Digital Parliamentary Data in Action (DiPaDA 2024)

Workshop co-located with the 8th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2024, May 27-31)

Reykjavík, Iceland, Tuesday 28 May 2024 at 9.00-16.00

https://dhnb.eu/conferences/dhnb2024/workshops/dipada/

Date: Tuesday, 28/May/2024

9:00am - 10:00am H-207	**WS2-1: FULL-DAY WORKSHOP (DiPaDA 2024)** Location: **H-207** **Full programme at: https://dhnb.eu/conferences/dhnb2024/workshops/dipada/** **09:00-09:10** Welcoming (Organizing Committee) **09:10-09:30** Transcriber effects in the Icelandic parliament corpus (*Anton Karl Ingason, Lilja Björk Stefánsdóttir*) **09:30-10:00** Augmenting the Analysis of Political Discourse: A Word Embedding and Context-sensitive Methodological Approach to the Swedish Parliamentary Corpus (*Lejf-Jöran Olson, Daniel Brodén, Mats Fridlund, Magnus P. Ängsal, Patrik Öhberg*)
10:30am - 12:00pm H-207	**WS2-2: FULL-DAY WORKSHOP (DiPaDA 2024)** Location: **H-207** **Full programme at: https://dhnb.eu/conferences/dhnb2024/workshops/dipada/** **10:30-11:00** Parliaments as Networks of Power: The Analysis of Power and Gender Relations in Selected European Parliaments (*Jure Skubic*) **11:00-11:20** "Matrikelmoderation", what else? Topic Modeling of the Holy Roman Empire ́s Imperial Diet Records of 1576 (*Roman Bleier, Florian Zeilinger*) **11:20-11:40** Reppin' your constituency? Geographical Representation in Swedish Parliamentary Speech (*Albert Wendsjö*) **11:40-12:00** The Relevance of AI: Perspectives from the British and Slovenian Parliament (*Ajda Pretnar Žagar, David Moats*)
1:00pm - 2:30pm H-207	**WS2-3: FULL-DAY WORKSHOP (DiPaDA 2024)** Location: **H-207** **Full programme at: https://dhnb.eu/conferences/dhnb2024/workshops/dipada/** **13:00-13:30** Decoding the parliamentary debate on marketization of education in Sweden through computational analyses (*Eric Borgström, Martin Karlsson, Christian Lundahl* **13:30-13:50** Bilingual Parliament? Functions of Swedish, English and Latin in the Parliament of Finland (*Anna Ristilä*) **13:50-14:10** Accessing nature before "allemansrätten"? Combining two national parliament datasets to study a tradition before it was named (*Matti La Mela*) **14:10-14:30** Finding Patterns across Multiple Time Series Datasets: Democracy in the Twentieth-century Political Discourses in the United Kingdom, Sweden, and Finland (*Risto Turunen, Hugo Bonin, Pasi Ihalainen, Jani Marjanen*)
3:00pm - 4:00pm H-207	**WS2-4: FULL-DAY WORKSHOP (DiPaDA 2024)** Location: **H-207** **Full programme at: https://dhnb.eu/conferences/dhnb2024/workshops/dipada/** **15:00-15:30** The Politics of Compound Neologisms: A Novel Methodology for Mining of Conceptual Transformations in Swedish Parliamentary Discourse and Data (*Daniel Brodén, Claes Ohlsson, Henrik Björk, Mats Fridlund, Leif-Jöran Olsson, Leif Runefelt, Shafqat M. Virk, Magnus P. Ängsal*) **15:30-16:00** Concluding discussion

WS2-1: FULL-DAY WORKSHOP (DiPaDA 2024)

Time: Tuesday, 28/May/2024: 9:00am - 10:00am · *Location:* H-207

Full programme at: https://dhnb.eu/conferences/dhnb2024/workshops/dipada/

09:00-09:10 Welcoming (Organizing Committee)

09:10-09:30 Transcriber effects in the Icelandic parliament corpus (*Anton Karl Ingason, Lilja Björk Stefánsdóttir*)

09:30-10:00 Augmenting the Analysis of Political Discourse: A Word Embedding and Context-sensitive Methodological Approach to the Swedish Parliamentary Corpus (*Lejf-Jöran Olson, Daniel Brodén, Mats Fridlund, Magnus P. Ängsal, Patrik Öhberg*)

Transcriber effects in the Icelandic parliament corpus

Anton Karl Ingason, Lilja Björk Stefánsdóttir

University of Iceland, Iceland

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. 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.

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 syntactic head (/word) or a phrase to the front of a clause that has a phonological subject gap (Maling 1980; Thráinsson 2007; Holmberg 2006)

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 introduces Icelandic relative clauses (e.g., by excluding elements other than non-finite main verbs):

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, we manually checked all the examples of Stylistic Fronting in our study of MP Ásmundur Einar Daðason (Stefánsdóttir and Ingason forthcoming). Each relevant token was annotated as 1 if an automatic extraction of Stylistic Fronting in the transcription matched the recorded speech and as 0 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 also excluded tokens that were unresolved because the recording was missing from the parliament website.

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. 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.
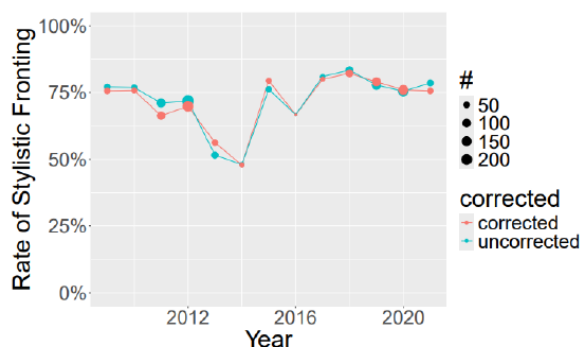

Figure 1: Error analysis in transcription by year

Conclusion

---

[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).

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.

Augmenting the Analysis of Political Discourse: A Word Embedding and Context-sensitive Methodological Approach to the Swedish Parliamentary Corpus

Leif-Jöran Olsson, Daniel Brodén, Mats Fridlund, Magnus P. Ängsal, Patrik Öhberg

University of Gothenburg, Sweden, Sweden

Introduction

Recent years have witnessed a rapidly expanding interest in data-driven research on parliamentary collections as well as a drive towards further development and standardisation of parliamentary infrastructures (La Mela et al. 2022: 3–4), including cross-national initiatives such as Parla-CLARIN (https://github.com/clarin-eric/parla-clarin) and its specialisation ParlaMint (https://www.clarin.eu/parlamint). As regards Sweden, although major work is carried out within the SWERIK infrastructure and the Welfare State Analytics project (WeStAc), research on the Swedish parliamentary datasets have, so far, been primarily exploratory and driven by broad conceptual issues and application of common statistical measurements (see Ohlsson et al. 2022; Brodén et al. 2023; Jarlbrink & Norén 2023), leaving a significant gap in the data modelling and the contextual understanding of the data.

