. BERTopic is a highly modular TM library that can be customised to suit specific project requirements (Grootendorst, 2022, n.p.). While we did not have the time to explore this in-depth, the potential for the processing of historical languages is promising.

When using LDA, it is essential to select a number of topics that should be found. Our initial tests with 20 and 50 topics showed a trend towards better results with a higher number, although in both cases clearly separable topics were sometimes mixed up and other known topics were not found. In contrast, 75 or 100 topics produce clearer thematic word groups and uncover previously overlooked subject areas. While 75 topics contain slightly fewer duplicate topics, some may not be present within them, as opposed to a list of 100 topics. Unlike Mallet or Gensim, BERTopic does not assume a specific number of topics in all documents; instead, it distils exactly one topic from each text (Grootendorst, 2022, n.p.). From the approx. 10,000 paragraphs used, BERTopic identified 103 topics. The specific topics found by the three methods and compared in the following section, are listed side by side in the appendix to this paper.

[7] https://mimno.github.io/Mallet/topics.html; https://radimrehurek.com/gensim/.

[8] We used the default setting and sentence-transformers. It is recommended to use paragraphs or sentences as input data: https://maartengr.github.io/BERTopic/getting_started/tips_and_tricks/tips_and_tricks.html.

[9] An example would be the Reference Corpora for Middle and Early New High German: https://www.linguistics.rub.de/comphist/projects/ref/.

6. Discussing the Results and Comparing Topics to Subject Terms

Primarily, the texts that have survived from pre-modern parliamentary assemblies are those that were used to create and/or document political and social realities (Stollberg-Rilinger, 1999, 17f.). By employing language-based LDA and BERTopic, we are close to the political language of the contemporaries in which they recorded their deliberations[10] (though human interpretation remains necessary). Appropriate indexing and visualising thus serve as heuristic tools for further research and distant reading. Unsurprisingly, the most frequent topics include typical but relatively general terms related to specific councils, procedural aspects, or the minutes of different meetings (Mallet: M5, M21[11]). This demonstrates that our method not only captures the content of the deliberations but also sheds light on procedural aspects, documentation practices, and the language employed by minute-takers—an aspect that has not previously been indexed to this extent so far. Additionally, some topics are composed of terms found exclusively in specific minutes (e.g., M30[12] and M93[13] for the Imperial Privy Council, or M33[14] for the Electoral Saxon minutes), enabling us to compare different recording practices.

Table 1

An exemplary comparison of topics generated using Mallet, Gensim and BERTopic, reveals topics that can be interpreted as very similar in content and aligned with the subject terms from the RTA 1576 edition. The words that form the basis of this interpretation and comparison are highlighted.

Subject Term	Mallet Topic(s)	Gensim Topic(s)	BERTopic Topic(s)
Ottomans, Border and Taxes (Tax Mode and Duration)	M32: ***hülff***, ***pfennig***, ***gemainen***, *underthonen*, ***jar***, ***gelt***, ***türggen***, ***mitl***, *gemaine*, ***rohmtzug***, ***bewilligt***, *zusamen, noth, hoch, vorrath,* ***römertzug***, *gros, armen, volckh, dargegen*; M47: ***monat***, ***hilff***, *zuerlegen,* ***eilenden***, *replic,* ***bewilligung***, ***bewilligt***, *martini,* ***zubewilligen***, ***jhar***, *nott, termin, ziel, bewilligen,* ***beharrlichen***, *not, laetare, erlegen, angelegt, ersten*; M79: ***pfennig***, ***gemeinen***, *underthanen, vota, session, modo, contribution, fridt,* ***romerzug***, *quantitate, abzulesen,* ***römerzug***, *hoch, specie, votum,* ***jhar***, ***gemeine***, *absondern, summa,* ***gemein***; M87: ***gemain***, ***pfenning***, ***hilff***, *besatzung, schwendi, schrifften,* ***grenitz***, ***romzueg***, *wesen, fall, ritterorden, ratsam, widerstandt, erhalten, ordinary, propositione, gaistlichen, proponirn,* ***beharliche***, *handlung*	G0: ***Trier***, ***hilff***, ***Pfaltz***, *Hetten, sachen,* ***rath***, ***Sachsen***, ***monat***, ***Meintz***, *caesari*; G30: ***monat***, *rath,* ***gemeinen***, ***pfennig***, *abzulesen, sache,* ***16***, ***hulffen***, ***monaten***, ***24***; G54: ***monat***, *concept,* ***eilenden***, ***Trier***, *stenden,* ***6***, ***60***, ***8***, *zustellen,* ***Sachssen***; G59: ***monat***, *termin,* ***bewilligung***, ***Brandenburg***, ***48***, ***6***, *stendt,* ***bewilligt***, *vergleicht, idem*; G91: ***hülff***, *usschus, crais, stendt,* ***jar***, *Reich,* ***pfennig***, *anno, herr, herrn*	B0: *stett, erbarn, erbarn stett,* ***hülff***, ***mitl***, ***gelt***, *werckh, specie,* ***pfennig***, *sachen;* B1: *stendt, stenden, sachen, stende,* ***grenitz***, *schrifft, versehen, hoch,* ***hulff***, *cron*
Procedure, Deliberating in the Council of Electors	M11: ***sachsen***, ***trier***, *caesar,* ***pfaltz***, *caesari,* ***Brandenburg***, ***umbfrag***, *geredt,* ***cöln***, *vernemmen,* ***zuerinnern***, *puncten, zureden,* ***maintz***, *werckh, wissen, caesarem,* ***bevelch***, *relation, wiß*;	G87: ***Trier***, ***Pfaltz***, ***Coln***, ***Sachssen***, *sachen,* ***Meintz***, *zuthun, dohin,*	B3: ***brandenburg***, ***brandenburgk***, ***bevelch***, *achten, sachen, underthanen,*

[10] Furthermore, we can see that some topics consist of words exclusively from particular minutes (like T26), which allows us to compare different recording practices.

[11] M5 words: crais, craisen, anno, bairischen, saltzburg, khündten, schwäbischen, zuwider, osterreichischen, … M21 words: trier, achten, sachen, sachssen, coln, meynunge, … zulassen, hern, anno, … schlissen, brandenburg, …

[12] M30 words (Latin phrases): propter, vel, dies, hoc, catholicis, festum, ante, … nihil, mainung, actum, …

[13] M93 words (members): harrach, trautson, weber, bedencken, ubergeben, hern, caesar, her, vihauser, …

[14] As the only ones using constantly the form "*steierischen*" instead of "*steirischen*".

	M76: *versehen, lenger, bevelch,* ***räth,*** *handlung,* ***meintzische, vicecantzler,*** *vernommen, allergnedigst, zeigt, ambt,* ***mittel,*** *underlassen, zuverrichten, wisten,* ***churfürsten, räthe,*** *thetten, schreitten, bitt*	*bedencken, meynunge*	*caesari, dohin,* ***pfaltz, zuerinnern***

As shown in Table 1, if the TM results are compared with one another as well as with the subject terms for 1576 (subsequently provided in double quotes), significant overlaps can be observed between certain topics and subject terms. The overlaps are often not immediately obvious through a direct string comparison—often only 10-20% identical terms—due to spelling variation, but they are much more apparent to a human reader: For instance, Mallet topics M8 and M24 both refer to "'*Landfrieden*'/Peace, Offences, War in the Netherlands". Particularly with Mallet, highly frequent subjects such as "Ottomans, Border and Taxes", aligning with the fiscal-military paradigm of state-building history, distributed across multiple topics (e.g., M32, M47, M79, M87). Furthermore, within the 100-topic model, specific text types no longer form standalone topics (in contrast: "Instructions" in the 75-topic model or the index of the RTA 1556/57) but instead appear in combination with procedure-related terms (e.g., M15 "*werbung*" [suiting] or M85 "Imperial Recess"). Nevertheless, Mallet identifies also a separate topic, M22, on (private) petitions referencing additional files ("*apud*" in phrases like "*apud acta*", "*apud supplicationes*"). Similarly, Gensim identifies one *apud*-topic and two *acta*-topics, differentiated by their respective deliberation contexts (G21, G23, G46). This demonstrates that the minutes served not only as records but also as tools for accessing supplementary files, underscoring their role as key sources (Vismann, 2000, 23, 170f.).

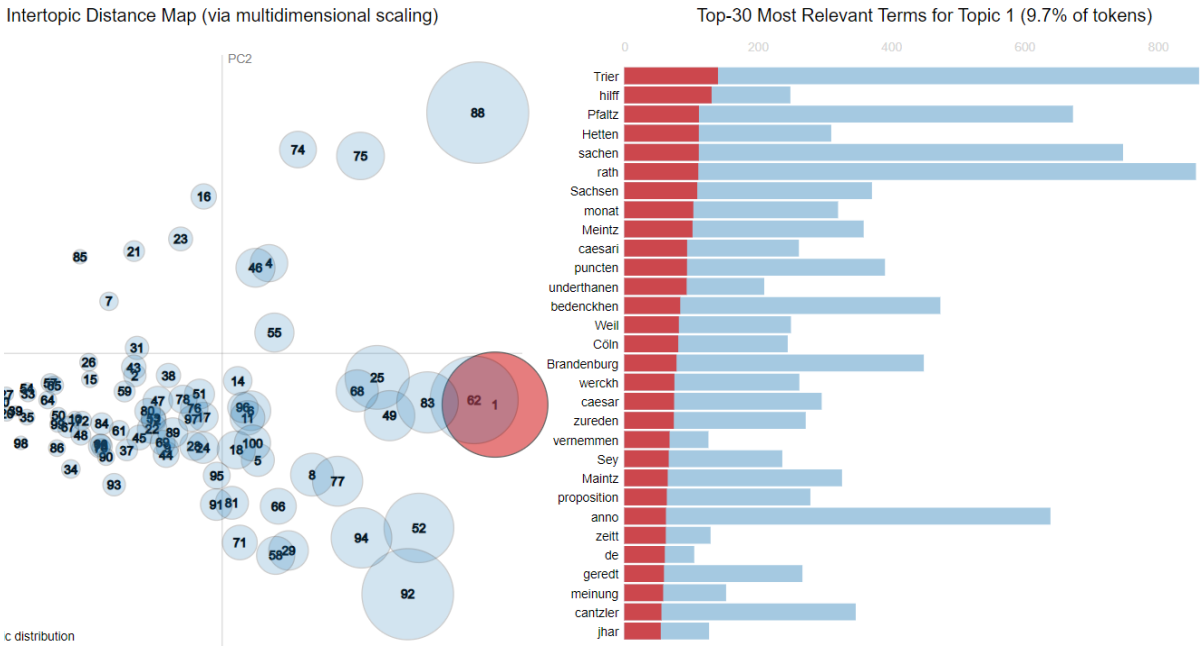


Figure 1: Part of a standard visualisation of the Gensim data (1 stands for G0)

To explore Gensim, the pyLDAvis visualisation library is frequently used. Figure 1 shows such a visualisation, where the left-hand side highlights the most representative topics (displayed as large, partially overlapping circles). On the right-hand side, the most important terms per topic (e.g., G0) are shown in a bar chart (blue: overall term frequency; red: term frequency within the topic). The most frequent topics again include "Ottomans, Tax Mode and Level"[15] (G0 containing 9.7% of all tokens, e.g., the electors "*Trier*", "*Pfaltz*" "*Maintz*" etc., "*hilff*" [taxation as an aid], "*pfennig*" [tax mode], "*monat*" [duration]), and "Procedure, Deliberating in the Council of Electors (on Ottomans)" (G87). There are internal overlaps, for instance, between G0, G59 and G87, as well as overlaps with Mallet.

[15] Corresponding to the granularity of the index for 1576: Leeb et al. 2023, Sachregister; in earlier subject indices logically subsumed under imperial taxes: Leeb 2013, Register: "*Reichssteuer*, *Türkensteuer*".

For instance, Gensim's broad G0 overlaps especially with Mallet's equally broad M11, M21, M36 and M46 ("Deliberating"). Like Mallet, Gensim also identifies previously overlooked topics, such as one concerning the imperial family, the empress and archdukes, Cardinal Morone, and divine services as rituals (M61, G33). Empress Maria, Emperor Maximilian's wife and cousin from Spain, was a strong advocate of the Catholic Church, in contrast to him, and maintained excellent relations with the papal curia, which also elevated two of her sons to the rank of cardinal (Koller, 2016, 86–97).

Regarding BERTopic, the language-based model enables the clustering of diverse pure Latin texts (B35) as well as various texts containing the name "Ferdinand" (former Emperor Ferdinand I, his declaration, Archduke Ferdinand II etc., B38[16]). The standard visualisations provided by BERTopic are quite helpful, such as a heatmap showing overlaps between different topics, which allows the researcher to manually merge some of them (e.g., B67 and B92, both concerning votes in favour of earlier proposals using "*placet*"). The largest topics are "Ottomans, Border and Taxes" (B0, B1), followed by "Religious Gravamina" (B2[17]), and "Procedure, Deliberating in the Council of Electors" (B3). Not only were the negotiations concerning the Ottomans the most important and extensive (top-ranked negotiation item), but two minutes of the Council of Electors and seven on religious issues were edited instead of just one.

References to previous RTT were regularly labeled with recurring terms: While Mallet and Gensim identify "*anno*" as a relevant term across various topics (appearing ten times: M5, M8, M21, M25, M38, M46, M50, M63, M74, M83; and twenty times: G4, G6, G13, G16, G23, G25, G39, G45, G52, G56, G58, G60, G67, G68, G73, G77, G91, G93, G96, G99), BERTopic identifies precisely one topic of Imperial Assembly references concerning the "*Reichsmatrikel*" negotiations (B23), illustrating the RT's long-term institutional memory (cf. Haug-Moritz, 2023, §106). However, BERTopic is unable to cluster the more disparate mentions of "*apud*" or those of the empress and her family into a single topic.

All three methods, Mallet, Gensim and BERTopic, identify function-related topics such as "Imperial Taxes" as well as those related to the functioning, the procedural form, and its verbal documentation, such as the names of council members. They also capture terms like "*ansagen*"/announcing of meetings ("Procedure, Begin": B65, G48, M30), "*concept ablesen*"/reading out drafts ("Procedure, Deliberating") e.g., B5, G95, M74), etc., which were previously marginalised in indices compared to classical function-related terms. For recent research, it is crucial to examine functioning (Neu, 2024, 19–24) in order to analyse both the performative-symbolic and instrumental aspects of the pre-modern process of representing, staging thematic positions, and establishing political order (Möllers, 2021, 126). For function-related issues, Mallet sometimes identifies multiple distinct topics rather than a single one. A similar pattern is observed with Gensim for procedural aspects. However, both Mallet and Gensim identify several topics for recurring terms such as *anno* or *apud acta*, which appear in various thematic contexts. In contrast, BERTopic, which only extracts one topic per text, identifies only one topic for some matters, but demonstrates greater sensitivity to other typical terms.

7. Conclusion

This experiment is the first in a series of automated TM studies on Early New High German texts, designed to explore the RTA. For the experiment, we compiled a small corpus of transcriptions from the minutes of the most recent RTA edition and compared different TM methods, including software packages based on Latent Dirichlet Allocation (LDA) and BERTopic. All methods delivered satisfactory results with approx. 100 topics.

Some key lessons learned for future research are as follows: (1) Pre-processing is crucial for all three TM tools. We are also considering incorporating a normalisation step in the future, as the linguistic variance in Early New High German significant, and even basic normalisation could improve results. For the LDA tools, lemmatisation would also be desireable. As no suitable tools for preprocessing Early New High German texts are currently available, we are considering using large language models (LLMs), as initial tests have shown promising results. However, it is remarkable how effectively BERTopic recognises semantic relations and words, even without orthographic normalisation. On the

[16] B38 words: ferdinandi, keiser, declaration, ertzhertzog, ferdinandt, …, 59, ferdinandi declaration, …, keiser ferdinandi.
[17] B2 words: catholischen, catholische, religion, augspurgischen, stendt, gravamina, …

other hand, BERTopic does not always select words that are truely "representative" of the content but sometimes focus on wording and practice of documentation, which enables analysing them too. (2) Automatically generated topics extend beyond the previous content- or function-related terms. Mallet, Gensim and BERTopic all identify various function-related topics as well as procedural aspects, both of which are central to current research. Nevertheless, they detect slightly different numbers of topics depending on whether they focus is on function-related themes, procedural aspects, or phrases used in multiple contexts.

8. Acknowledgements

The work on this project was funded by the Historical Commission at the Bavarian Academy of Sciences and Humanities and the Commission for Modern Austrian History (project title: "Geschichtswissenschaftliches Edieren in der digitalen Transformation: Eine Annäherung am Beispiel einer Themengeschichte der Reichstage der zweiten Hälfte des 16. Jahrhunderts").

References

Althage, Melanie. 2022. "Potenziale und Grenzen der Topic-Modellierung mit Latent Dirichlet Allocation für die Digital History." In *Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft* edited by Karoline D. Döring, Stefan Haas, Mareike König et al.: 255–277. Studies in Digital History and Hermeneutics 6. Berlin: De Gruyter Oldenbourg.

Bedos-Rezak, Brigitte M. 2011. *When Ego was Imago*. Signs of Identity in the Middle Ages. Visualising the Middle Ages 3. Leiden et al.: Brill.

Blei, David M. 2012. "Topic Modeling and Digital Humanities." In *Journal of Digital Humanities*. Vol. 2/1. https://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. 2003. "Latent dirichlet allocation." In *The Journal of Machine Learning Research*. Vol. 3: 993–1022. DOI: 10.5555/944919.944937.

Bleier, Roman, Zeilinger, Florian, and Vogeler, Georg. 2022. "From Early Modern Deliberation to the Semantic Web: Annotating Communications in the Records of the Imperial Diet of 1576." In *Proceedings of the Digital Parliamentary Data in Action (DiPaDa) Workshop Co-located with the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB) 2022*, edited by Matti La Mela, Fredrik Norén and Eero Hyvönen, 86–100. http://ceur-ws.org/Vol-3133/.

Blockmans, Wim. 2024. *The Voice of the People?* Political Participation before the Revolutions. London: Routledge.

Brantner, Elisabeth, and Rammer, Constanze. 2022. "Tagungsbericht: Neue Wege der Edition frühneuzeitlicher Ständeversammlungen. Aktuelle geschichtswissenschaftliche Konzeptualisierungen ständischer Teilhabe und digitale Methoden. 6.–8.4.2022." In *HSozKult*. https://www.hsozkult.de/conferencereport/id/fdkn-129909.

Gotthard, Axel. 2002. "'Gut so', aber nicht 'weiter so'! Die Edition der neuzeitlichen RTA – ein Zwischenresümee aus gegebenem Anlaß." In *Zeitschrift für Rechtsgeschichte. Kanonistische Abteilung* Nr. 88: 461–469.

Grootendorst, Maarten. 2022. BERTopic: "Neural topic modeling with a class-based TF-IDF procedure." arXiv:2203.05794v1.

Hartmann, Thomas F. 2017. *Die Reichstage unter Karl V.* Verfahren und Verfahrensentwicklung 1521–1555. Schriftenreihe der Historischen Kommission bei der Bayer. Akademie der Wissenschaften 100. Göttingen: Vandenhoeck & Ruprecht.

Haug-Moritz, Gabriele. 2020. "Arbeitspapier: Ständeversammlungen digital edieren. Ein neues Editionskonzept für den Reichstag 1576(–1662): Grundlagen, editorische Konsequenzen, praktische Umsetzung." In *Materialien zum Workshop des D-A-CH-Projekts „Der Regensburger Reichstag 1576. Ein Pilotprojekt zum digitalen Edieren frühneuzeitlicher Quellen"*, edited by id., Georg Vogeler, Roman Bleier et al.: 3–6. Graz. https://gams.uni-graz.at/o:rta1576.6540.

Haug-Moritz, Gabriele. 2021. "Deliberieren. Zur ständisch-parlamentarischen Beratungskultur im Lateineuropa des 16. Jahrhunderts." In *Historisches Jahrbuch*. Vol. 141: 114–155.

Haug-Moritz, Gabriele. 2023. "Historische Einführung: Der Reichstag des 16. Jahrhunderts als europäische Ständeversammlung. Zugleich eine Einführung in Schlüsselbegriffe der Reichstagsgeschichte." In *Der Reichstag zu Regensburg 1576* edited by Leeb, Neerfeld, Ortlieb et al. https://gams.uni-graz.at/o:rta1576.bt3564p3.

Hébert, Michel. 2014. *Parlementer*. Assemblées représentatives et échange politique en Europe occidentale à la fin du Moyen Age. Paris: Éditions de Boccard.

Kampmann, Christoph. 2023. "Immerwährender Reichstag und Türkengefahr im späten 17. Jahrhundert. Kommunikation – Konkurrenz – Konfrontation." In *Konkurrenzen in der Frühen Neuzeit. Aufeinandertreffen – Übereinstimmung – Rivalität*, edited by Jorun Poettering, Hillard von Thiessen and Franziska Neumann: 593–603. Göttingen: Vandenhoeck & Ruprecht.

Kewes, Paulina, Gunn, Steven, Pietrzyk-Reeves, Dorota, et al. 2022. "Early modern parliamentary studies: Overview and new perspectives." In *History Compass*: 1–16. DOI:10.111/hic3.12757.

Kieserling, André. 1994. "Interaktion in Organisationen." In *Die Verwaltung des politischen Systems. Neuere systemtheoretische Zugriffe auf ein altes Thema*, edited by Klaus Dammann: 168–182. Opladen: Westdeutscher Verlag.

Koller, Alexander. 2016. Maria von Spanien, die katholische Kaiserin. In *Nur die Frau des Kaisers?* Kaiserinnen in der Frühen Neuzeit, edited by Bettina Braun, Katrin Keller and Matthias Schnettger. Veröffentlichungen des Instituts für Österreichische Geschichtsforschung 64: 85–97. Vienna: Böhlau.

La Mela, Matti, Norén, Fredrik, and Hyvönen, Eero (eds.). 2022. *Proceedings of the Digital Parliamentary Data in Action (DiPaDa) Workshop Co-located with the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB) 2022*. http://ceur-ws.org/Vol-3133/.

Lanzinner, Maximilian. 2012/2023. "Der Gemeine Pfennig. Eine richtungsweisende Steuerform?" In Bayern – Heiliges Römisches Reich – Friedensstiftung. Ausgewählte Abhandlungen zur frühneuzeitlichen Geschichte, edited by Michael Rohrschneider and Arno Strohmeyer. Schriftenreihe zur Neueren Geschichte 42 (Neue Folge 5): 139–189. Münster: Aschendorff.

Leeb, Josef (ed.). 2013. *Der Reichstag zu Regensburg 1556/57*, edited by Maximilian Lanzinner. Munich: Oldenbourg. https://reichstagsakten.de.

Leeb, Josef, Neerfeld, Christiane, Ortlieb, Eva, Zeilinger, Florian, Bleier, Roman, Brantner, Elisabeth, and Rammer, Constanze (eds.). 2023. *Der Reichstag zu Regensburg 1576*. Digitale Edition, edited by Gabriele Haug-Moritz and Georg Vogeler. Deutsche Reichstagsakten, Reichsversammlungen 1556–1662. https://gams.uni-graz.at/context:rta1576.

Möllers, Christoph. 2021. *Freiheitsgrade*. Elemente einer liberalen politischen Mechanik edition suhrkamp 2755. Berlin: Suhrkamp.

Moraw, Peter. 2014, 1980. "Versuch über die Entstehung des Reichstags." In *Verfassungsgeschichte des Alten Reiches*, edited by Gabriele Haug-Moritz: 133–170. Stuttgart: Steiner.

Neu, Tim. 2024. "Teilhabegefüge, oder: Wie lässt sich die europäische Beratungskultur der Vormoderne konzeptionell fassen (und was könnte daraus für Editionsprojekte folgen)?" In *Digitale Edition und vormoderner Parlamentarismus/Digital Scholarly Edition and Pre-modern Parliamentarism. Eine interdisziplinäre Annäherung an frühneuzeitliche Quellen/An Interdisciplinary Approach to Early Modern Sources*, edited by Florian Zeilinger, Roman Bleier and Josef Leeb. Schriftenreihe der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften 114: 19–41. Göttingen: Vandenhoeck & Ruprecht.

Neuhaus, Helmut. 2006. "Der Reichstag als Zentrum eines „handelnden" Reiches." In *Heiliges Römisches Reich 962 bis 1806. Altes Reich und neue Staaten 1495–1806*. Vol. 2, edited by Heinz Schilling et al.: 43–52. Dresden.

Plener, Peter, Werber, Niels, and Wolf, Burkhardt. 2023. "Vor-Schrift." In *Das Protokoll*, edited by Peter Plener, Niels Werber and Burkhardt Wolf: V–VIII. Administudies 2. Berlin: Springer.

Ristilä, Anna, and Elo, Kimmo. 2023. "Observing political and societal changes in Finnish parliamentary speech data, 1980–2010, with topic modelling." In *Parliaments, Estates & Representation*: 1–28. DOI:10.1080/02606755.2023.2213550.

Stollberg-Rilinger, Barbara. 1999. *Vormünder des Volkes?* Konzepte landständischer Repräsentation in der Spätphase des Alten Reiches. Historische Forschungen 64. Berlin: Duncker & Humblot.

Stollberg-Rilinger, Barbara. 2005. "Was heißt Kulturgeschichte des Politischen? Einleitung." In *Was heißt Kulturgeschichte des Politischen?*, edited by id.: 9–24. Zeitschrift für Historische Forschung, Beiheft 35.

Vismann, Cornelia. 2000. *Akten*. Medientechnik und Recht. Frankfurt on the Main.

Zeilinger, Florian. 2023. "'...das wir nicht schuldigk sein, eines itzlichen gesandten affecten nachzuhangen.' Frühneuzeitliche Reichstags-Bevollmächtigte im Spannungsfeld von Repräsentation und Decision-Making am Beispiel der Kursächsischen Gesandten in Regensburg 1576." In *Frühneuzeitinfo*. Vol. 34: 37–65.

Appendix

The most important examples from the topics generated using the three different software packages—Mallet, Gensim, and BERTopic—mentioned and compared in the previous text are presented here side by side once again (with the words on which the interpretation and comparison are based highlighted). The dataset has been uploaded to Github (https://github.com/mining-rta/tm-rta-1576).

Table 2
Topics related to the subject term "*Landfrieden*/Peace, Offences, War in the Netherlands"

Tool	Topic(s)	Words
Mallet	M8	*anno, kreiß, abschidt,* ***vorbrecher****, genugsam, denen, maß,* ***straffe****,* ***unbillich****,* ***reuter****, gemacht,* ***bestallung****, sehe, zuberuhen, zuvorbessern, vorbesserung, vorabschidet, fernere, unnöttig, albereitt*
	M24	*placet,* ***ordnung****, gehandlet, quantum,* ***niderlandt****, deputationtag, pollicei, gut, primum, neuen, vertrauen, pleiben, gern, schickung, anttworten, articulus, turckenhilff, zwar, verabschidet, furgenommen*
Gensim	G96	*rath, caution, proposition,* ***schaden****, anno,* ***straff****,* ***vorbrecher****, churfursten, Julii, fursten*
BERTopic	B56	***gestrafft****,* ***verprecher****,* ***ordnung****,* ***acht****,* ***schaden****,* ***executions ordnung****,* ***executions****, ans,* ***execution****, vorfaren*

Table 3
Topics on (the empress,) the imperial family, Cardinal Morone and divine services

Tool	Topic(s)	Words
Mallet	M61	***ertzhertzogen****,* ***legaten****, getragen, hernacher,* ***legatus****,* ***junge****, tantum, heiligkeit,* ***keiserin****, geliebten, seitten, tragen,* ***sacrament****,* ***kertzen****,* ***procession****,* ***cardinalis****,* ***dhomstifft****,* ***königin****,* ***moronus****, sessel*
Gensim	G33	*usschus,* ***festum****,* ***legatus****, stenden, sachen,* ***junge****, anzal, begeben, Augusti, election*
BERTopic	-	-

Table 4
Topics on earlier imperial assemblies and the RTT's institutional memory ("*anno*")

Tool	Topic(s)	Words
Mallet	M5	*crais, craisen,* ***anno****, bairischen, saltzburg, khündten, schwäbischen, …*
	M8	***anno****, kreiß, abschidt, vorbrecher, genugsam, denen, maß, …*
	M21	*trier, achten, sachen, sachssen, coln, meynunge, dohin, sache, hette, zulassen, hern,* ***anno****, dinge, …*
	M25	*reichs, schreiben, muntzordnung, abschidt, executorn,* ***anno****, …*
	M38	*rethe,* ***anno****, augspurgischer, saxischen, stat, sachen, graven, confession, erfolgt, hessen, baden, wirtemberg, …*
	M46	***anno****, bedenckhen, rath, derwegen, meinung, pleiben, …*
	M50	*usschus, stendt,* ***anno****, reichs, abschidt, bedenckhen, puncto, puncten, fürgenommen, weiter, rath, …*
	M63	***anno****, matricul, moderation, sachen, deputation, ergentzung, …*
	M74	*concept, trier, coln, pfaltz, brandenburg, bleiben, sachssen, gemeß, abgelesen, zusetzen, meintz,* ***anno****, wortt, …*
	M83	*stifft, reich, closter, recht, herkommen, underthanen, potentaten, steirische, billich, anschlag, immediate, dupli,* ***anno****, …*
Gensim	G4	*Reichs, versehen, Meintz,* ***anno****, rath, tag, …*
	G6	*schrifft, rethe, concept, graffen, stende, bitten, gesandter, anhero, 2),* ***anno***
	G13	***anno****, 21, deputation, sachen, proceß, tag, matricul, …*

	G16	***anno**, anzeig, zulassen, 70, Lunenburg, supplication, …*
	G23	*zöll, **anno**, zoll, Revers, 70, würtzburgischen, abschidt, …*
	G25	*Lubeck, stende, Schweden, **anno**, rath, schiff, …*
	G39	*stende, müntzordnung, gar, herrn, müntzen, graven, müntz, crais, Bairn, **anno***
	G45	*sachen, zolle, **anno**, newe, Hagenauw, iustitia, personen, …*
	G52	*ad, sonderbare, ex, hoch, erhalten, versehen, gleich, her, sachen, **anno***
	G56	*müntz, kreis, müntzen, müntzedict, probation, Befinden, **anno**, Saltzburg, …*
	G58	*Brandenburg, Trier, decreto, **anno**, fiscall, stende, bezalt, …*
	G60	*stendt, **anno**, session, fiscal, proceß, deßhalber, matricul, …*
	G67	***anno**, legation, bedacht, Lifflandt, caesari, puncten, …*
	G68	*bedenncкhen, **anno**, fürstenraths, sequestration, gefallen, bedenckhen, stendt, restitution, …*
	G73	***anno**, Reichs, schreiben, zulassen, proceß, Schweden, Reich, …*
	G77	*personen, rath, zulassen, 6, **anno**, visitation, ad, assessores, 70, räth*
	G91	*hülff, usschus, crais, stendt, jar, Reich, pfennig, **anno**, …*
	G93	*sachen, herr, friden, vergleichen, stenden, **anno**, …*
	G96	*rath, caution, proposition, schaden, **anno**, straff, vorbrecher, …*
	G99	*religion, catholische, churfursten, schreiben, **anno**, weitters, …*
BERTopic	B23	***anno**, 70, **anno 70**, 66, 71, **anno 66**, **anno 21**, 21, matricul*

Table 5

Topics on the recording of files ("*apud acta*")

Tool	Topic(s)	Words
Mallet	M22	*zulaßen, sach, caesar, hielt, zuweisen, coln, contra, reich, zubitten, derwegen, billich, bericht, hierin, caesarem, bevelch, **apud**, alsdan, camera, decreto, hora*
Gensim	G21	*Burgundt, Hispanum, jar, Stabell, moderation, beschwerungen, **ad**, legation, **acta**, Session*
	G23	*zöll, anno, zoll, Revers, 70, würtzburgischen, abschidt, stendt, Lunenburg, **acta***
	G46	*Schencking, **ad**, papam, pro, capittell, 2, ritterschafft, **apud**, Munster, sach*
BERTopic	-	-

Table 6

Topics related to the subject term "Procedure, Begin" ("*ansagen*"/announcing of meetings)

Tool	Topic(s)	Words
Mallet	M30	*meintz, **ansag**, meintzischen, meintzische, **zettel**, rath, **erschienen**, **ansagen**, herrn, **zuerscheinen**, räthe, sechsische, cantzler, cantzley, sechsischen, reichs, **ansags**, marschalckh, churfürsten, fürsten*
Gensim	G48	kreis, rath, **zettel**, **ansag**, meintzischen, **ansagen**, sechsische, Sachsen, **erschienen**, Reichs
BERTopic	B65	*zettel, **ansagen**, **ansags**, **ansags zettel**, cantzley, **zuerscheinen**, reichs marschalckh, marschalckh, reichs, ghen*

Table 7

Topics related to the subject term "Procedure, Deliberate" ("*concept ablesen*"/reading out drafts)

Tool	Topic(s)	Words
Mallet	M11	***concept**, **abgelesen**, stettrath, saltzburg, nomine, **referirt**, abschiedt, bracht, similiter, östereich, uffs, papir, churfürstenrath, placet, beschluß, contra, gesetzt, **gefallen**, maintzischer, **berathschlagung***

	M74	***concept**, trier, coln, pfaltz, brandenburg, **bleiben**, sachssen, gemeß, **abgelesen**, zusetzen, meintz, anno, wortt, **gefallen**, genere, **beratschlagung**, osterreich, zumachen, kreiß, passiren*
Gensim	G6	*schrifft, rethe, **concept**, graffen, stende, bitten, gesandter, anhero, 2), anno*
	G8	*stendt, sachen, **concept**, puncto, sessionis, selbst, **pleiben**, verglichen, rath, ursachen*
	G54	*monat, **concept**, eilenden, Trier, stenden, 6, 60, 8, zustellen, Sachssen*
	G74	*rath, fursten, **bedencken**, cantzler, raths, churfursten, chur-, Meintzische, Fursten, **concept***
	G82	***concept**, nomine, fürstenraths, **referirt**, visitation, stendt, **zulassen**, fürstenrath, bedenckhen, rath*
	G95	*contra, puncto, decret, **concept**, **abgelesen**, **bedenckhen**, Furstenrath, rath, müntz, **Lassen***
BERTopic	B5	***concept**, trier, **abgelesen**, wissen, condition, **gefallen**, schrift, zuordnung, bericht, lisen*
	B40	***concept**, trier, sachssen, **concept gefallen**, zusetzen, **gefallen**, wortt, achten, **trier concept**, pfaltz*
	B51	*churfurstenrath, **concept**, churfürstenrath, **pleiben**, fürstenrath, müntz, **concepten**, befinden, churfursten, **abgelesen***

The Politics of Compound Neologisms: A Novel Text-Mining Approach for Tracing Conceptual Transformations in Parliamentary Discourse

Daniel Brodén[a], Claes Ohlsson[b], Magnus P. Ängsal[a], Henrik Björck[a], Mats Fridlund[a], Leif-Jöran Olsson[a], Leif Runefelt[c] and Shafqat M. Virk[a]

[a] *Gothenburg University*
[b] *Linnaeus University*
[c] *Södertörn University*

Abstract

This paper highlights the underutilized analytical potential of compounds and neologisms as indicators of discursive change in text mining applications, particularly in the study of parliamentary discourse and conceptual transformation. Drawing on results from two research projects, this project-wide paper discusses how compound neologisms function as markers of discursive change through case studies focused on the formation, frequency, and productivity of compounds related to the key concepts of 'market' and 'terrorism' in the Swedish Parliament. The analysis combines distant reading techniques to identify large-scale trends and close reading to examine the specific contexts of these compounds. By focusing on compound formation, we emphasize the analytical potential of basic linguistic features often overlooked in Digital Humanities research, offering a fresh perspective on large parliamentary datasets and their role in tracing conceptual transformations over time.

