Enhancing Swedish Parliamentary Data: Annotation, Accessibility, and Application in Digital Humanities

Shafqat Mumtaz Virk
University of Gothenburg, Sweden
shafqat.virk@svenska.gu.se

Claes Ohlsson
Linnaeus University, Sweden
claes.ohlsson@lnu.se

Henrik Björck
University of Gothenburg, Sweden
henrik.bjorck@lir.gu.se

Nina Tahmasebi
University of Gothenburg, Sweden
nina.tahmasebi@gu.se

Leif Runefelt
Södertörn University, Sweden
leif.runefelt@sh.se

Abstract

The Swedish bicameral parliament data presents a valuable textual resource that is of interest for many researches and scholars. The parliamentary texts offer many avenues for research including the study of how various affairs were run by governments over time. The Parliament proceedings are available in textual format, but in their original form, they are noisy and unstructured and thus hard to explore and investigate. In this paper, we report the transformation of the raw bicameral parliament data (1867-1970) into a structured lexical resource annotated with various word and document level attributes. The annotated data is then made searchable through two modern corpus infrastructure components which provide a wide array of corpus exploration, visualization, and comparison options. To demonstrate the practical utility of this resource, we present a case study examining the transformation of the concept of 'market' over time from a tangible physical entity to an abstract idea.

1 Introduction

In recent years, the digitization of historical and contemporary text has facilitated valuable research in text-based fields particularly in digital humanities. While newspapers and literature offer important avenues, parliamentary text are both complementary and important sources of knowledge. In this paper, we present the digitized Swedish bicameral parliamentary data annotated with various token and text level attributes. The corpus is made accessible through the modern corpus infrastructure of Språkbanken, the Swedish Language Bank[1]. By applying a range of token and text-level annotations, we aim to enhance the accessibility and usability of the parliamentary records, making them a valuable resource for a broad range of research inquiries.

The annotation process involves several layers, from basic tokenization to more advanced linguistic tagging. For this purpose, we rely on several in-house developed and external annotation tools. These annotations not only help in structuring the data but also enable researchers to extract meaningful patterns and insights that would otherwise remain obscured. Furthermore, by integrating these annotated records into a modern and well-established corpus infrastructure, we ensure that the data is both easily accessible and scalable for various computational analyses.

To illustrate the practical value of our enriched parliamentary data, we present a case study that explores the evolution of market-related language within legislative and political discourse. Today, the market is a ubiquitous concept in both professional and private matters. However, this was not always the case historically, and 'the market' has gradually developed into the central concept it is today. While conceptual history studies have examined the market, particularly in later periods (Leary, 2019), a continuous description of the market as a concept across various contexts and discourses remains lacking. In response to this gap, a project called "The Market Language"[2] aims to provide an empirical based analysis of the market concept and its usage in Sweden, utilizing available corpora (Ohlsson et al., 2022).

Historically, the concept of "the market" has undergone transformation from an original meaning as a tangible, physical and time-specific space for trade, to more abstract notions like the "iron market," and eventually to a role of an active agent with the potential for influencing professional or private daily life, as in expressions such as "the market reacted badly to the latest inflation news." This linguistic and semantic shift reflects broader socio-economic changes and raises numerous research

[1] https://spraakbanken.gu.se/

[2] https://www.gu.se/forskning/marknadens-sprak-studier-i-talet-om-marknader-fran-medeltid-till-nutid

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 280–288
November 16, 2024.

questions of interest to both linguists and historians and also scholars in several other disciplines. For instance, when and how did this conceptual shift in market language occur? Are there discernible linguistic patterns associated with different types of semantic change? Can linguistic changes in how "market" is used be related to social and historical development or changes and also vice versa?

The annotated parliamentary texts offer a rich and structured resource for investigating these questions. Through this case study, we demonstrate the potential of our annotated corpus not only as a tool for linguistic analysis but also as a gateway to deeper historical inquiries, as done in (De Bolla, 2023) as an example.