This paper will present the augmented methodological approach to analysing the Swedish parliamentary discourse on terrorism developed in the 'Terrorism in Swedish Politics' (SweTerror) project (2021–2025) (Edlund et al. 2022). Drawing upon a mixed methods approach, we will discuss a set of context-sensitive analyses of the Swedish parliamentary record, integrating language technology (LT) and contextualising methods from, among other research areas, political science, history of ideas and linguistics. Notably, the discussion connects to the current debate within digital humanities (DH) about the need for engaging with the contextual complexities of text mining large-scale archival collections. According to digital historian Jo Guldi (2023), dedication to the question of what makes text mining accurate and robust will only get datadriven analysis of large-scale collections so far, as without a contextual sensibility applied to the materials the results tend to raise more questions than they answer (see also Bode 2018).

We here outline SweTerror's enactment of a contextualised understanding of the text data through the development of a custom-made dataset and use of word embeddings (vectors) for analysing the framing of terrorism in the parliamentary debates, 1968–2018. The paper presents a LT approach firmly grounded in humanities and social sciences (HSS) research questions, and highlights its methodological and analytical potential by presenting results from two case studies of how the Swedish Parliament (riksdag) and the different political parties have engaged with the issue of terrorism.

The dataset and LT approach of SweTerror

The paper will focus on our work with developing the SweTerror corpus, using and adapting the Swedish Parliament Corpus of the edited transcripts of minutes that are currently being cleaned up, partly re-digitised and curated for research purposes (current version 0.14). The SweTerror corpus contains an adapted and enriched subset, 1958–1967 and 1968– 2018, of the Swedish Parliament Corpus with sync on specific versions (some features both ways). The dataset is longitudinal and encompasses both the bicameral and unicameral Parliament (1867–1970 and 1971–2018, respectively), consisting of roughly 4 M tokens per parliamentary year (see below). The structure of speeches is reintroduced with a correctness of 90+ percent. Notably, the dataset is annotated with metadata about Members of Parliament (MPs) concerning name, party affiliation, gender and regional representation. Furthermore, we describe the exchange between SweTerror and SWERIK, with SweTerror's LT analyst Olsson serving on the advisory board and technical advisory board of SWERIK and further enriching and curating the Swedish Parliament Corpus for the benefit of the infrastructure and our research purposes. This work includes contributing various forms of quality control; in this paper we will point at some issues of relevance for our analysis, including the identification of omissions in the dataset such as missing debate protocols.

From the infrastructure perspective, the paper will highlight the integration of workflows into the Språkbanken Text (SB Text) infrastructure, including the Korp tool (Borin et al. 2012) to avoid reprocessing in the SB Text infrastructure. In turn, the process will introduce more flexibility to Korp's word picture functionalities and feed into new Sparv plugins (Borin et al. 2016) as well as be accessible through APIs. In extension, this means that workflows and data will also be integrated into the CLARIN ERIC infrastructure.

Concerning contextualisation, writings in Digital Humanities on Swedish Parliamentary data have mostly focused on more technical and formalistic aspects of the documentary record, such as issues surrounding OCR quality and metadata as well as how the transcriptions of the minutes are the result of post-speech editing (see Norén & Jarlbrink 2024). However, SweTerror seeks to enact a more contextualising understanding of the data, a simple yet significant element being our choice to, contrary to most parliamentary datasets, group the debates by parliamentary year (autumn–summer), rather than calendar year, to distinctly represent the mandate period of the Parliament. An important rationale for this is that changes of government during election years (mid-calendar year) affect the political dynamics, and we have previously shown that governmental position is a major factor for MPs' motion writing on the topic of terrorism (Brodén et al. 2023). Following our lead, SWERIK in 2024 adopted this principle for their Swedish Parliament Corpus.

From the technical perspective, the paper will describe our workflows around the annotation pipelines, where the outputs are continuously aggregated and analysed in an iterative process with each layer of annotation having at least one manual evaluation. Specifically, we discuss our work with word vectors and metadata, respectively, for contextual readings in two case studies concerning the occurrence of terrorism discourse in the Swedish parliamentary debate transcripts.

Case study 1: Vectors of violence

A key part of the SweTerror project is applying word vectors, word embeddings (Mikolov et al. 2013) and vectors of different kinds, word2vec and fastext with different algorithms, e.g. skipgram and cbow, for wordlevel analyses, and, larger segments e.g. documents, paragraphs and sentences, ParagraphVectors and sentence-BERT, for longitudinal exploration and examination of conceptual changes and conceptual similarities related to the notion of terrorism (see Stampnitzky 2013; Ditrych 2014; Zoller

2021). The word vectors are used in combination with enriched document annotation and quality assessed metadata to create 'temporal lenses' to traverse our analytic universe. Furthermore, we highlight the work concerning document based analyses of classification and Named Entity Recognition (NER). Both of these enrichments are used for traversal or retrieval of similar sections and related activities based on network relations.

This case study explores the Swedish parliamentary speech on terrorism with regard to discourse semantic patterns and competing terms from the realm of political violence. We aim at comparing the usage of discourse-relevant lexical items such as the Swedish terms 'terror', 'terrorism', and 'våldsbejakande extremism' ('violence-affirming extremism'). To highlight this, we diachronically compare these units by means of their similarity and closeness, estimated through a plethora of word vectors, allowing us to trace the development of the parliamentary discourse on terrorism, with regard to continuities and discontinuities. Comparing different vectors of semantically related lexical units diachronically allows us to identify potential discursive shifts in the framing of terrorism over time. A key focus of the analysis is if, and then how, the vectors of 'terror' undergo a change when the modern usage in Swedish of the word 'terrorism' emerged in the early 1970s. Another point of interest is possible discursive shifts related to the establishment of the specific term 'våldsbejakande extremism' in the 2010s (Andersson Malmros 2022) and whether it had an impact on the vectors of 'terrorism'.

In our preliminary findings, the calculated similarity between 'terror' and 'terrorism' remains rather high over time, ranging from the lowest value of 0,73 (2015) to 0,93 (1974), with value 1,0 meaning identical embeddings in the data. Although there are noteworthy differences between separate years, a general decrease in similarity after 2001 is discernible, with 2005 as the first year when similarity drops below 0,8 and 2015 displaying the all-time lowest degree of similarity. This discursive change can likely be interpreted as a further specialisation of the term 'terrorism' post 9/11, meaning that it increasingly diverts from the more general term 'terror'. A further point of inquiry is examining different ways of wording political violence by way of the lexical items mentioned in relation to MPs party-affiliation. This methodical advancement enables the outlining of differences and similarities in framing terrorism with respect to lines of political-ideological differences in parliament diachronically.

Case study 2: Terrorism is not for beginners

Another integral part of the SweTerror project is combining word vectors and metadata (see above) for classifier tasks and traversal. The act of giving speeches holds significance for analysis, as it influences the discourse surrounding policies, but also serves to inform the public about position-taking and the law-making process. However, the allocation of speaking opportunities and debates on certain issues among legislators is not random. There are patterns. One such pattern is gendered speech behaviour. For instance, women often deliver fewer speeches (Bäck, Debus & Müller 2014) and exhibit more emotional (Dietrich, Hayes & O'Brien 2019) and less aggressive speech tendencies (Kathlene 1994). Moreover, women tend to speak less on subjects regarded as 'masculine' (Bäck & Debus 2019).