Keywords 1

Parliamentary data, Sweden, text mining, compound neologisms

1. Introduction

This paper stems from two distinct research projects in text mining on Swedish parliamentary data, both exploring the historical transformation of key concepts and shifts in their meaning and usage in public discourse. The project *The Market Language* (2022–2025, Ohlsson et al. 2022) investigates the discourse surrounding markets from the Middle Ages to the present, while *Terrorism in Swedish Politics* (*SweTerror*) (2020–2026, Edlund et al. 2022) examines the framing of political terror in parliamentary discourse from 1968 to 2018. Despite their different designs and focuses, both projects have produced interesting results, indicating that the word formation of compound words and compound neologisms in Swedish plays a role in shaping and developing political discourse.

The overarching aim of this project-wide paper is to highlight underexplored analytical potential of compounds and neologisms as 'indicators of discursive change' in text mining applications for studying parliamentary discourse and conceptual transformation. By indicators of discursive change, we refer to how compounds and neologisms can help identify patterns and shifts in word usage and discourse, which may warrant further investigation (Hornscheidt 2008). More specifically, we argue that combining corpus linguistics and conceptual history for text mining of compounds represents an

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, Reykjavik, Iceland, May 28, 2024.
EMAIL: daniel.broden@gu.se (D. Brodén); claes.ohlsson@lnu.se (C. Ohlsson); magnus.pettersson.angsal@gu.se (M. Ängsal), henrik.bjorck@lir.gu.se (H. Björck), mats.fridlund@gu.se (M. Fridlund), leif-joran.olsson@gu.se (L-J Olsson), leif.runefelt@sh.se (L. Runefelt), shafqat.virk@svenska.gu.se (S. Virk) ORCID: 0000-0002-5914-1516 (D. Brodén); 0000-0002-6252-6126 (C. Ohlsson); 0000-0001-5996-5067 (M. Ängsal), 0000-0003-2171-7361 (H. Björck), 0000-0002-5759-0027 (M. Fridlund), 0000-0001-7107-4101 (L-J Olsson), 0000-0003-2453-7040 (L. Runefelt), 0000-0002-5030-9191(S. Virk)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

overlooked and underdeveloped approach in data-rich discourse analysis. Theoretically, we situate this issue within the linguistic context of how multi-word phrases are expressed in different languages and how compounds constitute a particular morphological feature of Swedish (Finkbeiner and Schlücker 2019). In doing so, we also emphasize the analytical potential of leveraging linguistic features from less dominant languages in academic research for text mining purposes.

Building on results from previous studies in both projects, which combine distant and close reading of Swedish parliamentary datasets, this paper aims to illustrate and discuss the analytical significance of compound word formation in Swedish when tracing political discourse through two case studies – one on the transformation of the concept of the *market* and the other on *terrorism*. We examine how both concepts have expanded over time to encompass different meanings and topical issues by compound neologisms. We suggest that this opens a previously under-researched analytical avenue in text mining, offering a lens for examining discursive transformations through the specific linguistic phenomenon of compounds as manifested in language use.

1.1. Disposition

Following some theoretical and methodological reflections on compounds and the approaches of both projects to the Swedish parliamentary datasets, the paper first turns to the concept of the market. The findings from the Market language project focuses on the productivity of compound words over time and the specific discourses associated with these patterns of usage. We then turn to the second case and parliamentary debate on terrorism, illustrating how neologisms help trace various aspects of the historical transformation of the discourse on political terror. In this section, we discuss how the productivity of compounds reflects different historical contexts and discursive shifts deserving of further analysis. We conclude by summarizing our key points regarding the analytical and also potential methodological benefits of using compound neologisms to explore parliamentary discourse.

2. The analytical significance of compounds

The SweTerror and The Market Language projects are both multidisciplinary endeavors that combine analytical methods for processing large textual datasets with inquiries into actual language usage and the deployment of specific concepts in the parliamentary texts under study. In this regard, the two projects build on previous research that has explored the transformation of key concepts in Swedish parliamentary datasets through statistical analysis (Jarlbrink et al. 2022; Norén et al. 2022). However, in both projects we have identified and focused on the presence and analytical value of compound neologisms for tracing transformation in historical parliamentary discourse.

The two projects are grounded in distinct concepts and phenomena that may initially appear unrelated and therefore difficult to compare. However, they share several common features across different dimensions. One such area is the way concepts are expressed using different terms that nonetheless exhibit shared morphological and syntactic characteristics, with compound formation as the main feature. The use of compounds in text mining has been noted by Postiglione (2024), and this approach facilitates comparisons and serves as a central aspect of the case studies' investigations. Furthermore, words used to discuss topics like the market or terrorism over time are subject to discursive changes, both in general and specifically within the parliamentary discourse analyzed in both projects. The fact that politicians have debated various aspects of the market and terrorism for legislation through the use of compounds represents a unifying factor between the projects.

Text mining often involves analyzing linguistic patterns to uncover structures and relationships within language, including how languages form and use lexical units. N-grams are a widely used method for identifying recurrent multi-word combinations in natural language, both in spoken and written discourse (Lyse and Andersen 2012). While n-grams are a useful tool for mapping recurring multi-word combinations, they do not effectively capture the inherent capacity for neologism creation in certain languages.

Notably, the formation of new words through the synthetic amalgamation of different terms into cohesive lexical units – compounds – is a distinctive morphological feature of the Swedish language (Finkbeiner and Schlücker 2019). Although similar morphological patterns can be found in other

Germanic languages, including the Nordic languages and Dutch, this is not as prominent in English, the dominant language in academic research. English tends to blur the boundaries between compounds and multi-word expressions (Bauer 2019), as seen in the irregular spelling of compounds, which are often represented orthographically as separate lexical units (e.g., the English compound '*labor market'* and its Swedish equivalent *arbetsmarknad*).

While orthographic patterns of compound formation in English considerably limit the ability to detect compounds computationally in large datasets, Swedish presents an overlooked potential for discourse analysis. The latter language readily allows for and visibly reflects the creation of compounds by combining existing words into new morphologically coherent lexical units with two or more elements. This form of lexical composition is valuable both as a unit of analysis in computational linguistic studies and as a discursive phenomenon, offering concentrated semantic information. Unlike simplex nouns, which often require embedding in multi-word expressions to convey similar meaning, Swedish compounds encapsulate complex ideas within a single lexical unit

By examining how our two focal words, 'terrorism' and 'market', co-occur with other words in compounds, we can identify significant semantic patterns of usage that reflect evaluative approaches and attitudes, serving as indicators of discursive change. This is in line with the approach within conceptual history – or history of concepts, or *begreppshistoria* in Swedish – to perceive conceptual changes as indicators of and factors in processes of social transformation. Conceptual change is an indicator of social change but should not be reduced to a mere reflection of extra-linguistic forces. At the same time, an altered conception of one's current position and future possibilities becomes a factor in the course of events; people act differently when they perceive themselves in terms of "citizens of the republic" rather than "subjects to the king" (Kurunmäki & Marjanen 2018). Our analytical tools include word frequency analysis, keyword collocation examination and the exploration of multi-word expressions, such as phrases, involving our key terms. These often reveal entrenched phraseological relationships, providing insights into the lexical, syntactic, and semantic characteristics of the words (Koteyko et al. 2010). This is crucial for understanding the discursive use of concepts – how words are situated within specific contexts. By examining patterns of compound usage and their contexts, we can explore how compounds shape recurring perceptions of the phenomena they represent and serve as indicators of conceptual transformation.

3. Materials and Methods

Both projects utilize publicly available Swedish parliamentary datasets. The Market Language project's case for this paper draws from the complete dataset of texts from the Bicameral Parliament (1867–1970) (riksdagstryck.kb.se). The period of the Swedish Bicameral parliament from 1867 to 1970 is of particular interest to the project, as it encompasses a time of industrialization and economic transformation, laying the foundations for both the modern market economy and modern Swedish democracy. This dataset has been available in .pdf and .xml formats through the Swedish parliament's website for several years, but it has since been downloaded by the project's Language Technology (LT) analyst (Virk) processed, annotated, and included in the Språkbanken Text infrastructure as a sub corpus (*'Tvåkammarriksdagen'*), allowing for the use of sub-genre categories for specialized queries (Virk et al. 2024). The project initially focused on mapping how market-related terms in general have been used in the material, based on frequencies and collocation patterns. Early-stage text mining revealed that compounds with "market" as an element were notably frequent, varied, and, above all, productive. Based on these preliminary findings, we proceeded with more detailed analyses of such compounds, categorized by decade within the corpus (Ohlsson et al. 2022).

The work in the SweTerror project initially relied on the corpus from riksdagstryck.kb.se, but now draws on the Swedish Parliament Corpus of the minutes, developed by the SWERIK infrastructure (latest version 0.14). This infrastructure cleans, partially re-digitizes, annotates with metadata and curates the national parliamentary record for research purposes (https://github.com/swerik-project/swerik-project.github.io) (Yrjänäinen et al. 2024). SweTerror's LT analyst (Olsson) is also further enriching SWERIK's dataset to meet the specific research needs of the project. Notably, this draws on results from two prior studies within the SweTerror project: one based on the corpus from riksdagstryck.kb.se (Fridlund et al. 2022) and the other one on the improved SWERIK dataset (Brodén

et al. 2023). Together, these studies span both the Bicameral Parliament (1867–1970) and the Unicameral Parliament (1971–2018). While the datasets used in these studies are not identical, they cover distinct time periods, and we have not observed any discrepancies in terrorism-related word frequencies that would significantly affect our discussion here.

In this paper, both projects have continued the straightforward, yet productive approach used in our prior studies. While our methods differ in certain technical details, both involve querying the respective corpora for relevant Swedish lemmas ('market' and 'terror' and 'terrorism'). This process generated data on all compound words containing these lemmas, providing information about frequencies of the compounds and the 'productivity' of compound neologisms over time, measured in units such as years or decades. This data served as the foundation for our analysis. Importantly, while we primarily employed distant reading, we also incorporated close reading to a limited extent, examining the specific contexts of significant compounds and neologisms, with a particular focus on their first occurrences.

4. Case Study: Market

The Market Language project explores how the concept of the market has been used, developed, and transformed in Sweden, from the Middle Ages to the present. This involves examining various relevant corpora from different social domains during this long historical span. A unifying factor is the focus on text types that have existed over long periods with relative stability, such as newspapers, laws and other political texts. Given these starting points, the project aims to compare the occurrence of market-related terms in various forms across different types of material over time with awareness to the various practical definitions and uses for market or even markets (Frankel 2018).

A first step has been to map the frequency of both tokens and types in the material, with parliamentary texts serving as a key component. The results show that market words have been used consistently over an extended period, albeit with some differences in meaning and contextual setting, where the development from a concrete meaning to more abstract meanings is the most prominent. The same development is also a result in a previous study on market words in Swedish historical press corpora from 1820 to 1900 (Ohlsson 2020), which opens for an interpretation of market as a lock word as of Baker (2011). Words with lock word properties may change in meaning over time but are relatively stable in terms of frequency in different types of corpus material. The consistency in how market concepts appear as word or words in corpora over time stands in contrast with the second case study of this article (see below), where 'terror' is a concept expressed with narrower and more fluctuating frequency. However, in mapping the frequency of market terms, there is significant variation in the composition of specific compounds used across time periods and corpora as also reported in Ohlsson (2020) on market words in a historical press corpus.

We have therefore focused on the productivity of compounds, both in terms of their occurrence and frequency, as well as the emergence of new compound forms that begin to be used over time in the text material of the Bicameral Parliament. Some of us have previously discussed this productivity aspect of compositional forms in parliamentary data (Ohlsson et al. (2022)) and our results indicate that new compositional forms represent areas that gain prominence in political debates and also have the capacity to generate further compositional forms.

4.1. The productivity of compounds

A trend that we further explore in this paper, is then the consistent increase in new compositional forms featuring *marknad* (market) as an element. As shown in Figure 1, the number of new compounds in which market is either a prefix or suffix element increases over time. The development, spanning the period from 1867 to 1970 in the parliamentary texts at hand, shows an accelerating rate, evident from around 1920 onwards, and later also a more rapid increase in the post-World War II period and towards the end of the bicameral Parliament period in 1967, which can be aligned with the so-called Swedish model of combining a free market economy with redistributive politics in the 1950s.

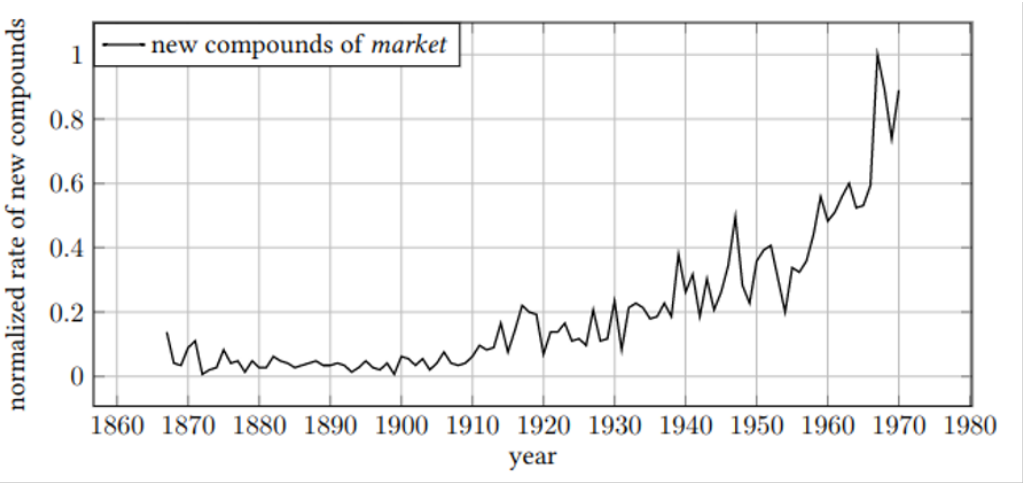


Figure 1: The productivity of new compound types with market as an element.

In Figure 1, we highlight all new compound types with *marknad* ('market') for every year but not appearing the previous year, thus all compound types in the data set. We define the total compound productivity of a word in a corpus as the number of compounds in the corpus formed from a word, e.g. market, and the compound productivity as the number of new compounds of this word per year. We have accounted for the increase in the number of texts and words within the corpus during the period from 1867 to 1970 by normalizing the number of new types in relation to the prevailing text volume. This count of new compound types reveals the general productivity of words with market accelerating over time and also indicates that this increase of types in numbers appears to be more and more varied in terms of content and discursive usage, even in a specific context such as the parliamentary texts of the bicameral Parliament. More and more compound variation appears over time, and this also means that market compounds in general become more frequent in the dataset.

For instance, the emergence of the compound *arbetsmarknad* ('labor market') leads to the creation of additional compositions, based on that compound. *Arbetsmarknad* makes its first appearances as a new compound in the bicameral parliament already in the 1860s but it increases in use by frequency in relation to the number of texts in the data set during the 1900s with a sharp acceleration from the 1950s as seen in Figure 2. The figure shows relative frequencies for both the indefinite and definite forms (*arbetsmarknad* and *arbetsmarknaden*), where the usages trail each other until the 1950s where the indefinite form increases in use.

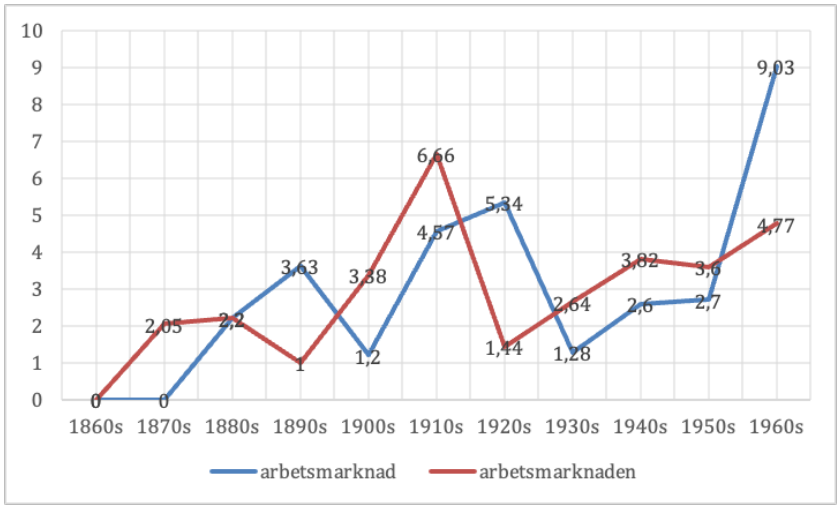


Figure 2: Relative frequency for occurrences of *arbetsmarknad* ('labor market') and *arbetsmarknaden* ('the labor market') in the Bicameral Parliament 1867–1970.

These patterns of compositional productivity serve as a foundation for discussing the growing utilization of the concept of market in political discourse in general and the attribution of new properties and roles to the concept itself.

4.2. A trajectory of compounds

Another example of the mapping of compound productivity in the bicameral parliament data makes use of semantic grouping or clustering of the most common market compound types during the whole period from the 1860s to the 1960s. We focused on the 20 most frequently occurring compound forms per decade, calculated based on their absolute and relative frequencies within each material period. These compounds were then manually grouped according to their semantic and contextual characteristics. Each semantically coherent group of compounds was defined by 3–4 examples for which the research team reached a consensus on the final interpretation. Seven examples of semantic groups are presented in Table 1, with a color code assigned to each group to highlight changes in their occurrence over time.

Table 1

Semantic groups for most frequent compounds (top 20) with *marknad* ('market') in the Bicameral Parliament 1867–1970.

1860s	1870s	1880s	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s
finance	finance	world	world	labor	labor	labor	labor	labor	labor	labor
market	market	market	market	market	market	market	market	market	market	market
world	world	finance	finance	world	world	world	world	housing	housing	finance
market	market	market	market	market	market	market	market	market	market	market
market	physical	market	labor	finance	finance	housing	finance	world	finance	housing
for goods	market	for goods	market	market	market	market	market	market	market	market
physical	market	physical	physical	market	housing	finance	housing	finance	world	world
market	for goods	market	market	for	market	market	market	market	market	market
			market	goods	market	market	market		market	
			for goods	market	value	for goods	for goods		value	
				value	market	market	market			
					for goods	value	value			

The 1800s decades show that the most common semantic groups for that period involve compounds relating to finance markets (e.g. money, credit, lending), geographically denoted markets (e.g. world, domestic), markets for goods (e.g. iron, butter, wood) and the physical market location (e.g. place, day, event). In the 1890s, labor market compounds make their first appearances among the 20 most common compounds, and this semantic group is from this time always the most common cluster until the end of the bicameral parliament in the 1960s. It is also visible that the physical market as a place or time lost its significance during the first half of the 1900s, which also seems to be the case for the group of markets for goods. New appearances during the decades of the 1900s are the two groups of market value and housing markets, while the groups for compounds relating to finance markets and geographically denoted markets remain among the 20 most common compounds over time.

An additional look at the example of the labor market group reveals an example of what can be called intra-compound productivity where the original bi-element compound has the capacity to produce new tri-element or even longer compounds. This is shown in Table 2, which shows this productivity pattern from the 1930s to the 1960s, since this is the period when labor market compounds accelerate in productivity and also are among the most common compounds with market in the dataset.

Table 2
Compound forms with *arbetsmarknad* ('labor market') among the most frequent compounds in the Bicameral Parliament data from 1930 to 1970.

1930–1939	1940–1949	1950–1959	1960–1969
(the) labor market	the labor market (the) labor market commission (the) labor market board (the) labor market situation	(the) labor market board (the) labor market (the) labor market situation	(the) labor market board (the) labor market labor market politics/political (the) labor market situation (the) labor market authorities (the) labor market agency

In the 1940s, the labor market compound was the base for new, frequently used compounds including examples of government authorities such as *arbetsmarknadsstyrelse* ('labor market board') or *arbetsmarknadskommission* ('commission'). There is also discussion on *arbetsmarknadssituationen* ('the labor market situation'). In the last the decade of the bicameral parliament in the 1960s, this multi-element compound productivity increases even more with tri-element forms for *arbetsmarknadspolitik* ('labor market politics') and further references to government authorities, aimed at regulating and managing the Swedish labor market situation, *arbetsmarknadsverket*. All in all, these patterns of usage indicate that a market for labor began to be used as an idea for political thought and action during the 1900s to become the dominant idea in the discourse of debating and legislating on conditions of labor in the bicameral parliament over time. The productivity of labor market compounds further indicates the strength of this idea.

5. Case Study: Terror

Our second case study focuses on the development of the closely related words 'terror' and 'terrorism' (same spellings in Swedish and English) when they appear as constituents in compounds within Swedish parliamentary discourse from 1867 to 2021. This is based on previous research within the SweTerror project, which has shown in two separate studies (Fridlund et al. 2022; Brodén et al. 2023), based on chronological extractions of compound neologisms containing the words 'terror' or 'terrorism', that the concept of terrorism first gained its more contemporary meanings in Swedish parliamentary discourse in the early 1970s. This aligns with results from previous international research (Stampnitzky 2013).

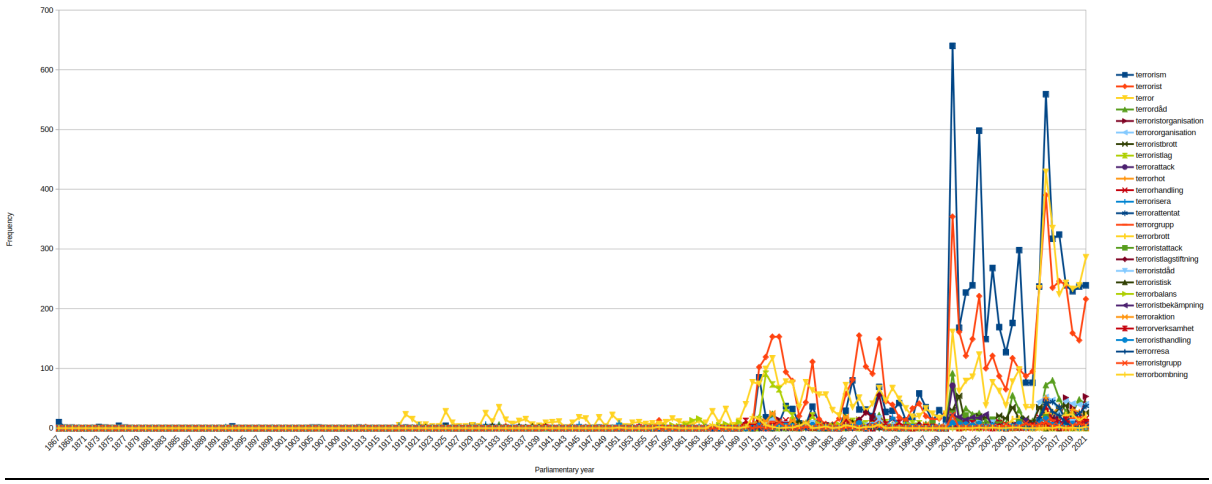


Figure 3: The frequency of compounds with 'terror' and 'terrorism' as an element in Swedish parliamentary debate 1867–2021 (lemmas with more than 100 occurrences in the corpus), including neologisms introduced prior to the period, such as the major trend lines of the 2000s 'terrorism', 'terrorist' and 'terror'. Note that the SweTerror project enact a contextualizing understanding of the parliamentary data by grouping debates by parliamentary year (autumn–summer) instead of calendar year, aligning with the Riksdag's mandate period.

5.1. Contextualizations and periodizations of compounds

In the first study (Fridlund et al., 2022), which focuses on debates in the Bicameral Parliament, we highlighted the complexity of the early usage of the words 'terrorism' and 'terror'. Although 'terrorism' appeared in parliamentary debate as early as 1867, terror-related words and compounds only began to gain traction from 1918 onwards. Furthermore, the early use of compounds featuring 'terrorism' reveals that the term was initially discussed metaphorically. For example, *valterrorism* ('electoral terrorism') was used in 1893 to describe perceived oppressive voting procedures within the Parliament itself. It was not until the early 1900s that MPs began using 'terrorism' to refer to political violence, though its usage remained scarce, with only 16 occurrences overall between 1900 and 1969. Instead, other terms, such as *anarkism* ('anarchism'), were sometimes used to describe acts of political violence that we now associate with terrorism. The study also showed that the usage of 'terrorism', both alone and in compounds, actually preceded that of 'terror', with the latter first introduced in Swedish parliamentary speech in 1918 in reference to the Finnish Civil War.(c.f. Fridlund et al. 2020).[2]

When analyzing the use of terror compounds between 1919 and 1970, besides simple terms (*terror, terrorisera, terrorism, terrorist, terroristisk, terroriserande, terrorisering*) only 8 compound words occurred more than 10 times: *terrorbalans* ('terror balance'), *terrorvapen* ('weapons of terror'), *terroranfall* ('terror attacks'), *blodsterror* ('blood terror'), *fackföreningsterror* ('trade union terror'), *terrorbombning* ('terror bombing'), *terrorregim* ('terror regime') and *terrordåd* ('act of terror'). Of those, only two compounds – *terrorbalans* and *terrordåd* – were productive enough to make it into the top 20 neologisms overall (see Table 3) indicating that many of these compounds were primarily useful in a particular historic context. This shows that not only the introduction of neologisms is worth studying but also their development and decline.

Figure 4: The production of compounds with ‘terror’ and ‘terrorism’ as an element in Swedish parliamentary debate 1919–1970. Detail of figure (Figure 1; Fridlund et al. 2022) showing the first occurrences of new terror compounds in the bicameral corpus.

Connected to this context-dependent use of neologisms, perhaps most significant in the context of this paper is that we were able to identify distinct periods of compound productivity, which appear to be connected to and reflective of specific historical contexts (Figure 4). For instance, terms like *fackföreningsterror* (‘labor union terror’) and *arbetsgivarterror* (‘employer terror’) were used between 1925 and 1935 to describe various disruptive actions by both labor unions and employers. Another example is *luftterror* (‘aerial terror’), in reference to the threat of wartime aerial bombings against civilian targets from 1936 to 1940. We also see *atomterror* (‘atomic terror’) and related terms during the period from 1948 to 1963, denoting the nuclear threat of the Cold War. This serves as another clear example of how compounds can function as indicators of discursive change.

Table 3
The 'top 20' neologisms with 'terror' or 'terrorism' as an element in Swedish Parliamentary debate 1867–2021. Note that the SweTerror project enact a contextualizing understanding of the parliamentary data by grouping debates by parliamentary year (autumn–summer) instead of calendar year, aligning with the Riksdag's mandate period.

Rank	Lemma	Parliamentary year	Occurrences
1	terrorism	1867	6348
2	terrorist	1912	5503
3	terror	1917	5357
4	terrordåd	1932	928
5	terroristorganisation	1971	531
6	terrororganisation	1953	469
7	terroristbrott	1988	466
8	terroristlag	1971	426
9	terrorattack	1976	320
10	terrorhot	1973	283

Rank	Lemma	Parliamentary year	Occurrences
11	terrorhandling	1943	277
12	terrorisera	1869	276
13	terrorattentat	2000	273
14	terrorgrupp	1938	222
15	terrorbrott	1976	206
16	terroristattack	1972	198
17	terroristlagstiftning	1972	193
18	terroristdåd	1934	185
19	terroristisk	1872	180
20	terrorbalans	1955	179

5.2. Emerging compound contexts

In our second study, focusing on the debate in the Unicameral Parliament (Brodén et al. 2023), we could observe how neologisms signaled the emergence of a novel legislative framing of terrorism (Figure 5; Table 3). Among the compounds introduced by MPs during this period were *terroristlag* ('terrorist law') in 1971, *terroristlagstiftning* ('terrorist legislation') in 1972 and *terrorist-bestämmelser* ('terrorist regulations') in 1975. While it is not surprising that the Parliament approached terrorism from a legislative perspective (given its core function of passing legislation), 'legislative' terror compounds were notably rare before the 1970s (see Figure 3). The sudden increase in legislative neologisms in the 1970s clearly suggests that terrorism had become to be seen as a domestic issue to be addressed through legislation. The killing of the Yugoslavian ambassador in Stockholm in 1972 and the Bulltofta hijacking in 1973, carried out by Croatian militants, led to the Swedish Parliament adopting the Terrorist Act in 1973. This law was designed to prevent politically motivated acts of violence with an international dimension, specifically targeting foreign nationals (Ribbing 2000; Hansén 2007). This development followed a broader trend in Western states toward criminalizing international terrorism (Stampnitzky 2017; Zoller 2021). In Sweden, parliamentary discourse on terrorism increasingly revolved around the controversial Terrorist Act, which was renewed annually and thus became the subject of debate. For instance, the use of the compound 'terrorist law' increased in the 1980s in connection with the controversies surrounding the application of the Terrorist Act against several Kurds suspected of involvement in the murders of two defectors from the PKK (Kurdistan's Workers Party) in 1984 and 1985, as well as the assassination of Prime Minister Olof Palme in 1986.

We also found that the production of terrorism compounds in the Parliament intensified and transformed following the 9/11 attacks in 2001 (Brodén et al. 2023; see also Ängsal et al. 2024). An overview of the most frequently used words incorporating terror and terrorism, including neologisms introduced prior to 1971 (Figure 3), shows how much the issue gained traction post 9/11. Whereas 129 neologisms appeared during the period between 1971/1972 and 2000/2001, just as many (134) were introduced in about half the time span of 2001/2002 to 2013/14.

From 2014 onwards, terrorism also became associated with specific groups in a way that had not been seen before, with 80 neologisms appearing from 2013/2014 to 2020/2021. Previously, various organisations had been mentioned in debates about terrorism, but not as part of compounds, with the rare exceptions of *ustasjaterrorister* ('Ustaše terrorists') and *arabterrorist* ('Arab terrorist') used once in 1971 and 1973, respectively. However, starting in 2014, *IS-terrorist* ('IS terrorist') was used 88 times in reference to the Islamic State (IS), the militant Salafist group proclaiming itself a worldwide caliphate. Additionally, the phenomenon of foreign fighters travelling abroad to participate in armed conflicts also became linked to IS. The term *terrorresa* ('terror travel') appeared in 2013, followed by *terroristresa* ('terrorist travel') in 2014, and *terrorismresa* ('terrorism travel') and *terroriststridande* ('terrorist combatant') in 2015.

In addition to use these neologisms to describe the new threat of the Islamic State, these compounds reflect a growing tendency to associate terrorism with warfare. While connections between terrorism and warfare were not entirely new per se – especially with metaphors and official government

designations like the 'Global War on Terrorism' coined by the Bush administration – Sweden had long maintained a less militaristic view of terrorism in its policymaking, compared to the US and other European countries. However, by the 2000s, the idea of terrorism as a military threat began to emerge in the Swedish parliamentary discourse, as evidenced by compounds like *terroristkrig* ('terrorist war') in 2005 and *terrorkrigsbrott* ('terror war crime') in 2015.

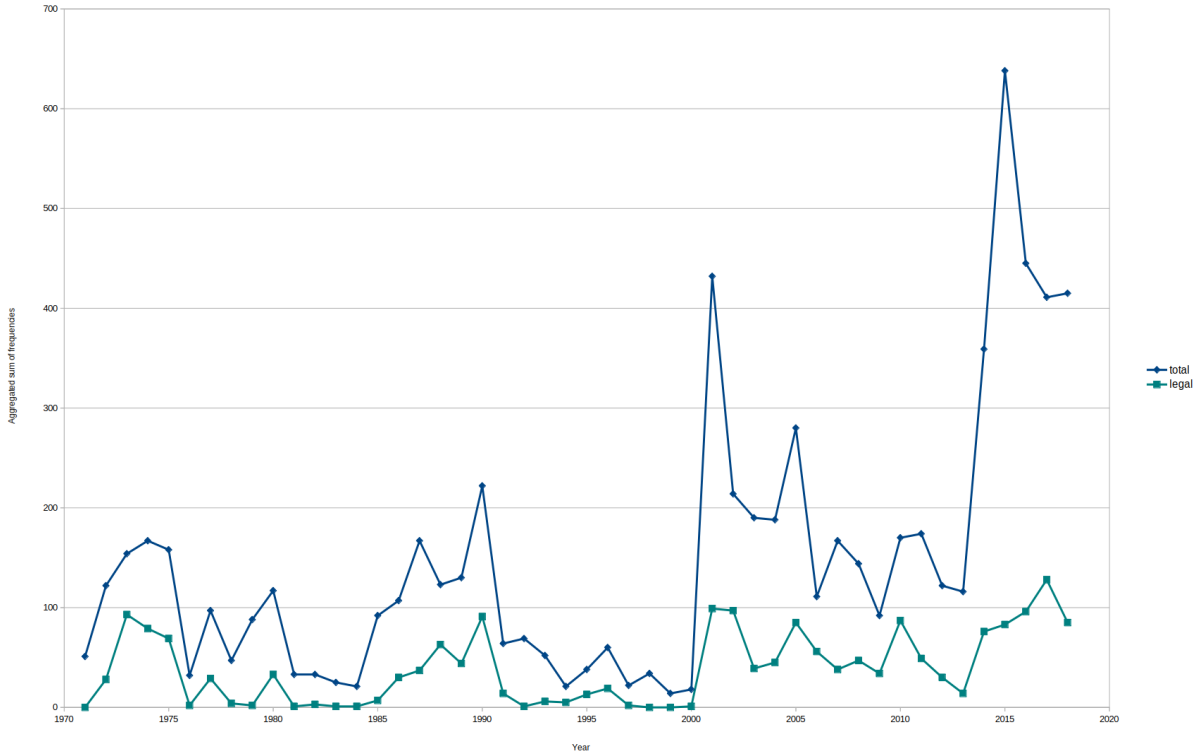


Figure 5: The frequency of compounds with 'terror' and 'terrorism' as an element in Swedish parliamentary debate 1971–2021, focusing on the relation between the total number of compounds (blue line) and 'legalistic' compounds (green) (Brodén et al. 2023). Note that the SweTerror project enact a contextualizing understanding of the parliamentary data by grouping debates by parliamentary year (autumn–summer) instead of calendar year, aligning with the Riksdag's mandate period.