By enabling researchers to trace the usage, frequency, and context of market-related terms across time, our corpus can provide insights into the factors driving these linguistic transformations. Studying the change in meaning of the concept of "market" in the Swedish language offers several advantages as an example of how parliamentary data has been incorporated into the corpus infrastructure. The word "market," or "marknad" in Swedish, has a consistent lexical form over time, making it a good candidate for annotation. Additionally, "market" is a concept that appears across various types of texts or discourses. This allows for comparisons between its use in political or legislative discourse and its use in other corpora within the infrastructure, such as media or literary texts. Furthermore, existing research and studies have shown examples and parts of the development of the word "market," from the primarily concrete meaning that is still in use, to become increasingly abstract. The possibility to use a large set of discourse-specific data such as the parliamentary corpus, opens for showing a fuller picture of this development over time.

2 The Swedish Parliament Data

The history of the Swedish parliament or *Riksdag* starts in the 15th century when the Riksdag of the Estates was formed with roots in parliamentary gatherings of Swedish noblemen that span much longer. In the early modern era, this type of parliament was replaced with the bicameral parliament in 1867, which was in function until 1970 when the current one chamber parliament of the modern Riksdag was installed. The bicameral Riksdag existed at a time when Sweden as a society was subject to many changes and went from being a poor, war-ridden and mostly rural country to becoming a modern and industrialized nation with a strong focus on democratic and emancipated values. The fundamentals of today's democratic form of government were essentially laid out, debated, and tried during the periods of the late 1800s and first half of the 1900s when the bicameral parliament was in function. This makes the texts of the bicameral Riksdag important for researchers and scholars from several disciplines and there are other projects focusing on making other features than purely linguistic ones more available such as the The Open Parliament Laboratory (OPaL)[3] at Örebro University in Sweden and also the SWERIK[4] project. Similar efforts to make parliamentary data more available for especially research purposes can be found in the ParlaMint project that encompasses comparable corpora of parliamentary debates from 29 European countries and regions. This is a valuable resource for synchronic data, but the inclusion of the Swedish bicameral parliament data also makes older, historical datasets available for research.

The texts of the bicameral parliament of 1867 to 1970 were scanned by the Royal Library and are available in pdf format via the Riksdag website[5] and the Royal Library of Sweden[6]. The downloadable data files contain some metadata regarding type of text and year but are otherwise unstructured and difficult to use for large-scale searches or for linguistic analysis purposes. The files are also presented per year, which makes the total sum of text files that are accessible relatively high. This makes the data set problematic to handle as a bundle for both researchers, and particularly, an interested public. The interface where the original pdf files can be reached is not developed for quick access to the material in full.

To overcome some of these obstacles, we have transformed the bicameral parliament data into a structured lexical resource by enriching it with a set of word and text level attributes. We have also made the data searchable though *Korp* and *Strix*, which are modern corpus infrastructure tools developed and maintained at Språkbanken Text [7] (the Swedish Language Bank) at the University of

[3] https://www.oru.se/english/research/research-environments/hs/opal--/
[4] https://swerik-project.github.io
[5] https://www.riksdagen.se/en/
[6] https://www.kb.se/in-english.html
[7] https://spraakbanken.gu.se/

Gothenburg. These tools provide a range of options to easily access, explore, visualize, and compare the data that were not previously available.

3 The Corpus Categories and Statistics

The Swedish bicameral parliament data encompasses a variety of text types, including protocols, debates, and motions from individual members of parliament. The type categorization of the material is the original structure of the scanned text file and is used also for our integration of the parliamentary data into the corpus infrastructure, since the categories are based on the different text types of the parliamentary practices. The use of type categorization opens for more specialized searches and also for comparisons of different text types within the data set as well as with texts from both previous and contemporary parliamentary resources. Table 1 shows the list of corpus categories, and statistics on number of documents, sentences and tokens in each category. As can be noted, there are roughly 190k, 46M and 0.9B number of documents, sentences, and words respectively in total. The data contains some metadata as aforementioned (type and year information). These categories follow the rationale for parliamentary work over the period 1867 to 1970, with focus on the process of proposing and debating legislature. 10 different categories are distinguished (as shown in Table 1) and these are now searchable through the interfaces Korp and Strix.