Since the parliamentary speeches in our dataset are automatically assigned a Persistent Identifier (PID) for the speakers that can be used in connection to other metadata and calculated metadata. This allows for an analysis of structural differences in the debates on terrorism at party level, including government versus opposition, speech volume measured as token percentage, differences between men and women. Furthermore, we will integrate 'seniority' as an analytical factor through measurements based on the speaker's age, years in parliament, position (Minister or not, membership in Parliamentary Committees, committee chairs, governmental party, etc). Our temporal periods are dynamic in the sense that they can encompass, among other things, parliamentary year, government period, and eras defined by MPs that have been particularly influential in the debate on terrorism, all of which will be explored for both continuities and discontinuities. Networking inside parliament is explored and workflows have been established for visualising and comparing extraparliamentary and intra-parliamentary networking.

This case study also provides an overview of women's speech participation over an extended period of time and to explore which MPs who have discussed terrorism. Our preliminary findings suggest that female speech percentages have risen slowly during the early 20th century, levelling out close to 50 percent in the last decades (since 1988). Nowadays, women speak about as much as men do. Whether that pattern is also reproduced in the terrorism debate we will present and discuss when we have produced more finished results for our conference presentation.

Conclusions

We conclude by drawing together our lines of thought about the augmented methodological approach of SweTerror to analysing Swedish parliamentary discourse and how it can enrich a contextual understanding of parliamentary data, in Sweden and beyond.

WS2-2: FULL-DAY WORKSHOP (DiPaDA 2024)

Time: Tuesday, 28/May/2024: 10:30am - 12:00pm · *Location:* H-207

Full programme at: https://dhnb.eu/conferences/dhnb2024/workshops/dipada/

10:30-11:00 Parliaments as Networks of Power: The Analysis of Power and Gender Relations in Selected European Parliaments (*Jure Skubic*)

11:00-11:20 "Matrikelmoderation", what else? Topic Modeling of the Holy Roman Empire´s Imperial Diet Records of 1576 (*Roman Bleier, Florian Zeilinger*)

11:20-11:40 Reppin' your constituency? Geographical Representation in Swedish Parliamentary Speech (*Albert Wendsjö*)

11:40-12:00 The Relevance of AI: Perspectives from the British and Slovenian Parliament (*Ajda Pretnar Žagar, David Moats*)

Parliaments as Networks of Power: The Analysis of Power and Gender Relations in Selected European Parliaments

Jure Skubic

Institute of Contemporary History, Ljubljana, Slovenia, Slovenia

Parliamentary debates and parliamentary discourse are an integral part of a nation's political power because parliaments, as representative institutions, are responsible for shaping the legislation that affects people's daily lives. Thus, parliaments are a source of power for Members of Parliament (MPs) and other politicians (Bischof and Ilie 2018). Parliamentary discourse is at the heart of political decision-making and is an expression of political power – an important and widely theorized concept especially in cultural and social studies (Simon 1952; Parsons 1963). Parliamentary debates are therefore an important source of highly relevant data not only for the social sciences and humanities, but also for computer science, making parliamentary discourse interesting for both qualitative (van Dijk 2000; Bayley 2004; Ilie 2015) and quantitative (Abercrombie and Batista-Navarro 2020; Rheault et al. 2016; Cherepnalkoski and Mozetič 2016) and multidisciplinary research (Andrushchenko et al. 2022; Blaxill 2013).

We analyzed gender representation and power relations in the national parliaments of three European countries – Spain, Slovenia and the United Kingdom. The aim of the work was twofold: first, we analyzed the differences in the political representation of women in national parliaments and examined different manifestations of power within parliamentary discourse through the analysis of parliamentary speeches. With our analysis, we wanted to examine the differences in women's political representation and show that high political representation alone does not necessarily warrant the actual power of women MPs in the respective parliaments. Our second, equally important goal was to show that the intertwining of data science and social science research can generate meaningful results that might otherwise be overlooked.

To gain insight into the distribution of power of parliamentarians in the different national parliaments in Europe, we used one of the most comprehensive parliamentary datasets available – the ParlaMint dataset (Erjavec et al. 2023). This dataset contains session transcripts from more than 20 European national parliaments and spans several legislative periods with transcripts of speeches made between 2009 and 2022 (depending on the country). The corpora are uniformly encoded, linguistically annotated using the Universal Dependencies Standard and contain extensive and informative metadata of more than 11 thousand speakers. In addition, the named entities are annotated, which was very important for our analysis.

We analyzed the lower houses of the national parliaments of three European countries: the House of Commons of the United Kingdom, the Congreso de los Diputados in Spain and the Državni zbor in Slovenia. We analyzed the last completed parliamentary term (UK 2017 – 2019, Spain 2016 – 2019, Slovenia 2014 – 2018) in order to have a fully balanced group of parliamentarians. We were interested in three things in particular: first, we examined the overall gender representation in all three parliaments so that we could analyze the proportion of female MPs in national parliaments. Secondly, we analyzed the argumentative power of MPs and focused our analysis on how MPs' speeches and mentions of MPs can shed light on the power of MPs in political debates. We inspected how argumentative power becomes visible through parliamentary discourse and how the speeches and mentions shape the power relations in parliament. Last but not least, we focused on the structural power of MPs within the respective parliaments. We were interested in how the speaking practices of male and female MPs relate to certain topics and the distribution of power. We approached this question by referring to the division of topics within parliamentary debates into 'hard' topics (i.e. topics that are more likely to be discussed by male MPs) and 'soft' topics (i.e. topics that are more likely to be discussed by female MPs) (Baeck et al. 2014). We focused on five key policy issues, namely energy, finance ('hard' topics), education, healthcare ('soft' topics) and immigration, which we treated as an ambiguous topic. In doing so, we wanted to investigate whether this pattern of division could be observed in selected parliaments. To better understand and visualize the power relations between MPs, we created directed networks in which the nodes represented the mutual mentions of MPs and the weights were the number of mentions. For each parliament, we created both general and topic-specific networks, which helped us to better understand the speech dynamics on a particular topic.

Our results show that the political representation of women in all three national parliaments is quite high (between 32% and 41%), but still below the expected and desired level (50%). This suggests that although the implementation of gender quotas has been rather successful, political participation still remains highly gendered and there is still room for improvement. Furthermore, we found that despite the relatively high level of political representation, women still have much less power of argumentation than men. In all three parliaments, the active and passive importance of women was lower than their share of political representation, with Spain showing particularly negative values. This shows that although the Spanish parliament has the highest proportion of female MPs, we were able to prove that the mere presence of female MPs in parliament does not justify their participation in debates. Although they hold the same position, female MPs have fewer opportunities to express their opinions.

We also show that the dichotomy between 'hard' and 'soft' topics is not universal, but depends largely on the political and cultural context of the parliament under study. The fact is that structural power in parliaments varies thematically and that even on issues that are generally considered 'soft', women do not necessarily have more say than men. Our networks have visualized the distribution of speeches and mentions of male and female MPs well, showing that despite the presence of some prominent female politicians, men often still dominate parliamentary discourse. This visualization, combined with statistical and social network analysis suggests that the gender of MPs may have a significant impact on their argumentative and structural power.

Although many of the findings collected in our study deserve further analysis and explanation, we have scratched the surface of an important social issue of gender inequalities in politics and, most importantly, shown that multidisciplinary research can provide meaningful results and uncover constructions that serve as harmful blockages to the realization of gender equality in politics.