Furthermore, these compounds reflect a growing tendency to associate terrorism with warfare. While connections between terrorism and warfare were not entirely new per se – especially with metaphors and official government designations like the 'Global War on Terrorism' coined by the Bush administration – Sweden had long maintained a less militaristic view of terrorism in its policymaking, compared to the US and other European countries. However, by the 2000s, the idea of terrorism as a military threat began to emerge in the Swedish parliamentary discourse, as evidenced by compounds like *terroristkrig* ('terrorist war') in 2005 and *terrorkrigsbrott* ('terror war crime') in 2015.

Moreover, the use of compounds in the 2000s indicates a strengthened counter-terrorism discourse (Figure 6). The first term appears to be *motterror* ('counter-terror') that appeared in 1969 followed by *terrorbekämpning* ('terror combatting', approx. 'anti-terror measures') and *terroristbekämpning* ('terrorist combatting', approx. 'anti-terrorist measures') that both first appeared in 1975, there has been growing productivity in counter-terrorism related compounds since the late 1980s. Unlike many other Western countries, Sweden was initially reluctant to establish a national counter-terrorist strike force. Such initiatives only gained traction after the Palme killing, with the neologism *terroriststyrka* ('terrorist force', approx. 'counter-terrorist force') that emerged in 1989, when Parliament approved the creation of a national counter-terrorist unit. The rise of a more distinct counter-terrorism discourse can also be traced through neologisms referring to professionalized forms of expertise, such as *terrorexpert* ('terror expert') appearing in 2008.

Overall, all these neologisms, and particularly these later compounds, reflect an emerging 'terrormindedness', an integrated perception and practice that treats and domesticates terrorism as an

ever-present threat (Fridlund 2011). While the compound *terrorhot* ('terror threat') was introduced as early as 1974, indicating an awareness of terrorism as a potential threat, a number of neologisms in the 2010s indicate the development of a more sophisticated terrormindedness. These include *terrorhotbedömning*('terror threat assessment') in 2011, *terrorhotnivå* ('terror threat level') and *terrorbekämpningsförmåga* ('counter-terror response capability') in 2016.

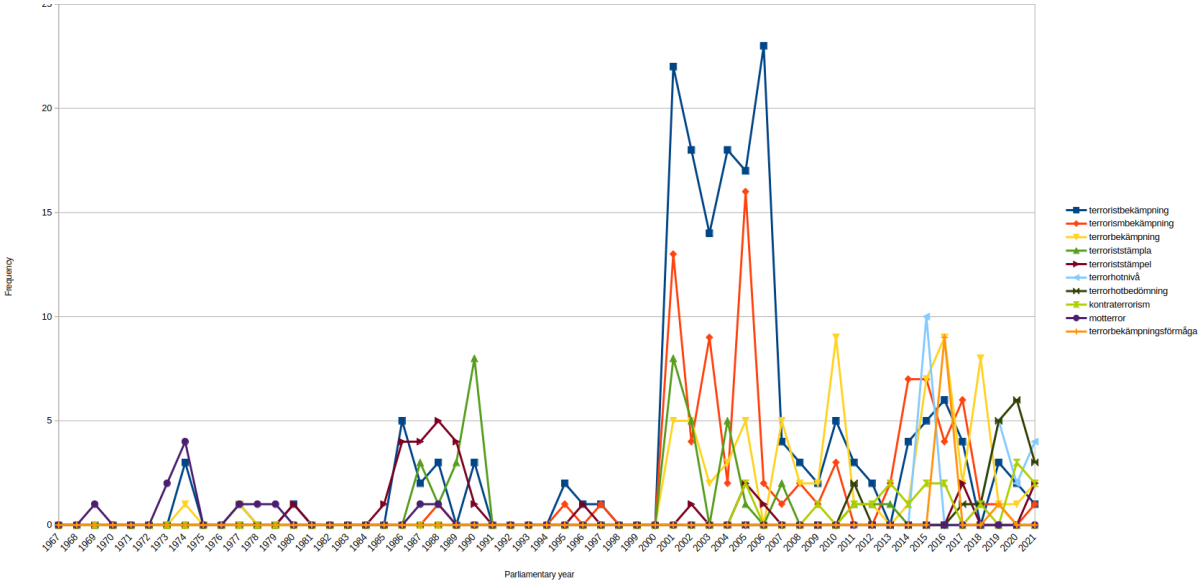


Figure 6: The frequency of the compounds related to counter-terrorism practices in Swedish parliamentary debate 1968–2021 (lemmas with more than 10 occurrences in the corpus) (Brodén et al. 2023). Note that the SweTerror project enact a contextualizing understanding of the parliamentary data by grouping debates by parliamentary year (autumn–summer) instead of calendar year, aligning with the Riksdag's mandate period.

6. Conclusions

In this paper, we have explored the analytical potential of compound word formation in Swedish when text mining parliamentary discourse, focusing on two case studies concerning the transformations of the concepts of the market and terrorism. By highlighting how both concepts have expanded over time to encompass different meanings and topical issues, we have proposed that compounds and neologisms can serve as indicators of discursive change. The results from the case study on compound productivity related to the concept of the market reveal significant variation in the usage of compound neologisms within political and legislative discourse over time, showing how these compounds introduce new perspectives and phenomena such as the emerging labor market. The second case study demonstrates how compound neologisms involving the words terror and terrorism as elements emerge during specific periods in Swedish parliamentary discourse, reflecting broader discursive transformations and the rise of the modern conception of terrorism.

We have argued that the morphological feature of forming compound words with two or more elements in Swedish and several other Germanic languages opens an analytical window for tracing discursive transformations. Similar to other areas, political language is adaptive, responding to new political, social and cultural contexts as well as to shifts in public discourse.

While our focus on tracing compound word formation may seem somewhat 'blunt' – as word formations alone cannot fully capture the complexities of discursive transformations – and methodologically 'unsophisticated' from a methodological DH perspective, relying primarily on common statistical measurements, it nonetheless offers an elemental approach to exploring discursive patterns. As we have shown, compound neologisms can reflect the emergence of new policy issues or broader societal concerns. Although we have not pursued this analytical avenue, it seems plausible that tracing these linguistic transformations could, to some extent, facilitate engagement with the

relationships between parliamentary discourse and public discourse as represented in, for instance, newspaper or other media sources.

While it would be overstating our case to claim that compound neologisms provide a 'comprehensive' reflection of discursive change, this method suggests promising ways for engaging with parliamentary datasets and other large corpora in languages with similar morphological traits. Furthermore, our approach underscores the potential of leveraging features of less dominant languages than English in DH research. We hope this paper serves as both inspiration and a call for further empirical and theoretical investigations into similar materials and methodologies.

Acknowledgments

The SweTerror project (https://sweterror.se) is funded by the research program DIGARV (supported by the Swedish Research Council, Riksbankens Jubileumsfond, and the Royal Swedish Academy of Letters, History and Antiquities). The work presented in this paper also ties into the national research infrastructures funded by the Swedish Research Council, Huminfra (contract no. 2021-00176), and Swe-Clarin and the National Language Bank (contract no. 2017-00626). The authors from both projects would like to express their gratitude to the Marcus and Amalia Wallenberg foundation for funding. We would also like to thank Victor Wåhlstrand Skärström who contributed to both projects in his previous role as research engineer at the Gothenburg Research Infrastructure in Digital Humanities (GRIDH).

References

Baker, Paul. 2011. "Times may change, but we will always have money: diachronic variation in recent British English." *Journal of English Linguistics*, 39:1, 65–88.

Bauer, Laurie. 2019. "Compounds and multi-word expressions in English." In B. Schlücker (ed.). *Complex Lexical Units. Compounds and Multi-Word Expressions*. Berlin/Boston: de Gruyter.

Brodén, Daniel, Mats Fridlund, Leif-Jöran Olsson, Magnus Ängsal and Patrik Öhberg. 2023. "The Diachrony of the New Political Terrorism: Tracing Neologisms and Frequencies of Terror-related Terms in Swedish Parliamentary Data 1971–2018." In *DHNB 2022: Proceedings*, CEUR-WS.

Edlund, Jens, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus P. Ängsal and Patrik Öhberg. 2022. "A multimodal digital humanities study of terrorism in Swedish politics: an interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018." In K Arai (ed), *Intelligent Systems and Applications: IntelliSys 2021. Lecture Notes in Networks and Systems* (Vol. 295). Cham: Springer.

Finkbeiner, Rita and Barbara Schlücker. 2019. "Compounds and multi-word expressions in the languages of Europe." In B Schlücker (ed.). *Complex Lexical Units. Compounds and Multi-Word Expressions*. Berlin/Boston: de Gruyter.

Frankel, Christian. 2018. "The 's' in markets: mundane market concepts and how to know a (strawberry) market." *Journal of Cultural Economy*, 11:5, 458-475, DOI:10.1080/17530350.2018.1502677.

Fridlund, Mats. 2011. "Buckets, Bollards and Bombs: Towards Subject Histories of Technologies and Terrors." *History and Technology.* 27:4, 391–416.

Fridlund, Mats, Daniel Brodén and Victor Wåhlstrand Skärström. 2022. "The diachrony of political terror: Tracing terror and terrorism in Swedish parliamentary data 1867–1970." In E Volodina, D Dannélls, A Berdicevskis, M Forsberg & S Virk (eds). *Live and learn: Festschrift in honor of*

Lars Borin, Gothenburg: Research Reports from the Department of Swedish, Multilingualism, Language Technology.

Fridlund, Mats, Leif-Jöran Olsson, Daniel Brodén, Lars Borin. 2020. "Trawling the Gulf of Bothnia of News: A Big Data Analysis of the Emergence of Terrorism in Swedish and Finnish Newspapers, 1780–1926." *Proceedings of CLARIN Annual Conference 2020*, CLARIN, 61–65.

Hornscheidt, Antje. 2008. "A concrete research agenda for critical lexicographic research within critical discourse studies: an investigation into racism/colonialism in monolingual Danish, German, and Swedish dictionaries." *Critical Discourse Studies*, 5:2, 107-132.

Jarlbrink, Johan, Fredrik Norén and Robin Saberi. 2022. "Contextual modeling of 'propaganda', 'information' and 'upplysning' in Swedish parliamentary speeches, 1920–2019." In *Digital parliamentary data in action* (DiPaDA 2022) workshop, Uppsala University, Sweden, March 15, 2022.

Koteyko, Nelya, Thelwall, Mike, and Nerlich, Brigitte. 2010. "From carbon markets to carbon morality: Creative compounds as framing devices in online discourses on climate change mitigation." *Science Communication*, 32:1, 25-54.

Kurunmäki, Jussi and Jani Marjanen. 2018. "Begreppshistoria" In K Boréus and G Bergström (eds.), *Textens mening och makt: Metodbok i samhällsvetenskaplig text- och diskursanalys*, Studentlitteratur, 2018.

Lyse, Gunn Inger and Gisle Andersen. 2012. "Collocations and statistical analysis of n-grams." *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, 49, 79.

Norén, Fredrik, Johan Jarlbrink, Alexandra Borg, Erik Edoff and Måns Magnusson. 2022. "The transformation of 'the political' in post-war Sweden." In E Bunout, M Ehrman & F Clavert (eds). *A new Eldorado for historians? Reflections on tools, methods and epistemology*. De Gruyter.

Ohlsson, Claes. 2020. "Market Language Over Time. Combining Corpus Linguistics and Historical Discourse Analysis in a Study of Market in Swedish Press Texts." In: J. Hansson & J Svensson (eds.). *Doing Digital Humanities: Concepts, Approaches, Cases*. Växjö: Linnæus University Press. 199–218.

Ohlsson, Claes, Victor Wåhlstrand Skärström and Henrik Björck. 2022. "The market as a concept in Swedish parliamentary records from 1867 to 1970: A mixed methods study." In *Digital Parliamentary Data in Action* (DiPaDA 2022) workshop, Uppsala University, Sweden, March 15, 2022.

Postiglione, Alberto. 2024. "Finite State Automata on Multi-Word Units for Efficient Text-Mining". Mathematics, 12, 506. https://doi.org/ 10.3390/math12040506.

Stampnitzky, Lisa. 2013. *Discipling terror: How experts invented 'terrorism'*, Cambridge: Cambridge University Press.

Virk, Shafqat M., Claes Ohlsson, Nina Tahmasebi, Henrik Björck and Leif Runefelt. 2024. "Enhancing Swedish Parliamentary Data: Annotation, Accessibility, and Application in Digital Humanities". *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*. 280–288.

Yrjänäinen, Väinö, Fredrik Mohammadi Norén, Robert Borges, Johan Jarlbrink, Lotta Åberg Brorsson, Anders P. Olsson, Pelle Snickars and Måns Magnusson. 2024. "The Swedish Parliament Corpus 1867–2022." *LREC-COLING 2024*: 16100–16112.

Ängsal, Magnus P, Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson and Patrik Öhberg. 2024. ”Terrorism som tolkningsram: en diskurssemantisk studie av svensk riksdagsdebatt 1993–2018”. In D Jansson, I Melander, G Westberg & D Yassin Falk (eds.), *Svenskans beskrivning. Förhandlingar vid trettioåttonde sammankomsten Örebro 4–6 maj 2022,* 194–210.

Functions of English and Latin in the Parliament of Finland 1970-2020

Anna Ristilä[a]

[a] *University of Turku*

Abstract

The Finnish parliament, *eduskunta*, operates in two national languages, Finnish and Swedish. Other languages, such as English and Latin, are also present in the plenary speeches in trace amounts. This paper explores the functions English and Latin may have in *eduskunta*, and contemplates whether they are used as elite closure, and if so, in what contexts. The questions are approached with quantitative and qualitative methods: first the use cases of English and Latin in the plenary speeches of the Finnish parliament between 1970 and 2020 are quantified, then ten possible intended functions for English and Latin use are identified by close reading the immediate contexts of each use case. Finally, the distributions of the use cases and their functions along topics are examined, and the possibility that English and Latin could be used as tools of elite closure is discussed.

Keywords

English, Latin, *eduskunta*, parliament of Finland, multilingualism, elite closure

1. Introduction

Different languages and language varieties used in a society have different amounts of political capital. The languages/varieties with the most political capital are those used in government, business, and in higher education (Myers-Scotton, 2002: 35). If such varieties are only available to some people, it can practically exclude people without such linguistic skills from positions of power. For the elite who do have the skills it can be an advantageous strategy to maintain the linguistic differences; this type of strategy has been named *elite closure* by Myers-Scotton (1993; 1997; 2002).

Latin used to be a powerful tool for elite closure during its time as the lingua franca of the learned Europe. Today Latin is practically a dead language since no-one speaks it as their mother tongue, but it has some active L2 speakers (Engelsing, 2017; Coffee, 2012). There are few studies concerning contemporary uses of Latin, but the main areas where Latin is still used are science (Roelli, 2021), law (Gałuskina & Sycz, 2013), Catholic Christianity, and the occult (Banner, 2021; Piętka, 2016). In the modern world English has taken more than just the place of Latin; it has become the closest to a true global lingua franca than any language before. It also seeps widely into other languages through borrowing and codeswitching phenomena (e.g. Furiassi et al., 2012). However, even this new lingua franca is still often the language of the learned and the elite, albeit much less exclusively so than what Latin used to be. For example, in many African and Oceanian countries and in India English is the official language despite not being the majority language. Also, varieties containing English codeswitching have been reported as acts of upper-class identity performance (e.g. Sanei, 2022; Jahan & Hamid, 2019; Matus-Mendoza, 2002).

In Finland there are two national languages, Finnish and Swedish. Finnish is the majority language, while Swedish speaking Finns are a historical minority currently estimated at 5.6% of the population (Statistics Finland 11rm, 2023) and with their own political party, the Swedish People's Party of

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, Reykjavik, Iceland, May 28, 2024.
EMAIL: ankrris@utu.fi
ORCID: 0000-0003-0236-5401

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

Finland. English has no official status and only 0.7% of the population state English as their native language (Statistics Finland 11rm, 2023). However, the use of English in the Finnish society has steadily increased in recent decades and English functions in many ways as a lingua franca (Peterson, 2022; European Commission, 2012). In a 2024 EU Barometer survey 81% of Finnish respondents claimed they could have a conversation in English, which is an 11%-unit increase from the previous survey in 2012 (EC, 2024; EC, 2012). The number will probably keep increasing as English continues to spread outside the traditional contexts of media, business and education (Leppänen & Nikula, 2007).

The Finnish parliament, *eduskunta*, operates in the two national languages, Finnish and Swedish. However, many languages are still present in parliamentary speech in small quantities, often as rhetorical devices. In this paper the occurrences and functions of English and Latin, the current and former *lingua francae*, in the Finnish plenary speeches between 1970 and 2020 are studied. The results will be interpreted from the point of view of a prestige language as political capital and as possible elite closure. The research questions are: 1) What functions do English and Latin play in *eduskunta*? 2) Are they used as elite closure and if so, in what contexts?

Studying how much, with what intentions and in which topics the current and former lingua franca languages are used in the parliament may help us better assess the process of how English is spreading and how it contributes to elite closure. This study will also give a better understanding of English and Latin use specifically in a political context, and of how and why these languages can be used as political tools and devices. As far as the author's knowledge, linguistic elite closure has not been studied with quantitative methods from parliamentary speeches before.

2. Background

2.1. Elite closure

Elite closure denotes "a type of social mobilization strategy by which those persons in power establish or maintain their powers and privileges via linguistic choices" (Myers-Scotton, 1993: 149). There are two forms of elite closure, strong and weak. The strong form is typically exercised through official language policies and very effectively excludes non-speakers of a language from political access; it is typically present in countries where universal education of the elite variety is not available (Myers-Scotton 2002: 35). This is the case e.g. in many African countries and India (Norro, 2022; Hickey, 2019; Bedi 2019). The weak form is more subtle, informal, and common since it is expressed through linguistic repertoires (e.g. sociolects) and use patterns (e.g. code switching) that can potentially be learned by anyone through universal education (Myers-Scotton, 1993:151-152). Weak elite closure is more or less present everywhere, especially in the western countries (*ibid*.), but it is more difficult to address because of its subtlety and links to identity building.

One way to study weak elite closure is to examine speech patterns, such as foreign language use, in contexts where elite identity performance is expected, such as in the parliament. Parliaments are public stages where elite identity performance can help legitimize and elevate the speaker's authority not just in front of their peers but also of the media and the public. On the other hand, parliaments are formal institutions with specific norms, rules, and expectations that shape how identity can be performed. Identity performance in a parliamentary context is very strategic, as the speaker must conform to institutional expectations while still distinguishing themselves for political purposes. Elite identity performance through speech patterns can impact how the speaker is perceived in terms of credibility, leadership, and trustworthiness, but at the same time it may work as a tool of exclusion.

2.2. Foreign language use cases

There are many kinds of linguistic use patterns that could be mapped to study elite closure. Many previous studies have focused strictly on code switching. The focus of this article is more widely on (extra- and intra-sentential) code switches, loanwords, quotes, or any kind of embedded language use clearly distinguished from the matrix language, or the language that provides most of the semantic and morphosyntactic procedures in a bi- or multilingual setting (Myers-Scotton, 1997[1993]: 75). These will be collectively referred to as *embedded language use cases*.

Recognizing what counts as embedded language is not always straightforward. Quotes are easy to recognize, most code switches as well, but loanwords can be more or less integrated into the borrowing language. A helpful classification is to see how fully they have been translated (Haugen, 1950: 214-215): fully, partially, or not at all (loanwords proper). Partial and untranslated loans were evaluated on a case-by-case basis and counted if they clearly expressed 'foreignness' in the context in question. The feature for identifying 'foreignness' was orthography and/or pronunciation: a loanword that has not been adapted and still retains the original orthography has been defined as a foreignism (Haspelmath & Hapdor, 2009: 42) (e.g. *high-tech*, compare to e.g. *skuutti*, "electric/kick scooter"). Foreignisms are close to single-word code switches but are more established (*ibid*.: 43), which means they are easily omitted from studies that focus solely on code switches.

2.3. Functions of use cases

Every embedded language use case has some function – or multiple functions – that it performs in the context it appears in. Because the parliament is a venue for extremely strategic language use, the functions identified in this paper have been formulated from the point of view of the speaker's probable *intention*, either conscious or subconscious, rather than from a general usefulness perspective. In other words, the functions should always answer the question "what seems most likely to be the speaker's intended outcome for this use case" instead of a more general "why is the use case here". Many sources that list functions for any type of embedded language use have not always formulated the functions with (conscious or subconscious) intention as the centre. For example, Hockett (1958 [1964]: 404-405) gives two well-known motives for borrowing: *prestige* and *need-filling*. If rephrased with intention, they would be something akin to *to sound more profound* and *to fill a lexical gap*. Gumperz (1982:75-80) lists five conversational functions for code switching: *quotations*, *addressee specification*, *reiteration*, *message qualification*, and *personalization versus objectivization*, some of which lack an explicit intention part. For example, Gumperz's *addressee specification* does include an explicit intention: the speaker intends to focus their message to a certain addressee, so they use an embedded language for the sake of that focus. But as for Gumperz's *quotations*, which could be either direct quotes or reported speech, the definition as-is does not reveal any specific purpose for embedded language use. A probable intended function for any quote or reported speech could be *clarification*, another *appealing to authority*. Romaine (1995 [1989]: 161-164) elaborates on Gumperz with more attention on intention and adds especially the concepts of topic and comment to explain what Gumperz called *message qualification*. Gumperz's categories with Romaine's commentary were used as a starting point when identifying functions of embedded language use cases in the parliamentary context. The process of defining the function categories is explained in the Materials and methods section, and the final categorization is introduced in the Results section.

3. Materials and methods

All the plenary speeches in the Finnish parliament since 1907 have been collected and made available online and in computer readable format in the Semantic Parliament project (Parlamenttisampo, Hyvönen et al., 2024). The dataset used in this paper consists of years 1970-2020 and includes the actual speeches – born-digital after 2010, converted from originals with optical character recognition before that – and metadata which contains e.g. date and speaker information. The used subset included approximately 650'000 speeches and over 100 million words. English and Latin use cases were detected from the subset, and the most probable intended functions were defined for each detected use case.

3.1. Language use case detection

In order to analyse the embedded language use cases, they first had to be identified. Both automatic detection and manual verification was used for the detection of English and Latin, but in different ways. The automated tool used for English detection was a Python library called Lingua (Lingua 2.0.2, 2023). Since the embedded language could be either Finnish (ca. 97% of the corpus) or Swedish, a separate

Swedish sentence corpus was created along the way. First, speeches were iterated one sentence at a time through a language detector object that could differentiate between Finnish, Swedish and English. If the confidence value of a sentence for Swedish was over 0.5 it was accepted as Swedish. Then the collected sentences were manually browsed through and the few ones that were clearly not Swedish were removed. This resulted in 8695 Swedish sentences.

Since English appeared typically in smaller units than sentences, to automatically catch as many use cases as possible a two-word sliding window was used to feed word pairs to Lingua's language detector object. Again, the differentiation was made between Finnish, Swedish and English, but this time if the language with the highest confidence value was English, the word pair was accepted as English. This resulted in approximately 1000-2000 hits per year which were browsed through manually. Only about 3% of these were considered as truly English use cases and were picked out separately (N=2337). Typical false positives from Lingua were either Swedish language or contained inflected versions of Finnish words *komissio* (“commission”) or *oppositio* (“opposition”). Setting a detection probability threshold for English in Lingua could have improved the low percentage of true positives and sped up the process, but any threshold would have potentially increased false negatives. Since English use cases were dispersed very sparsely, it was not possible to reliably control for false negatives. Because of this, and because it was considered important to catch as many use cases as possible, the manual reading was seen as necessary to ensure accuracy.

In case of Latin the process was very different due to automatic language detection not working well with Latin. A list of 181 common Latin phrases was compiled and these phrases were searched with a simple regex search with Python (see Appendix 1 for the list of searched and found Latin phrases). Only 59 of these phrases appeared in the corpus at least once and all found instances (N=795) were collected. Since the list of searched phrases was finite, it is possible that some Latin use cases were not caught, but again due to extreme sparsity it was not possible to approximate how many.

3.2. Topic modeling

The topics present in the corpus were defined with an LDA (latent Dirichlet allocation) topic model created in a previous study (Ristilä & Elo, 2023). Since the details of the model creation have been described in the aforementioned article, only the method and the main characteristics of the model are explained here.

Topic modeling is an unsupervised machine learning method that uses corpus vocabulary to define 1) what topics are present, 2) how probable each topic is in a text passage, and 3) how probably each word is related to any of the topics. Simply put, the model sees all topics in any given text, just in different proportions. The topics provided by the model are abstract collections of words that computationally "go together" and a human needs to interpret what the topic is actually about. For example, one of the topics defined was interpreted as *education* as it was strongly associated with words such as "university", "education" and "school".

The original topic model was trained with the same parliamentary speeches as used in this study but only from years 1980-2010. A longer time span was not modeled because creating a model from a large dataset is very time consuming and the supercomputer used for the training had rather strict time constrictions. Also, since the topics of the chosen model were very broad, they were considered to be similarly present also in the preceding and succeeding decades of 1970s and 2010s in a way that the model would still provide reliable results for texts from these decades. The model defined 26 topics: *administration, agriculture, budget, commerce*, *crime, democracy, development cooperation, education, employment, energy, foreign and security policy, general, housing, legislation, law proposals, parliamentary factions, pensions, public sector, question time, regionality, social and health care, social benefits*, *social problems*, *taxation*, *traffic and transport*, and *voting*.

In order to accurately define the topics where an embedded language use case appeared, some amount of text from around every use case had to be fed into the topic model. The topics were pre-defined for entire speeches, but since speech lengths varied a lot (from single word remarks to multi-page monologues), a five-sentence passage (later "context passages") – the sentence where the use case appeared, plus two sentences from both sides of the use case sentence – was decided to be enough context to represent the immediate surroundings of any use case. If the use case appeared in the

beginning or end of a speech, five sentences worth of immediate context was included. Since the topic model was only trained in Finnish, any Swedish sentences were first translated with the European Commission's online machine translation tool (eTranslation v13.3, 2024, with "General Text" as the setting) and lemmatized with Turku Neural Parser Pipeline (Kanerva et al., 2018), before being processed through the monolingual topic model. Machine translation has been shown to be a viable method to enable use of multilingual texts with topic models (e.g. Maier et al., 2022).

Both machine translation and lemmatization were affected by errors in optical character recognition of the original paper documents. Especially older documents had suffered from low quality paper (data provided by the Semantic Parliament project). This resulted in vocabulary mistakes that could potentially affect the topic model. However, this was mostly the case before 2011, after which the data was digital-born and significantly less mistakes were noticed.

3.3. Defining functions

Only the most probable intended function was decided for each identified embedded language use case of English (N=2337) and Latin (N=795). This was done by close reading the context passages surrounding the use cases and considering what the speaker probably wanted to signal with the language choice. The focus was on the embedded language vocabulary and the immediate context, but metadata (e.g. speaker, party affiliation, year) was also taken into consideration when applicable. Since no existing categorization was fully compatible with the focus on intention, Gumperz's system (described in Gumperz 1982 and Romaine 1995) was used as a basis for a new one that was constructed by close reading and suggesting categories that would suit the context of the dataset. Assigning the categories for each use case was a semi-iterative process. If a use case could be categorized with an existing label (Gumperz's original categories or modified with Romaine's commentary), it was labeled as such, otherwise a new label would be thought of. Whenever at a later stage a use case was found not to fit in any of the previous labels, the categorization system was revised and either a new label was added or a previously used label was renamed, in which case the use cases with the redundant labels were re-evaluated and re-labeled.

It should be underlined that the functions represent the author's *interpretation* of the most probable functions intended by the speakers in the context of the parliament, not actual functions. It would not be possible to be certain of the actual intended functions, as they may be subconscious as well as conscious and even interviewing the speakers would not give fully reliable answers. Any listener's interpretation would also be different, subjective, and perhaps even more difficult to average out. Furthermore, usually multiple parallel functions could have been designated for any given use case, but only the most probable ones were marked, which was likely to oversimplify the situation. Lastly, though suitable category labels were discussed with other researchers, since the author was the sole annotator the resulting categorization of the data is still highly subjective. Planned subsequent studies with similar data will use multiple annotators.

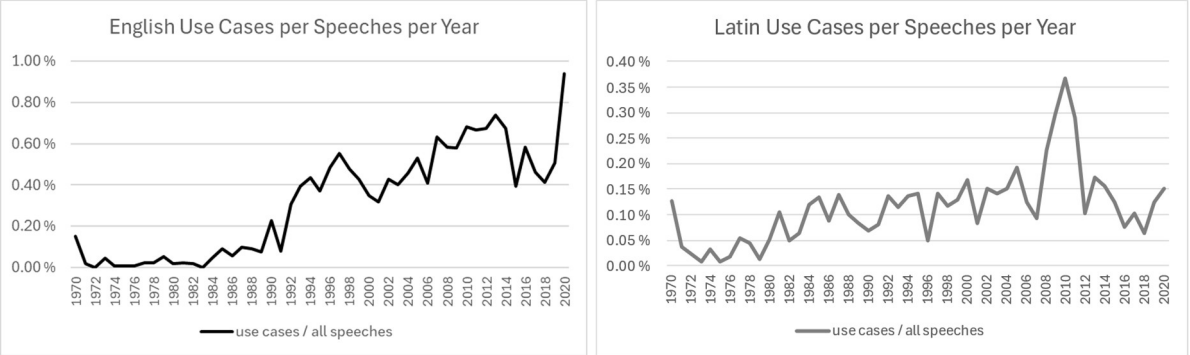
4. Results

After identifying the use cases of English and Latin, interpreting their functions, and running the five-sentence context passages through the topic model, distributions of the use cases, functions and topics were examined. In the following subsections the main findings will be presented: First, the use case frequencies of English and Latin; Second, the topic distributions around English and Latin use cases; Third, the function definitions and the distributions of the functions for each language; Fourth, the topic distributions around different functions.

4.1. Use case frequencies

English was used much more than Latin, which was unsurprising. Yearly differences can be seen from figures 1a and 1b. The most striking points for English are the steep increase in use around the 1990s, when Finland joined the EU (1995) and the Internet became more easily available to the public. The

impact of EU may be seen I the use of phrases such as “moral hazard”, “peace making”, and “joint implementation”, whereas the impact of the Internet may show in phrases such as “high-tech” and “free net”. For Latin, 2010 is a clear peak year which an be explained with heated discussions about *de minimis* EU aid.



Figures 1 a and b. English and Latin use cases per number of speeches per year. Notice the difference in scale between y axes.

Table 1.
The most used English phrases in select peak years.

1993	1997	2010	2013	2020
moral hazard (9)	high tech (14)	no bail out (12)	cleantech (21)	take away (36)
high tech (7)	peace making (4)	must carry (7)	pooling and sharing (10)	no bail-out (14)
free net (4)	working poor (3)	not in my backyard (4)	private label (6)	carry back (6)
fifty-fifty (3)	masking factor (3)	so what (3)	total return swap (4)	label (5)
we cannot safely leave politics to politicians (2)	joint implementation (3)	fresh start (3)	no bail-out (3)	end of waste (4)

Table 2.
The most used Latin phrases in select peak years.

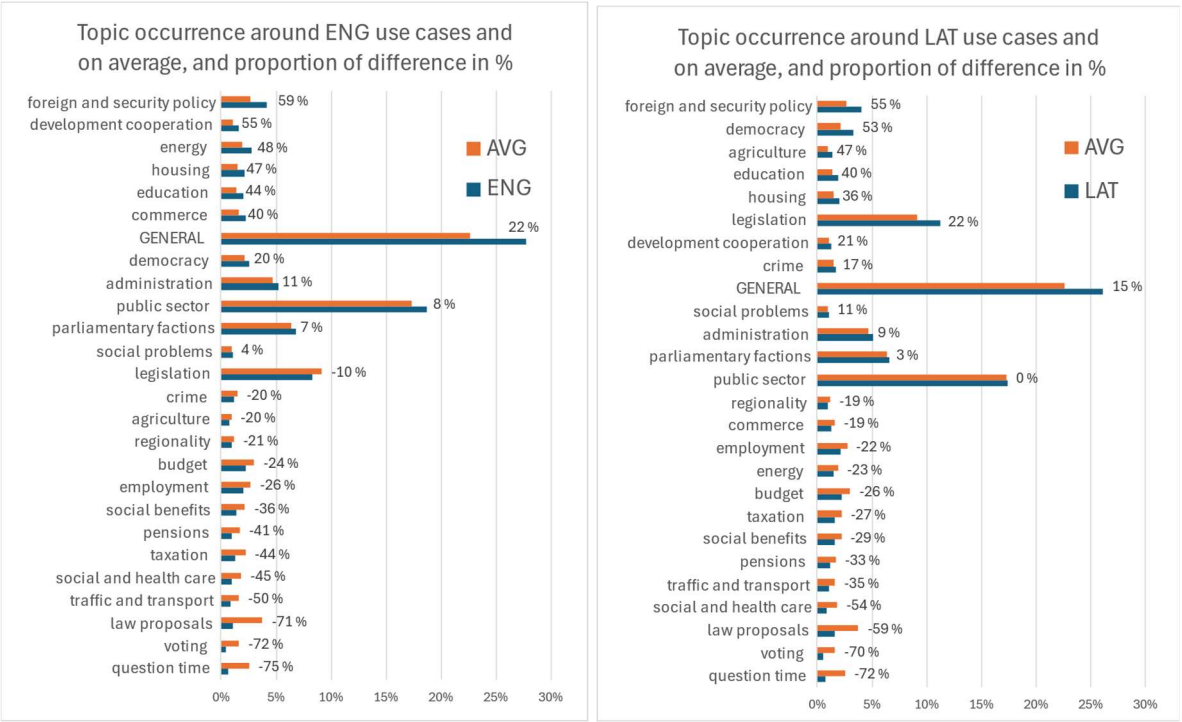
1993	1997	2010	2013	2020
de facto (7)	de facto (13)	de minimis (31)	ne bis in idem (9)	de facto (13)
sui generis (4)	bona fide (3)	de facto (12)	de facto (7)	de minimis (5)
in dubio pro reo (4)	vestigial terrent (1)	lex in casu (4)	status quo (2)	sui generis (1)
bona fide (1)	nomen est omen (1)	ad hoc (4)	de minimis (2)	ad hoc (1)
ad hoc (1)	ad hoc (1)	bona fide (2)	ad hoc (1)	ceterum censeo (1)

Making clean frequency lists of the found English phrases was difficult because of various forms the phrases could take: the words could appear with (English or) Finnish inflection (e.g. *take awayn*, *take awaysta*), be used as compound words (e.g. *fifty-fifty-periaatteella*), and also exhibit variation in spelling (e.g. *know how*, *know-how*, *knowhow*). Removing variations is not an easy task, so it was done only for five peak years of 1993, 1997, 2010, 2013 and 2020. The top five most frequent use cases for each peak ear are shown in table 1, and similar frequencies for Latin are given in table 2.

To give an idea of typical English phrases found without addressing variation, the ten most frequent English phrases with all different forms left as-is are listed: *cleantech* (43), *cleantechin* (26), *fifty-fifty* (22), *so what* (14), *high techin* (14), *know-how* (11), *masking factor* (10), *high tech* (9), *fair play* (9), *deadline* (9). The total frequencies of found Latin phrases can be found in Appendix 1, the most frequent overall being: *de facto* (329), *de minimis* (130), *ad hoc* (52), *ex tempore* (32), *status quo* (27), *ne bis in idem* (25), *sui generis* (20), *bona fide* (14), *nomen est omen* (9), *pacta sunt servanda* (9).