4 The corpus infrastructure

In recent years, there has been a remarkable surge in the production of digital textual data, i.e., corpora, and the conversion of non-digital texts into digital formats. This has simultaneously driven the need for the development of efficient methods for storing and exploring these extensive datasets. Consequently, technology has evolved from basic string-matching search approaches to the creation of advanced corpus infrastructures that offer query-based search, comparison, and visualization capabilities. In the following sections, we will briefly introduce two such tools in the corpus infrastructure domain: Korp and Strix. These tools offer a wide array of options for exploring, comparing, and visualizing corpus and related statistics at word, sentence, and document levels.

4.1 Korp

Korp[8] (Borin et al., 2012) is one of the key components of the corpus infrastructure developed and maintained by Språkbanken[9] (the Swedish language bank). It comprises separate backend and frontend components designed for corpus storage and exploration. The backend is used for importing data into the infrastructure, annotating it, and converting it to various formats for downloading. Korp provides an annotation pipeline for adding a range of lexical, syntactic, and semantic annotations to the corpus, utilizing both internal and external annotation tools. On the other hand, the frontend offers a variety of search options, including basic, extended, and advanced search functionalities, enabling users to extract and visualize search results, annotations, statistics, and comparison between corpora. Section 5 provides some practical examples of these capabilities.

4.2 Strix

Strix[10] is another tool within Språkbanken's corpus infrastructure. While it shares similarities with Korp by offering search and exploration capabilities for text collections and their annotations, it distinguishes itself from Korp in several ways. Most notably, in Strix, a search result corresponds to a document rather than an individual occurrence. Furthermore, Strix includes additional features such as support for metadata filtering, text similarity analysis, and a reading mode with annotation highlighting.

5 Utilizing the Corpus Infrastructure

This section provides a detailed account of the data annotation process and the subsequent steps taken to ensure its accessibility through Korp and Strix. As previously mentioned, Språkbanken's corpus infrastructure utilizes a pipeline architecture, which is known as Sparv. This pipeline encompasses a variety of both external and internal annotation tools designed for a wide range of word, structural, and text-level annotations. A comprehensive list of available annotations and the corresponding annotation tools are listed in the Sparv user manual[11].

[8] https://spraakbanken.gu.se/korp/#?stats_reduce=word

[9] https://spraakbanken.gu.se/

[10] https://spraakbanken.gu.se/strix

[11] https://spraakbanken.gu.se/sparv/#/user-manual/available-analyses

Category	# Documents	# Sentences	# Words
Berättelser och redogörelser (narratives and accounts)	803	3,041,471	61,348,401
Betänkanden, memoria, och utlåtanden (reports, memorandums and opinions)	49,919	7,767,619	195,467,124
Motioner (motions)	77,555	3,324,913	73,189,180
Propositioner och skrivelser (propositions and letters)	19,761	13,516,498	319,201,218
Protokoll (protocols)	10,653	13,092,772	327,554,657
Register (register)	1,262	1,847,991	23,323,395
Reglementen (regulations)	133	129,154	2,628,009
Riksdagens författningssamling (the constitution of the Riksdag)	52	3,880	83,964
Riksdagsskrivelser (letters of the Riksdag)	30,383	1,359,009	29,775,566
Statens offentliga utredningar (government official investigations)	701	2,369,348	59,266,835
Total	191,222	46,452,655	904,138,844

Table 1: Statistics on the number of documents, sentences, and words category-wise and in total in the corpus collection.

	Annotation		Annotation
Word-Level Annotations	sense: marknad	Text-Level Annotations	Blingbring:innanhav
	compound word forms:[empty]		Swedish FrameNet: Amounting_to
	compounds:[empty]		readability index:50.09
	universal features: Case:Nom Definite:Ind Gender:Com Number:Sing		word variation index:66.63
	dependency relation: Complement of preposition		nominal quote:2.21
	sentiment:neutral		file name: kombet_1911____6_5-D8E06b5b1.txt
	Swedish FrameNet: [empty]		category:utredningar-kombet-sou
	baseform:marknad		date:1911
	lemgram:marknad (noun)		
	msd (Morpho-Syntactic Analysis): NN.UTR.SIN.IND.NOM		
	part-of-speech:noun		

Table 2: List of word and text level annotations.