Keywords: parliamentary discourse, network analysis, argumentative power, structural power, relevance

"Matrikelmoderation", what else? Topic Modeling of the Holy Roman Empire´s Imperial Diet Records of 1576

Roman Bleier[1], Florian Zeilinger[2]

[1]University of Graz, Austria; [2]Historical Commission at the Bavarian Academy of Sciences and Humanities, Germany

During the 16th century the Imperial Diets, or Reichstage, of the Holy Roman Empire of the German Nation were convened by the emperors of the House of Austria at intervals of one or more years in changing locations. They were events in which the head of the empire called the aristocratic and urban estates to him to discuss and decide on political topics and thus the fate of Central Europe. Imperial Diets were the–though differently shaped–forerunners of modern parliaments and an important part of the constitutional history of the Empire. Therefore, albeit being an institution becoming more and more organisational yet still without prescriptive rules of procedure, Imperial Diets can be seen as proto-parliamentary legislative assemblies following recent endeavours such as "Recovering Europe's Parliamentary Culture, 1500-1700". Their subjects of deliberation are diverse and touch on such different and central areas of social coexistence, such as, in modern terms, questions of internal security, the judicial system, taxes, defense against external danger, monetary policy and the shaping of social coexistence in the face of confessional plurality, to name just the most important.

Imperial Diet research is primarily based on the "German Imperial Diet records" edition series of the Historical Commission at the Bavarian Academy of Sciences and Humanities, in which records for each individual Diet have been collected and published in four print series since the 19th century. They are provided with general indices of persons, places and subject terms. While this enterprise is still going on, the Commission has begun the retro-digitisation of existing volumes. However, without technical tools and the resulting search options as well as maintaining the often hermetic, source-oriented keyword terminology, it is almost impossible to gain an overview of this data.

Over the last several years a team of researchers at the University of Graz and the Historical Commission created the first genuinely digital edition of the Imperial Diet of 1576. In this context it also experimented for the first time with methods of information extraction. A follow up project that is currently still in the application phase will focus on the topic history of the Imperial Diets in the 16th century. In a small project in spring 2024 the authors have tested a frequently used tool for topic modeling with Latent Dirichlet Allocation, Mallet (https://mimno.github.io/Mallet/topics.html).

In order to contextualise the results, the digital edition on 1576 was compared with a second edition: The Imperial Diet records from 1556/57 (published 2013), which are available in retro- 2 digitised form (https://reichstagsakten.de/), contain different structured texts (general index with links to PDFs of individual pages). The heterogeneous edited, transcribed and annotated texts are segmented and have human assigned keywords for negotiation topics linked to the according index. These keywords follow the traditional indices from previous editions of the respective series established in 1988, but were slightly modified for the latest edition of 1576 (more abstract analytical terms; no complex, nested general registers of people, places and topics). It was necessary to manually extract the subject terms from the general index and link them to specific text passages. Then they were compared with computer generated topics to critically reflect on the editorial keywording and indexing of contemporary source texts that document the political discourse. We started with the minutes, which can be found in each of the two editions and which are relatively structured texts documenting the entire duration of the Diet. The keywords assigned by editors are collected in a cross-edition database, as are the topics automatically created on the basis of the bundles of terms from the sources. This means that the topic modelling approach uses contemporary political language as a starting point and reflects on it. Especially the automated generation and historiographical interpretation of topics makes it possible to practise data and editorial criticism. For example the automated topic modeling shows the importance of procedure related source terms that have not been accordingly indexed in previous editions following their focus on deliberation contents.

This experiment provides first insights into the potentials of topic modeling for the Imperial Diet records. It also highlights issues that will need to be addressed in the new project focusing on topic modeling and in which the assignment of negotiation topics in several editions can be checked and the keywords can be merged into a cross-edition index.

Reppin' your constituency? Geographical Representation in Swedish Parliamentary Speech

Albert Wendsjö

University of Gothenburg, Sweden

In representative democracy, parliamentarians are elected to represent their voters. That they succeed in doing this is one, if not the most, important goal in representative democracies (e.g. Pitkin 1967), and a lot of research has studied to what extent this is achieved. For example, to what extent do parliamentarians represent people of different genders, ethnicities, or ages (see e.g. Elsässer and Schäfer 2023, Wägnerud 2009, Persson et al. 2023). However, one important aspect of representation is geographical representation, namely that parliamentarians represent the interests of all geographical regions of the polities. It is to this end that many representative democracies practically elect parliamentarians through local constituencies that are geographically spread throughout the country. However, to what extent geographical representation is substantially achieved is a question that has partially been overlooked. The purpose of this paper is to show how geographical text analysis (GTA) (see e.g. Porter et al. 2015) can be used to study geographical representation using parliamentary debates and to study to what extent geographical representation is achieved in the Swedish Riksdag. In other words, this paper studies which places are more frequently mentioned in politicians' parliamentary speeches.

When talking about representation, one often makes the distinction between substantial and descriptive representation (Pitkin 1967). While descriptive representation refers to whether parliamentarians share characteristics with those they represent,

substantial representation refers to whether parliamentarians represent their political preferences. In the case of gender, descriptive representation refers to whether the share of female parliamentarians is proportional to the share of females in the population, whereas substantial representation asks whether parliamentarians represent their interests. In the case of geographical representation, the equivalent questions become whether or not parliamentarians are elected from all geographical regions, and whether they represent their interests. While the first question is ensured through the electoral system, the second question is less certain.

Prior research has mainly focused on substantial representation in terms of gender, ethnicity, or class and less so in terms of geography. Prior studies in Sweden have shown that parliamentarians adapt their language on immigration based on local conditions (Olander 2018), but we know little of how this generalizes to other forms of substantial representation. On a more granular and local level, Folke et al. (2021) found that local politicians' neighborhoods are often associated with reduced likelihoods of public "bads", but we don't know if parliamentarians at the national level act the same for those regions they are elected to represent. Beyond Sweden, there are a few studies that have started to study geographical representation in a broader sense, often through mentions of local regions. These studies have found several explanations of local representation, for example, electoral volatility, politicians' background, and electoral system (Schürmann and Stier 2023, Zittel et al. 2021, Russo 2021, Nagtzaam and Louwerse 2023). However, most studies have been limited in temporal coverage, and carry methodological limitations.

Adding to this literature, our work studies geographical representation in the Swedish Riksdag from 1864-2022 using the SWERIK dataset (Swerik 2023). To measure geographical representation, previous studies have mainly relied on dictionary approaches, which force the researcher to identify all possible references to local areas. To move beyond this, this study uses what can be referred to as geographical text analysis (see e.g. Porter et al. 2015). The study uses named entity recognition (NER) to extract all geographical mentions in the parliamentary data, the study then geotags all places to create a dataset of geographical representation over time.

More specifically, we use a Swedish-cased BERT model pre-trained by KBLab and fine-tuned for NER classification by Recorded Future. In their domain data, the model reports an F1 score of 0.91, indicating both high precision and recall. However, to minimize the risk of missing certain places, we take a two-stage approach where we first find all possible mentioned places using a low classification threshold, which increases recall at the cost of precision. In the next stage, we feed all these mentions to an API connected to Google Maps to find the coordinates of each place, which should exclude potential misclassifications from the previous stage.

Using this data, the study examines two aspects of geographical representation. First, the study explores to what extent geographical representation is achieved, that is to what extent parliamentarians talk about a region proportional to how many people live there. Second, the study explores to what extent parliamentarians talk more about their constituency compared to other constituencies. Overall, the study finds that speech as well as descriptive representation have varied over time, and seen an increase during the second half of the 20th century.