4.2. Topics around languages

After defining the topic distributions around every use case, the distributions were summed together per language to form two comprehensive distributions that would represent the English (ENG) and Latin (LAT) "subcorpora". These distributions were then compared to the average topic distribution of the full Finnish-language corpus of years 1970-2020. In figures 1a and 1b the sub and full corpora distributions and their proportional differences are given. Positive difference means the topic is more likely to appear around English or Latin than in average text, negative difference means the topics is less likely to appear around the studied languages than in average text. For example, the topic *energy* is proportionally almost 50% more likely to appear around English use cases than it would appear on average, though its average prevalence is only around 2%.



Figures 2a and 2b: Topic distributions around English and Latin use cases and on average. The x axis shows percentages for the orange and blue topic bars, the numbers after each bar pair are the differences between the bars (ENG or LAT and AVG) as proportions from the average (AVG).

The topics that mostly preferred English were related to international affairs: *foreign and security policy* (59%) and *development cooperation* (55%). Other topics that also clearly preferred English (>40%) were all fundamental pillars of societal infrastructure and development: *energy*, *housing*, *education* and *commerce*. The topics that preferred Latin the most were probably connected through juridical discussions regarding them: *foreign and security policy* (55%) and *democracy* (53%). *Agriculture* (47%) appeared strongly with Latin due to one extremely frequently used term, *de minimis*, referring to certain type of EU aid. Three topics, *agriculture*, *legislation* and *crime*, that avoided English, preferred Latin. On the other hand, topics *energy* and *commerce* which preferred English avoided Latin. The three topics that strongly avoided both English and Latin represented quite formal

contexts: *question time* (ENG -75%, LAT -72%), *voting* (ENG -72%, LAT -70%), and *law proposals* (ENG -71%, LAT -59%).

4.3. Functions of languages

For both English and Latin ten functions were identified during the close reading of the use cases and their surrounding context passages. These were divided into two categories: practical functions, which represent intention to communicate efficiently and in a way that most people can understand, and rhetorical functions, which represent intention to accent and enhance expression. It was assumed that practical functions did not take as much part in elite closure as the rhetorical ones, since a willingness to be practical does not logically match well with a willingness to differentiate from others, while intentional linguistic (self-)enhancement can be more easily interpreted as a willingness to differentiate and exclude. Explanations and examples of functions are given below but, again, it is important to remember that any given use case usually employs several functions at once.

Practical functions included four functions:

- TRAN; to discuss **translations** or meanings of foreign words.
  Example: *Oli muuten vaikea löytää sopivia suomenkielisiä vastineita sellaisille sanoille kuin ruotsin "*lagarbete*" ja englannin "*team work*"*.
  "Let me tell you, it was difficult to find suitable Finnish equivalents for words such as 'lagarbete' of Swedish and 'team work' of English."
- CLAR; to **clarify**, often hinting that the English or Latin version is better known or at least may help better understand the concept or phenomenon.
  Example: *Tätä kutsutaan tieteessä ja ekologiassa* masking factor *-ilmiöksi*.
  "In science and ecology this is called the 'masking factor' phenomenon."
- ESTA; **established** use, often no Finnish equivalent in common use.
  Example words: *skinhead*, *hightech*, *cleantech*.
- EFFI; to be **efficient**, when the concept in English is shorter or quicker to say than the Finnish equivalent.
  Example: *Siitä tehdään parhaillaan virkamiesten* think tank *–työskentelyä*.
  "Civil servants are currently doing 'think tank' work on it."
  (Finnish equivalent for 'think tank' is *ajatushautomo*, which is six syllables, while the English term is only two syllables.)

Rhetorical functions included six functions, as shown below:

- ATTI; to express **attitude** or emotion.
  Example: *…tutkaillaan, sopiiko se sen rahoituksen piiriin,* all right*, ei sovi, nakataan jälleen seuraavaan…*
  "…let's see if it is suitable for this funding, 'all right', no it isn't, let's throw it to the next one…"
  (Here the English part is taken as an expression of 'someone doesn't care', or 'someone shows disregard', so this use case also functions as DIST; see below.)
- EMPH; to **emphasize** the contents of the message.
  Example: *Kun hän on hyvässä kunnossa,* in good order*, kun hän on eläväinen, henkisesti valveutunut ja ruumiillisesti hyvässä kunnossa…*
  "When they are in good condition, 'in good order', when they are lively, spiritually enlightened and physically in good condition…"
- HUMO; to add **humour**.
  Example: *USA:ssa on Bill Clinton, Bob Hope ja Johnny Cash, Suomessa taas Esko Aho,* no hope and no cash.
  "In the USA there is Bill Clinton, Bob Hope and Johnny Cash, in Finland there is Esko Aho, 'no hope and no cash'."
  (It should be noted that it is irrelevant whether the audience thinks of the use cases as funny or not, just that the speaker intends their utterance to be funny.)

- SELF; to **self-emphasize**, attempting to emphasize personal linguistic and cultural knowledge, to be trendy or to show off.
  Example: *Think globally, act locally*.
- AUTH; to appeal to **authority**, either by quoting/paraphrasing a person or by using a proverb or saying.
  Example: *...englantilaiset ovat useissa yhteyksissä toistaneet vanhaa brittiläistä viisautta ″*right or wrong - my country*″*…
  "The English have in many contexts repeated an old British wisdom 'right or wrong – my country'".
- DIST; to **distance** oneself from the topic or reported speaker, usually in negative contexts.
  Example: *...kun matkustivat Kiinaan, ilmoittivat, että* business is business and human rights are human rights…
  "When they travelled to China, they stated that 'business is business and human rights are human rights'".

The time difference between the original utterances and the time this paper is written raises some small issues, since the examined period reaches more than 50 years into the past from the time of writing. Many foreign language elements may have become practical necessities over time through original unwillingness to find or use suitable domestic language variants due to elite closure. Diachronic changes in word uses are also constantly taking place, meaning that an annotator making an interpretation about a use case today might have made a completely different interpretation 50 years ago. It was not possible to address these issues in detail, but they have been considered when interpreting especially the older speeches.

Figures 3a and 4a show that English was used mostly for practical functions and that there has been a clear rise in the use of English since the 1990s, which is the decade when Finland joined the EU and when internet connections started to become common in households. There is also a steep spike in 2020, with "take away" being the most popular phrase during that time (i.e. take-away food portions during the Corona pandemic, see example (a)) but not enough to fully explain the spike. English's functions spread more evenly than Latin's (2a), but still concentrated more on EFFI, CLAR, ESTA and SELF. The most used function for English was EFFI, but CLAR was also surprisingly prominent, hinting that English is often seen as the better-known version for many concepts and thus given as a clarifying aid.

Latin use concentrated strongly on two functions, EFFI and SELF (Figure 3b). *De facto* was the most prominent Latin phrase overall (329 use cases) and was typically categorized as either EFFI or SELF. Both functions almost always overlapped, which is demonstrated in examples (b) and (c). The first example (b) was categorized as SELF, on the grounds that though *de facto* has probably replaced the Swedish word *faktiskt* making the sentence that much easier to pronounce (EFFI), the sentence overall is quite long and "inefficient", which supports the interpretation that Latin was used more for the sake of raising the register and consequently the status of the speaker. In example (c) *de facto* was used the "correct" way as an antonym to *de jure*, both of which could be said in Finnish terms but using the well-known Latin terms is more accurate, quicker, and thus more efficient (EFFI).

(a): *Ravintolat pystyvät jossain määrin sopeutumaan tilanteeseen* **take away** *-myynnillä ja kotiintoimituksilla. Kotiintoimitukset eivät kuitenkaan ole mahdollisia alkoholijuomien osalta (...)*

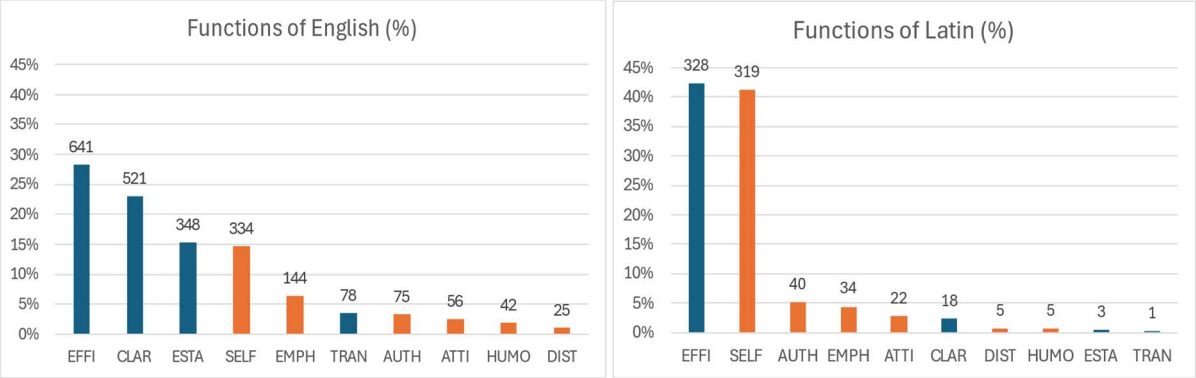
" Restaurants can to some extent adapt to the situation with **take away** sales and home deliveries. However, home deliveries are not possible for alcoholic beverages (...)"

(b) *Fler och fler konsumenter har dock sakta men säkert börjat ifrågasätta och fråga sig vad maten som finns att köpa i dagligvaruhandeln innehåller, om den* **de facto** *är så trygg, hälsosam och välsmakande som det sags*.

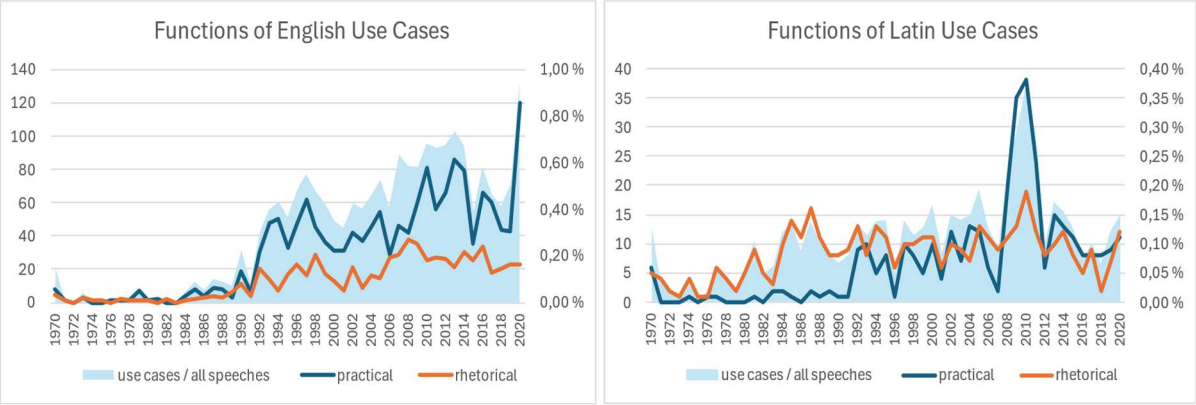
"However, more and more consumers have slowly but surely begun to question and wonder what the food available for purchase in the grocery store contains, if it is ***de facto*** as safe, healthy and tasty as it is said."

(c) *Krimin niemimaa on tällä hetkellä* **de facto** *venäläisten joukkojen valvonnassa. Venäjä piirittää alueella olevia Ukrainan tukikohtia ja rajavartioasemia.*
"The Crimean peninsula is currently ***de facto*** under the control of Russian forces. Russia is besieging Ukrainian bases and border guard stations in the region."

Diachronic use of Latin has been rather steady for both practical and rhetorical functions (Figure 4b), except for the spike around 2010 which can be explained with intensive discussions about *de minimis* aid, a type of European Union funding. The small rise in Latin practical functions during the 1990s can be in part explained with the same phenomenon. It is also possible that Latin use has increased because of English use, since English texts often employ a lot of Latin.



Figures 3a & 3b: Proportions of functions of English and Latin. Practical functions are in blue, rhetorical functions in orange.



Figures 4a & 4b: Frequencies of practical and rhetorical functions of English and Latin use cases per year (blue and orange lines). Note the difference in scale between the left y axes. The light blue area shows the proportion of use cases per all speeches given every year and it uses the secondary y axis on the right.

Three different translation strategies were observed for functions: either a use case had some kind of a translation or explanation (after the use case), had no translation, or was used as a translation for a concept or idea which was first explained in Finnish or Swedish. Example (d) shows how a single word equivalent is given as a translation for an English term, and (e) exemplifies a longer explanation. Example (f) shows a use case of Latin with no translation, suggesting that the speaker either thinks everyone knows the concept or deliberately wants to speak abstrusely. Example (g) shows an English (ATTI) use case with no translation, and (h) shows how an English concept is used as a translation or explanation for a concept first given in Finnish.

(d) *Siellä sillä on aivan oma terminsä. Se on* **looting** *elikkä ryöstö, ihan puhtaasti sanottuna ryöstö.*

"There it has its own term. It's ***looting***, that is, robbery, clearly said, robbery."

(e) *Meillä on tiukemmat säännöt euroalueella, meillä on* **fiscal compact** *eli tämmöinen taloussopimus, jossa on muitakin kuin euromaita (...)*
"We have stricter rules in the euro area, we have a **fiscal compact**, i.e. an economic agreement which also includes non-euro countries (...)"

(f) *Kristillinen liitto vaatii 0,0 promillen rajaa tieliikennelakiin. Lopuksi, arvoisa puhemies, kristillisen liiton* **ceterum censeo***: Alkoholijuomat on poistettava elinkustannusindeksistä.*
"The Christian alliance demands a limit of 0.0 per mille in the road traffic law. Finally, honorable Speaker, ***ceterum censeo*** of the Christian alliance: Alcoholic beverages must be removed from the cost-of-living index."

(g) *Det finns nu ett* **window of opportunity** *till en fredlig lösning enligt det som kallas medelvägskonceptet, inte självständighet utan självstyrelse, och (...)*
"There is now a **window of opportunity** for a peaceful solution according to what is called the middle way concept, not independence but self-government, and (...)"

(h) *Erityisesti minua entisenä neuvottelijana kiinnostaa, pystytäänkö tässä poistamaan niin sanottuja näkymättömiä kaupan esteitä,* **invisible barriers to trade***, ja sitten sellaisia, jotka liittyvät tavalla tai toisella turvallisuuspolitiikkaan (...)*
"In particular, as a former negotiator, I am interested in whether it is possible to remove so-called invisible barriers to trade, **invisible barriers to trade**, and then those that are related in one way or another to security policy (...)"

The strategies were marked for each use case, and the results are given in figures 5a and 5b. Overall, Latin use cases were explained less than English use cases, though close reading revealed that rare cases were usually translated. ESTA cases typically had no translations, as expected, since established words and phrases, e.g. *high tech*, were well-known. EFFI and ATTI also usually had no translations, but probably for very different reasons; many use cases what were typically used for the sake of efficiency, such as *win-win* or *de facto*, are also usually well-known, but explaining an expression of attitude would water down the effect on those who understood the use case (example f). Similarly, HUMO in English also had little translations probably for the same reason as ATTI, as explaining a joke takes away the humour. Latin only had five cases of HUMO, three of which came with translations, which reflects the fact that Latin as a little-known language may not work well as humour. SELF and DIST had mostly no translations but were also used a little bit with and as a translation. AUTH had almost half of the cases translated in English, probably because the function included longer quotes; most Latin AUTH cases also had translations. CLAR, EMPH and TRAN were the most evenly distributed functions among the three translation strategies for both English and Latin (except for TRAN which only had one occurrence in Latin).

English				
Function	AT%	NT%	WT%	total # of use cases
ATTI	0 %	95 %	5 %	57
AUTH	4 %	49 %	47 %	75
CLAR	28 %	25 %	47 %	521
DIST	4 %	84 %	12 %	25
EMPH	20 %	58 %	22 %	144
ESTA	1 %	96 %	3 %	348
HUMO	2 %	83 %	14 %	42
EFFI	2 %	90 %	8 %	641
SELF	7 %	75 %	18 %	334
TRAN	15 %	28 %	57 %	79
total	10 %	68 %	22 %	2266

Latin				
function	AT%	NT%	WT%	total # of use cases
ATTI	0 %	95 %	5 %	22
AUTH	10 %	23 %	68 %	40
CLAR	39 %	28 %	33 %	18
DIST	0 %	80 %	20 %	5
EMPH	35 %	47 %	18 %	34
ESTA	0 %	100 %	0 %	3
HUMO	0 %	40 %	60 %	5
EFFI	0 %	92 %	7 %	328
SELF	3 %	88 %	9 %	319
TRAN	100 %	0 %	0 %	1
total	12 %	83 %	4 %	775

Figures 5a and 5b: Placement of explanations or translations around use cases of English and Latin. AT = as translation, NT = no translation, WT = with translation (or explanation).

Finally, a brief estimation of phrase occurrence rates per function for both languages was also conducted. Latin was easier to estimate because the searched phrases were predefined: EFFI use cases included dominating cases such as “de minimis” (121), “de facto” (58), and “ad hoc” (48), the other functions mostly had more than five occurrences each. SELF was dominated by “de facto” (256). English was more difficult to estimate because the use cases varied from single words to multi-sentence quotes and had varying orthographies. By browsing an alphabetically ordered list of all use cases most English functions (CLAR, TRAN, EMPH, AUTH, ATTI, HUMO and DIST) seemed to have almost exclusively use cases with only 1-5 occurrences each. SELF use cases had mostly 1-5 occurrences each, with some exceptions of around 6-10 occurrences, but “fair play” alone had 20 occurrences. EFFI had a large portion of use cases with 5-20 occurrences, the most frequent being “no bail-out” with 70 occurrences (all after 2010 and related to the global financial crisis). ESTA was dominated by few extremely frequent use cases, especially “cleantech” (91 occurrences), “high tech” (89), and “know-how” (40).

4.4. Topics around functions

To see the differences between functions more clearly, the summed topic proportions of each function class were subtracted from the average topic proportions. The remainder showed in which contexts the functions were more or less likely to be used. A one-sample t-test was used to identify differences that were statistically significant ($p<0.05$, degrees of freedom for each function were $n_f - 1$ where n_f was the total number of use cases in each function given in figures 5a and 5b), and the significant results are presented for English in figure 6 and for Latin in figure 7. It should be noted that all the differences listed are still quite small, and since the point of view given by the used topic model is only one of many possible ones, these results should be taken as preliminary and possible trends only. To have more reliable results would require the inspection of results from several different models.

The most noticeable feature of English was that the *general* topic was the least likely context for functions EFFI and ESTA, which are practical functions, while it was the most likely for all the rhetorical functions plus TRAN. The division along the *general* topic is interesting because it is the largest topic in the corpus overall and indicates a meaningful division. It seems that English is used for practical reasons mostly in very specific contexts, while its rhetorical use is more freely spread. Other points of notice were that ESTA function favoured topics *public sector* and *energy*, meaning that these contexts probably have accumulated established English vocabulary without good Finnish counterparts, while EFFI concentrated on topics *legislation* and *public sector*, hinting that perhaps the vocabulary used for its efficiency in these contexts is in a process of becoming established. An example of such concept is “no bail out”, which has been used especially during the financial crisis of Greece around 2010 to refer to an idea that no EU Member State is liable for the debts of other Member States, which in turn comes from the common way to refer to article 125 of the Treaty on the Functioning of the European Union as the “no bailout clause”. The concept re-emerged after 2020, meaning that in the past decade no Finnish equivalent has replaced the English concept.

Latin use was expected in topics that in some way correspond to the known use contexts of science and law. Since science did not emerge as its own topic in the model, the focus here was on the topic *legislation*. Also, only EFFI and SELF had had a large number of occurrences, and only EFFI seemed to favour *legislation* as a context while SELF avoided it. This suggests that legislative speech may actually require Latin to be efficient, and does not go well together with rhetorical uses which could add ambiguity. The lack of instances made it difficult to generalize the results in many cases, but the most significant findings are presented in figure 7.

Practical functions								Rhetorical functions											
EFFI (n=641)		**CLAR** (n=521)		**ESTA** (n=348)		**TRAN** (n=78)		**EMPH** (n=144)		**SELF** (n=334)		**AUTH** (n=75)		**ATTI** (n=56)		**HUMO** (n=42)		**DIST** (n=25)	
topic	diff	topic	diff	topic	diff	topic	diff	topic	diff	topic	diff	topic	diff	topic	diff	topic	diff	topic	diff
legisl	0,9 %	for & sec p	1,0 %	pub sec	3,2 %	GEN	4,1 %	GEN	2,0 %	GEN	3,3 %	GEN	3,1 %	GEN	3,0 %	GEN	3,3 %	GEN	3,9 %
pub sec	0,8 %	legisl	0,9 %	ener	3,1 %	educ	2,9 %	taxation	0,6 %	parl fac	1,3 %	parl fac	2,0 %						
hous	0,4 %	dev coop	0,5 %	comm	1,8 %	legisl	2,5 %												
comm	0,2 %	democ	0,4 %	emplo	1,0 %	admin	1,8 %												
				soc prob	0,4 %														
				traf & tran	0,3 %														
						traf & tran	-0,2 %												
						soc & heal	-0,3 %												
						law prop	-0,3 %												
						agric	-0,4 %												
						soc prob	-0,5 %												
				crime	-0,6 %	taxation	-0,5 %					pensi	-0,3 %						
				dev coop	-0,7 %	crime	-0,6 %					educ	-0,6 %						
				democ	-0,8 %	emplo	-0,8 %					emplo	-0,6 %						
voting	-0,1 %			parl fac	-1,2 %	budget	-0,8 %			comm	-0,7 %	hous	-0,9 %	voting	-0,2 %	agric	-0,3 %	democ	-0,9 %
soc prob	-0,3 %	educ	-0,3 %	for & sec p	-1,6 %	comm	-1,3 %			ener	-0,8 %	comm	-1,0 %	emplo	-0,5 %	law prop	-0,4 %	comm	-1,2 %
emplo	-0,4 %	comm	-0,4 %	legisl	-1,8 %	ener	-1,4 %	for & sec p	-0,9 %	legisl	-1,2 %	ener	-1,3 %	dev coop	-1,0 %	comm	-1,0 %	ener	-1,3 %
GEN	-1,1 %	parl fac	-0,8 %	GEN	-3,5 %	pub sec	-5,7 %	ener	-1,2 %	pub sec	-1,3 %	pub sec	-2,8 %	pub sec	-2,8 %	for & sec	-1,7 %	pub sec	-3,9 %

Figure 6. Topics around functions of English use cases, significant differences from average.

Practical functions				Rhetorical functions							
EFFI (n=328)		**CLAR** (n=18)		**EMPH** (n=34)		**SELF** (n=319)		**AUTH** (n=40)		**ATTI** (n=22)	
topic	diff	topic	diff	topic	diff	topic	diff	topic	diff	topic	diff
legisl	2,4 %					pub sec	1,4 %	parl fac	3,9 %	GEN	5,9 %
emplo	0,7 %										
		voting	-0,3 %					region	-0,4 %		
		soc & hel c	-0,4 %					comm	-0,5 %	soc ben	-0,6 %
		dev coop	-0,7 %	tra & tran	-0,4 %			ener	-0,6 %	budg	-0,7 %
		law prop	-0,9 %	crime	-0,6 %			emplo	-0,8 %	ener	-0,9 %
		soc ben	-1,0 %	agric	-0,6 %			agric	-0,8 %	law prop	-0,9 %
voting	-0,1 %	agric	-1,1 %	emplo	-0,8 %	emplo	-0,6 %	tax	-0,9 %	agric	-1,0 %
parl fac	-0,9 %	parl fac	-2,0 %	tax	-0,9 %	legisl	-1,9 %	pub sec	-2,7 %	legisl	-2,5 %

Figure 7. Topics around functions of Latin use cases, significant differences from average. The functions that had less than 10 occurrences (DIST, HUMO, ESTA and TRAN) have been left out.

5. Discussion

The English and Latin use cases examined in this paper demonstrate how the parliamentary speakers utilize linguistic resources to fulfil communicative needs and enhance expression. Certain features of the observed patterns imply weak elite closure while others hint at the opposite: inclusion.

English use was found to have increased since the 1990s, especially the practical uses. The topic/function distribution for English suggested that practical uses of English were more common in specific contexts (avoiding the *general* topic), while rhetorical English use was more common in very general contexts. The contexts where English use was the most prominent were all important areas of societal infrastructure and development: *energy*, *housing*, *education* and *commerce*. Practical Latin use also increased somewhat after the 1990s, probably backed by the prominence of English. The topics that preferred Latin the most were probably connected through juridical discussions regarding them: *foreign and security policy* (55%) and *democracy* (53%). However, Latin use concentrated on two functions, EFFI and SELF, which were strongly divided along the topic *legislation*: efficiency was preferred while self-emphasis was avoided in speeches regarding legislation. Both English and Latin use cases included some extremely frequent ones, but also several rare occurrences, which seemed more likely candidates for elite closure especially when left untranslated or unexplained.

5.1. Roles of English and Latin

Based on the findings, Latin seems to have two main roles in the parliament. First, it appears as a practical means in Latin's traditional use context of law. Second, Latin is used to underline the speaker's own eloquence and to achieve a higher register, but this seems less likely to happen in legislative

contexts. The results reflect a perhaps apparent dichotomy for modern use of Latin: there are very specific contexts where Latin use is almost a necessity, and in most other contexts its use appears to be out of place – in an elitist kind of way. The situation of English is a bit similar, since two main roles can be deduced: practical English use was also more common in specific contexts (avoiding the *general* topic), while rhetorical English use was more common in very general contexts. The roles of English are surprisingly close to Latin's roles, though the division is clearer with English and the rhetorical uses of English do not seem nearly as much elitist as Latin's.

Like mental biases, elite closure can be conscious or subconscious, intentional or unintentional. Though field-specific jargon is usually seen as a classic example of elite closure, the practical use cases of Latin in the context of law seem to arise from a need to speak in precise terms, rather than from an intentional attempt at elite closure. Examples of such use are concepts like *de minimis* (a type of EU aid) or *ne bis in idem* (a criminal law principle under which a person cannot be punished and be subject to several procedures twice for the same facts), both of which would take long to explain without Latin every time and might lose some important connotations if translated. Overall, since "legalese" is so well established and deep-rooted, it is probably very difficult to talk about certain subjects without using at least some Latin terminology. Because of this, even without intentional exclusion, the result may still be an exclusive variety that clearly marks an elite identity. Similar process may be taking place with English in its specific practical use contexts. Especially established words and phrases such as *cleantech* or *hands-free* are hard to circumnavigate, and practical choices matter in the parliament where one is under pressure to speak clearly, effectively, and sometimes even with a time limit. Another possibility is that a speaker can misjudge the fluency levels of their listeners and chooses not to explain, which may lead to unintentional elite closure; the small number of translated English use cases supports this possibility, but more detailed and qualitative studies would be required to verify.

There is also a difference between an elite variety that one cannot fully understand, and an elite variety that one can understand but cannot produce. The former is clearly a stronger form than the latter. Given this premise, both extensive Latin and English uses in the parliament mostly fall into the latter category, and only very few use cases can fulfill both roles. Some distinction between cases can be seen from the occurrence rates of the use cases and the amount of explanation surrounding them. Surprisingly, Latin use cases were explained less than English use cases though Latin is a less known language in Finland. In case of Latin, rare phrases left unexplained would be clear attempts at elite closure, since only few people with good knowledge of Latin could fully understand. However, close reading of the excerpts revealed that most of the rare or infrequent Latin use cases *were* often explained, both practical and rhetorical cases. This is probably a strategy of experienced speakers: the speaker is able to mark themselves with an elite identity while avoiding being accused of being too abstruse. Latin use is not easy to replicate for a person who is not used to hearing it and does not know how and where such phrases should be used, so the elite variety remains exclusive even when the use cases are translated. English, on the other hand, is universally taught and acts as a common *lingua franca* for most Finns (EC, 2024), so despite the large variation in fluency levels it generally takes heavier use and more disregard to explanations to produce English use cases that could not be understood by most listeners in the parliament.

English is commonly linked with an international outlook, and similar tendencies were observed in the rapid rise of English use since the 1990s, as Finland joined the EU, and the world started to become even more interconnected via the internet. Interestingly, the last year of observation (2020) hints at a similar elbow in the future frequency of English use. Also, practical Latin use seems to have increased together with the use of English in the 1990s. Integrating Latin words and phrases into English is quite common especially in higher registers, so it could be that the overall increased use of English in the society has also strengthened the role of Latin. Especially the most common Latin phrases with practical functions, i.e. *de facto* and *ad hoc*, may have become more accepted because people have heard them used so many more times because they hear so much more English. Again, replicating features of elite English in Finnish may be another way to mark an elite variety of Finnish.

For both English and Latin the occurrence rates hint at certain words and phrases being popular and widely used during certain times (e.g. "no bail out" during financial crises) and others being rare and used only very few times (e.g. *si vis pacem, para bellum*; "if you want peace, prepare for war"). Especially the rare cases seem likely candidates to be attempts of elite closure. Some of the mid-rate cases may even be political attempts at redefining a concept (e.g. discussions about the differences

between “peace keeping”, “peace making” and “peace enforcement”). But if a phrase is used often by many different people, it is probably well known enough (in the parliamentary setting at least) that it is not possible to see such use as elite closure. But could it be the opposite? Cases where the speaker gives foreign language (usually English) terms for concepts they already explained in Finnish or Swedish is an interesting phenomenon (see example g earlier). The two most typical interpretations for this type of use are that either the speaker thinks that the foreign language term is better known than the Finnish term and wishes to clarify what is being spoken of (CLAR), or the speaker is condescendingly flaunting their language skills (SELF). The CLAR type is basically inclusive, as it can firstly help the listeners to see the spoken topic from multiple angles by providing a concept from another language, and secondly help the listener to follow if the speaker is to use the term more frequently.

5.2. Future trends

As stated by Taavitsainen and Pahta (2008: 3), "forces of globalisation, growing economic interdependence and the ensuing social and demographic shifts, have had a deep impact on the patterns of language use in the world". The changes in the language use of the Finnish parliament are perhaps subtle but still noticeable even in this brief study. For example, practical uses of Latin increased at the same time with English in the 1990s, which was very likely a by-product of increased English use. The fact that the majority of functions of English use were practical reflects the situation where English skills are really becoming a necessity even in the Finnish society. The use of English in Finland has long been seen both as a threat and as a possibility (e.g. Laitinen et al., 2023; SKL, 2018; Leppänen & Pahta, 2012). The results support the idea of English as a language of international communication, but also as a marker of elite or upper-classness in general, not just in specific contexts. There are similarities between the roles of English and Latin, as both can be used as upper-class and elite identity performance. The present study showed that some of this performance can be effectively exclusive.

The present paper leaves many directions for further studies. The most interesting direction in terms of elite closure would be to examine the used vocabulary more closely and with a discourse analytic perspective. Since English use cases varied from single words to multi-sentence quotations and were not pre-defined, which was the case with Latin, phrase or vocabulary rates for English were not easy to estimate. Same phrases can also be used in various contexts, so a deeper qualitative examination would reveal more about the way English (and Latin) is used to either unite or to differentiate. Another avenue for further examination lies in the question of *who* more specifically are using English or Latin in the Finnish parliament, which will also shed light on the question of for who it will be a necessity to use English (or Latin) in the future.

Acknowledgments

I want to thank my supervisors Veronika Laippala and Kimmo Elo for their guidance. I also thank the Finnish IT Center for Science (CSC) for providing computational resources and my colleagues at TurkuNLP, the University of Turku, and Langnet network for support.

References

Banner, Nicholas. 2021. “Do Not Read the Latin: Latin as Satanic Signifier in Supernatural Horror Cinema”. Classical Receptions Journal 13 (3): 399–415. https://doi.org/10.1093/crj/claa033.

Bedi, Jaskiran. 2019. English language in India: a dichotomy between economic growth and inclusive growth. New York: Routledge.

Coffee, Neil. 2012. “Active Latin: Quo Tendimus?” Classical World 105 (2): 255–69. https://doi.org/10.1353/clw.2012.0007.

Engelsing, Eduardo. 2017. “Census Latinus 2009: Goals, Data Collected, Importance, Perspectives”. Classical World 110 (3): 399–421. https://doi.org/10.1353/clw.2017.0023.

“eTranslation v13.3”. 2024. European Commission, Directorate-General for Translation. https://commission.europa.eu/resources-partners/etranslation_en.

[EC] European Commission. 2012. “Europeans and Their Languages”. 386. Special Eurobarometer. https://op.europa.eu/en/publication-detail/-/publication/f551bd64-8615-4781-9be1-c592217dad83,.

[EC] European Commission. 2024. “Special Eurobarometer 540 - Europeans and Their Languages”. 540. Special Eurobarometer.

Furiassi, Cristiano, Virginia Pulcini, Félix Rodríguez González, and European Society for the Study of English, eds. 2012. The anglicization of European lexis. Amsterdam ; Philadelphia: John Benjamins Pub. Co.

Gałuskina, Ksenia, and Joanna Sycz. 2013. “LATIN MAXIMS AND PHRASES IN THE POLISH, ENGLISH AND FRENCH LEGAL SYSTEMS – THE COMPARATIVE STUDY”. Studies in Logic, Grammar and Rhetoric 34 (1): 9–26. https://doi.org/10.2478/slgr-2013-0020.

Gumperz, John J. 1982. Discourse strategies. Studies in interactional sociolinguistics 1. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Haugen, Einar. 1950. “The Analysis of Linguistic Borrowing”. Language 26 (2): 210–31.

Hickey, Raymond, ed. 2019. English in multilingual South Africa: the linguistics of contact and change. Studies in English language. New York, NY: Cambridge University Press.

Hyvönen, Eero, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2024. “Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland”. Semantic Web. https://www.semantic-web-journal.net/system/files/swj3605.pdf.

Jahan, Iffat, and M. Obaidul Hamid. 2019. “English as a Medium of Instruction and the Discursive Construction of Elite Identity”. Journal of Sociolinguistics 23 (4): 386–408. https://doi.org/10.1111/josl.12360.

Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. “Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task”. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics.

Laitinen, Mikko, Sirpa Leppänen, Paula Rautionaho, and Sara Backman. 2023. ”Englanti Suomen kansalliskielten rinnalla – Kohti joustavaa monikielisyyttä”. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 2023 (59).

Leppänen, Sirpa, and Tarja Nikula. 2007. “Diverse Uses of English in Finnish Society: Discourse-Pragmatic Insights into Media, Educational and Business Contexts”. Mult 26 (4): 333–80. https://doi.org/10.1515/MULTI.2007.017.

Leppänen, Sirpa, and Päivi Pahta. 2012. “Finnish Culture and Language Endangered — Language Ideological Debates on English in the Finnish Press from 1995 to 2007”. In: Dangerous

Multilingualism, edited by Jan Blommaert, Sirpa Leppänen, Päivi Pahta, and Tiina Räisänen, 142–75. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137283566_7.