Leveraging Sparv, we have substantially enhanced our dataset with a diverse array of annotations. Table 2 provides an overview of various word and text-level attributes and their respective values illustrated through an example search term, 'marknad,' within the sentence: *Den mesta strömmingen föres saltad till marknad i Linköping*. (Most of the herring is brought salted to market in Linköping)

Some of these attributes are self-explanatory, while others need a brief explanation. The 'universal features' attribute reveals essential linguistic details such as case, definiteness, gender, and number associated with the selected word. The 'dependency relation' attribute provides the dependency relationship of the selected word with the head in the sentence. The 'sentiment' indicates the emotional tone. More details can be found in the Sparv user manual as mentioned above.

Text-level annotations encompass critical metadata, including the corpus category, file name, date, and some other text-level attributes.

Regarding corpus search functionality, Korp offers three distinct types of searches: 'simple,' 'extended,' and 'advanced.' The 'simple' search is the most straightforward free-text search, allowing users to query the corpus for any given string. Figure 1 illustrates the search hits in the Keyword In Context (KWIC) view when searched for the term 'marknad' (ENG *market*). The left-hand pane displays sentences retrieved from all documents within selected corpora containing the search term, while the right-hand pane presents both text-level and word-level attributes (for the selected string highlighted with a black background) as explained above.

The search can be refined or broadened by utilizing the 'Extended' search tab, which permits filtering based on various word and text-level attributes, and combining various attributes with logical AND and OR operators. For instance, one can filter the sentences where the term 'marknad' occurred as a noun only (or any other part-of-speech tag and other attributes).

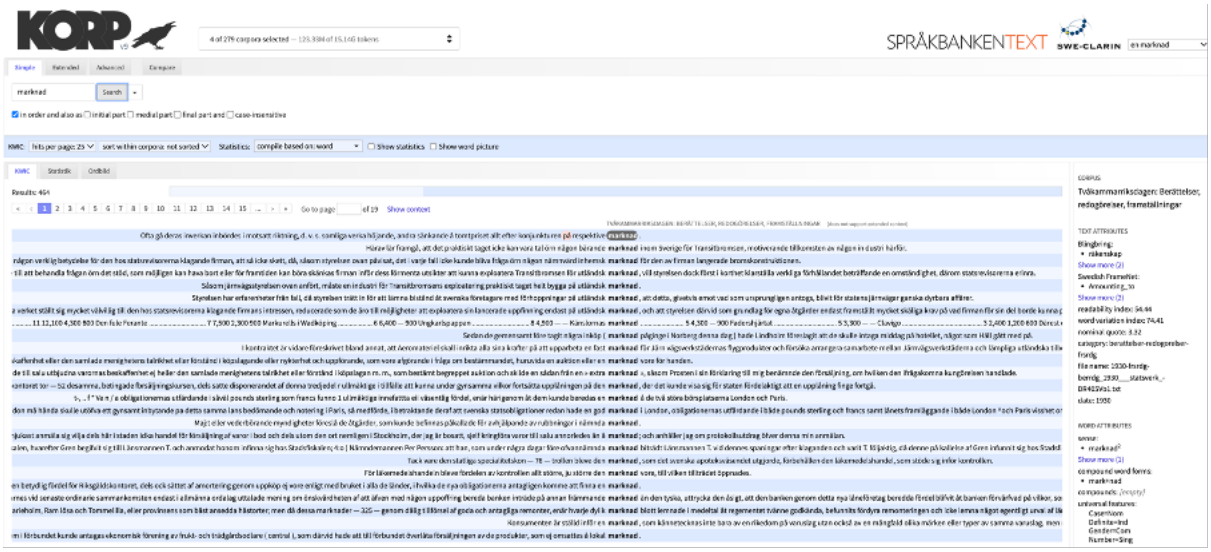

Figure 1: Screenshot of Korp frontend 'Basic' search

Additionally, for more specialized requirements, the frontend offers an 'Advanced' search option, enabling users to design search queries using the CQP query language (Christ, 1994). Beyond these functionalities, the frontend comes with various other compelling features, including the ability to display context (a few sentences before or after the searched term), which can be invaluable during corpus exploration.