Overall, the contribution of this study is twofold. Firstly, this explores how geographical text analysis can be used to study parliamentary speech, to answer substantial questions of representation. Through exploratory analysis, the study also points to several avenues for future research. For example, geographical text analysis can be further used to study how globalization emerges through parliamentary speech. Secondly, this study contributes substantively to the literature on representation, by showing how local representation in Sweden has developed over time.

Perspectives on AI in the British and Slovenian parliament

Ajda Pretnar Žagar[1], David Moats[2]

[1]Institute of Contemporary History, Slovenia; [2]Faculty of Social Sciences, University of Helsinki, Finland

Parliamentary debates illustrate which legislative topics are relevant and how these topics shift over time. When observing how a specific topic is handled in the parliament, it is possible to pinpoint tangential topics and focal points. The identification of policy foci is particularly interesting in a cross-country comparison. In this contribution, we analyse and compare the results from seven years of parliamentary debates (2015-2022) from the British and Slovenian ParlaMint corpus (Erjavec et al. 2023). While previous research in comparative computational linguistics covered debates on migration (Blaette et al. 2020, Navarretta et al. 2022), attitudes to the EU (Hörner 2013), and right-wing populism (Schwalbach), we focus on the debates about the artificial intelligence (AI) to investigate how AI was discussed in the two countries.

The main research question is how the debates differ in the "industry leader" country, such as the UK, versus the "industry follower" country, such as Slovenia. To answer this question, we employ computational and close reading techniques. We used collocation networks to visualize the semantic relationships between words within the debates. By identifying frequently co-occurring word pairs, we discern key themes that centre around AI. We used word enrichment analysis to identify lexico-semantic patterns of each parliamentary discourse. This allowed us to pinpoint significant AI-related topics, revealing the unique emphases of British and Slovenian legislators. We used semantic document maps (Godec et al. 2021) to further elaborate on the topics pertinent to artificial intelligence. This provided a landscape of policy foci for further cross-country comparison. We supplemented the findings by close reading. We also analysed how the debate is characterised by parties and different MPs.

Initial observations reveal markedly different topics, reflecting the distinct political landscapes and socio-cultural contexts of the two nations. Both debates started to take off after 2016. However, the UK debate is much more prominent, while in Slovenia, there are more than a handful of mentions of AI only in the late 2019. The UK debate centres around protecting national interests, mostly in relation to company takeovers such as DeepMind. A prominent topic in the British subcorpus is also how AI can be implemented in healthcare. In Slovenia, on the other hand, the debate revolves around general digital transformation.

The presented research is part of a larger project mapping public values in algorithmic systems. Further research will include the analysis of additional national parliamentary debates from Denmark, Finland, and Sweden.

WS2-3: FULL-DAY WORKSHOP (DiPaDA 2024)

Time: Tuesday, 28/May/2024: 1:00pm - 2:30pm · *Location:* H-207

Full programme at: https://dhnb.eu/conferences/dhnb2024/workshops/dipada/

13:00-13:30 Decoding the parliamentary debate on marketization of education in Sweden through computational analyses (*Eric Borgström, Martin Karlsson, Christian Lundahl*

13:30-13:50 Bilingual Parliament? Functions of Swedish, English and Latin in the Parliament of Finland (*Anna Ristilä*)

13:50-14:10 Accessing nature before "allemansrätten"? Combining two national parliament datasets to study a tradition before it was named (*Matti La Mela*)

14:10-14:30 Finding Patterns across Multiple Time Series Datasets: Democracy in the Twentieth-century Political Discourses in the United Kingdom, Sweden, and Finland (*Risto Turunen, Hugo Bonin, Pasi Ihalainen, Jani Marjanen*)

Decoding the parliamentary debate on marketization of education in Sweden through computational analyses

Eric Borgström, Martin Karlsson, Christan Lundahl

Örebro University, Sweden

During the last four decades, the Swedish education system has gone through a transformation from one of the most centrally planned and uniform to one of the most marketized and decentralized in the OECD area (cf. Fredriksson, 2010; Lundahl et al., 2013). Though part of a global trend of marketization of welfare services (Fuller, 2019), the development of the Swedish education system is more radical in comparison both with other welfare areas in Sweden, as well with eduction reforms in other countries (Lundahl, 2016). As such, the trajectory of Swedish education policy from the 1980s until today can be described as a shift between two extreme positions seldom witnessed in neither welfare policy nor education policy research (Fredriksson, 2010).

While a large number of studies have produced important insights into the causes, processes and consequences of this development (cf. Fredriksson, 2010; Ringarp, 2011 & Hultén, 2019), the complexity, fragmentation and extension of this process have made comprehensive and systematic analyses difficult. The task of analyzing this policy process expands beyond what is possible in traditional qualitative policy analysis. This paper aims to illustrate the potential in utilizing novel computational methods and open parliamentary data to systematically investigate the complex processes of educational reform. The paper makes use of the recent development in Natural Language Processing (NLP) to analyse the parliamentary debate on school marketization in Sweden.

Aim and research questions

The study draws on the comprehensive and systematized record of the Swedish parliament's current and historical work, made public as open resource (data.riksdagen.se). Using a combination of computational techniques – including sentiment analysis, and topic modeling – as well as qualitative text analysis, this study aims to map the dominant arguments in parliamentary debates surrounding the marketization of the Swedish education system, between 1993 and 2023. The study is guided by the following research questions:

1. What the most frequent types of arguments made in the parliamentary debate on school marketization in Sweden, 1993-2023?

2. How does the prevalence of these different types of arguments change over time?

3. How does the prevalence of these different types of arguments vary across parties?

4. What MPs and political parties are most active in the debate on school marketization across the time period, 1993-2023?

Taken together, answering these research questions will lay the foundation for creating a comprehensive picture of the structure and development of argumentation in the parliamentary debate related to school marketization in the Swedish parliament.

Theory

The marketization of the Swedish education system can be said to represent a paradigm shift in education policy (Lundahl et al., 2013; Alexiadou & Lundahl, 2016). A policy paradigm can be defined as "a framework of ideas and standards that specifies not only the goals of policy and the kinds of instruments that are used to attain them, but also the very nature of the problems they are meant to be addressing" (Hall, 1993: 278). In very general terms, the transformation of the swedish education system in a marketized direction represent a shift in understanding of the central problems of education policy, from a problem of attaining equity and equality, towards a problem of insufficient efficiency and flexibility in the school system (Börjesson, 2016: 77 & 144). This shift in definition of the central problems of education policy was followed by a transformation of the policy instruments used as well as the goals of those instruments.

A paradigmatic view of policy change is thus a suitable theoretical framework for analyzing the marketization of the Swedish education system. Such a perspective puts focus on the link between ideational dynamics, the ebb and flow of normative ideas about education, and the policy instruments designed and implemented. Policy paradigm theory suggests that the understanding of policy development requires analysis of the problems and goals identified in the policy debate.

This paper focuses on understanding the ideational development in the parliamentary debate on school marketization. The empirical basis is open parliament data, that form an ample basis for empirically analyzing policy actors' normative ideas, not least through transcripts of parliamentary speeches and debates.