Maier, Daniel, Christian Baden, Daniela Stoltenberg, Maya De Vries-Kedem, and Annie Waldherr. 2022. “Machine Translation Vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections”. Communication Methods and Measures 16 (1): 19–38. https://doi.org/10.1080/19312458.2021.1955845.

Matus-Mendoza, Maríadelaluz. 2002. “The English Lexical Loan: A Class Marker”. Journal of Hispanic Higher Education 1 (4): 329–37. https://doi.org/10.1177/153819202236977.

Myers-Scotton, Carol. 1993. “Elite closure as a powerful language strategy: the African case”. International Journal of the Sociology of Language 103 (1). https://doi.org/10.1515/ijsl.1993.103.149.

———. 1997. Duelling Languages: Grammatical Structure in Codeswitching. Oxford: Clarendon Press.

———. 2002. Contact Linguistics. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299530.001.0001.

”Parlamenttisampo”. 2023. https://parlamenttisampo.fi.

Peterson, Elizabeth. 2022. “The English Language in Finland: Tool of Modernity or Tool of Coloniality?” In: Finnishness, Whiteness and Coloniality, edited by Josephine Hoegaerts, Tuire Liimatainen, Laura Hekanaho, and Elizabeth Peterson, 267–89. Helsinki University Press. https://doi.org/10.33134/HUP-17-11.

Piętka, Radosław. 2016. “A Thrill for Latinists: Latin Language in Contemporary Horror Films”. Teoksessa Antiquity in Popular Literature and Culture, edited by Konrad Dominas, Elżbieta Wesołowska, and Bogdan Trocha, 255–66. Newcastle upon Tyne: Cambridge Scholars Publishing.

Ristilä, Anna, and Kimmo Elo. 2023. “Observing Political and Societal Changes in Finnish Parliamentary Speech Data, 1980–2010, with Topic Modelling”. Parliaments, Estates and Representation 43 (2): 149–76. https://doi.org/10.1080/02606755.2023.2213550.

Roelli, Philipp. 2021. Latin as the Language of Science and Learning. De Gruyter. https://doi.org/10.1515/9783110745832.

Romaine, Suzanne. 1995. Bilingualism. 2nd ed. Language in society 13. Oxford, UK ; Cambridge, Mass., USA: Blackwell.

SKL. 2018. ”Suomi tarvitsee pikaisesti kansallisen kielipoliittisen ohjelman. Suomen kielen lautakunnan kannanotto 26.10.2018”. Oct 26, 2018. https://www.kotus.fi/ohjeet/suomen_kielen_lautakunnan_suosituksia/kannanotot/suomi_tarvitsee_pikaisesti_kansallisen_kielipoliittisen_ohjelman.

Stahl, Peter. 2023. ”Lingua 2.0.2”. Python. https://github.com/pemistahl/lingua-py/releases/tag/v2.0.2.

Statistics Finland. 2023. 11rm: “Language according to sex by municipality, 1990-2023”. https://pxweb2.stat.fi/PxWeb/pxweb/fi/StatFin/StatFin__vaerak/statfin_vaerak_pxt_11rm.px/.

Taavitsainen, Irma, and Päivi Pahta. 2003. ”English in Finland: Globalisation, Language Awareness and Questions of Identity”. English Today 19 (4): 3–15. https://doi.org/10.1017/S0266078403004024.

Appendix 1: Searched Latin phrases

Found:

a priori (7)
ad hoc (52)
alea iacta est (1)
alma mater (1)
ave caesar, morituri te salutant (1)
bona fide (14)
carpe diem (1)
ceteris paribus (5)
ceterum censeo (8)
de facto (329)
de jure (5)
de minimis (130)
deus ex machina (2)
dies irae (1)
ecce homo (1)
errare humanum est (3)
ex cathedra (6)
ex officio (3)
ex oriente lux (3)
ex post (3)
ex tempore (32)
festina lente (5)
finis finlandiae (3)
homo homini lupus (1)
ille faciet (1)
in absentia (1)
in dubio pro reo (7)
in memoriam (1)
ius sanguinis (1)
ius soli (1)
lex in casu (6)
liberum veto (1)
mare nostrum (3)
mea culpa (1)
mens sana in corpore sano (2)
mutatis mutandis (2)
navigare necesse est, vivere non est necesse (2)
ne bis in idem (25)
nomen est omen (9)
non scholae sed vitae discimus (1)
numerus clausus (5)
o sancta simplicitas (2)
o tempora, o mores (8)
pacta sunt servanda (9)
panem et circenses (1)
pater familias (3)
per se (1)
persona non grata (1)
primus inter pares (6)
pro forma (3)
pro patria (3)
quis custodiet ipsos custodes (1)
si vis pacem, para bellum (2)
status quo (27)
sui generis (20)
tabula rasa (2)
veni, vidi, vici (1)
vestigia terrent (8)
vox populi (4)

Not found:

a posteriori
ab ovo
ab urbe condita
acta est fabula, nunc plaudite
ad acta
ad maiorem dei gloriam
ad usum delphini
alterego
amicus curiae
amicus plato, sed magis amica veritas
amor patriae
an nescis, mi fili, quantilla prudentia mundus regatur
annus mirabilis
argumentum ad hominem
argumentum ad misericordiam
argumentum ad nauseam
ars gratia artis
ars longa, vita brevis
auri sacra fames
casus belli
causa sui
caveat emptor
citius, altius, fortius
cogito ergo sum
corpus delicti
cui bono
cuius regio, eius religio
cum hoc ergo propter hoc
damnatio memoriae
de dicto
de lege ferenda
de lege lata
de mortuis nil nisi bene
de profundis
de re
de se
delictum
deus vult
dicto simpliciter
dis manibus
e pluribus unum
ecclesiola in ecclesia
et al.
et nunc et semper
et tu, brute
etiamsi omnes, ego non
ex more
ex nihilo nihil fit
ex nihilo
ex post facto
falsa demonstratio non nocet
filioque
habeas corpus
hannibal ad portas
homo sovieticus
ignoramus et ignorabimus
in hoc signo vinces
in loco parentis
in medias res
in pectore
in situ
in vino veritas
incertae sedis
inter alia
ipso facto
jure uxoris
larvatus prodeo
magnum opus
memento mori
modus operandi
naturalia non sunt turpia
ne plus ultra
nemo me impune lacessit
nemo propheta in patria
noli me tangere
nomen nescio
non compos mentis
non praevalebunt
non sequitur
novus ordo seclorum
obiter dictum
omnia mea mecum porto
otium
passim
pax vobiscum
pecunia non olet
penitenziagite
per aspera ad astra
per capsulam
persona grata
plenum plenorum
plurale tantum
post hoc ergo propter hoc
pro bono
pro hac vice
puer aeternus
puer robustus sed malitiosus
pulvis et umbra sumus
quid pro quo
quo vadis, domine
quod licet iovi, non licet bovi
requiescat in pace
salva veritate
semper fidelis
sic semper tyrannis
sic transit gloria mundi
sine qua non
stare decisis
sub rosa
summa cum laude
terra australis
terra incognita
terra nullius
ultima ratio regum
ultra vires
urbi et orbi
uti possidetis
vade retro, satana
vagina dentata
vanitas vanitatum et omnia vanitas
veto exclusionis
vicarius filii dei
vidi aquam

Perspectives on AI in the British and Slovenian parliament

Ajda Pretnar Žagar[1], David Moats[2]

[1]Faculty of Computer and Information Science, University of Ljubljana
[2]Center for Consumer Society Research, University of Helsinki
[2]Digital Humanities, Kings College London

Abstract

Parliamentary debates illustrate relevant legislative issues and how they change over time. The paper presents debates on artificial intelligence in the British and Slovenian parliaments. We analyze ten years (2015-2024) of parliamentary debates, observing semantic networks and significant words of artificial intelligence debates. We conduct a descriptive comparative analysis, showing the differences in the debates between a non-EU country with a strong AI sector (the UK) and an EU country with a weak AI sector (Slovenia). We conclude that in the former, the debate is more sectoral and focused on supporting businesses, while in the latter, the debate is linked to EU policies and focused on education and digitization. A comparative analysis of the two approaches can serve as a basis for diversifying AI policies.

Keywords

parliamentary debates, comparative analysis, discourse analysis, artificial intelligence

1. Introduction

On May 21, 2024, the European Union enacted the first comprehensive legislative framework on artificial intelligence (AI) in the world. The European Commission proposed the legislative framework in 2021, while the draft legislation was submitted to the parliament in 2023 (Madiega 2021). Concurrently, the AI Regulation Bill was proposed in the United Kingdom on March 18. In light of these significant shifts in policy responses to AI, we conducted an analysis of parliamentary debates on AI in two European countries: the United Kingdom and Slovenia. The rationale behind selecting these two countries is not merely contingent on our proficiency in their respective languages. When qualitatively examining the parliamentary discourse surrounding AI in a multitude of European nations, it became evident that the British and Slovenian debates exhibited a striking divergence.

Two main differences between Slovenia and the United Kingdom are anticipated to influence the debates and policy-making. One key distinction is that Slovenia is a member state of the European Union, whereas the United Kingdom ceased to be a member on February 1, 2020.

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, Reykjavik, Iceland, May 28, 2024
*Corresponding author.
ajda.pretnar@fri.uni-lj.si (A. P. Žagar); david.moats@helsinki.fi (D. Moats)
https://github.com/ajdapretnar (A. P. Žagar)
0000-0002-5927-4538 (A. P. Žagar); 0000-0001-9622-9915 (D. Moats)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

Consequently, Slovenia is obliged to adhere to EU policy, whereas the UK is not. A second point of contrast pertains to the relative strength of the national AI sector. The United Kingdom has the third largest national AI sector, preceded only by the United States and China (Szczepański 2024). In contrast, Slovenia is not referenced in the majority of AI monitoring reports[1]. The EU membership and the strength of the AI sector both result in the interplay between innovation and regulation, with the EU favoring the ”top-down”, more regulatory approach, and the UK favoring the ”bottom-up”, less regulatory approach Chan, Papyshev, and Yarime 2024; Carvão 2024.

These differences are evident in the parliamentary debates. Based on the distinction between ”top-down” and ”bottom-up” regulatory approaches, the UK debates should be more concrete, legislation-oriented, and focused on the present, while in Slovenia, the debates should be more abstract, strategy-oriented, and focused on the future. To elicit differences in parliamentary speeches on AI, we compare the debates from the British and Slovenian parliaments from the ParlaMint corpora (Erjavec et al. 2023), which we manually supplemented with debates from mid-2022 to 2024. We employ computational and close reading techniques to determine the differences and similarities of the debates. We relate the debates to legislative acts and other parliamentary documents to establish a relation between the debates and their application through legislation. The descriptive analysis serve to support the ”top-down” vs. ”bottom-up” distinction.

2. Related work

Analyzing AI policy debates is critical for understanding regulatory frameworks, which should foster public trust and sustainable societal integration of AI. While the proclaimed goal is a responsible and equitable AI-driven future, much of the social science literature analyses how the framing of the debate impacts policy choices. For example, AI policy is frequently presented in terms of competition between nations.

Cave and ÓhÉigeartaigh (2018) argue that portraying AI as a “race for strategic advantage” or a “winner takes all” scenario may encourage the circumvention of safety and governance measures. Furthermore, it assumes a competitive or even aggressive stance when, in fact, cooperation is essential for the development of effective AI governance. Bareis and Katzenbach (2022) show that China, US, France and Germany all present AI as an inexorable and disruptive force. In their analysis, AI is presented as inevitable by drawing on significant technological transformations in the past. International competition around AI is used to presume the necessity of AI for economic health, but paradoxically, uncertainty about such future is deployed to argue for leadership or action. Imbrie et al. (2020) argue that framing AI as an “arms race” has implications for policy directions. Although we find this work helpful and examine the role of competition narratives in greater detail below, we are also interested in how perceived competition affects the nature of the debates themselves.

Analysis of parliamentary debates can also inform the development of AI governance frameworks. A comparative analysis of legislation in the European Union, Brazil and Canada investi-

[1]Many statistics are reported as total counts rather than per capita, which may result in an unfavorable representation of Slovenia (and potentially the UK).

gates how each AI governance framework involves the citizens in AI governance and finds that long-lasting inclusion is precluded by the lack of dedicated participatory mechanisms Unver 2024. Another study uncovered how MEPs perceive and comprehend bias and discrimination in AI and their stance on regulatory measures by studying hearing transcripts of the Special Committee on Artificial Intelligence in a Digital Age (AIDA). Both studies highlight the importance of citizen participation in AI regulatory processes.

Studies in AI governance explore the distinction between top-down and bottom-up approaches in AI policy-making. The EU focuses heavily on comprehensive AI legislation, while the UK is leaving AI-specific regulation to the remit of existing regulators Hadfield and Clark 2023. The global economic competition, institutional structure, and policy preferences of domestic actors all heavily shape AI policy-making in the EU Justo-Hanani 2022. The UK, on the other hand, is not required to comply to a supranational institutional structure, enabling their policy-making to be less incremental and more flexible. Brexit certainly had a dampening effect on the EU competitiveness in AI, further making the EU a laggard in the "AI race" compared to the United States and China (Csernatoni 2019). There are also major downsides to the bottom-up policy-making. (Papyshev and Yarime 2024) present the case of Russia's ethics-based approach and argue how such approaches quickly become unenforceable, thus relying on the morale of the AI companies for self-regulation. Cultural context thus plays a crucial role in how values are translated into policies Feijóo et al. 2020; Lewis et al. 2020.

In continuation, we observe AI framing in the UK and Slovenia, to see how they compare with each other. We lean on the distinction between bottom-up and top-down approaches and see how they manifest in the rhetoric on AI.

3. AI policies in the UK and Slovenia

A comparison of the timeline of policy and legislative events with the technological progress of AI (Figure 1) reveals that policy-making frequently lags behind technological development or is even unrelated to it. It is evident that certain technologies are constrained by domain limitations, such as AlphaGo (board games), AlphaFold (biology) and Transformer models (linguistics). In such cases, the impetus for legislative action is relatively low. Conversely, the testing of autonomous vehicles has a profound impact on public life, prompting the UK to expeditiously adopt the Automated and Electric Vehicles Act.

A comparison of the timeline reveals a necessity for technology to mature. The term "mature" is not used in the sense of technological maturity, but rather in the context of societal awareness. It is essential that the general public and legislators recognize the potential impact of specific technologies and the domains for their application. Once there is sufficient general awareness of a technology, there is often a subsequent push for legislation to be enacted. Therefore, the timeline on the left and right of the axis is not directly related, but it illustrates how politicians have become increasingly involved in AI policy-making.

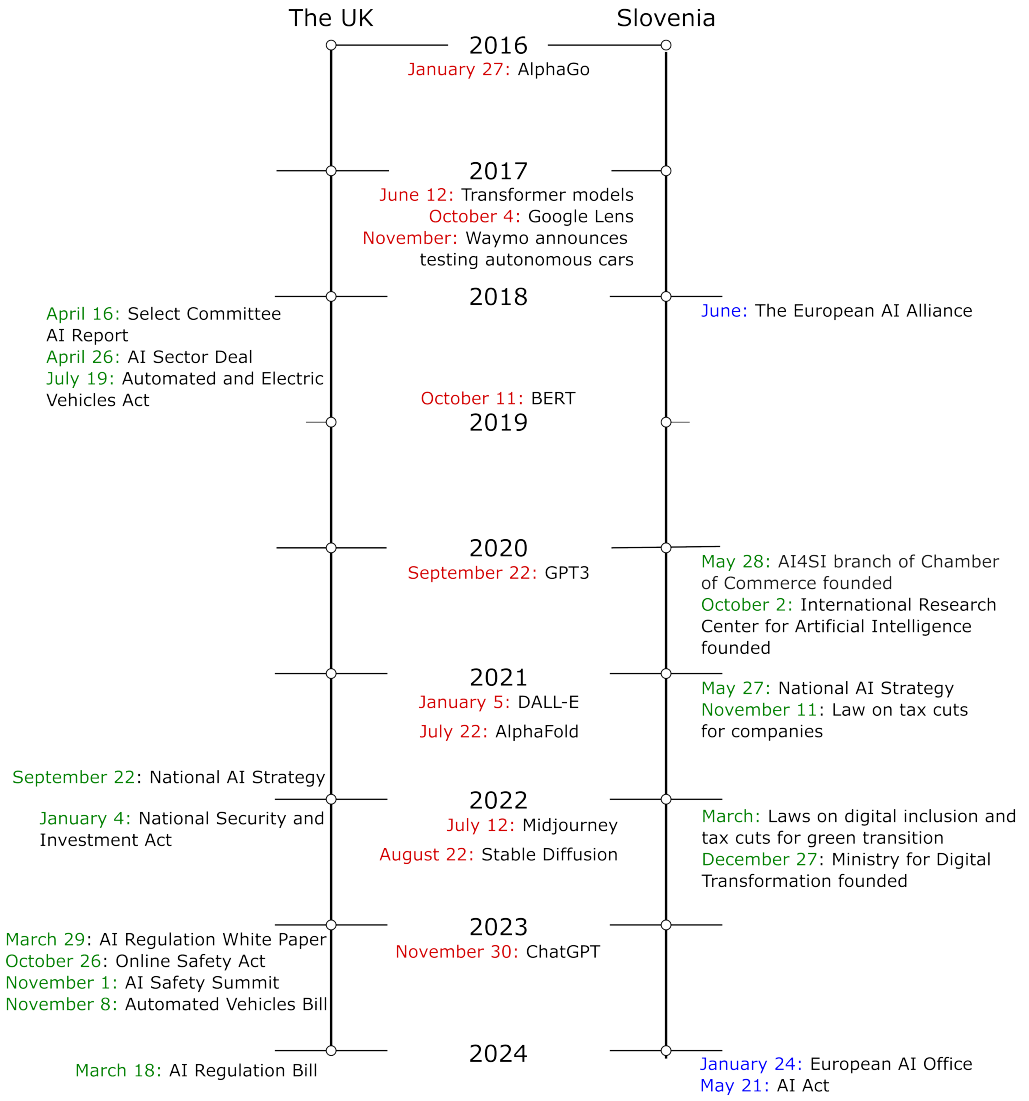


Figure 1: Timeline of technological development and the countries' AI policy response. The UK is on the left, and Slovenia on the r ight. The blue color in Slovenia's timeline corresponds to the EU acts, while the green is national legislation in both images.

4. Methodology

We used a combination of text mining and close reading approaches to answer the following question: "How are parliamentary debates on artificial intelligence framed in EU and non-EU countries?" We employed co-occurrence networks and word enrichment to determine how the parliamentary debates on AI evolved over time, which topics arose in the AI debates, and whether the debates reflected technological progress in AI.

Table 1
Comparison of the two subcorpora.

Subcorpus	N	From	To
British	1406	2015-09-08	2024-04-17
Slovenian	120	2016-07-15	2024-03-28

5. Data

We used the British and Slovenian debates from the Parlamint 4.0 corpus (Erjavec et al. 2023) and supplemented them with the debates from mid-2022 to 2024[2]. The data contains official transcripts of parliamentary proceedings. For the UK, the data includes both the House of Lords and the House of Commons. We retrieved the debates mentioning artificial intelligence with a regular expression *"\bartificial intelligence\b|\bAI\b"* for British and *"\bumetn\w* inteligenc\w*\b"* for Slovenian debates. This resulted in 1406 and 120 documents respectively (Table 1).

Given the large difference in the size of the two subcorpora, we can assume that British MPs talk more about the AI than their Slovenian counterparts. However, we must also take into account the size of the parliament and the number of speeches given. The following statistics are limited to the Parliamint 4.0 part of the subcorpus[3] due to a lack of metadata for manually added debates.

In relative terms, the British MPs talked about AI in 0.17% of debates in a given period, while Slovenian MPs talked about AI in 0.1% of debates. The ratio of speakers mentioning AI to the total number of speakers was 22% and 10%, respectively. This shows that the proportion of AI speeches is similar in both parliaments. However, in the British Parliament the debate is distributed among more speakers than in the Slovenian Parliament. The speaker with the most speeches mentioning AI in the UK, Timothy Francis Clement-Jones, represents only 4% of speeches, while in Slovenia, Franc Breznik represents 19%. 180 speakers (46%) mention AI more than once in the UK, while 15 speakers (40%) do so in Slovenia. In summary, AI is a fairly niche topic in Parliament, with only a few MEPs going beyond a fleeting mention. The UK shows a slightly more mature debate in this respect than Slovenia.

6. Results

We used two text mining techniques to approach the main research question from different perspectives. Co-occurrence networks reveal closely related terms that suggest underlying themes. Word enrichment reveals characteristic terms for a given subcorpus and allows to observe the evolution over time with temporal subsets. Results are verified and substantiated by close reading, policy comparison, and technological timeline. Slovenian results are manually translated. Full-size figures, stopword lists, and data are available on GitHub[4].

[2]Data was retrieved from Hansard (https://hansard.parliament.uk/) for the UK, and Parlameter for Slovenia (https://-parlameter.si/).

[3]1025 speeches on AI in the UK, and 77 in Slovenia.

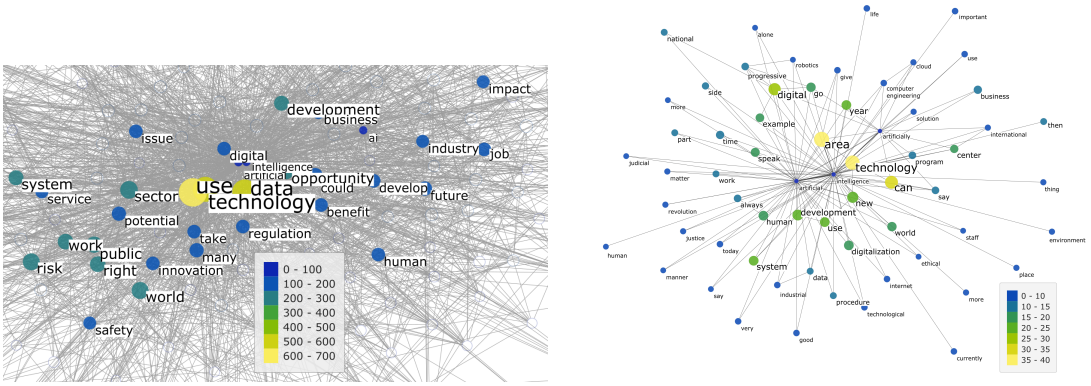
[4]https://github.com/ajdapretnar/AI-perspectives

6.1. Co-occurrences

We computed co-occurrence networks for the terms ”AI” and ”artificial intelligence” (Figure 2). We removed NLTK stopwords and additional parliamentary stopwords. We also removed tokens appearing less than 10 times and those appearing in a single document. The window size for observing co-occurrences was 5.

The UK network was much denser, due to the larger number of documents. However, besides the difference in network density, there is also a difference in central terms.

The words most commonly associated with the terms ”AI” and ”artificial intelligence” in the UK parliament reflect a focus on AI usage (*use*, *technology*, *data*), business (*world*, *work*, *sector*, *opportunity*) and public impact (*risk*, *public*, *right*). In Slovenia, the focus is on AI potential (*can/could*, *new*, *development*), legislative aspects (*year*, *use*, *human [rights]*), and digital transformation (*digital*, *system*, *digitalization*).



(a) The UK (b) Slovenia

Figure 2: A co-occurrence network for the AI terms in the parliamentary debates. The size of the node corresponds to the frequency of the word. ”AI”, ”artificial” and ”intelligence” are intentionally de-scaled due to their disproportionate frequency. For the UK, the top 30 terms are shown.

6.2. Significant words

We extracted significant words for each year to further characterize the content of the speeches. We used word enrichment, a technique that finds statistically significant words in a subset of documents. The technique compares the subset to the entire set, thus overcoming the ubiquitous parliamentary stop words in the corpus.

We performed the analysis on sentences mentioning AI. Tables 2 and 3 show the top five significant terms for each year. The words are ranked by p-value of the hypergeometric test, with the cut-off at 0.05. Thus some fields are lacking entries. Note that there was not much data for the 2015 British debates and the 2017 Slovenian debates, which makes the results for these two years unreliable. There is only one document for the 2016 Slovenian debates, so we excluded it from the list.

In both countries, the initial conversation started with talk about the potential of AI and how

2015 (4)	2016 (41)	2017 (145)	2018 (280)	2019 (146)
admiration	robotic	car	industrial	250
carney	internet	driverless	data	clinician
permanently	revolution	robotic	health	last
currency	baker	broadband	nh(s)	topol
radically	driverless	botnet	centre	hospital
2020 (183)	**2021 (158)**	**2022 (80)**	**2023 (679)**	**2024 (191)**
19	defence	clearview	summit	developer
covid	cyber	enforcement	safety	label
list	aukus	fought	risk	regulator
coronavirus	undersea	newcastle	regulation	safety
cryptographic	amendment	equipment	already	model

Table 2
Word enrichment of British parliamentary debates by years.

2017 (5)	2018 (41)	2019 (22)	2020 (280)
data	revolution	preparation	stefan
point_V	industrial	solution	innovative
can_V	half	next	institute
aggression	finance	still_ADV	jožef
insane	employ	blockchain	agreement
2021 (24)	**2022 (14)**	**2023 (29)**	**2024 (11)**
ethical	skill	engineering	information
autonomous	digital	artificial	intellectual
speech	young	big data	modern
investment	free	study_V	decision
”ministering”	programming	-	contemporary

Table 3
Word enrichment of Slovenian parliamentary debates by years. A dash indicates no word fits the cut-off criteria.

”revolutionary” it is. However, the debate in the UK became much more concrete in 2017 with the discussion about driverless cars, while in Slovenia the term ”revolution” did not appear until 2018. By then, the UK was discussing the role of AI in healthcare, followed by the response to the COVID pandemic, defense, legal challenges in AI solutions (Clearview), and finally the path to safe and responsible AI. In Slovenia, on the other hand, the debate did not focus on a specific topic until 2020, with the establishment of the International Research Center for Artificial Intelligence (IRCAI) under the auspices of UNESCO. In 2021, the only reference to AI was the nomination of the new Minister of Justice, whose campaign included a commitment to responsible and ethical AI. Later, the focus was on education and the implementation of the National Program to Promote the Development and Use of Artificial Intelligence.

7. Qualitative perspective

Finally, we conducted close-reading to identify specific aspects of national AI policies. Consistent with the literature, we found that both countries framed the issue in terms of competition. For

example:

> China dominates on a scale that we simply cannot comprehend over here. Its technological capabilities and its investment in quantum computing, and so on, mean it already owns 40% of the world's data, and it is moving further afield. Once a country moves into the Chinese way of thinking — Huawei, and so on—it is very difficult to get out It is only a matter of time before countries that are already financially compelled or obliged to support Chinese methods and systems will have to move over to China's global positioning system, and so on. (Tobias Martin Ellwood, British MP, 2019-10-14)

While the UK seemed mostly concerned about the shifting relationships between China, America and the EU, Slovenia mentioned other countries less frequently, mainly referring to the EU, on which it depends for regulation.

The UK has a very clear AI policy, formally set out in its National AI Strategy. It focuses on encouraging innovation and investment while protecting core values. It aims to position itself carefully between innovation and security. The UK has the third largest national AI sector in the world. It aims to protect its competitive advantage and businesses, while ensuring the safety of its citizens.

> I am often asked at technology events, which I attend assiduously, what the Government's policy is on artificial intelligence. On the one hand, there is an important focus on safety for artificial intelligence to make it as safe as possible for consumers, which in itself begs the question of whether that is possible; on the other, there is a need to ensure that the UK remains a wonderful place for AI innovation. (Lord Vaizey of Didcot, British MP, 2024-03-22)

The policy is markedly sector-specific. It covers health, business, the military and even the creative industries. In all areas, the emphasis is on a careful balance between two policy aspects: innovation and security.

> For creatives, the risk of AI-generated material flooding the market gives rise to significant regulatory and ethical challenges, but these can be overcome, or at least mitigated, with well thought out and considered policy that balances the legitimate concerns of creatives with the need to foster digital innovation. (Sarah Olney, British MP, 2023-01-02)

In Slovenia, the debate about AI technology and the legal framework that needs to be adopted for its proper use is still overshadowed by party divisions. In some debates, AI is mentioned fleetingly as a rhetorical device, completely outside of any concrete legislative or policy proposals.

> I apologize because I have to admit that I did not write this text myself, but it was written from the first to the last letter, period, and comma by the artificial intelligence program Open Chat GPT. The point is that the whole world, including artificial intelligence and Martians, understands that Slovenian independence was a collective action. Here in our small Slovenia, one person and one party are claiming these merits. (Jonas Žnidaršič, Slovenian MP, 2023-04-21)

Another salient topic is the lack of digital readiness in the public and private sectors. In particular, the healthcare system lacks the appropriate digital infrastructure to enable the electronic transfer of records.

> At a time when artificial intelligence is entering our lives through wide-open doors, as we have heard many times today, unfortunately, we still have to carry paper reports and other documentation with us when we visit a general practitioner, a specialist, or for diagnostic tests elsewhere. (Felice Žiža, Slovenian MP, 2023-06-21)

However, healthcare is not the only example of inadequate infrastructure, and Slovenian lawmakers see digitalization and upskilling as a crucial step forward. Artificial intelligence is just one area of digital transformation, along with cloud computing, big data, blockchain, and quantum computing.

Slovenia is tied to the EU in terms of legislation. While it can certainly adopt its own legislation, it would make little sense to prepare legislation that would not be in line with EU legislation in a few months. So it tends to wait for the supranational body, while the UK is free to adopt its own legislation quickly.

To ascertain geopolitical orientations, we additionally examined the results of named entity recognition. The most common are self-mentions (50% of documents contain self-mentions for the UK and 84% for Slovenia). In addition to self-mentions, the UK refers to the USA (15% of documents), European countries (13%) and China (9%), while Slovenia refers to the EU (38%), European countries (22%) and the USA (10%). Qualitative researchers argue that country comparisons reveal important framings of the AI debate (Bareis and Katzenbach 2022; Imbrie et al. 2020). However, a country's framing of the AI debate also encapsulates its geopolitical focus. The UK sees the US as its biggest reference point and competitor (followed by China). Slovenia, on the other hand, as part of the EU, is deeply intertwined with its politics, which is also evident in the debates.

8. Conclusion

Computational text analysis helped us to analyze the AI policy debate in the UK and Slovenia. The analysis shows that the way the policy is discussed in the two countries is quite different. In the UK, the debate was mainly about protecting national interests and maintaining a strong position in the AI sector, while in Slovenia, the focus was on general digital transformation.

In the UK, the policy aims to balance innovation with AI security. The careful alignment reflects the national business landscape, which is heavily influenced by AI te chnology. The parliamentary debate on AI is fairly structured, with several committees dedicated to working out the details of specific sector legislation. In Slovenia, on the other hand, policy sees AI as just one part of the overall digital transformation. The focus is on digital literacy, infrastructure and education. However, the parliamentary debate lacks focus and lumps AI together with other technologies, such as blockchain and cloud computing.

Our computational and qualitative analysis has shown that the UK is more focused on specific policy instruments and sectors, and justifies investment in AI with references to competition between countries. Slovenia, on the other hand, speaks more vaguely about future opportunities. These differences in the character of the debate are likely a result of the two countries' geopolitical situations and existing technology sectors. We deem both dimensions, the EU membership and the AI sector, equally relevant for influencing the political debate on AI. EU membership influences the debate in terms of political agility, where the UK is able to respond quicker to events and implement legislation, while the EU Member States have to coordinate their policies to comply with the EU legislation. The strength of the AI sector influences the balance between innovation and regulation. Too stringent regulation inhibits innovation, thus the UK, with its strong AI sector, has to be laxer with regulation than the EU, whose AI sector

is not as strong.

In both countries, there has been a recent shift towards AI safety, with the adoption of the AI Act, which has had an impact even on the non-EU UK. Slovenia will duly comply with EU legislation, and some delay in its policy-making is due to the need for coordination with the EU. It will be interesting to monitor the development of policy after the AI Act comes into force on August 1, 2024. However, in the context of the pre-AI Act period, we argue that the UK's legislation is more progressive and specific to AI, while Slovenia's is more conservative and broad. In the future, we would like to extend our analysis to other EU countries and also make explicit links between dominant discourses and arguments and concrete policy initiatives.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency research program P6-0436: Digital Humanities: resources, tools and methods (2022-2027) and the Reimagine ADM project (GA number 101004509), funded by European Union's Horizon 2020 Research and Innovation Programme (Chanse).

References

Bareis, Jascha, and Christian Katzenbach. 2022. "Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics." *Science, Technology, & Human Values* 47 (5): 855–881.

Carvão, Paulo. 2024. *The Dual Imperative: Innovation and Regulation in the AI Era.* arXiv: 2407.12690 [cs.CY]. https://arxiv.org/abs/2407.12690.

Cave, Stephen, and Seán S. ÓhÉigeartaigh. 2018. "An AI Race for Strategic Advantage: Rhetoric and Risks." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society,* 36–40. AIES '18. New Orleans, LA, USA: Association for Computing Machinery. https://doi.org/10.1145/3278721.3278780.

Chan, Keith Jin Deng, Gleb Papyshev, and Masaru Yarime. 2024. "Balancing the tradeoff between regulation and innovation for artificial intelligence: An analysis of top-down command and control and bottom-up self-regulatory approaches." *Technology in Society* 79:102747.

Csernatoni, Raluca. 2019. "An Ambitious Agenda or Big Words? Developing a European Approach to AI." *Security Policy Brief* 117.

Erjavec, Tomaž, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Manex Agirrezabal, Tommaso Agnoloni, José Aires, et al. 2023. *Multilingual comparable corpora of parliamentary debates ParlaMint 4.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1859.

Feijóo, Claudio, Youngsun Kwon, Johannes M. Bauer, Erik Bohlin, Bronwyn Howell, Rekha Jain, Petrus Potgieter, Khuong Vu, Jason Whalley, and Jun Xia. 2020. "Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy." *Telecommunications Policy* 44 (6): 101988.

Hadfield, Gillian K, and Jack Clark. 2023. "Regulatory markets: The future of ai governance." *arXiv preprint arXiv:2304.04914*.

Imbrie, Andrew, James Dunham, Rebecca Gelles, and Catherine Aiken. 2020. *Mainframes: a provisional analysis of rhetorical frames in AI.* Issue brief. Center for security and emerging technology, August.

Justo-Hanani, Ronit. 2022. "The politics of Artificial Intelligence regulation and governance reform in the European Union." *Policy Sciences* 55 (1): 137–159.

Lewis, Dave, Linda Hogan, David Filip, and P. J. Wall. 2020. "Global Challenges in the Standardization of Ethics for Trustworthy AI." *Journal of ICT Standardization* 8, no. 2 (April): 123–150.

Madiega, Tambiama. 2021. *Artificial intelligence act.* Briefing PE 698.79. EPRS | European Parliamentary Research Service, March.

Papyshev, Gleb, and Masaru Yarime. 2024. "The limitation of ethics-based approaches to regulating artificial intelligence: regulatory gifting in the context of Russia." *AI & SOCIETY* 39 (3): 1381–1396.