In addition to presenting the search hits in the KWIC view, the statistics on total number of occurrences for each matched word in the selected corpora overall and in each of the sub-corpora separately are also available under the 'statistics' tab (see Figure 1. Another useful view is the 'word-picture' view which shows other words associated with the searched word based on certain dependency relations such as subject, object, preposition, pre and post modifiers etc. Figure 3 shows word picture of the word 'marknad' within the 'Riksdag' corpora. Such a view can come handy while analyzing a search word within a corpus by looking at associated words or by comparing it to another corpora (or sub-corpora). For example, in Korp we also have another corpora of the recent Swedish parliament proceedings (Riksdagen-open-data). The word 'marknad' can be compared in these two collections through the word-picture view. Figure 4 present this comparison, and as can be noted, the separate list of associated words in two collections can reveal certain comparative aspects of the corpora. For example, while the bicameral data speaks of the market as Joint, Nordic, domestic, European, foreign, open, large and free, the modern parliamentary data speaks of it with a different focus on free, joint, global, open, black, digital, international and lucrative.

Due to space constraints, we cannot delve into all of Korp's features here, so we encourage readers to visit the https://spraakbanken.gu.se/korp/ to explore the corpus and experiment with the many search options available.

As can be seen in the given screenshots, in Korp each search hit is restricted to 'a sentence' (or a few sentences if the context visualization is turned on). An alternative is to return the documents containing the searched terms as search hits (as opposed to sentences), and then provide an option to view the full document in reading mode. This is exactly, what the Strix tool and interface is designed for. If we search for the term 'marknad' through the Strix interface, a list of documents from the collection containing the search term will be displayed as shown in Figure 5.

This list can be filtered further based on various text-level attributes (e.g. document type, document title, year etc.) using the given metadata filtering options in the left-hand side pane.

Clicking on any document will open the full document in text mode as shown in Figure 6.

Various text and word-level attributes of the selected text are displayed in the right-hand side, while the document itself is displayed on the left