Data and materials

The data utilized in this study comes from the open data of the Swedish parliament, available at https://data.riksdagen.se/. This data-base includes: parliamentary speeches, government bills, motions, government commission reports, voting records, and records of appointments. The sample selected for for analysis consists of all parliamentary speeches identified using the search term "friskol*" (i.e. independent schools with alternative word endings) in the open data of the Swedish parliament (using the GUI *riksdagssök – riksdagsdata.oru.se*), between 1993 and 2023. The sample consists of 2538 speeches made by 1279 members of parliament. Apart from a text transcript of the speech the data consists of the time of the speech as well as party affiliation, name and identification number of each speaker.

The search term chosen to delimit the sample, *friskola*, [independent school], is a central concept in the debate on school marketization. The central reform creating a quasi-market among education providers did so by allowing privately operated and owned schools to receive equivalent public funding per student as public schools. However, the sample used is not comprehensive as parts of the debate about school marketization can consist of speeches not mentioning independent schools.

Research methods

The first stage of the analysis consisted of applying automated NLP-technologies (using Dcipher Analytics, https://www.dcipheranalytics.com) to extract arguments in the data set. First, a rough set of arguments were identified using GPT-3. Each argument was classified based on its sentiment (positive, negative, neutral) using sentiment analysis. The arguments were then plotted on a two-dimensional vector space, and clustered into distinct topics using topic detection (Figure 1). Finally, clusters were aggregated, yielding a preliminary compilation of 88 different machine-identified arguments, each with aggregated information including 15 debate excerpts illustrative of the particular argument.

In the second stage of the analysis, each of the arguments, were validated, refuted or refined through qualitative analysis of the debate excerpts as data. In each excerpt collection, arguments (i.e. claims and premises, see Rocha et.al. 2022) were manually identified, categorized as pro or contra educational marketization and labelled inductively. The automated sentiment analysis of stage one proved to be inadequate in identifying pro and contra arguments (an argument in favour of a particular view may well be stated in a negative sentiment, e.g. *independent schools are not inferior schools*). The qualitative analysis resulted in a tentative, high-level typology comprising four (4) main topics and nine (9) argument types for and against independent schools was developed and used to code the data set.

Preliminary results

- The arguments put forward in the parliamentary debate on school marketization a are characterized by a broad diversity of topics. However, four high level topics can be distinguished that encompass a substantial share of the arguments made. These are:
    - *Meta-debate on school marketization:* As any other policy debate, the parliamentary debate on school marketization has a substantial element of meta-debate or a debate on the subject of debating school marketization. Meta-debate arguments are concerned with the debate itself (e.g. is school marketization an important issue to debate), or on the actors of the debate (e.g. debating the credibility of MPs or political parties).
    - *Organization of the school market:* The organization of the school market is one prevalent topic in the parliamentary debate on school marketization. Rather than question or defend the existence of a school market the arguments made within this topic of the debate focus on how such a market should be organized.
    - *Existence of the school market:* While arguments in this category often resemble those in other categories, the difference lies in their starting point. What is debated in this theme is the existence of the marketized school system in itself, rather than its organization or regulation. Hence, the debate focuses on whether the system should prevail or be abandoned.
    - *Quality of independent schools:* The last theme of arguments in the debate on school marketization concerns the quality of independent schools. While there are arguments put forward on both sides, it is important to point out the strong skewness towards pro-marketization arguments within this theme.
- There has been a gradual shift in issue ownership in the parliamentary debate on school marketization over time, from the ideological right to left. At the start of the investigated time-period, the political right dominated the debate. However, gradually over time MPs from the parties on the left have become more active in the debate on school marketization.
- The distribution of parliamentary speeches among Swedish MPs follows a clear power law, or long-tail distribution. A small minority of MPs have conducted a majority of the parliamentary speeches mentioning independent schools. The large majority of MPs that have mentioned independent schools in parliamentary speeches have done so in few or single speeches, while 11% (n=114) of the MPs stand for a majority of speeches mentioning independent schools.

Bilingual Parliament? Functions of Swedish, English and Latin in the Parliament of Finland

Anna Ristilä

University of Turku, Finland

According to the Finnish law (Constitution of Finland §51) only the two official languages of Finland – Finnish and Swedish – are allowed to be used in the Finnish parliament (*Eduskunta*) but fragments of other languages are sometimes present in the discussions as well. However, the presence of these fragments has not been extensively studied, e.g. are certain topics more prone to include fragments of foreign languages. The speeches given in the Finnish parliament have been topic modelled before (Loukasmäki & Makkonen 2019, Ristilä & Elo 2023) but the distribution of languages over topics has not been studied.

This study attempts to fill this gap by examining the distributions of Swedish, English and Latin across topics identified in the Finnish parliament 1970-2020 and discussing the different languages' functions in the parliamentary context. These three languages have very different popularity trends: Swedish is an official language in Finland but its use has been in decline for a long time, English is not an official language in Finland but generally used as *lingua franca,* and Latin is a dead language but has a strong footing in especially legal contexts. Comparing the topic distributions of these languages gives us better understanding of their functions in the political context, and of how and why different languages can be used as political tools and devices.

The materials used in this study are the plenary speeches given in the Finnish parliament between 1970 and 2020. The speeches have been made computer readable and enriched with metadata (Hyvönen *et al* 2024).

This study built on the topic model by Ristilä and Elo (2023) and used language detection (Python's Lingua library) to define what languages are present in which topics. Language use occurrences and, when necessary, their surrounding context was processed through the topic model to get topic distributions for each language. The topic model was monolingual, and since there were many Swedish passages, all Swedish contents were machine translated into Finnish (eTranslation). English and Latin only appeared as small fragments, so their close context was used to define the topics.

Functions of language use were defined with close reading. Ten function categories were defined for all languages, and additional two for Swedish. These were divided into two groups: functional and rhetoric. Kari Palonen's four concepts of politics (Palonen 2003, 1993) – policy, polity, politicking and politicization – were used to get a better understanding of how the functions work in a political context: policy entails the regulation aspect of politics, polity can be understood as the sphere of established norms and procedures, politicking is the performative aspect of politics, and politicization makes something political. Especially the politicization function of language choice, or the making of something political just by using a marked (nontypical) language, was of interest.

The preliminary results indicate that Swedish and Latin usages have very similar topic distributions. Both were used significantly more than average around topics named *public sector* and *legislation*, and significantly less in topics *foreign and security policy* and *traffic and transport*. The distribution of English usage, on the other hand, closely followed the average topic distribution, reflecting its commonplace role.

Accessing nature before "allemansrätten"? Combining two national parliament datasets to study a tradition before it was named

Matti La Mela

Uppsala University, Sweden

Sweden, Finland, and Norway share a longstanding tradition of public access rights known as allemansrätten, which grants access to nature for activities such as camping and foraging on both public and private lands. What is interesting, however, is that allemansrätten, although considered an age-old Nordic custom, was not named until the 1930s and only gained common usage after World War II. This has led some scholars to view allemansrätten as a political construct and to challenge the prevailing narrative about the historical roots of these public access rights (see e.g. Wiktorsson 1996; Valguarnera 2016). The later history of the term allemansrätten has been explored through parliamentary debates in Finland (Kettunen & La Mela 2022), and in Sweden through selected Swedish parliamentary motions, protocols, and official inquiries (Sténs & Sandström 2014). This paper focuses on the roots of allemansrätten before the term was coined, through parliamentary debates in Sweden and Finland. It asks whether such practices of access rights to nature were recognized in Sweden and Finland before they received the name "allemansrätten"; in other words, to what extent can we identify the concept or idea before it was formally defined?