Szczepański, Marcin. 2024. *The United Kingdom and artificial intelligence.* Briefing PE 762.285. EPRS | European Parliamentary Research Service, April.

Unver, Mehmet B. 2024. "AI governance: Compromising democracy or democratising AI?" In *TPRC 2024 The Research Conference on Communications, Information & Internet Policy.*

How You Sample Determines What You Find: Investigating Bias in Parliamentary Data Sampling Methods

Martin Karlsson[a], Eric Borgström[a] and Christian Lundahl[a]

[a] *Open Parliament Laboratory (OPaL), Örebro University*

Abstract

This study addresses the issue of sampling error within research on subsets of parliamentary text corpora. Two samples of parliamentary speeches relating to the marketisation of the Swedish education system, drawn through different sampling techniques, are analysed and compared. The analyses find that diverging sampling methodologies can be complementary as each method adds substantial quantities of unique documents to the dataset. Further, the diverging sampling methodologies employed produce documents with similar semantic content. However, analyses of the distribution of speeches between party affiliations and speakers indicate vast differences between the two samples. These results indicate that sampling frames can substantially influence the findings of parliamentary text analyses. We conclude that combining different sampling techniques can be a way to reduce the risk of sampling error, which in turn can have a strong influence on the conclusions drawn from analyses of parliamentary texts.

Keywords

Parliamentary data, sampling error, school marketisation

1. Introduction

Research utilising parliamentary text data often focuses on the general dynamics of how parliaments operate (Slapin & Proksch, 2008; Rheault et al., 2016; Rheault & Cochrane, 2020). However, in recent years, a growing number of studies have utilised such data to investigate how specific policy issues develop and are debated in parliament (Magnusson et al., 2018; Isohao et al., 2019; Müller-Hansen et al., 2021; Voigt et al., 2024). Such research holds great potential to contribute to understanding policy discourses' development over more extended time periods and systematically map position-taking and policy frames in policy debates.

Investigating specific policy issues, however, requires creating subsamples of parliamentary text data that relate to a specific policy issue. As with all forms of data sampling, this process introduces the risk of sampling errors that can lead to biased data and, through this, misleading conclusions. Therefore, it is important to consider and investigate possible sources of sampling error.

The most frequently utilised sampling methodology in earlier research has arguably been term-based sampling. That is, identifying terms that are expected to be necessary and sufficient for the inclusion of documents within a specific policy issue or policy discourse. Such terms are ideally topical words exclusively used concerning a particular policy issue. These words are then used to search within or filter parliamentary text databases, creating a sample with documents that include one or more of the chosen terms.

To reduce sampling error, researchers often conduct further analyses to identify and exclude false positives, i.e. documents that include the search terms without belonging to the sub-population of documents that the sampling method is designed to identify. These analyses can be qualitative as well

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, Reykjavik, Iceland, May 28, 2024.
EMAIL: martin.karlsson@oru.se (A. 1); eric.borgstrom@oru.se (A. 2); Christian.lundahl@oru.se (A. 3)
ORCID: 0000-0002-5485-8577 (A. 1); 0000-0002-0126-0416 (A. 2); 0000-0001-8173-7474 (A. 3)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

as quantitative and are not seldom based on topic modelling of the sample to identify and remove remote topics among its documents (see Voigt et al., 2024 for an example).

However, we argue that earlier research has been insufficiently concerned with another source of sampling error: the possibility of false negatives. Such false negatives are documents relevant to the policy issue in the focus of the analysis, without including the search terms used in the sampling frame. In this study, we will demonstrate this potential issue through analyses of two separate samples of parliamentary speeches concerning the marketisation of education in Sweden.

One of these is a term-based sample (see the method section for further details on the sampling technology), including parliamentary speeches that mention independent schools (fristående skolor) or free schools (friskolor). This school form has developed as a result of this marketisation process. The second sample is based on the headlines the parliament has given for sessions of plenary debates. All parliamentary speeches in debates that are given headlines that pertain to school marketisation are included in this sample. Our analyses of these samples seek to answer four research questions:

1. Do these sampling methodologies produce distinct samples - i.e. samples containing different parliamentary speeches?
2. Do these sampling methodologies produce samples that share topical similarities?
3. To what extent does the distribution of speakers from different parties vary between the two samples?
4. To what extent does the distribution of speeches between speakers vary between the two samples?

RQ1 and RQ2 focus on the potential added value of additional sampling techniques besides term-based sampling. RQ1 focuses on whether the headline-based sample adds additional documents to the analysis. Suppose the two samples are nearly identical (i.e. containing few or no documents unique to any of the two samples). In that case, the added value of additional sampling techniques is negligible. RQ2 instead focuses on whether the contents of the two samples are similar enough to contribute to the understanding of the same policy issue. If the two samples are too distinct in terms of the topics discussed in the speeches in each sample, there is reason to believe that they are, in fact, samples of different populations.

RQ3 and 4 instead focus on investigating the potential bias of the samples. Two comparative analyses are conducted. The first (RQ3) focuses on the distribution of party affiliations among speakers in the two samples, and the second (RQ4) investigates the general distribution of speeches among MPs (speakers).

The remainder of this paper has the following disposition. First, we present the policy development analysed in our samples: school marketisation in Sweden. After that, the two sampling methodologies utilised in this study are presented together with the resulting samples (answering RQ1). The following section presents the results of a semantic clustering analysis to answer RQ2. Subsequently, the analysis of the distribution of speeches in the two samples is presented, addressing RQ3 and 4. In the paper's final section, our analyses' conclusions are presented and discussed.

2. The case of school marketisation

Sweden has often been described as a prime example of a social democratic welfare-state regime characterised by de-commodification, universalism and egalitarianism (Esping-Andersen, 1990:28). Nevertheless, during the last four decades, the Swedish school system has gone through a transformation from one of the most centrally planned and uniform to one of the most marketised and decentralised school systems in the OECD area (cf. Fredriksson, 2009; Lundahl et al., 2013). This transformation coincides with a global trend of marketisation of welfare services in general and education in particular (Fuller & Stevenson, 2019). However, the Swedish school system's development is more radical compared with other welfare areas in Swedish society and school system reforms in different countries (Lundahl, 2016). As such, the trajectory of Swedish education policy from the 1980s until today can be described as a policy paradigm shift (Hall, 1993) seldom witnessed in welfare policy and education policy research (Fredriksson, 2009).

School marketisation is here defined as the sum of reforms of the Swedish education system that induced (1) internal marketisation, i.e., de-regulation and devolution of the management and control of

education (Lundahl et al., 2013), and (2) external marketisation, i.e. the creation of a quasi-market of schools competing for students and public funding through school choice institutions, a voucher system and profit motives (Lundahl, 2002). Important reforms in this process concern the decentralisation of education from the state level to local governments (Åhlin & Mörk, 2008) and the introduction of an educational voucher system that made it possible for parents to choose among public and independent schools (Carnoy, 1998).

The parliamentary debate surrounding school marketisation concerns a broad range of topics, including segregating effects of the marketised education system, profit margins of private companies running independent schools, and grading disparities between public and independent schools.

3. Method

3.1. Sampling methodology

Both samples are drawn from the open parliamentary data of the Swedish parliament[2]. This database contains text transcriptions of parliamentary speeches from the autumn of 1993. The term-based sample was created by filtering the parliamentary speeches in the open data of the Swedish parliament only to include speeches with occurrences of the words independent school (with alternate endings) and free school (with alternate endings)[3]. This filtering produced a dataset of 2875 parliamentary speeches given between 1993 and 2024. One potential strength of this sample is that the inclusion criteria guarantee that the speech concerns the main object of the debate on school marketisation, the free (or independent) schools. A corresponding drawback is that the debate in focus has a broader scope, encompassing also internal marketisation (as mentioned above). Therefore, the term based sample does not cover the whole parliamentary debate on school marketisation in Sweden.

The headline-based sample utilised the headlines created for sessions of plenary debates in the parliament. These headlines are transcribed from the agenda of the parliament and added to the parliament's protocol by the parliamentary administration. All headlines for parliamentary debate sessions 1993-2023 were collected. After that, headlines that did not regard education were filtered out[4]. Headlines relevant to school marketisation were identified by qualitatively analysing the remaining headlines. Examples of headlines deemed as relevant are:" Regulation of Public Grants to Independent Schools", "Education on Contract in Primary School", and "Increased School Choice - Expanded Options for Upper Secondary Education".

All speeches in debate sessions with relevant headlines were compiled in a dataset (N=1374). The benefit of the headline-based sample is that it can (and does) include speeches pertinent to the topic but does not mention the terms used in the term-based sample. As the criteria of inclusion/exclusion in this sample are made on the group level (the debate headline), the sample allows more lexical variations in the speeches. However, one clear disadvantage of this sampling methodology is that speeches made in general political debates in the parliament that are not given a topical headline are neglected.

The two datasets were then compared to identify potential overlaps regarding speeches appearing in both samples. These analyses utilised the unique ID attributed to each speech in the database. The overlap between the two samples was limited as only about a fifth (22%, N=777) of the speeches collected appeared in both samples. This indicates that the headline-based sample was reasonably distinct from – and contributed with additional data to – the term-based sample.

3.2. Semantic clustering

To determine the extent to which the two samples share topical similarities, we conducted a semantic clustering analysis using the licensed service Dcipher Analytics (www.dcipheranalytics.com) to compare the two datasets. This method has the advantage of generating consistent results and avoiding the problems of variability that traditional topic modelling such as Latent Dirichlet allocation suffers

[2] https://data.riksdagen.se/
[3] The exact search phrase was: "friskol* OR" fristående N1 skol*"
[4] Filtering based on the search terms: pupil, teacher, school, education, learning

from (e.g. Agrawal et al 2018). The datasets were pre-processed by merging the two samples into a combined dataset while retaining identifiers through tagging. Texts in the combined dataset were tokenised, with preprocessing steps including removing stop words, punctuation, and words shorter than four characters.

The first step in the pipeline was vectorisation, which was performed using the text-embedding-3-small model from OpenAI (https://openai.com/index/new-embedding-models-and-api-updates/). This process generates high-dimensional vectors within a 1,536-dimensional space. The model is pre-trained and deterministic, producing identical results for identical inputs across multiple runs.

The second step involved reducing the high-dimensional vectors to a lower-dimensional space to facilitate analysis. We used UMAP (Uniform Manifold Approximation and Projection; McInnes et al., 2018) for this. UMAP reduces noise and enables the efficient — and sometimes even feasible — identification of clusters in the data. Conceptually, the UMAP algorithm aims to preserve the structural relationships of the high-dimensional space in the resulting low-dimensional representation. Empirical evidence suggests that the optimal trade-off between structure preservation and usability is achieved in a 5-dimensional space (ibid.). The algorithm was trained on our dataset.

The third and final step was hierarchical, centroid-based clustering (Ding & He, 2002) performed in the 5-dimensional space to identify semantic clusters in the data. In the Dcipher tool, this process is implemented as a customised variant of agglomerative clustering, incorporating a feature called the "broadness level." This parameter automatically determines the optimal breadth of the cluster hierarchy, ensuring a balanced and meaningful data segmentation. In our analysis, we set the number of levels in the hierarchy to one (1).

3.3. Analysis of the distribution of speeches

Simple descriptive analysis was used to investigate (1) the distribution of party affiliations of speakers in the two samples and (2) the distribution of speeches made among speakers in the samples. These analyses drew on party affiliation and speaker identity metadata from the original parliamentary database. Further, linear regression analyses were conducted to estimate trends over time in party affiliation of speakers as well as debate intensity (number of speeches per year).

4. Results

4.1. Semantic clustering

To what extent is the content of parliamentary speeches identified through the two sampling techniques similar? First and foremost, the distribution of the identified topics across the two samples shows the topical outlay to be generally similar (see Tables 1 and 2 below). As an illustration, only 11% of the speeches in the word-search-based sample dealt with topics exclusively found in speeches from that particular sample. In other words, just shy of 90% of the speeches treat topics that are common to the two samples.

The emphasis or distribution of the content is also similar between the two samples. Tables 1 and 2 below present the ten most prominent topics for each sample (concerning the number of speeches identified to belong to the topics). For each sample, only two of the ten most prominent topics are exclusive to that sample. These topics are signified with an asterisk in the tables below.

The semantic clustering analysis indicates that despite limited overlap (22% of speeches) between the two samples, the clusters identified in the samples are overwhelmingly similar. We interpret these results as indicating that the parliamentary speeches identified through these diverging sampling methodologies do regard the same policy debate. In other words, we see these results as reassuring that we compare two samples of the same population.

Table 1.
The ten most prominent semantic clusters: Term-based sample

Topic	%	Cluster label	Distinct words
1	15.5	Pupils	teachers, the school, students, school, the students, the government
2	12.9	Marketization of welfare	welfare, private, companies, healthcare, quality, welfare, elderly, public
3	8.7	Type of school	independent schools, municipal schools, the school, students, think
4*	6.8	Labor market	Sweden, people, jobs, politics, Göran Persson, need, the government, speaker, country
5	6.2	Profit	profit, money, schools, activity, run, profits, think
6	5.2	Religious free schools	religious, denominational, independent schools, religious schools, focus, Muslim, religion, Sweden
7	5.2	Financing of education	independent, schools, the schools, the National Agency for Education, school, the municipality, students, grants
8	4.6	Disability	independent, the schools, municipal, children, the Green Party, choose, parents, disabilities
9*	4.4	Transparency of free schools	the principle of public access, Lotta Edholm, transparency, information, introduce, schools, the Liberals, school information
10	3.4	Immigration	students, school choice, newly arrived, children, school, choose, Swedish, schools, parents

Table 2.
The ten most prominent semantic clusters: Headline-based sample

Topic	%	Cluster label	Distinct words
6	9.5	Religious free schools	religious, denominational, independent schools, religious schools, focus, Muslim, religion, Sweden
7	9.2	Financing of education	independent, schools, the schools, the National Agency for Education, school, the municipality, students, grants
1	8.8	Pupils	teachers, the school, students, school, the students, the government
11*	7.3	Segregation	segregation, Ulf, Inger, Nilsson, choose, Lundberg, children, Bengt Silfverstrand
3	7.1	Type of school	independent schools, municipal schools, the school, students, think
10	7.1	Immigration	students, school choice, newly arrived, children, school, choose, Swedish, schools, parents
2	6.4	Marketization of welfare	welfare, private, companies, activity, healthcare, speaker, quality, welfare, elderly, public
5	6.2	Profit	profit, money, the activity, schools, school, the school, activity, run, profits, think
8	5	Disability	independent, the schools, municipal, children, school, schools, the Green Party, choose, parents, disabilities
12*	4.8	School voucher	school voucher, school, the municipalities, the Social Democrats, the Swedish National Audit Office, responsibility

4.2. Issue ownership

A political party holds ownership of a policy issue to the extent that it is perceived as most competent to handle it (Stubager, 2018). Further, the general expectation is that MPs from parties that perceive ownership over a specific issue engage in efforts to make that issue more salient in the media and parliamentary debates (Ivanusch, 2024).

In the following analysis, we will use the concept to denote a related but distinct property; issue ownership here means the extent to which the MPs of a party participate in the parliamentary debate on a specific issue. Thus, we simply analyse what parties dominate the parliamentary debate on school marketisation in Sweden. These analyses are performed separately on the two samples of parliamentary speeches to reveal if, and to what extent, there are differences in the depiction of issue ownership between the two.

The results of these analyses are presented in Table 3 below. Each party's MPs are compared across the period 1993-2024. However, two of the parties (the Green Party and the Sweden Democrats) have not been represented in parliament for the entire period (the first year of parliamentary representation is noted in the table). Further, the current parliamentary coalitions[5][6] are used as reference points for comparisons across the whole period, even though there has been substantial volatility in coalition formation during the period investigated. However, it is vital to utilise time-invariant categorisations to compare the two samples. This way, variations in coalition formation over time do not influence the comparison results. Also, this categorisation roughly resembles a division of members of parliament at the ideological median position throughout the period.

The findings indicate substantial differences in the engagement of MPs between different parties in the two samples. Within the term-based sample, MPs from the centre-right parties have made a majority (53.1 percent) of the parliamentary speeches in the debate on school marketisation. In comparison, MPs from the parties in the centre-left coalition have made a minority of the speeches (46.9 percent). In the headline-based sample, however, the tables are turned as MPs from the centre-left coalition parties dominate the debate on school marketisation, having made almost two-thirds (62.9 percent) of the speeches. This difference is not least driven by vast differences in debate presence of MPs from the largest parties in each coalition, the Social Democrats and the Moderate party. MPs from the social democrats made only 22.9 percent of the speeches in the term-based sample but 37.3 percent in the headline-based sample (+14.4 percentage points). On the other hand, MPs from the Moderate party made 24.3 percent of the speeches in the term-based sample but only 14.1 percent of the speeches in the headline-based sample (-10.2 percentage points).

Overall, the analyses indicate vast differences in their depiction of issue ownership in the parliamentary debate on school marketisation. These differences are crucial as they depict what coalition "owns" the debate. The two samples together indicate an evenly distributed debate activity across the left-right divide. However, interpreting each sample in isolation risks leading to faulty conclusions about the partisan dynamics of this policy debate.

Table 3

Party affiliation of speakers in the parliamentary debate on school marketization 1993-2024

Party/coalition	Headline	Term	Difference[7]
Left party	12.1%	8.7%	+3.4%
Social democrats	37.3%	22.9%	+14.4%
Green party (1994-)	5.9%	8.6%	-2.7%
Centre party	7.6%	6.7%	+.9%
Liberal party	11.4%	16.5%	-5.1%
Moderate party	14.1%	24.3%	-10.2%
Christian democrats	9.1%	7.3%	+1.8%
Sweden democrats (2010-)	2.5%	5%	-2.5%
Centre-left coalition	62.9%	46.9%	+16%
Centre-right coalition	37.1%	53.1%	-16%

[5] Center-right coalition: Liberal party, Moderate party, Christian democrats and Sweden democrats
[6] Center-left coalition: Left party, Green party, Social democratic party and Centre party
[7] Share of speeches in term-based sample subtracted from the share of speeches in headline-based sample

Issue ownership over time

Thus far, we have analysed the ownership issue of the totality of the parliamentary debate on school marketisation between 1993 and 2024. However, parliamentary debates on issues spanning long periods should be expected to have temporal dynamics. According to Hall (1993), policy issues follow paradigmatic patterns. As new ideas (frames of interpretations, solutions, knowledge) arise, old ones come out of fashion. Further, parliamentary politics follow temporal patterns, not least related to electoral cycles. Therefore, it is also necessary to investigate how issue ownership developed over time in the debate on school marketisation. Again, these analyses are conducted with the two samples separately to disclose potential variations in the results.

The analyses are presented in Figures 1 and 2 below. These scatter plots indicate the number of speeches within the debate on school marketisation conducted by MPs from the centre-left and centre-right coalition per year. Further, linear regression calculates the linear trends in issue ownership over time.

Starting with the term-based sample (Figure 1, below), the analyses indicate that MPs from the centre-right coalition parties dominated the debate in the first part of the investigated time-period. However, over time, the debate engagement increased among centre-left MPs to overtake the centre-right MPs in terms of debate activity in recent years. Lastly, the analysis indicates that debate activity generally increases over time, regardless of party coalition. Hence, the term-based sample paints the picture that the parliamentary debate on school marketisation becomes more intense towards the end of the time-period.

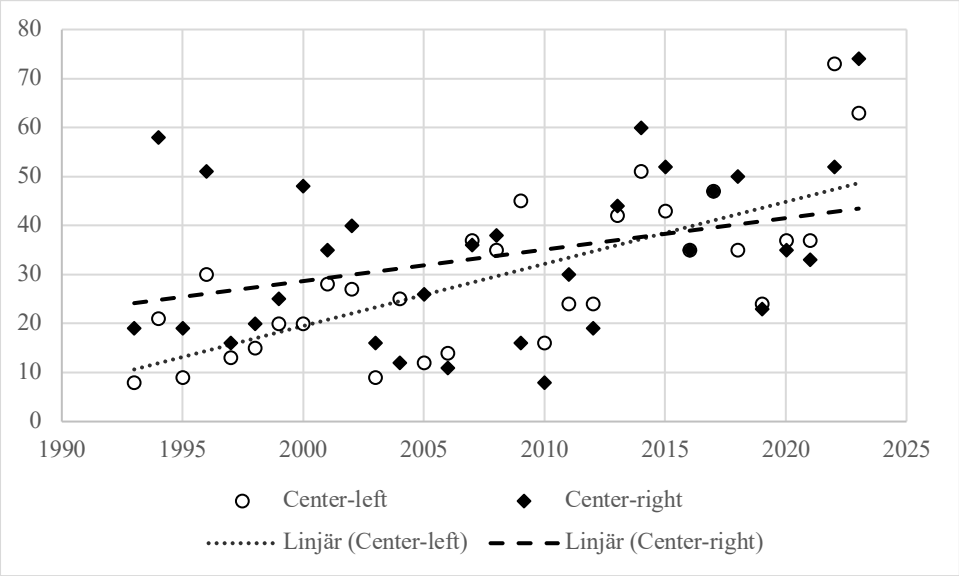


Figure 1. Number of speeches made by MPs in each coalition 1993-2024 - Term-based sample

Turning to the headline-based sample (Figure 2, below), we find a fundamentally different depiction of the parliamentary debate on school marketisation. This analysis indicates that early issue ownership among MPs from the centre-left coalition has decreased over time as MPs from the centre-right coalition have engaged more frequently in this debate. Moreover, the findings indicate that MPs from the centre-left coalition engage in this debate with decreasing frequency over time. Lastly, we find no indication of a general increase in debate activity over time (as in the term-based sample). Instead, the frequency of debate is found to be stable over time.

In summary, these analyses indicate vast differences between the two samples. On many accounts, analyses of these samples in isolation lead to contrasting, if not opposing, conclusions about the dynamics of the parliamentary debate regarding school marketisation in Sweden. These differences regard what coalition of MPs are most active in the date, how the dynamic of issue ownership changes across the investigated time-period, and how the activity of the debate develops over time.

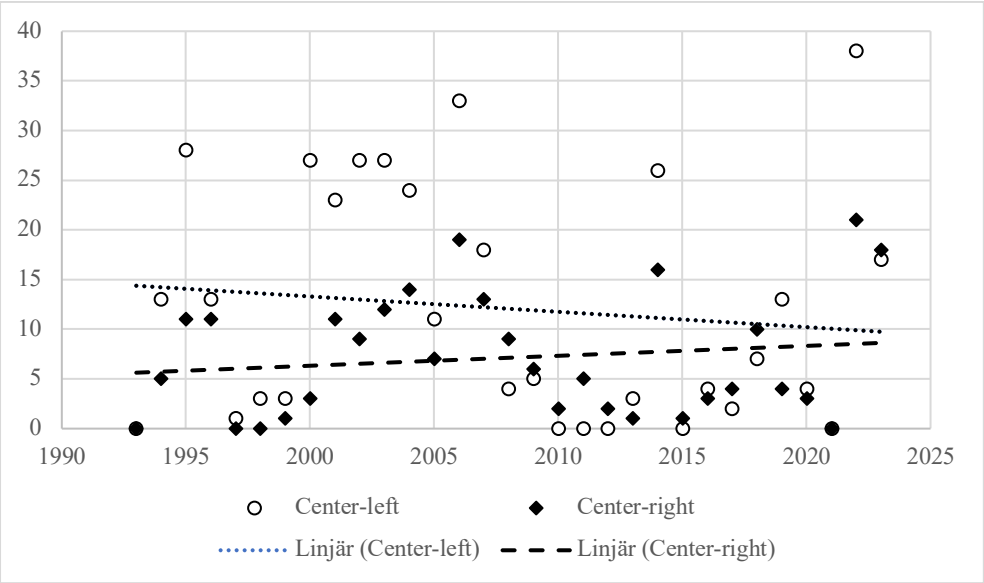


Figure 2: Number of speeches made by MPs in each coalition 1993-2024 – Headline-based sample

Speaker distribution

Parliamentary speeches among MPs tend to follow a power law distribution (Çöltekin et al., 2024). A small subset of MPs, often members in senior positions such as party leaders, tend to be highly active in parliamentary debates, while most MPs engage in parliamentary debates sparsely. Unsurprisingly, this is also the case for the debate on school marketisation. However, as we will demonstrate below, the two samples of parliamentary speeches relating to school marketisation give diverging indications of how strong this power-law distribution is, or in other words, how skewed the distribution of speeches per speaker is within this debate. The results of this analysis are presented in Figure 3 below.

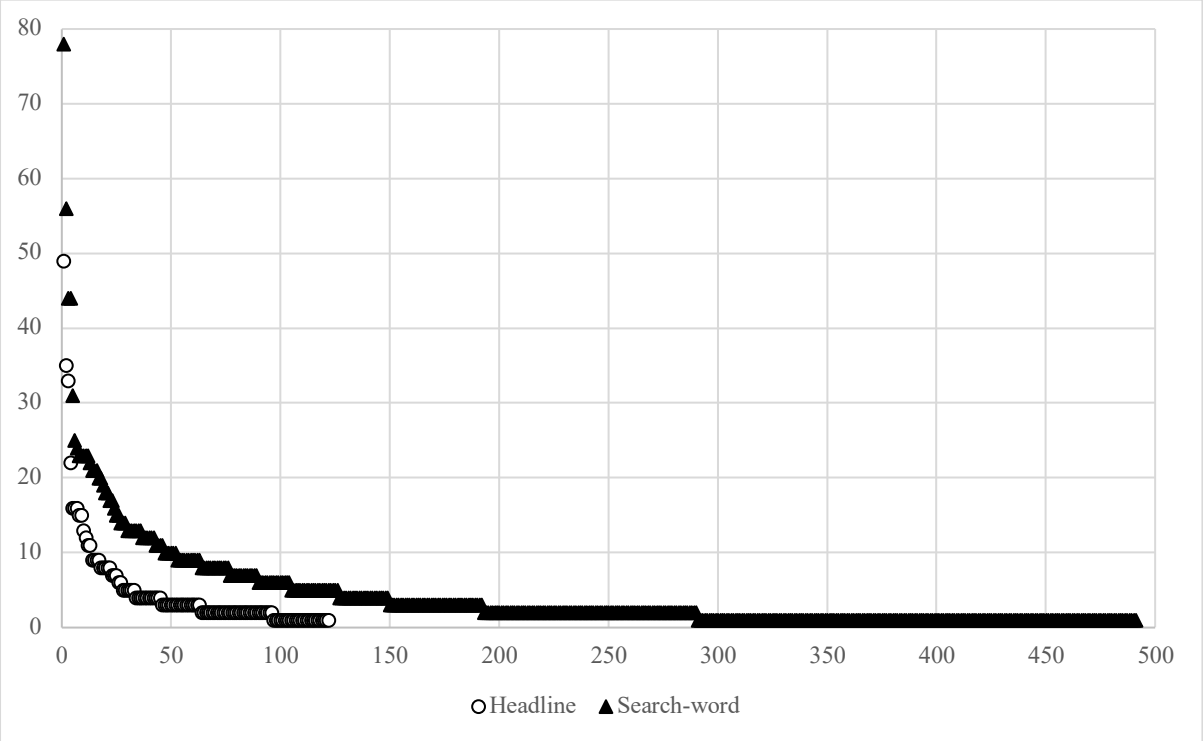


Figure 3: Distribution of speeches on school marketization between MPs.

The term-based sample is substantially more skewed regarding the speech distribution among speakers than the headline-based sample. In the term-based sample, 5 percent of the speakers make a majority of the speeches, while in the headline-based sample, 14 percent make a majority.

Within the term-based sample, the distribution of parliamentary speeches among Swedish MPs follows a clear power law or long-tail distribution. A small minority of MPs (5%) have conducted most of the parliamentary speeches within the sample, while the overwhelming majority of MPs have made limited contributions to this debate.

The headline-based sample shows a somewhat different picture of the distribution of speeches in the parliamentary debate on school marketisation. Within this sample, a majority of the speeches are given by 14% of the MPs. While still indicating a significantly skewed distribution of speeches, the analysis of the headline-based sample still signifies more broadly distributed engagement in the debate on school marketisation within parliament. Further, it is not the same MPs that make up the small group that is most actively participating in the debate on school marketisation in the two samples.[8] Rather, the two samples to a great extent identify different "elite debaters". This result marks another aspect where analyses of the respective samples would lead to diverging conclusions.

5. Discussion

In this study, we have addressed the potential issue of sampling error in research on subsets of parliamentary text corpora. We suspect that the risk for such sampling errors is more significant in studies of the parliamentary debates around specific policy issues. Our study has focused on illustrating the potential biases associated with particular sampling techniques through analysing different samples of parliamentary text data (parliamentary speeches) relating to the marketisation of the Swedish education system.

Our analyses find that diverging sampling methodologies can be complementary because each method adds substantial quantities of unique documents to the dataset. We found limited overlap between the two samples that we analysed. In more concrete terms, this indicates that a term-based sample on its own excludes substantial quantities of parliamentary speeches that are relevant to the topic (in this case, school marketisation), but that does not include the specific terms used in the sampling frame (in this case independent schools and free schools). The same is true for the headline-based sample. Using such a sampling methodology in isolation would exclude a significant number of speeches relating to school marketisation, which are given in debates not topically tied to this policy issue (e.g., general political debates).

Further, our results indicate that the diverging sampling methodologies produced documents with similar topical content. The semantic clustering analyses identified overwhelming similarities in the semantic content of the two samples. This result leads us to conclude that diverging sampling methodologies can be used in tandem without resulting in a too heterogeneous dataset.

Lastly, our analysis of the distribution of speeches between parties and speakers identifies substantial differences between the two samples. Our analysis of the term-based sample finds that MPs from the centre-right parties have been most active in the debate on school marketisation. On the contrary, the analysis of the headline-based sample finds that MPs from the centre-left parties dominate this debate. Further, the analyses of the two samples differ regarding the development of issue ownership and debate activity over time, as well as how uneven activity in this debate is distributed among all MPs in parliament.

Overall, we illustrate that these samples paint remarkably different images of the parliamentary debate on this issue. This leads us to conclude that decisions about drawing samples from parliamentary text corpora can fundamentally influence the conclusions of subsequent research.

One can draw a parallel between the challenges of sampling data from corpora of parliamentary databases and what Labov (1972) termed the Observer's Paradox in linguistics. Labov noted that in observing language, researchers inevitably influence the naturalness of the speech they wish to study. Similarly, data selection methods—whether through specific keywords, subject terms, or filters—shape

[8] There is a 52% overlap in the identity of the most active MPs across the two samples. Conversely, 48% of the most active MPs are divergent between the two samples.

the very representation of the policy area we aim to examine. The dataset we construct is not a neutral reflection of the political discourse but is shaped by the tools we use to extract it. This resonates with the Observer Effect in physics, where measuring a system alters its state—such as how measuring tire pressure changes the pressure. In the case of our study, each selection method inherently introduces biases or filters that affect the dataset and, consequently, the conclusions drawn about the policy domain.

Acknowledging these effects is crucial, as it underscores the importance of methodological transparency and critical reflexivity when conducting research using large-scale text data. Further, we recommend that researchers tasked with drawing a sample from parliamentary corpora reflect the parliament's dealings with a specific issue and utilise multiple sampling techniques in tandem. Similar to a repeated measurements approach to validate the reliability of a particular measure (De Vet et al., 2006), a diversified sampling methodology can compensate for potential sampling errors in each methodology. Further, such an approach creates opportunities to validate interpretations and conclusions across diverging sub-samples.

Acknowledgements

This work was funded by the Swedish Research Council (Grants 2022-04606 and 2023-04477). We thank Tomas Larsson at Dcipher Analytics for his generous support of the analyses presented in this paper.

References

Agrawal, A., Fu, W., Menzies, T. (2018) What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, 98, 74-88

Åhlin, Å., & Mörk, E. (2008). Effects of decentralization on school resources. Economics of Education Review, 27(3), 276-284.

Carnoy, M. (1998). National voucher plans in Chile and Sweden: Did privatization reforms make for better education?. *Comparative education review*, 42(3), 309-337.

Çöltekin, Ç., Kopp, M., Meden, K., Morkevicius, V., Ljubešić, N., & Erjavec, T. (2024). Multilingual Power and Ideology Identification in the Parliament: a Reference Dataset and Simple Baselines. arXiv preprint arXiv:2405.07363.

De Vet, H. C., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. Journal of clinical epidemiology, 59(10), 1033-1039.

Ding, C. & He, X. (2002). Cluster Merging and Splitting in Hierarchical Clustering Algorithms. Proceedings - IEEE International Conference on Data Mining, ICDM. 139- 146. 10.1109/ICDM.2002.1183896.

Esping-Andersen, G. (1990). The three worlds of welfare capitalism. Princeton University Press.

Fredriksson, A. (2009). On the consequences of the marketisation of public education in Sweden: For-profit charter schools and the emergence of the 'market-oriented teacher'. *European Educational Research Journal*, *8*(2), 299-310.

Fuller, K., & Stevenson, H. (2019). Global education reform: understanding the movement. *Educational Review*, 71(1), 1-4.

Hall, P. A. (1993). Policy paradigms, social learning, and the state: the case of economic policymaking in Britain. *Comparative politics*, 25(3): 275-296.

Ivanusch, C. (2024). Issue competition in parliamentary speeches? A computer-based content analysis of legislative debates in the Austrian Nationalrat. *Legislative Studies Quarterly*, 49(1), 203-221.

Isoaho, K., Moilanen, F., & Toikka, A. (2019). A Big Data View of the European Energy Union: Shifting from'a Floating Signifier'to an Active Driver of Decarbonisation?. *Politics and Governance*, 7(1), 28-44.

Labov, William (1972). Sociolinguistic Patterns. Philadelphia: University of Pennsylvania. p. 209. ISBN 0631177205.

Lundahl, L. (2002). Sweden: decentralization, deregulation, quasi-markets-and then what?. *Journal of education policy*, 17(6), 687-697.

Lundahl, L., Arreman, I. E., Holm, A. S., & Lundström, U. (2013). Educational marketization the Swedish way. *Education inquiry*, 4(3), 22620.

Lundahl, L. (2016). Equality, inclusion and marketization of Nordic education: Introductory notes. *Research in comparative and international education*, 11(1), 3-12.

Magnusson, M., Öhrvall, R., Barrling, K., & Mimno, D. (2018). Voices from the far right: a text analysis of Swedish parliamentary debates.

McInnes, L. & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 10.48550/arXiv.1802.03426.

Müller-Hansen, F., Callaghan, M. W., Lee, Y. T., Leipprand, A., Flachsland, C., & Minx, J. C. (2021). Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling. *Energy Research & Social Science*, 72, 101869.

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12), e0168843.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1), 112-133.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3), 705-722.

Stubager, R. (2018). What is issue ownership and how should we measure it?. *Political behavior*, 40(2), 345-370.

Voigt, J., Litvyak, O., & Lampoltshammer, T. J. (2024). Analysing Parliamentary Discourse on Forestry Management and Timber Industries in Austria and Germany using BERTopic.