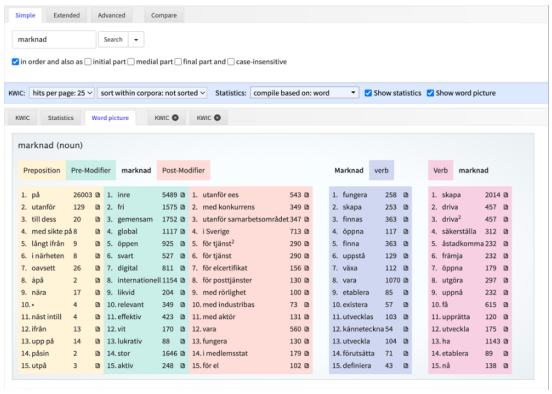

Figure 2: 'marknad' in the modern corpus

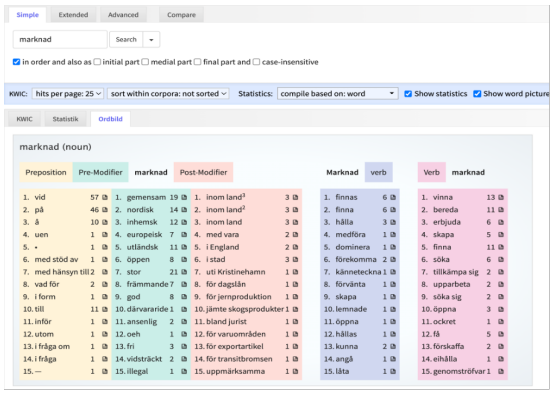

Figure 3: 'marknad' in the bicameral corpus

Figure 4: The word-pictures view in Korp

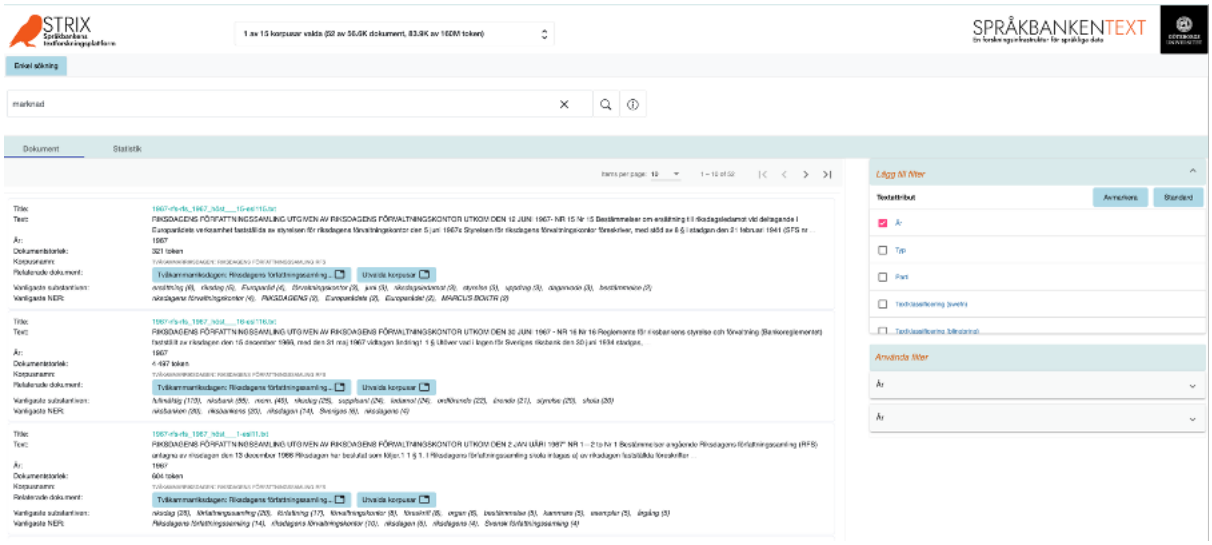

Figure 5: Screenshot of Strix

and side pane. Also note that the selected document can be further searched using the 'Search the current document' search box on top. Again, due to the space limitations, it is not possible to explain all searching and exploring options provided by Strix, and we refer the reader to Språkbanken for further details.

5.1 Resource URL's

The following url can be used to reach the Korp:

https://spraakbanken.gu.se/korp/

After switching to the 'English' mode, a particular corpus (or sub-corpus) can be selected using the drop-down list of available corpora before making the search as shown in Figure 7. Note, the 'bicameral Riksdag' corpus is placed under the 'Governmental texts' category.

Use the following url to access the Strix and then select the 'bicameral Riksdag' corpus from the drop-down list:

https://spraakbanken.gu.se/strix/

Once opened, add filters (if any) in the right-hand side pane, and then make the search as explained above.

6 The Market Language

The research project "The Market Language" has particularly benefited from the inclusion of material from the bicameral parliament in the Swedish Language Bank collection of corpora. The project's need for corpus data spanning different time periods and sources has also played a role in the decision to process this material.

"The Market Language: Studies in the Discourse about Markets from Medieval to Modern Times"