The data used in the paper is linked open parliamentary datasets published in both countries: in Sweden in the SWERIK project for the years 1867-2022 (https://swerik-project.github.io/), and in Finland in the Semantic Parliament project from 1907 until today (Hyvönen et al. 2023). The paper builds on the approaches in digital conceptual history that employs parliamentary debate data (see e.g. Jarlbrink & Noren 2023; Ihalainen & Sahala 2020; Elo 2022). As the concept of allemansrätten (in Finnish, jokamiehenoikeus) becomes a shared concept between Sweden and Finland after the 1940s, provides the study of the national parliamentary debates from these countries a way to follow and contrast the trajectories in how the term became articulated. At this first stage of research, the paper focuses on the years where the two datasets converge, thus on the years from 1907 onwards. Moreover, the focus is on the practices of access of nature by studying debates about foraging for the resources of wild nature. This is done because the (universal) right to forage berries, mushrooms, and non-protected plants, is at the core of today's allemansrätten.

The paper presents the preliminary results where the the access rights are traced in the debates at two levels. First, the paper examines the debates where wild berries and mushrooms are mentioned, and investigates the word contexts around them. Second, the paper employs topic modeling in the Swedish parliamentary debates to identify access rights-related debates with a particular focus again on foraging and the use of wild resources as key terms. For this, the paper applies the Swedish BERTopic implementation and the BERT language models developed by the Swedish National Library (KBLab). The topic modeling is guided with other seeds too that are key terms from contemporary debates on allemansrätten, such as access right, outdoor recreation, that allows to converge the models towards topics where the concept appears. During both steps, the identified debates are classified manually based on their legislative context and read more closely to study the views of the members of parliament.

The paper contributes to offering new knowledge about the early roots of allemansrätten. It also provides an example of how to identify and study a concept through the semantic content in a bilingual parliament dataset, rather than approaching it by the term that identifies the concept.

Finding Patterns across Multiple Time Series Datasets: Democracy in the Twentieth-century Political Discourses in the United Kingdom, Sweden, and Finland

Risto Turunen[1], Hugo Bonin[1], Pasi Ihalainen[1], Jani Marjanen[2]

[1]University of Jyväskylä, Finland; [2]University of Helsinki, Finland

This paper analyzes the contextual variation of nouns and adjectives related to democracy in the United Kingdom, Sweden, and Finland in the twentieth century. We compare parliamentary data (Hansard, Riksdag, and Eduskunta) against press data (UK: Guardian and Times; Sweden: Dagens Nyheter and Svenska Dagbladet; Finland: Helsingin Sanomat and Suomen Kuvalehti). By including both liberal and conservative newspapers as well as parliamentary debates, our study offers a fresh perspective on the relation between democratic discourses articulated by politicians on the one hand and journalists on the other. Given that the concept of democracy is at the heart of politics in the twentieth century and has undergone shifts in both evaluation (to have become near-universally positive) and in scope (to not being limited to the sphere of elite politics and encompassing more areas of life), it is crucial to understand whether or not the changes have been driven by parliamentary discourse, other public discourse, or both.

The approach includes visualizing the main similarities and differences in the use of the terms 'democracy' and 'democratic' between multiple historical time series datasets, as well as applying cross-correlation analysis to automatically find identical patterns between parliament and media or across different nations. The similarity of various word frequency time series charts is evaluated using the Pearson correlation coefficient (PCC), which can vary from -1 to 1. When two time series display simultaneous increases and decreases, the PCC value is nearer to 1 (Derrick & Thomas 2004). The strengths of the PCC are its mathematical simplicity, easy interpretability, and tolerance for noise, while its main limitation is sensitivity to extreme outliers, which can be mitigated by using sliding windows to analyze segments of the time series instead of the whole.

Our findings indicate that the cross-correlation is strongest between similar political terms in the same dataset, e.g., the relative frequency of 'democracy' and 'democratic' over time in a national parliament, as expected (in Hansard 0.91, Riksdag 0.76, and Eduskunta 0.65). Another strong set of cross-correlations can be observed when the same political term appears in different datasets from the same country, e.g., the frequency of 'democracy' in liberal and conservative papers (in the UK 0.87, in Sweden 0.82, and in Finland 0.61). Transnational correlations of political terms were not as strong as intra-national correlations, but they were clearly evident in the PCC values, e.g., for the frequency of 'democracy' they varied from 0.58 to 0.68 between three parliaments under investigation. The shared patterns include general increase in the use of 'democracy' over time in each parliament, with notable peaks in the 1930s as a reaction to totalitarianism, around the year 1968 related to the rise of social movements calling for participation side by side with 'representative democracy', and in the 1990s, with the expansion of digital communication that has changed the relationship between the representatives and the represented (Ihalainen et al. 2022). We ensured that our results were not due to intrinsic structural properties of the chosen datasets by calculating the PCC values also for non-political terms, which showed weak or non-existent correlation between political and non-political terms.

Methodologically, our contribution introduces time series methods to digital humanities, a field which has mostly focused on the manual examination of time series visualizations, with only a few exceptions (Wevers, Gao & Nielbo 2020). From the humanities perspective, we empirically demonstrate the strong linkage between the political discourses in parliament and the press, challenging the often-presented claim of parliamentary speech as distinct from debates in broader society. At least in the case of democracy, discourses in parliament and outside it seem to be in sync.

WS2-4: FULL-DAY WORKSHOP (DiPaDA 2024)

Time: Tuesday, 28/May/2024: 3:00pm - 4:00pm · *Location:* H-207

Full programme at: https://dhnb.eu/conferences/dhnb2024/workshops/dipada/

15:00-15:30 The Politics of Compound Neologisms: A Novel Methodology for Mining of Conceptual Transformations in Swedish Parliamentary Discourse and Data (*Daniel Brodén, Claes Ohlsson, Henrik Björk, Mats Fridlund, Leif-Jöran Olsson, Leif Runefelt, Shafqat M. Virk, Magnus P. Ängsal*)

15:30-16:00 Concluding discussion

The Politics of Compound Neologisms: A Novel Methodology for Mining of Conceptual Transformations in Swedish Parliamentary Discourse and Data

Daniel Brodén[1], Claes Ohlsson[2], Henrik Björk[1], Mats Fridlund[1], Leif-Jöran Olsson[1], Leif Runefelt[3], Shafqat M. Virk[1], Magnus P. Ängsal[1]

[1]University of Gothenburg, Sweden; [2]Linnaeus University, Sweden; [3]Södertörn University, Sweden

Introduction

This paper derives from two distinct research projects in text mining on Swedish parliamentary data, each exploring the historical transformation of key concepts and shifts in their meaning. 'Terrorism in Swedish Politics' (SweTerror) (2020–2026, Edlund et al. 2022) examines the framing of political terror in parliamentary discourse 1968–2018, and 'The Market Language' (2022–2025, Ohlsson et al. 2022), in turn, investigates the discourse surrounding markets from the Middle Ages to contemporary times. Despite having different designs and focuses, both projects have yielded results that attest that compound words and compound neologisms play a part in the formation and development of political discourses.