Augmented Analysis of Parliamentary Debates: The Word Embedding and Context-sensitive Approach of the SweTerror Project

Leif-Jöran Olsson[a], Daniel Brodén[a], Magnus P. Ängsal[a], Mats Fridlund[a] and Patrik Öhberg[a]

[a] *University of Gothenburg*

Abstract

This paper delves into the SweTerror project's use of word vectors to enhance the analysis of parliamentary debates concerning terrorism in Sweden during the electoral periods from 1968 to 2018. We focus on how word embeddings capture semantic shifts and the evolving context of key concepts like terror, terrorism, and extremism over time. By combining these computational tools with enriched metadata and document annotation as well as a mixed-methods and context-sensitive approach, we trace temporal changes in parliamentary discourse. The study demonstrates how generating vectors for distinct periods, such as electoral periods or parliamentary years, provides nuanced insights into conceptual transformations, including the introduction of the modern use of the concept of terrorism in the 1970s and the impact of the term violence-affirming extremism in the context of Islamism in the 2010s. We conclude by stressing that this approach allows for a more sophisticated analysis of linguistic and discursive patterns within Swedish parliamentary discourse.

Keywords

Terrorism, parliamentary data, research infrastructure, mixed methods, text mining

1. Introduction

In recent years, there has been a rapidly growing interest in data-intensive research on parliamentary collections and a push towards further development and standardisation of parliamentary infrastructures (La Mela et al. 2022: 3–4). This includes a range of national and cross-national initiatives, including the Pan-European Parla-CLARIN initiative (https://github.com/clarin-eric/parla-clarin). In Sweden, significant work is being done within the SWERIK infrastructure, which is developing an annotated corpus of Swedish parliamentary debate since 1867. However, research on Swedish parliamentary datasets can still be characterised as largely exploratory and de-contextualised. When it comes to Natural Language Processing (NLP) studies of terrorism they have been central to developing the field of Information Extraction through the DARPA financed Message Understanding Conferences (MUC) that in 1991 and 1992 used news reports of Latin America terrorism as material for developing improved methods for information extraction (Grishman & Sundheim 1996; Conlon, Abrahams, & Simmons 2015). However, in general when it comes to such studies employing NLP and machine learning methods they have primarily been focused on methods development and generally not been informed by political domain knowledge, while historical research has been primarily driven by the application of common statistical measurements rather than enhanced through

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, Reykjavik, Iceland, May 28, 2024.
EMAIL: leif-joran.olsson@gu.se (L-J Olsson), daniel.broden@gu.se (D. Brodén); magnus.pettersson.angsal@gu.se (M. Ängsal), mats.fridlund@gu.se (M. Fridlund), patrik.ohberg@gu.se (P. Öhberg)
ORCID: 0000-0001-7107-4101 (L-J Olsson), 0000-0002-5914-1516 (D. Brodén); 0000-0001-5996-5067 (M. Ängsal), 0000-0002-5759-0027 (M. Fridlund), 0000-0002-7433-8552 (P. Öhberg)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

data modelling and contextual understanding of the datasets (see Rouces et al. 2019; Ohlsson et al. 2022; Brodén et al. 2023; Jarlbrink & Norén 2023).

Similar to previous historical research on Swedish parliamentary datasets, this paper focuses on conceptual history (see Ohlsson et al. 2022; Brodén et al. 2023; Jarlbrink & Norén 2023), but its primary aim is not historical but to provide a methodological contribution to Digital Humanities (DH) by presenting the augmented methodological approach to analysing Swedish parliamentary discourse, developed within the Terrorism in Swedish Politics project (2021–2025, SweTerror) (Edlund et al. 2022). Drawing on a mixed methods approach, we discuss a set of context-sensitive analyses of the Swedish parliamentary record, integrating Language Technology (LT) and methods from political science, history of ideas and linguistics, among other disciplines. Our discussion, therefore, connects to the broader debate within DH) about the critical need to engage with the contextual complexities of text mining large-scale archival collections. As digital historian Jo Guldi (2023) argues, a focus on what makes text mining accurate and robust can only take data-driven analysis of large-scale collections so far. Without applying a contextual understanding of the materials, the results are likely to raise more questions than they answer (see also Bode 2018).

1.1. The complex notion of terrorism

In the following we outline our augmented methodological approach to the study of Parliamentary discourse through a case study of terrorism in Swedish parliament debates 1967–2021. The choice of terrorism was not arbitrary but was constitutive of the SweTerror-project as it was chosen as the project's focus of empirical research as it was a central domain expertise of several of the humanities and social science scholars who took part in the project. That said, we argue that our methodological approach developed within this terrorism-focsued project is not particular to the Swedish Parliament and the study of terrorism discourse but should apply equally well to other national parliaments and to other topics of parliamentary debate such as conservatism, feminism, taxes, war, health care, democracy, etc. etc.

However, before we proceed, it is important to address the complexity of the concept of terrorism, which lacks a generally accepted definition and can be described as an 'essentially contested concept' (Weinberg et al. 2004). The term has been used by various actors in conflicting ways, often to label their opponents' actions as illegal or illegitimate, and it can be understood, in part, as a social and cultural construct (Brulin 2015).

Terrorism acquired much of its contemporary meaning in the late 1960s and early 1970s. While the term dates back to the French Revolution and the Reign of Terror, the modern conception of terrorism became associated with the emergence of a transnational threat, often referred to as 'international terrorism', which included activities of Palestinian and left-wing militants, among others. This term brought together various violent tactics – such as assassinations, bombings and hostage-takings – that had previously been used in armed insurgencies, categorising them under a single label: 'terrorism'. Rather than being viewed as political violence confined to local conflicts, particularly in the Global South, terrorism began to be interpreted as a specific transnational threat, epitomised by Palestinian aircraft jackings starting in 1968 and the attack against the Israeli Olympic team at the 1972 Summer Games in Munich. International terrorism was seen as undermining the foundations of modern societal order, affecting not only the safety of authorities and civilians but also diplomatic relations and global communication networks, especially air traffic. As Stampnitzky (2013: 27) writes, "The concern was with violence out of place – spilling over from local conflicts into the international sphere". In the background were policy efforts in the 1960s by the U.S. government to curb guerrilla warfare in Latin America by criminalising and delegitimizing political violence, which were later expanded to a broader range of actors (Zoller 2018: 26–71).

To navigate this complexity, the SweTerror project employs historical contextualization and partly views terrorism as a product of discursive practices. We understand discourse as a virtual collection of intertextually linked utterances related to a macro-topic and as a social practice that shapes and is shaped by societal structures of power and agency (Reisigl & Wodak 2009). Specifically, we approach the discursive formation of terrorism in the Swedish Parliament through the

lens of 'framing', that is conceptual, habitually formed frameworks of meaning that are constructed through concrete linguistic acts (Entman 1993).

Although Sweden had experienced a few isolated acts of politically motivated violence before the early 1970s, it was around that time that the word 'terrorism' began to appear more frequently in debates within the Swedish Parliament (Riksdag) and in broader public discourse (Hansén 2007; Brodén et al. 2023). At that time, the concept of terrorism first became associated with political acts of violence carried out by Croatian separatists, including the killing of the Yugoslav ambassador in 1971 and the hijacking of a domestic flight in 1972. This was followed by two major incidents linked to the Red Army Faction (RAF) – the West German embassy siege and a botched kidnapping plan targeting former Social Democratic Minister Anna-Greta Leijon. In the 1980s, terrorism remained a concern, partly due to the controversy surrounding Kurdish nationalists associated with the PKK (Kurdistan Workers' Party) and the unsolved murder of Prime Minister Olof Palme in 1986. Notably, in Sweden, terrorism was primarily associated with foreign groups for a long time, and the far-right and racist attacks perpetrated by domestic Swedes in the 1990s was not necessarily framed as terrorism. After 2001, Sweden, like many other countries, experienced several acts of political violence connected to Islamist extremism, including the truck attack in downtown Stockholm in 2017 that left five people dead.

1.2. Purpose and aims

This paper outlines the SweTerror project's contextualised analysis of parliamentary debates on terrorism during the electoral periods covering our time period 1968 to 2018, utilising a custom-made dataset and applying word vectors in the form of word embeddings. It presents a LT approach firmly grounded in research questions and domain expertise from the humanities and social sciences (HSS), drawing on fields such as terrorism studies, political science, sociolinguistics, digital humanities and digital history.

The paper has a two-step structure. First we focus on how the SweTerror project combines enriched document annotation and word vectors to trace various aspects of the discourse on terrorism in the Swedish parliamentary debates during the electoral periods 1968–2018. We discuss our customisation of the SWERIK dataset and utilisation of word embeddings as a methodological approach, emphasising how we apply a contextualising understanding of parliamentary data. We then highlight the analytical potential of our approach to unpacking the framing of terrorism in the Riksdag through a case study of two key semantic shifts in the dominant Swedish discourse on terrorism: first, the shift from 'terror' to 'terrorism' in the early 1970s, and second, the shift in 2014–2015, when the novel Swedish expression *våldsbejakande extremism* ('violence-affirming extremism') was introduced in the context of radical Jihadism (Wahlström 2022).

Deploying word vectors (Yao et al. 2017), which represent the contextual closeness of words (semantically similar words represented by numerically similar vectors), we can track changes in word usage over time. A key question is to what extent, and by whom – Members of Parliament, MPs, and political parties – the aforementioned terms have been used to denote different, similar or related, activities or stances. By comparing similar parliamentary debate contexts, we can predict words with similar meanings and identify which related words were used in similar contexts earlier or later. Combining this approach with aggregated metadata on the gender and party affiliation of MPs allows us to examine sociolinguistic variables related to the framing of terrorism in the Riksdag.

2. Parliamentary dataset and context

The SweTerror project utilises an enhanced and curated subcorpus of the corpus from the parliamentary minutes provided by the SWERIK infrastructure. (latest version 1.2.0 is being processed and analysed, but the presented results are mainly for version 0.14.0 which then became public release v1.0.0). (https://github.com/swerik-project/swerik-project.github.io). SWERIK unifies and structures the Swedish parliamentary records 1867–2023 into a single corpus with a standardised data format. The speech records are delivered in the ParlaClarin format (Erjavec & Pančur 2021),

which is a specification of the Text Encoding Initiative (TEI) XML format (TEI Consortium 2007)). The SWERIK infrastructure employs computational and machine-learning approaches to segment, correct, enrich and annotate the data. It also compiles and curates information on MPs, ministers, speakers of the Riksdag and governments from multiple sources into a structured metadata database. Additionally, it implements a process to iteratively revise and improve the corpus over time and provides statistical estimates of multiple quality dimensions of the corpus and gold standard test sets. The corpus is version controlled using semantic versioning principles (Yrjänäinen et al. 2024: 2–3, 4, 8).

The longitudinal corpus used in the SweTerror project consists of approximately 4M tokens per parliamentary year. SweTerror's LT expert, Olsson, has customised and enhanced the SWERIK dataset to better align with our research needs. This work involves performing additional quality control tasks, such as identifying missing entries and debate protocols. The process is iterative with ongoing communication with the SWERIK team (Olsson being a member of its advisory board). This collaboration allows SWERIK to address gaps identified by our team and update their dataset accordingly. As of the latest version (1.2.0), the SWERIK corpus is, as far as can be determined, complete in this respect. (From an infrastructure perspective, the SweTerror corpus, including data and workflows, is integrated into the workflows of the national research infrastructure Språkbanken Text (SB Text) and the European CLARIN ERIC infrastructure.) However, ongoing revisions of the SWERIK corpus mean that there are still some structural issues with the text data. Notably, the structure of the speeches in SWERIK is not yet reintroduced with 100 percent accuracy. Furthermore, the SWERIK corpus – and by extension, our corpus – contains not only debate speeches but also 'secondary' debate text, such as headings, voting results and comments. For instance, the term 'terrorist organisation' (*terroristorganisation*) appears 566 times in the corpus (version 1.2.0), but only 461 times in the actual speeches. Similar issues also affect the metadata. While these discrepancies have a slight impact on our high-level analysis, they do not disrupt the analytic workflows we are developing in the project.

SweTerror aims to apply a contextualising understanding of parliamentary data. A key aspect of this approach, distinct from many other parliamentary datasets, is our decision to group the debates in the SweTerror corpus by parliamentary year (autumn–summer), rather than calendar year, in order to accurately reflect the Riksdag's mandate period (with SWERIK adopting the same approach). One important rationale for this is that changes of government during election years (mid-calendar year) impact the political dynamics of parliamentary activities. For instance, we have previously shown that governmental position plays a major role in MPs' motion writing on the topic of terrorism (Brodén et al. 2023). Moreover, as the speeches in our dataset are automatically assigned a Persistent Identifier (PID) for each speaker, we utilise SWERIK's metadata on MPs (e.g. name, party affiliation, gender, regional representation) in conjunction with other calculated metadata. This enables analysis of structural differences in the debates on terrorism at the party level, including government versus opposition, speech volume measured as token percentage and gender differences. Additionally, we are in the process of integrating 'seniority' as an analytical factor, using measurements based on the MP's age, years in parliament, position (e.g. ministerial roles, parliamentary committee membership, committee chairs). It is also worth noting that our dynamic time periods (see "Word vectors and methodological approach" below) incorporate a range of factors, such as parliamentary year, electoral periods and eras defined by influential MPs in the terrorism debate. These periods can be explored for continuities and discontinuities in the discourse.

Furthermore, SweTerror places greater emphasis on the genre of parliamentary debates than previous studies on Swedish parliamentary data. They have typically focused on technical, formalistic and infrastructural aspects of the documentary record, such as issues related to OCR quality, metadata, and the post-speech editing involved in the transcription of minutes (see Rødven-Eide 2020; Norén & Jarlbrink 2024). Parliamentary debates are widely regarded as a crucial arena for position taking by parties and parliamentarians (Proksch & Slapin 2012). In the Swedish Parliament, which since long has been characterised by a left-right dimension, MPs are free to speak in the parliament. There are no formal rules that prevent them from taking part in debates. The parliamentary floor is therefore a vital forum for MPs to present their own and their parties' positions on current issues (Bäck & Debus 2016). Moreover, MPs can submit an interpellation, a detailed question directed at a Minister responsible for a specific policy area, who must then provide a public response, and every

Thursday there is 'question time' in the Parliament where four Ministers visit the parliament to answering question (www.riksdagen.se). These tools give MPs, from both the government and the opposition, ample opportunities to take stand on different topics. While the government and its MPs present their policies as beneficial, the content of the debates from the opposition tend to revolve around efforts to reframe the government's narrative rather than reinforcing it.

3. Word vectors as methodological approach

In the SweTerror project, we combine enriched document annotation with word embeddings to create 'temporal lenses' on the Swedish parliamentary debates during the electoral periods during our period of study 1967–2021. By incorporating word vectors, primarily in the form of word embeddings (Mikolov et al. 2013), we enhance our text mining capabilities, allowing us to go beyond simple keyword searches or frequency counts. This approach enables us to identify linguistic patterns and key topics without the need for predefined categories. Word vectors provide a robust and nuanced way to explore conceptual similarities and transformations over time by embedding words in a high-dimensional space. In this space, the relative positions of different words reflect their semantic relationships, with words that have similar meanings being represented by vectors that are closer together in the multidimensional space. This approach of ours aligns itself with similar uses of word embeddings to study political change within recent research in conceptual history (Wevers & Koolen 2020; Verheul et al. 2022) some of which also includes terrorism as part of their study objects (Marjanen et al. 2018; Lorella & Verheul 2020; Eijnatten & Ihalainen 2022).

Since word vectors require simplification for visualisation and computational efficiency, we employ t-distributed stochastic neighbour embedding (t-SNE), as well as umap, umap2, and Principal Component Analysis (PCA) for certain applications, including sentiment analysis, to reduce their dimensions. For vector normalisation, we ensure that the word vectors have unit length, which aids in comparing similarities between vectors. Based on our iterative experiments with the parameters for generating usable models from our lemmatized datasets (all parliamentary years and electoral periods) including a baseline of the 10 parliamentary years preceding our datasets, we settled on reducing the vector dimensions to 300. This choice strikes a balance between capturing sufficient semantic information and managing the computational resources available to us. Our experiments demonstrated that higher dimensions led to overfitting and increased processing times, while 300 dimensions preserved meaningful relationships between words without making the model unnecessarily complex or slow.

Turning to our analytic workflows surrounding the annotation pipelines, we continuously aggregate and analyse the outputs in an iterative process, with each layer of annotation undergoing at least one manual evaluation. Our word2vec pipeline enables us to generate word embeddings that capture the specific language, topics and context unique to parliamentary discourse. This pipeline also supports subsequent analysis and applications, such as detecting linguistic patterns in the debates and analysing shifts in parliamentary language over time, using quality assessed metadata about the speakers. A key feature of the word2vec pipeline is that our models learn word embeddings by both predicting the surrounding context of a word (skip-gram) and by predicting a word given its surrounding context (Continuous Bag of Words, CBOW). The FastText pipeline, where we also use both skip-gram and CBOW, on the curated SWERIK corpus. In terms of fine-tuning the analytic workflows, the evaluation results are analysed after each iteration. While we experimented with adjusting model parameters during earlier iterations, these adjustments and retraining on more refined data were primarily aimed at improving the quality of the word vectors. Regarding continual learning, all pipelines are designed to continually update the models as new versions of the SWERIK parliamentary data become available.

The trained word vectors are used for various LT tasks, such as clustering topics in speeches, identifying key terms in specific debates and tracking changes in language use over time across multiple protocols. Notably, the generation of word vectors is performed for all dynamic time periods, not just once. This means that word vectors are recalculated for each distinct time period (such as electoral periods and government periods), allowing the models to capture shifts in language use and

context over time, rather than relying on a single, static set of word vectors. This approach ensures that temporal changes in meaning and usage are accurately reflected in the analysis.

Part of our work involves the intrinsic evaluation and validation of how well our word2vec models perform in tasks such as generating accurate word vectors and predicting word associations. We use cosine similarity to compare the similarity between selected word vectors, assessing how well the model captures semantic relationships. For instance, terms like 'terrorism' and 'terrorist' should be expected to be very close in vector space. We have also tested the model's performance on analogy tasks, such as schematically determining how analogous 'left-wing terrorism' is to 'terrorism' (estimated by subtracting the 'right-wing terrorism' vector from the 'terrorism' vector). In addition to intrinsic evaluation, SweTerror focuses on extrinsic evaluation, relying on the domain expertise of the HSS research team members to assess how well the word vectors perform in real-world application. We also continuously evaluate the usefulness of the word vectors in downstream LT tasks, such as text classification, Named Entity Recognition (NER) and topic modelling on documents in our corpora. Our results have been integrated into the general annotation pipeline Sparv (https://spraakbanken.gu.se/en/tools/sparv (Borin et al. 2016) of the Nationella Språkbanken. In the specific parliamentary context, we also evaluate how well the word vectors assist with tasks such as political stance detection, speech classification and sentiment analysis. Tested against other measures of, e.g. bias (Islentyeva 2021). Once trained, the word vectors are stored in text format, accessed via API with the SWERIK corpus, and imported into the SweTerror vector database with additional metadata for deployment and use.

In our document-based analysis, an important task is to move beyond word-level semantics and model and compare specific debate protocols. At this analytical level, instead of using vectors for individual words, we employ paragraph vectors (Mikolov et al. 2013) to generate fixed-length vector representations for entire paragraphs or documents. This approach allows us to perform classification and Named Entity Recognition (NER) at the paragraph level, enabling us to navigate, identify and retrieve similar sections or related content across different paragraphs based on their textual similarity and network connections.

For the metadata associated with the word and paragraph vectors, we have added rich metadata for storage in a vector database. This enriched metadata can be divided into direct and indirect categories. The direct metadata includes the version of SWERIK used for each model, the date intervals covering the utterances used for the dynamic time period (such as government period, electoral period and pre-election periods), the algorithm employed (skip-gram or CBOW), the type of word vector (Word2Vec or FastText), classification (such as NER and sentiment analysis), word type/token count and percentage for the base form within the chosen time period. The indirect metadata includes linking to protocol ID, utterance ID, classification text type, sentiment and links to sentence, paragraph or document vectors where a token occurs. Using the utterance ID, we can link to all information associated with persons identified by the corresponding speaker ID, such as name, age, political party, ministerial role, chair position and government status (with relevant dates). Other important indirect metadata related to party affiliation includes government constellation, opposition status and party coalitions blocks (such as political agreements between different parties).

4. Case Study: Vectors of violence

To illustrate our analytical implementation of word vectors we turn to a case study focused on the complex notion of terrorism, as the term 'terrorism', along with its derivatives and compounds, tends to exhibit considerable fluidity in meaning. This ambiguity is a central theme in both orthodox and critical terrorism studies, which explore how to define terrorism and distinguish it from related concepts such as political violence, extremism, terror and insurgency. Notably, this case study is based on an analysis of our SweTerror corpus based on version 0.14.0 of the SWERIK protocols. This SWERIK version later became public release version v1.0.0 (the differences primarily concern metadata). Version 1.2.0 is currently being implemented, and we have not observed any significant discrepancies regarding terrorism-related words between the different versions that should affect our analysis.

A key part of the SweTerror project is analysing the conceptual changes and continuities surrounding the term terrorism and related nouns, as reflected in parliamentary discourse. This case study aims to map out such discourse semantic patterns using the SweTerror data-set. Specifically, we seek to compare the usage of the lexical terms *'terror'*, *'terrorism'* and '(*våldsbejakande) extremism*' ('(violence-affirming) extremism'). To highlight their discourse semantic properties in contrast, we perform a diachronic comparison of these units by analysing their semantic similarity and discursive closeness within the data, using word vectors. By comparing semantically related vectors over time, we can identify potential discursive shifts in the framing of terrorism. A key focus is whether, and how, the vector for *terror* changes when the modern usage in Swedish of the word *terrorism* emerged and stabilised in the early 1970s.

4.1. Similarity of word pairs

As for the measurement of similarity, we first examine it by comparing the diachronic similarity of selected word pairs. In the second step, we compare the top 5 most similar words related to the base forms *terror*, *terrorism* and *extremism*, segmented by the terms of office of Swedish governments. Our findings indicate that the calculated similarity between *terror* and *terrorism* remains relatively stable over time. However, the similarity ranges significantly, from a low of 0,05 (in 1988) to a high of 0,49 (in 2016), with 1,0 representing identical semantic embeddings in the data. These fluctuations reflect noteworthy highs and lows between both individual parliamentary years, governmental periods and electoral periods. Despite these variations, (Figure 1), there is a general trend of increasing similarity between *terror* and *terrorism* over the time studied period. By comparing CBOW and skip-gram algorithms we control if term frequency is affecting similarity.

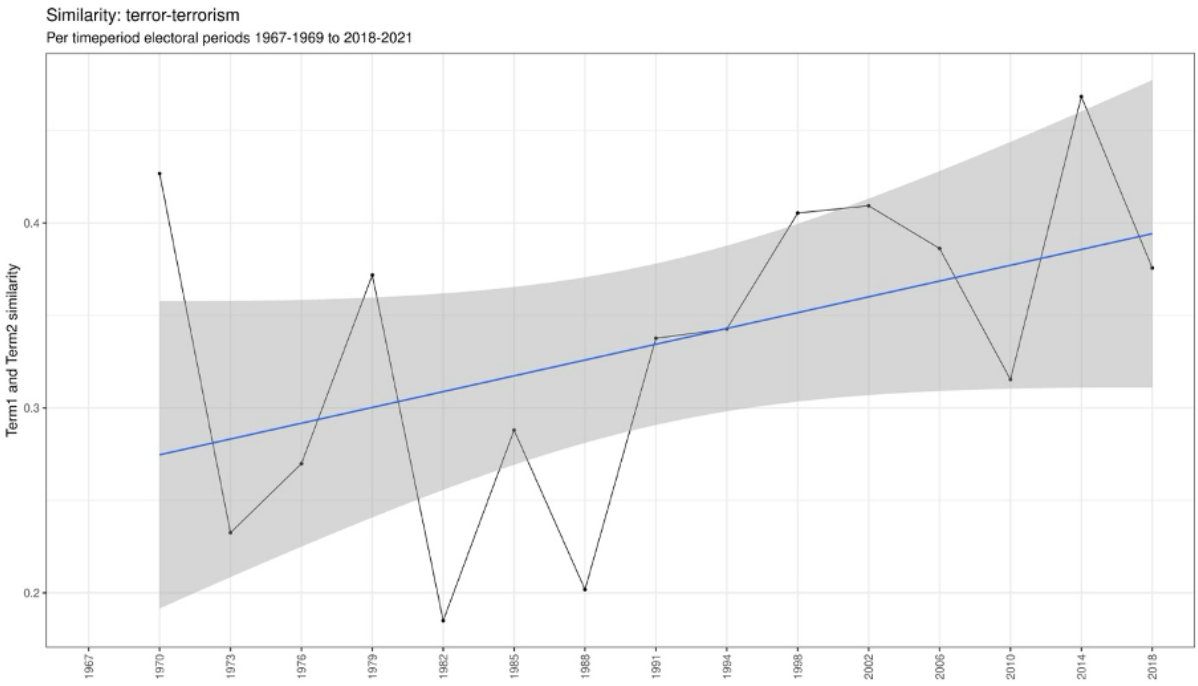


Figure 1: Similarity between the word vectors for *terror* and *terrorism* in the Swedish parliamentary debates across electoral periods from 1967–2021, based on the SweTerror corpus version 0.14. The blue regression line indicates a steady increase in similarity, while the black line connects the data points to make it easier to identify where uninterrupted data points occur for the electoral periods. The grey area represents the 95 percent confidence interval.

It is worth noting a few sudden drops and increases in the data. Interestingly, until 1970 the similarity calculation generates no comparison, because the lexical unit *terrorism* had not yet significantly

entered parliamentary discourse. While the similarity rate sets relatively high, it subsequently drops before rising again. The sharp rise in 1973 can likely be attributed to the beginning of parliamentary debates on the first Swedish Terrorism Act. As seen in Figure 1, the similarity between *terror* and *terrorism* reaches a relatively high level (0,42 in 1973). Another notable rise occurs in 2001, when the similarity jumps from 0,09 in 2000 to 0,45. This shift can likely be linked to political developments in the form of the 9/11 attacks in the U.S. When considering the total number of occurrences of these lexical units, 2001 shows the highest number of hits for *terrorism* with 626 (followed by 544 in 2015), while *terror* appears 1360 times, surpassed only by 2015 (with 1520 hits). The period from 2014 to 2021 shows the highest average similarity levels, likely reflecting the significant focus related to the rise of Daesh, the phenomenon of so-called *terrorresor* ('terror travels') to the Middle East (Ängsal et al. 2024) and terror attacks across Europe, including the 2017 truck attack in Stockholm.

Regarding the semantic similarity between the word pairs *terrorism* and *extremism*, a similar trend to the one described above appears in Figure 2. What stands out as different, however, is that the number of hits for each term is fewer than five in some cases, resulting in zero values. Nevertheless, the overall trend of increasing similarity over time is clear. Notably, from 2006 onwards, the similarity increases to levels above 0,30, with top values exceeding 0,40 in 2010, 2011, 2014, 2015 and 2016. This increase is steeper than that observed for the word pair terror and terrorism over the same period. The sharp rise in similarity likely corresponds to the introduction and institutionalisation of the term *våldsbejakande extremism* ('violence-affirming extremism') in the policy-making discourse (Andersson 2018; Andersson Malmros 2022). This two-word unit carries semantic overlap with 'terrorism' but is more closely tied to the discourse around counter-radicalization efforts, serving as a discursive node for preventive measures against terrorism: "Typically, the 'soft side' of counter-terrorism gathers its objectives and measures under the policy banners 'Preventing Violent Extremism' (PVE) or 'Countering Violent Extremism' (CVE)" (Andersson Malmros 2022, p. 290).

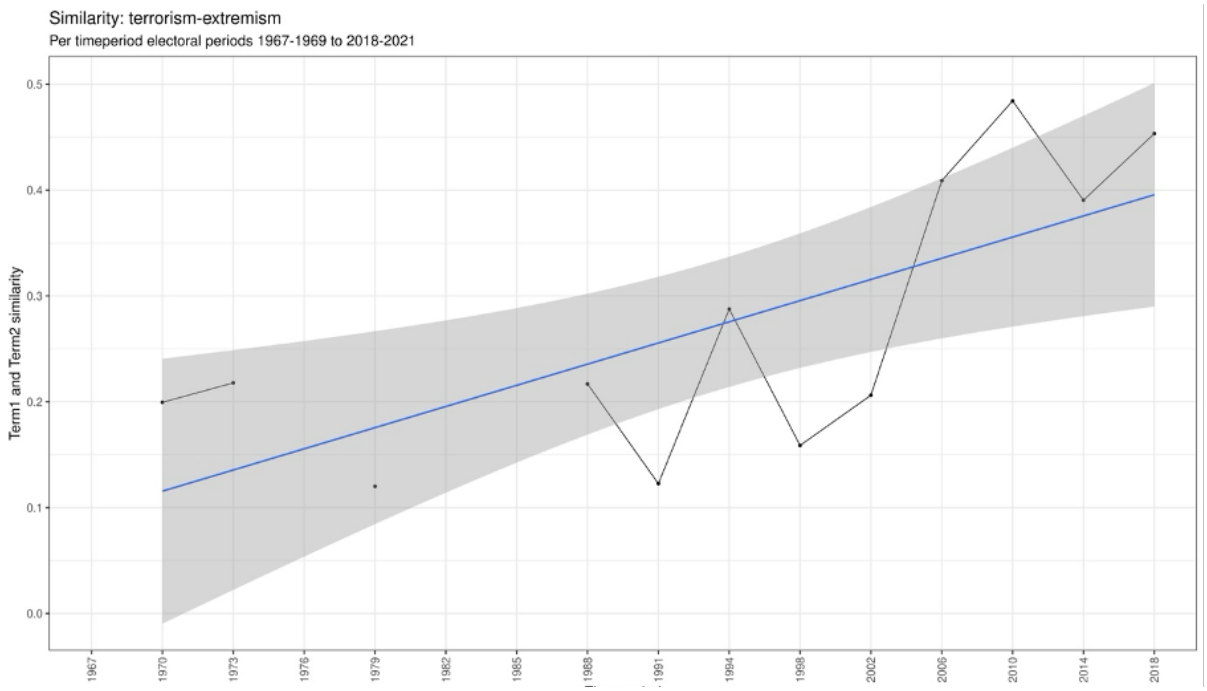


Figure 2: Similarity between the word vectors for *terrorism* and *extremism* in the Swedish parliamentary debates across electoral periods from 1967–2021, based on the SweTerror corpus version 0.14. The blue regression line indicates a steady increase in similarity, while the black line connects the data points to make it easier to identify where uninterrupted data points occur for the electoral periods. Note the break in data between 1976 and 1987. The grey area represents the 95 percent confidence interval.

4.2. Similarity of terms

The next step in divulging vectors of similarity involves examining the most synonymous terms for *terror*, *terrorism* and *extremism*. Figure 3 shows the top five words most similar to *terrorism* in the data, segmented by terms of office, with the similarity rate scaled from highest at the bottom to lowest at the top (with the strongest similarities at the bottom). Similarly, Figure 4 presents the top five most similar words to *terror*, following the same measurement principles.

Beginning with *terrorism*, it is clear that most of the strongest similarity terms are other "t-words", i.e. lexemes containing the base form *-terror-* such as *terrorist*, *terrordåd* ('terror deed') and *terror* (1970–1972) as well as *terrorhot* ('terror threat') (2010–2013) and *terroristorganisation* ('terrorist organisation') (2014–2017). Other terms with strong similarity include *våldsdåd* ('act of violence') (1970–1972) and *flygplanskapning* ('aircraft hijacking') (1973–1975), both of which designate acts often framed within a context of terrorism. These similarity terms can be further analysed in terms of semantic generality and specificity. While some, like *terrorist*, *terror*, *våldsdåd*, represent general concepts in terrorist discourse, others reflect specific political developments or acts of (alleged) terrorism debated by MPs.

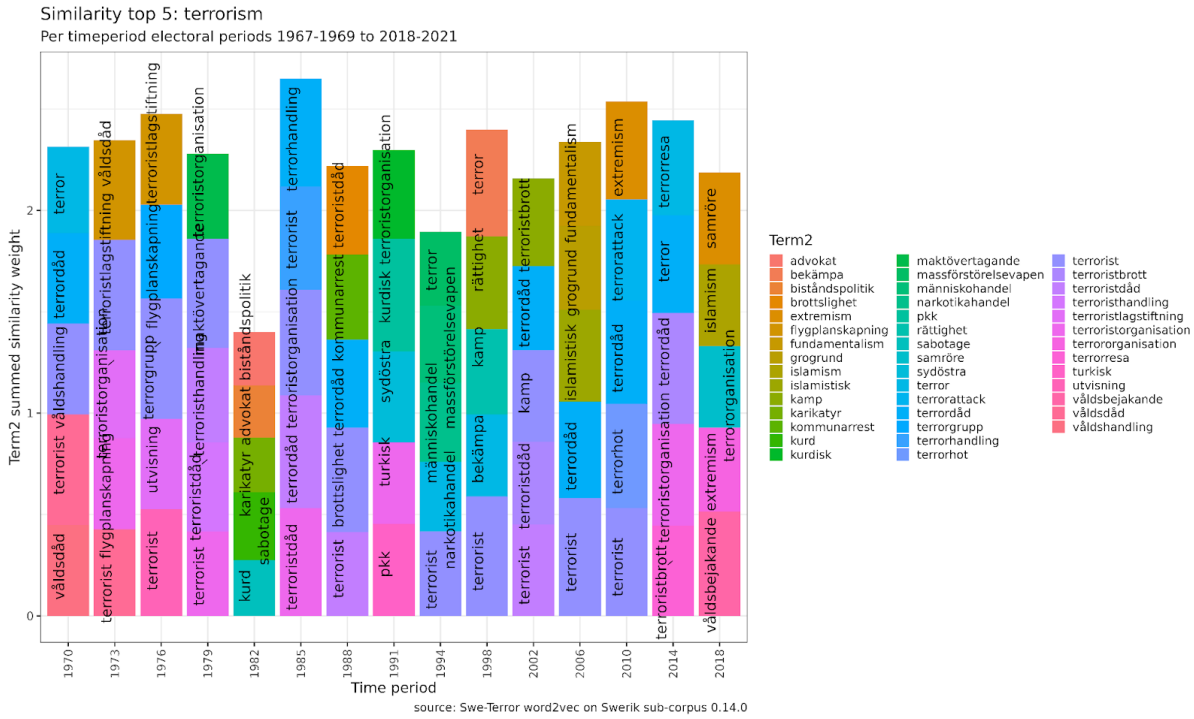


Figure 3: Top 5 lemmas most similar to the lemma *terrorism* in the Swedish parliamentary debate 1967–2021, based on the SweTerror corpus version 0.14.