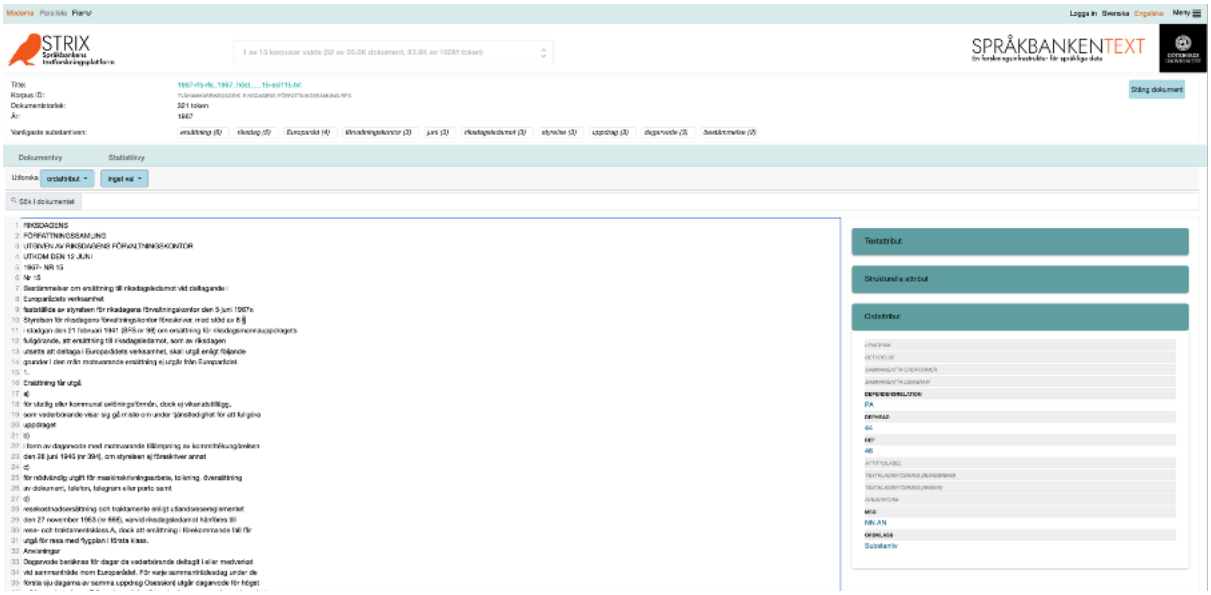

Figure 6: Strix Document View

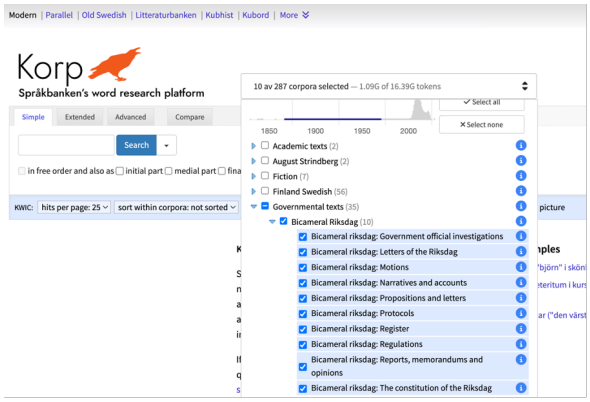

Figure 7: List of available corpora through Korp

is a four-year multidisciplinary project involving scholars from history of ideas, economic history, discourse-oriented linguistics, and computational linguistics. A central goal of the project is to deepen our understanding of the historical development that has led to the market becoming such a dominant concept today. This is achieved by analyzing how the word "market" has been used in the Swedish context over an extended period. A key objective of the project is to describe how expressions associated with the concept of "market" have been employed in Swedish language from the Middle Ages to the present. From the outset, the project aimed to combine qualitative methods from conceptual history and linguistic discourse analysis with primarily quantitative corpus linguistic analysis.

Making the complete records of the bicameral parliament from 1867-1970 available in the Language Bank's frontend has given the project's collaborators numerous new opportunities to conduct searches without needing specific computational linguistic expertise. A notable advantage is that the material can now be used and processed for searches based on linguistic form or language functions in text. While searchable text files were available previously, their format severely limited the ability to observe how specific lexical items, such as words or phrases denoting the market phenomenon, were expressed over time. This is particularly important because the material reflects political discourse at a time when both domestic and international economies were becoming increasingly interconnected, and the meaning of "market" was expanding.

One focus of the project has been to examine how the original, concrete meaning of "market" as a place or location in time has increasingly been replaced or complemented by abstract meanings of "market." By being able to search the bicameral parliament material in the Korp interface, all researchers in the project team, with different levels of computer linguistic training, have been able to confirm working hypotheses and develop new searches about how the market as an abstract concept has been expressed. For instance, a simple comparison of the occurrences of "market" in its indefinite form (marknad) with its definite form (marknaden) reveals a tendency for the definite form to be used more frequently over time. This

suggests that "market" began to be referred to as an entity with increasing agency during this period, which is interesting when compared with its usage in modern corpus material, where the definite form is used to a higher degree. This indicative result can be followed up by finding cases where "market" is used in the subject role in a sentence as resulting from an intra-team discussion and collaboration based on the initial search results. Figure 9 shows an increased use of market as a subject, both in the indefinite and definite forms with a sharper increase of the definite form *marknaden*.

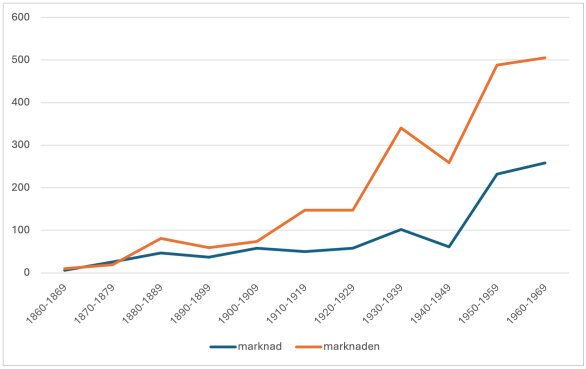

Figure 8: distribution of indefinite and definite forms of ”market” over time 1860–1970 in absolute frequency

In this way, the parliamentary material from 1867-1970 serves as a basis for observing the evolution of currently established meanings.