The overarching aim of this project-wide paper is to provide insights into the less explored significance of compounds when text mining conceptual transformation in parliamentary discourse. We argue that employing corpus linguistics in conjunction with conceptual history for text mining compounds is an underdeveloped approach in analysis of discursive change over time, framing the issue in the context of how multi-word phrases are expressed in different languages and how compound neologisms constitute a specific morphological feature of Swedish.

Drawing upon a combination of distant and close reading of Swedish parliamentary datasets, the paper illustrates such an approach through two case studies of the transformations of the concepts of the market and terrorism, respectively. We highlight how both these concepts have been expanded to include different topical phenomena and also 'colonised' different political domains through the use of new types of compounds. Thus, the paper opens an analytical window to following discursive transformations as anchored through a delimited linguistic phenomenon (compound word formation) explicitly manifested in language use.

Analytical significance of compounds

The SweTerror and The Market Language projects are multidisciplinary endeavours that integrate analytical methods for processing large textual datasets with inquiries into actual language usage and the utilisation of particular concepts in the political texts under examination. In this sense, the two projects follow previous research that have explored the development of key concepts in Swedish parliamentary datasets, drawing upon statistical measurements (Jarlbrink et al. 2022; Norén et al. 2022). However, in both projects we have turned our attention to the analytical usefulness of compound neologisms for tracing transformation in historical parliamentary discourse.

Using N-grams is a widely employed method for describing recurrent multi-word combinations in natural language, both in spoken and written discourse (Lyse & Andersen 2012). However, while N-grams are a useful tool to map recurring multi-word combinations, they do not seem to capture the inherent possibility to create neologisms in some languages. Notably, the creation of new words through the synthetic amalgamation of different terms into compounds, functioning as cohesive lexical units, is a distinctive morphological trait of the Swedish language (Finkbeiner and Schlücker 2019). While similar morphological patterns can be observed in other Germanic languages, including the Nordic languages and Dutch, this is not to the same extent the case in English (the dominant language within academic research), which is characterised by fuzzy boundaries between compounds and multi-word expressions (Bauer 2019). This is reflected in irregularities when it comes to the spelling of English compounds; these are often represented orthographically as separate lexical units. (c.f., the English compound labour market and its Swedish equivalent arbetsmarknad.)

While orthographic patterns of compound production of this kind considerably narrow the potential of detecting compounds computationally based on large datasets in English, there is arguably an overlooked potential in discourse analysis based on how Swedish allows for and visualises the creation of compounds by combining existing words to form new morphologically coherent lexical units. This particular form of lexical composition holds relevance as a unit of analysis in computational linguistic studies as well as a discursive phenomenon, offering concentrated semantic information compared with simplex nouns that often necessitate the embedding in multi-word expressions to convey similar meaning.

By examining how our words of interest ('terrorism' and 'market') co-occur with other words in compounds, we will describe significant semantic patterns of word usage that includes evaluative approaches and attitudes. Our analytical tools include word frequency analysis, examination of keyword collocations, and exploration of multi-word expressions, such as phrases, involving the focal keywords. The latter often reveal entrenched phraseological relations for a word, providing insights into its lexical, syntactic, and semantic characteristics (Koteyko et al. 2010). This is a crucial part in understanding the discursive use of concepts – how words are situated within specific contexts – and by examining patterns of compound use and the context of these compounds, we can turn attention to how compounds contribute to the creation of a recurring perception of the phenomena they represent and their role as indicators for conceptual transformation.

Materials

Both projects utilise publicly available Swedish parliamentary datasets. The Market Language project case for this paper draws from the complete data set of texts from the Bicameral Parliament 1867–1970 (riksdagstryck.kb.se)). This data set has been available through the Swedish parliament website in .pdf and .xml formats for some years but have been downloaded by the

project's language technology analysts and processed, annotated and included in the Språkbanken Text infrastructure as a subcorpus ('Tvåkammarriksdagen'), with the possibility to employ sub-genre categories for specialised searches. The period of the Swedish Bicameral parliament 1867–1970 is of particular interest for the project since it covers a time dominated by industrialization and economic changes where the foundations for a modern market economy and also the modern Swedish democracy are laid out.

The work in the SweTerror project, in turn, was at an earlier stage been based on the corpus from riksdagstryck.kb.se, but now departs from one provided by the SWERIK infrastructure of the corpus of the minutes (latest version 0.14) that is currently being cleaned up, partly redigitised, annotated with metadata and curated for research purposes (https://github.com/swerik-project/swerik-project.github.io). SweTerror's language technology analyst Olsson is also further enriching SWERIK's dataset for specific research purposes. In this presentation, we draw upon results from studies based both on the corpus from riksdagstryck.kb.se (Fridlund et al. 2022) and on the SWERIK dataset (Brodén et al. 2023) that together cover both the Bicameral and the Unicameral Parliament periods. The rationale for this is that the concept of terrorism went through drastic changes from the mid 1960s to the early 2000s (Stampnitzky 2013).

Case study: The Market

The results from the Market Language Project so far are primarily based on the productivity of compounds, both in terms of their occurrence and frequency in the current material, as well as the emergence of new compound forms that begin to be used over time in the text material of the Bicameral Parliament. We have previously discussed this productivity aspect of compositional forms in Ohlsson et al. (2022). Our results indicate that new compositional forms exerting significant influence in terms of usage represent areas that gain prominence in political debates and also have the capacity to generate further compositional forms.

A trend that we further explore in this paper, is the consistent increase in new compositional forms featuring "market" as an element, spanning the period from 1867 to 1970 in the parliamentary texts at hand, with an accelerating rate particularly evident from around 1920 onwards, and later also an increase in the post WW2 period aligned with the so called Swedish model of combining a free market economy with redistributive politics. For instance, the emergence of the compound arbetsmarknad leads to the creation of additional compositions, based on that compound. These patterns of compositional productivity serve as a foundation for discussing the growing utilisation of the concept of market in political discourse in general and the attribution of new properties and roles to the concept itself.

Case study: Terrorism

The second case study will focus on the development of the closely related words 'terror' and 'terrorism' when they appear as constituents in compounds, as manifested in the Swedish parliamentary discourse, 1867–2018. We have previously shown that although the word terrorism has been used since 1867, terror-related words and compounds first gained traction from 1918 and onwards with the word terrorism gaining its modern meaning in the early 1970s (Fridlund et al. 2022). Specifically, we found that a distinct legislative framing of the issue of terrorism emerged in the Swedish Parliament in the 1970s and that the 9/11 attacks in 2001 served as a watershed moment for the rise of a stronger counterterrorism discourse and an intensified production of compound neologisms (Brodén et al. 2023; see also Ängsal et al. i.p).

Besides chronologically tracing compound neologisms, we will apply word vectors to examine discursive transformations more deeply, exploring other terms that carry similar meaning as our neologisms. By also considering meta-data about party affiliation of the MPs who use neologisms, will be able to map the extent to which different political parties have used different compounds, thus allowing for a more multidimensional perspective on the role of neologisms in the development of the Swedish parliamentary discourse on terrorism.

Conclusions

We conclude by pointing out how the empirical findings from the two projects provide support for an alternate approach to tracing discursive transformation through text mining compound neologisms and the patterns of compositional productivity. We also note how our focus on the use of the morphology of Swedish, rather than English, directs attention to the potential of drawing upon features of less dominant languages in academic and digital humanities research for text mining purposes.