Beginning with *terrorism*, it is clear that most of the strongest similarity terms are other "t-words", i.e. lexemes containing the base form *-terror-* such as *terrorist*, *terrordåd* ('terror deed') and *terror* (1970–1972) as well as *terrorhot* ('terror threat') (2010–2013) and *terroristorganisation* ('terrorist organisation') (2014–2017). Other terms with strong similarity include *våldsdåd* ('act of violence') (1970–1972) and *flygplanskapning* ('aircraft hijacking') (1973–1975), both of which designate acts often framed within a context of terrorism. These similarity terms can be further analysed in terms of semantic generality and specificity. While some, like *terrorist*, *terror*, *våldsdåd*, represent general concepts in terrorist discourse, others reflect specific political developments or acts of (alleged) terrorism debated by MPs.

There are more similarity terms that position terrorism within specific political contexts. A notable example from the mid-1980s is *kurd* (1982–1984), which should be understood in the context of the ongoing debates on the PKK (The Kurdistan Workers' Party) and different acts of political violence in Sweden associated with Kurdish guerilla organisation during the period. Towards the end of the

studied period numerous strong similarity words increasingly frame terrorism as connected with radical Islam. Already in the 2006–2009 term, i.e. the post-9/11 era, some of the strongest similarity terms include *islamistisk* (‘Islamist’) and *fundamentalism*, both of which are discursively linked to organised radical political interpretations of Islam. By the end of the period, *extremism* and *terrorresa* (‘terror travel’) rank among the strongest similarity terms (2014–2017). The latter, as previously discussed in Ängsal et al. (2024) is closely associated with discourses on Swedish citizens or residents travelling to join the Jihadist group Daesh, either to support it or to commit terrorist crimes.

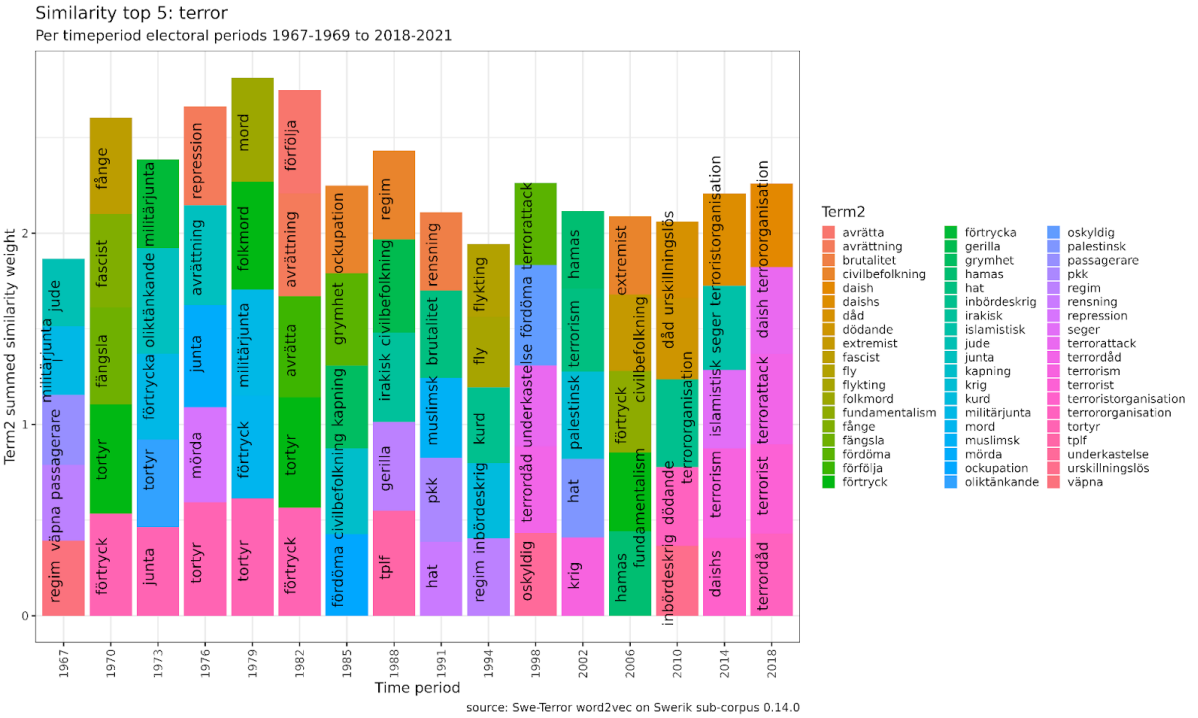


Figure 4: Top 5 lemmas most similar to the lemma *terror* electoral periods 1967–2021, based on the SweTerror corpus version 0.14.

Turning to *terror*, numerous similarity terms, particularly from the early 1970s trough the mid-1980s, indicate acts of violence associated with state terror: *tortyr* (‘torture’) (1970–1972), *junta* (1973–1975), *militärjunta* (‘military junta’) (1973–1975), *repression* (1976–1978), *förtryck* (‘oppression’) (1979–1981), *ockupation* (‘occupation’) (1985–1987). Over time, while words from this conceptual domain persist, they become less frequent among the top five similarity terms. In their place, terms more closely linked to clandestine terrorism gain prominence: *terrorattack* (‘terror attack’) (1998–2001), *terrorism* (2002–2005), *extremist* (2006–2009), *terrororganisation* and *dåd* (‘deed’) (2010–2013). Notably several actors from the Middle East appear: *PKK* (1991–1993), *kurd* (1994–1997), *palestinsk* (‘Palestinian’) (2002–2005), *Hamas* (2006–2009), and *Daish* (2014–2017). Additionally, religiously derived terms like *muslimsk* (‘Muslim’) (1991–1993) and *islamistisk* (‘Islamist’) (2014–2017) also surface among the top five similarity terms. This indicates a conceptual shift in terror, as it increasingly encompasses discourse related to the Middle East and Islamist actors. In this sense, *terror* aligns more closely with *terrorism* on a discourse semantic level, a trend further supported by the increasing similarity between the two terms, as shown in Figure 1.

Lastly, we turn to the similarity vector words for *extremism*. As shown in Figure 5, *extremism* generates zero results for a few terms of office, which can be attributed to the low frequency of the word during these periods. However, there are some notable patterns. First, *extremism* appears closely linked to conflicts in the Middle East, as indicated by similarity terms such as *sionism* (‘Zionism’) (1973–1976), *PLO*, *palestinier* (‘Palestinians’) (1979–1981) and *palestinsk* 1994–1997). High-ranking terms from the religious domain – *islam* (1991–1993); *religiös* (‘religious’) 2002–2005 – further support this discourse semantic pattern. Second, toward the end of the period, *extremism* became increasingly linked to discussions around preventing political violence, particularly from

2006 onwards. Terms such as *radikalisering* (‘radicalisation’) (2006–2009), *våldsbejakande* (2010–2013, 2014–2017) and *extremistmiljö* (‘extremist environment’) (2010–2013) reflect this shift.

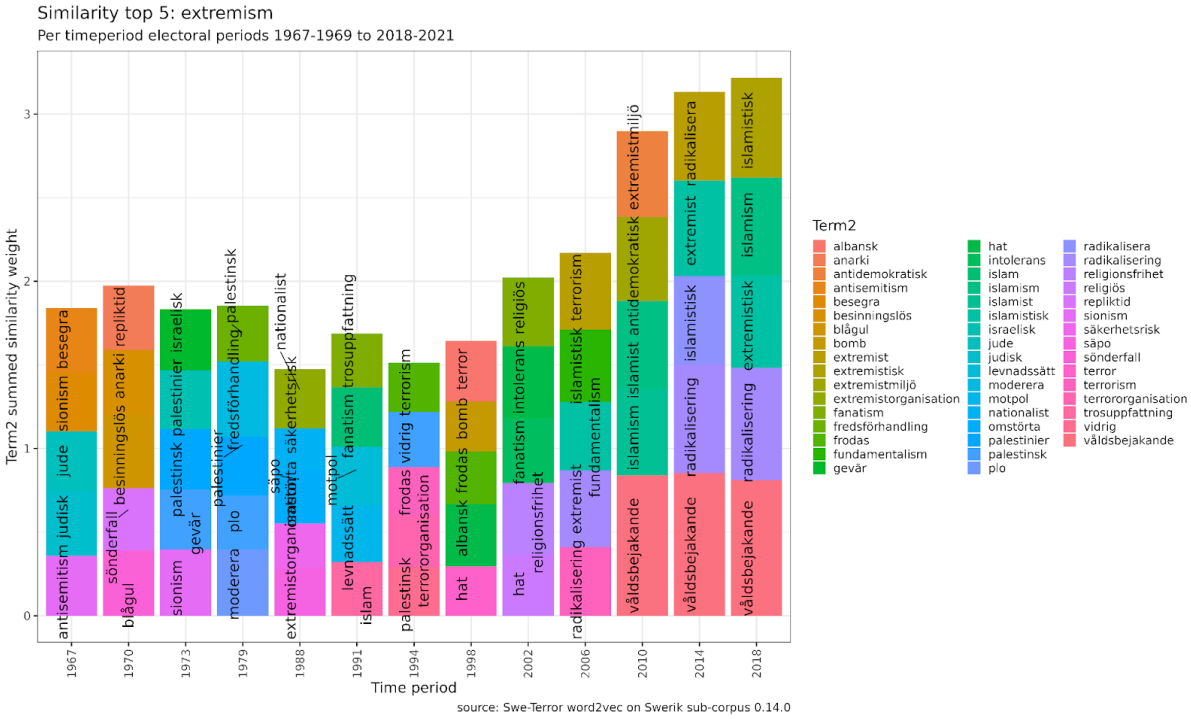


Figure 5: Top 5 lemmas most similar to the lemma *extremism* electoral periods 1967–2021, based on the SweTerror corpus version 0.14.

In summary, striking similarities as well as divergences can be observed between the three terms of political violence – *terrorism*, *terror* and *extremism* – based on the top five similarity words generated by the vectors, as shown in Table 1.

A first observation is that *terror*, *terrorism* and *extremism* appear to converge in similarity over time. Toward the end of the period, all three increasingly generate similarity words linked to Jihadist political violence, though to varying degrees. Comparing the stability and fluidity of these base forms reveals that terrorism remains the most semantically stable, while terror and extremism have undergone significant shifts. Early in the period, terror is largely associated with state terror and acts of oppression, but it gradually evolves into a term more integrated with (clandestine) terrorist discourse. Extremism, on the other hand, starts by being closely linked to conflicts in the Middle East, eventually developing semantic ties to radicalization and violence-affirming stances, embedding the word within a specifically Swedish context. One final observation is the tendency for *terrorism* to link with *terror* and terms involving the prefix *terror-*, whereas the reverse is not as strong – *terror* does not attract *terrorism* to the same extent. This likely reflects *terror* as a more general term encompassing various forms of violence, war and oppression, while *terrorism*, with its legalistic framing, holds a narrower and more specific meaning within parliamentary debate.

5. Conclusions

In this paper, we have described how the mixed-methods SweTerror project contributes to the methodological advancement of the study of Swedish parliamentary debates by augmenting the analysis through a combination of word embeddings and context-sensitive approaches. Focusing on the Swedish parliamentary discourse on terrorism during the electoral periods 1968–2018, the project integrates LT with domain expertise in political science, history, and linguistics, among other things, to perform a text mining of how terrorism has been framed in parliamentary debate. Our approach combines enriched corpus annotation with advanced word vector techniques, allowing for a detailed

Table 1

Top five lemmas (displayed in order of significance) most similar to the lemmas *terror*, *terrorism* and *extremism* in comparison, 1967–2021, based on the SweTerror corpus version 0.14. If there were not enough occurrences of a term during a period, the cell has been left blank.

Term of office	Terror	Terrorism	Extremism
1967-1969	regim, väpna, passagerare, militärjunta, jude		
1970-1972	förtryck, tortyr, fängsla, fascist, fånge	våldsdåd, terrorist, våldshandling, terrordåd, terror	blågul, sönderfall, besinningslös, anarki, repliktid
1973-1975	junta, tortyr, förtrycka, oliktänkande, militärjunta	terrorist, flygplanskapning, terroristorganisation, terroristlagstiftning, våldsdåd	sionism, gevär, palestinsk, palestinier, israelisk
1976-1978	tortyr, mörda, junta, avrättning, repression	terrorist, utvisning, terrorgrupp, flygplanskapning, terroristlagstiftning	
1979-1981	tortyr, förtryck, militärjunta, folkmord, mord	terrorist, terroristdåd, terroristhandling, maktövertagande, terroristorganisation	moderera, plo, palestinier, fredsförhandling, palestinsk
1982-1984	förtryck, tortyr, avrätta, avrättning, förfölja	kurd, sabotage, karikatyr, advokat, biståndspolitik	
1985-1987	fördöma, civilbefolkning, kapning, grymhet, ockupation	terroristdåd, terrordåd, terroristorganisation, terrorist, terrorhandling	
1988-1990	tplf, gerilla, irakisk, civilbefolkning, regim	terrorist, brottslighet, terrordåd, kommunarrest, terroristdåd	extremistorganisation, omstörta, Säpo, säkerhetsrisk, nationalitet
1991-1993	hat, pkk, muslimsk, brutalitet, rensning	Pkk, turkisk, sydöstra, kurdisk, terroristorganisation	islam, levnadssätt, motpol, fanatism, trosuppfattning
1994-1997	regim, inbördeskrig, kurd, fly, flykting	terrorist, narkotikahandel, människohandel, massförstörelsevapen, terror	palestinsk, terrororganisation, frodas, vidrig, terrorism
1998-2001	oskyldig, terrordåd, underkastelse, fördöma, terrorattack	terrorist, bekämpa, kamp, rättighet, terror	hat, albansk, frodas, bomb, terror
2002-2005	krig, hat, palestinsk, terrorism, hamas	terrorist, terroristdåd, kamp, terrordåd, terroristbrott	hat, religionsfrihet, fanatism, intolerans, religiös
2006-2009	hamas, fundamentalism, förtryck, civilbefolkning, extremist	terrorist, terrordåd, islamistisk, grogrund, fundamentalism	radikalisering, extremist, fundamentalism, islamistisk, terrorism
2010-2013	inbördeskrig, dödande, terrororganisation, dåd, urskillningslös	terrorist, terrorhot, terrordåd, terrorattack, extremism	våldsbejakande, islamism, islamist, antidemokratisk, extremistmiljö,
2014-2017	daishs, terrorism, islamistisk, seger, terrororganisation	terroristbrott, terroristorganisation, terrordåd, terror, terrorresa	våldsbejakande, radikalisering, islamistisk, extremist, radikalisera
2018-2021	terrordåd, terrorist, terrorattack, daish, terrororganisation	våldsbejakande, extremism, terrororganisation, islamism, samröre	våldsbejakande, radikalisering, extremistisk, islamism, islamistisk

analysis that offers insights into how the issue of terrorism has been framed over time. By using word embeddings to create temporal lenses on the debates, our approach enables tracking linguistic patterns and semantic shifts across decades. Additionally, the SweTerror project's custom-made dataset, integrated with enriched metadata, including speaker gender and party affiliation, further enhances its analytical potential.

Our case study specifically highlights the semantic shifts in the use of the key concepts of 'terror', 'terrorism' and 'extremism', revealing how these terms have developed over time. We also tested with other terms with known frequency spikes, e.g. 'refugee' and 'asylum' (Gries & Durrant 2020). For comparison, we also calculated TF-IDF both separately and together with word2vec models with generation of word rains (Skeppstedt et al. 2024) The use of word vectors for distinct time periods (parliamentary years, electoral periods and government periods) provides insights into the changing Swedish parliamentary context regarding terrorism. In particular, we have shown how the term terrorism began to appear more frequently in debates in the Riksdag in the 1970s and the similarity between 'terror' and 'terrorism' increased significantly following the 9/11 attacks in the U.S. Furthermore, we have outlined how the term violence-affirming extremism emerged and became established in the parliamentary discourse from 2010 onward. Our study has shown a growing association between terrorism and extremism, particularly after 2010, when violence-affirming extremism became a key term in policy discussions, primarily connected to discourses on countering terrorism, especially Islamist violence. Essentially, by integrating quantitative and qualitative methods, the project contributes to the methodological advancement of Swedish digital parliamentary studies while offering a deeper understanding of how terrorism is framed in political debate.

Acknowledgments

The SweTerror project (https://sweterror.se) is funded by the research program DIGARV (supported by the Swedish Research Council, Riksbankens Jubileumsfond, and the Royal Swedish Academy of Letters, History and Antiquities). The work presented in this paper also ties into the national research infrastructures funded by the Swedish Research Council, Huminfra (contract no. 2021-00176), and Swe-Clarin and the National Language Bank (contract no. 2017-00626).

References

Andersson, Dan-Erik. 2018. 'Från terrorism till våldsbejakande extremism: Att institutionalisera ett nytt begrepp i svensk politik'. In M. Arvidsson, L. Halldenius, & L. Sturfelt (eds.), Mänskliga rättigheter i samhället. Bokbox.

Andersson Malmros, Robin. 2022. 'Prevention of terrorism, extremism and radicalisation in Sweden: a sociological institutional perspective on development and change', European Security, 31:2.

Bode, Katherine 2018. A world of fiction. Digital collections and the future of literary history, Ann Arbor, MI: University of Michigan Press.

Borin, Lars, Markus Forsberg, Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In Proceedings of LREC 2012, Istanbul. ELRA.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. The Sixth Swedish Language Technology Conference (SLTC), Umeå University, 17–18.

Brodén, Daniel, Mats Fridlund, Leif-Jöran Olsson, Magnus P. Ängsal, Patrik Öhberg. 2023. The Diachrony of the New Political Terrorism: Tracing Neologisms and Frequencies of Terror-

related Terms in Swedish Parliamentary Data 1971–2018. In DHNB 2022: Proceedings, CEUR-WS.

Brulin, Remi. 2015. 'Compartmentalization, contexts of speech and the Israeli origins of the American discourse on 'terrorism''. Dialectical anthropology, 39.

Bäck, Hanna & Marc Debus. 2016. Political parties, parliaments and legislative speechmaking. Houndmills: Palgrave Macmillan.

Conlon, Sumali J., Alan S. Abrahams, and Lakisha L. Simmons. 2015. 'Terrorism information extraction from online reports'. Journal of Computer Information Systems, 55:3.

Ditrych, Ondrej. 2014. Tracing the discourses of terrorism, Palgrave Macmillan.

Edlund, Jens, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus P. Ängsal, Patrik Öhberg. 2022. A multimodal digital humanities study of terrorism in Swedish politics: an interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018. In K Arai (ed), Intelligent Systems and Applications: IntelliSys 2021. Lecture Notes in Networks and Systems (Vol. 295). Cham: Springer.

Eijnatten, Joris van & Pasi Ihalainen. 2022. Ecumene Redefined: Concepts of (Inter)National Religious Unity in British, Dutch and Swedish Parliamentary Debates, 1880–2020. In Pasi Ihalainen & Antero Holmila (eds.) Nationalism and Internationalism Intertwined: A European History of Concepts Beyond the Nation State. European Conceptual History 7. New York: Berghahn Books. https://doi.org/10.1515/9781800733152-012

Entman, Robert M. 1993. 'Framing'. Journal of Communication, 43:4.

Erjavec, Tomaž & Andrej Pančur. 2021. The PARLA-Clarin Recommendations for Encoding Corpora of Parliamentary Proceedings. Journal of the Text Encoding Initiative, 14.

Gries, Stefan Th., and Philip Durrant. 2020. Analyzing Co-Occurrence Data. In A Practical Handbook of Corpus Linguistics, edited by Magali Paquot and Stefan Th. Gries. Cham: Springer International Publishing.

Grishman, Ralph & Beth Sundheim. 1996. Message Understanding Conference- 6: A Brief History. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

Guldi, Jo. 2023. The dangerous art of text mining: A methodology for digital history, Cambridge: Cambridge University Press.

Hansén, Dan. 2007. Crisis and perspectives on policy change, diss., FHS.

Islentyeva, Anna. 2021. Corpus-Based Analysis of Ideological Bias: Migration in the British Press. Routledge Applied Corpus Linguistics. London; New York: Routledge, Taylor & Francis Group.

Jarlbrink, Johan & Fredrik Norén. 2023. 'The rise and fall of "propaganda" as a positive concept: a digital reading of Swedish parliamentary records, 1867–2019', Scandinavian Journal of History, 48:3.

La Mela, Matti, Fredrik Norén, Eero Hyvönen. 2022. Digital Parliamentary Data in Action 2022: Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop co-located with 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022) Uppsala, Sweden, March 15, 2022, Ceur-WS: 0074-3133-0.

Viola, Lorella, & Jaap Verheul. 2020. 'One hundred years of migration discourse in the times: A discourse-historical word vector space approach to the construction of meaning'. Frontiers in Artificial Intelligence, 3.

Marjanen, Jani, Jussi Kurunmäki, Lidia Pivovarova, & Elaine Zosa. 2020. 'The expansion of isms, 1820-1917: Data-driven analysis of political language in digitized newspaper collections.' Journal of Data Mining & Digital Humanities. 2020 Dec 18.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean. 2013. 'Distributed representations of words and phrases and their compositionality'. Advances in Neural Information Processing Systems. arXiv:1310.4546. Bibcode:2013arXiv1310.4546M.

Norén, Fredrik, Johan Jarlbrink. 2024. 'The Stenographic Bias: Shaping Formulaic Language in the Swedish Parliament 1920–2020', Formulaic Language in Historical Research and Data Extraction, Huygens Institute, Royal Netherlands Academy of Arts and Sciences, Amsterdam – 7-9 February 2024.

Ohlsson, Claes, Viktor Wåhlstrand Skärström, Henrik Björck. 2022. The market as a concept in Swedish parliamentary records from 1867 to 1970: A mixed methods study. In Digital Parliamentary Data in Action (DiPaDA 2022) workshop, Uppsala University, Sweden, March 15, 2022.

Oscarsson, Henrik, Felix Bäckstedt, Anna Cederholm Lager, Richard Karlsson, Maria Solevid. 2024. Väljarna och valet 2022. Göteborgs universitet: Valforskningsprogrammet.

Proksch, Sven-Oliver & Jonathan Slapin. 2012. 'Institutional foundations of legislative speech'. American Journal of Political Science, 56:3.

Reisigl, Martin & Ruth Wodak. 2009. The discourse-historical approach. In Wodak, Ruth, Michael Meyer (eds). Methods of critical discourse analysis. Sage.

Skeppstedt, Maria, Magnus Ahltorp, Kostiantyn Kucher, Matts Lindström 2024. 'From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts', Information Visualization 23:3, doi:10.1177/14738716241236188

SOU 2019:49. En ny terroristbrottslag: Betänkande av Terroristbrottsutredningen.

Stampnitzky, Lisa. 2013. Discipling terror, Cambridge University Press: Cambridge.

The Swedish Parliament Corpus, https://swerik-project.github.io

Rødven-Eide, Stian. 2020. Anföranden: Annotated and Augmented Parliamentary Debates from Sweden, in Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse, 11–16 May 2020.

Verheul, Jaap, Hannu Salmi, Martin Riedl, Asko Nivala, Lorella Viola, Jana Keck, & E. J. L. Bell. 2022. 'Using word vector models to trace conceptual change over time and space in historical newspapers, 1840–1914.' Digital Humanities Quarterly 16:2

Wahlström, Mattias. 2022. 'Constructing 'violence-affirming extremism': Swedish social problem trajectory'. Critical Studies on Terrorism, 15:4.

Weinberg, Leonard, Ami Pedahzur, Sivan Hirsch-Hoefler. 2004. 'The challenges of conceptualizing terrorism'. Terrorism and Political Violence, 16:4.

Wevers, Melvin, & Marijn Koolen (2020). 'Digital begriffsgeschichte: Tracing semantic change using word embeddings'. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53:4.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, Hui Xiong. 2017. Dynamic word embeddings for evolving semantic discovery, International conference on web search and data mining, WSDM.

Yrjänäinen, Väinö, Fredrik Mohammadi Norén, Robert Borges, Johan Jarlbrink, Lotta Åberg Brorsson, Anders P. Olsson, Pelle Snickars, Måns Magnusson. 2024. The Swedish Parliament Corpus 1867–2022, LREC-COLING 2024.

Zoller, Silke. 2021. To deter and punish: Global collaboration against terrorism in the 1970s, Columbia University press.

Transcriber effects in the Icelandic parliament corpus*

Lilja Björk Stefánsdóttir[1,*,†], Anton Karl Ingason[2,†]

[1]University of Iceland

Abstract

The Icelandic parliament corpus is being used to study individual lifespan change in sociolinguistic style-shift. We report on how the word order effect in question is affected by decisions made by those who transcribe the speeches and show that while some changes are made by the transcribers, the overall pattern of linguistic usage is not substantially altered. Ideally, each recording is manually checked by an annotator, but automatic annotation can be used with the understanding that quantitative findings are subject to minor errors.

Keywords

sociolinguistics, style-shift, Icelandic, error analysis, transcription, parliament speeches

1. Introduction

In the ERC-funded project Explaining Individual Lifespan Change (EILisCh), a key goal is to investigate how individual Members of Parliament change their linguistic behavior with respect to sociolinguistic style-shift over time. Sociolinguistic style-shift is the way in which speakers alter the rate at which they use variable linguistic phenomena depending on the context in which the speaking takes place. For example, some variants are used more frequently in formal speech and other more frequently in informal speech. Using the Icelandic Parliament Corpus (Steingrímsson, Barkarson, and Örnólfsson 2020) and automatic extraction of relevant examples, this kind of an analysis can be carried out very rapidly. However, in order to verify that the automatic methods are justified, it is necessary to evaluate how their quality compares with manually checking the data. This study adds that crucial step.

2. Stylistic Fronting in Icelandic Parliament Corpus

As in other language variation and change studies, this project focuses on certain grammatical variables. The variables monitored in the study are all stylistic indicators, with Stylistic Fronting (SF) as the study's primary variable. Stylistic Fronting is an optional movement in Icelandic of a

Digital Parliamentary Data in Action (DiPaDA 2024) workshop, Reykjavik, Iceland, May 28, 2024

*We would like to thank the European Research Council (ERC) for funding this research as part of ERC Project ID: 101117824, EILisCh (Explaining Individual Lifespan Change).

*Corresponding author.

†These authors contributed equally.

lbs@hi.is (L. B. Stefánsdóttir); antoni@hi.is (A. K. Ingason)

https://linguist.is (A. K. Ingason)

0009-0001-2407-2661 (L. B. Stefánsdóttir); 0000-0002-2069-5204 (A. K. Ingason)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

syntactic head (/word) or a phrase to the front of a clause that has a phonological subject gap (Maling 1980; Thráinsson 2007; Holmberg 2006). An example of SF is given in (1) where the non-finite verb *lesnar* 'read' is moved in front of the finite auxiliary *eru* 'are' in a relative clause where the grammatical subject has been extracted. Example (2) represents a parallel sentence in which SF has not been applied.

(1) Bækur [$_{CP}$ sem lesnar eru til skemmtunar] eru bestar.
books [$_{CP}$ that read are for entertainment] are best
'Books that are read for entertainment are the best ones.'

(2) Bækur [$_{CP}$ sem eru lesnar til skemmtunar] eru bestar.
books [$_{CP}$ that are read for entertainment] are best
'Books that are read for entertainment are the best ones.'

SF has no effect on truth-conditional meaning, and its only clear meaning component is a sociolinguistic one; the movement is associated with formal style. SF is found in both main clauses and subordinate clauses, as long as the subject is not phonologically overt. To control for factors that can condition the use of SF [1] we limit the scope of the study to the following word orders involving the complementizer *sem* that intriduces Icelandic relative clauses (e.g., by excluding elements other than non-finite main verbs).

We choose SF as a phenomenon to be studied because the rate of SF is a good indicator of the level of formality with a sample of speech data and therefore it is a good variable to study style shift. It is also convenient to use because it can be relatively reliably extracted with automatic methods, although this paper is an inquiry into just how reliably such automatic methods are. The choice of the parliament corpus as opposed to other corpora stems from the fact that by using speeches given from the same spot and social context, we can control for other contextual factors, and we have found in our previous research that fluctuations in the use of SF in parliament speeches reflect in an interesting way what is going on in the professional lives of the parliament members. Although we currently limit the scope of our inquiry in this way, our findings will provide a useful point of comparison when studying SF in other corpora in future work.

3. Transcriber error analysis

In order to estimate the effect of transcriber error, i.e. differences between what the Member of Parliament said and what was transcribed by the employees of the parliament, we manually checked all the examples of Stylistic Fronting in our study of MP Ásmundur Einar Daðason (Stefánsdóttir and Ingason 2024). In terms of sampling, we included all speeches given by this MP during the period in question and all relative clauses of the type described above. Each relevant token was first automatically annotated as 1 for Stylistic Fronting and 0 for an absence of Stylistic Fronting where it could have applied. Then the examples were all manually annotated as well so that we could check whether SF was correctly annotated by the automatic

[1] Wood's (2011) study showed that there are some prosodic factors that affect the use of SF, such as the constituent's number of syllables, building on earlier work that suggests that optionality in Icelandic syntax may be sensitive to favoring a regular trochaic stress pattern (Ingason 2008).

process or if the transcriber had reversed the word order, i.e. inserted Stylistic Fronting where there was none or removed Stylistic Fronting that was present in the speech. We excluded tokens that were unresolved because the recording was missing from the parliament website. The transcribers in this case are the employees of the parliament who transcribe the speeches based on audio and video recordings from parliament sessions. All speeches delivered in the Icelandic Parliament are documented and published in open access on the parliament's website. The transcribing process is relatively simple and relies mainly on recordings from a recording room above the parliamentary chamber. The transcribers' general procedure is to transcribe verbatim after the recordings, except in cases of obvious grammatical errors and unfinished utterances (Skrifstofa Alþingis 2017). The following is one example of an utterance that had to be corrected by our annotators:

(3) fjármagnið sem er verið að veita í landsbyggðarháskólana fjóra
funds that are being to allocated in countryside universities four
'... funds that are being allocated to the four countryside universities ...'

This transcription from the parliament had to be corrected because the MP did in fact say *verið er* 'being are' with a Stylistic Fronting word order.

We found that the rate of Stylistic Fronting was 73.3% (770/1050) in the automatic annotation that was directly based on the transcript and the rate was 72.9% (642/881) in the data that had been manually corrected against the recordings. This means that we looked at 1050 data points and after excluding examples that are not potential cases of Stylistic Fronting, we had 881 one examples left to carry out the analysis. We found 75 examples where an apparent Stylistic Fronting instance in the transcription had been introduced by the transcriber and 23 examples where there was no Stylistic Fronting in the transcript, but the MP had in fact uttered one as evidenced by listening to the recording. We find that this difference is not significant (X-squared = 0.031105, df = 1, p-value = 0.86).

As we are interested in studying changes in the rate of Stylistic Fronting over time for individual MP's, we furthermore plot the uncorrected and corrected rates for Ásmundur Einar Daðason throughout his political career as shown in Figure 1. It is obvious from the graph that the main trends found in the uncorrected data are replicated in the corrected data. As discussed by Stefánsdóttir and Ingason (2024), the lower rate of SF in 2013–14 and again in 2016 correlates with periods during which Daðason was going through crises in his career. He changed parties in 2011, became very unpopular and there were calls for his resignation, following which he lowered his use of SF in 2013–14. In 2015 he was reported to have been intoxicated and vomited on another passenger on an airplane, after which he lowered his use of SF in 2016. We hypothesize that the lower rate of SF during these periods corresponds to laying low and becoming more informal as a reactive response to a crisis. The wiggles in the data are almost identical in the corrected and uncorrected data.

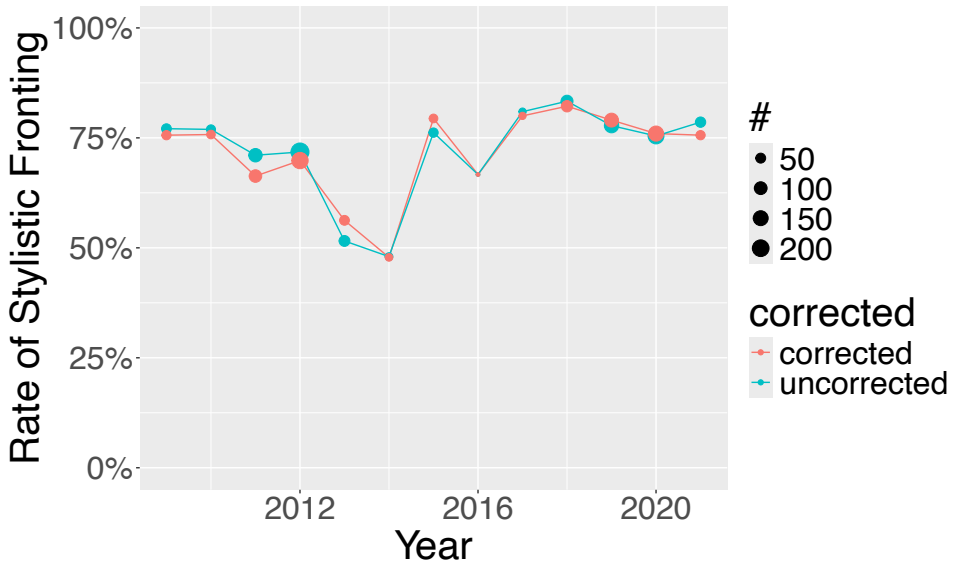


Figure 1: Error analysis in transcription by year

4. Conclusion

The fact that the rate of Stylistic Fronting is very similar in the automatically extracted findings and the manually checked ones is reassuring and suggests that in cases where resources are not available to check every token, it is valid to use the automatic analysis as a proxy for what a more detailed study would find. However, the automatic annotation does not match the manually checked one exactly, and therefore it is also the case that manual checking is valuable and preferable whenever this is a feasible option.

This means that for future research, one can remain optimistic that automatic methods can speed up the process of big data humanities approaches to studies of parliament speeches. However, any given project that focuses on a new variable must verify the accuracy of the automatic coding that is extracted based on Natural Language Processing and we must stress that our finding that the rate of Stylistic Fronting can be reliably extracted does not extend to other objects of study without further verification.

References

Holmberg, Anders. 2006. "Stylistic fronting." *The Blackwell companion to syntax,* 532–565.

Ingason, Anton Karl. 2008. *Hrynkerfi íslensku í bestunarkenningu. [Icelandic prosody in Optimality Theory.]* B.A. Thesis. University of Iceland.

Maling, Joan. 1980. "Inversion in embedded clauses in Modern Icelandic." *Íslenskt mál* 2:175–193.

Skrifstofa Alþingis. 2017. *Háttvirtur þingmaður – handbók um þingstörfin.* Published on the website of the Icelandic parliament.

Stefánsdóttir, Lilja Björk, and Anton Karl Ingason. 2024. "Wiggly lifespan change in a crisis. Contrasting reactive and proactive identity construction." *U. Penn Working Papers in Linguistics* 30 (2): 119–125.

Steingrímsson, Steinþór, Starkaður Barkarson, and Gunnar Thor Örnólfsson. 2020. "IGC-parl: Icelandic corpus of parliamentary proceedings." In *Proceedings of the Second ParlaCLARIN Workshop,* 11–17.

Thráinsson, Höskuldur. 2007. *The syntax of Icelandic.* Cambridge: Cambridge University Press.

Wood, Jim. 2011. "Stylistic fronting in spoken Icelandic relatives." *Nordic Journal of Linguistics* 34 (1): 29–60.