Another clear example of the benefits for the project of including the bicameral parliament material in the corpus infrastructure is the ability to map and analyze compounds containing "market." Forming multi-element compound words is a typical feature of word formation in Swedish and other Germanic languages with English as a notable exception. Through annotation work and the various functions of the Korp interface, it is easy to list all compounds with "marknad" as an element, both when it serves as a prefix and as a suffix. The project team has conducted many such mappings, and these searches reveal a great variation in the meaning and discursive semantics of these compounds. This type of search can now easily be conducted by all team members in the Korp interface, which facilitates initial interpretation and discussion of results. One more specific result is that new compounds with "market" appear to have been continuously formed over the period covered by the material, often in relation to the political changes discussed in parliament at different times. Compounds with concrete connotations (like "cattle market" or "butter market") are common in the mid 1800s but are gradually replaced by increasingly abstract constructions that indicate the Swedish economy’s connection to other economies (such as "world market" or reference to a "domestic market" in contrast to “foreign markets). Furthermore, the concept of "market" also extends into other political areas (such as "labor market" or "housing market") during the 1900s. Figure 9 shows the increase of the compound word "arbetsmarknad" (labor market) in relative frequency in the parliamentary texts over time. The compound is most frequent towards the end of the bicameral parliament in Sweden in the 1950s and 1960s.

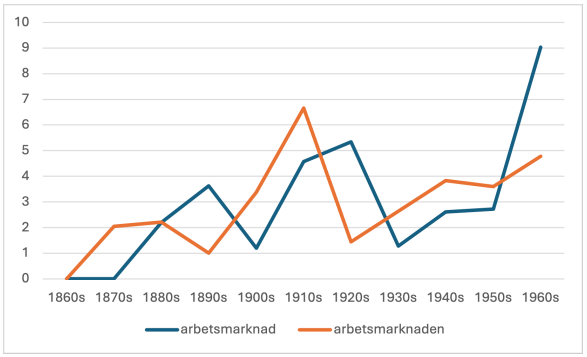

Figure 9: distribution of relative frequency (per million) occurrences of forms for “labor market” over time

7 Conclusions

We have annotated the Swedish bicameral parliament data with text-level as well as token-level attributes to make searching, filtering, and exploration much easier and more useful for broader groups of potential users with different level of technological knowledge. The Market Language project has incentivized the inclusion of the dataset into the infrastructure. The research team is now able to work on the project’s original research questions and can also address new questions that arise from the results with the help of the infrastructure interface. This makes it easier for team members without computational linguistics training and also improves the quality of the output data. Further, the inclusion of bicameral parliament data in the Språkbanken environment helps complete the picture of making public texts in Swedish available,

both for the research community and for other public purposes.

We believe and hope that this collection of texts will be a valuable resource for deeper analysis of the Swedish political discourse during the period when the bicameral parliament was in function. Together with existing corpora of other parliamentary data and government texts, the bicameral parliament corpus will also enable more informed cross-corpora searches. This benefits scholars in historical studies as well as researchers from many other disciplines, not least in studies of language. The data of the bicameral parliament has existed in other formats for a long time, but the inclusion of this material into Språkbanken makes it accessible and also adds a valuable new component to this infrastructure.

Acknowledgements

This work has in part been funded by the research program *Change is Key!* supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp — the corpus infrastructure of språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 474–478, Istanbul, Turkey. European Language Resources Association (ELRA).

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. *ArXiv*, abs/cmp-lg/9408005.

Peter De Bolla. 2023. *Explorations in the Digital History of Ideas: New Methods and Computational Approaches*. Cambridge University Press.

John Patrick Leary. 2019. *Keywords: The new language of capitalism*. Haymarket Books.

Claes Ohlsson, 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, Sweden, March 15, 2022*, pages 22–34. CEUR-WS. org.