4 Applied NLP for humanities research

Gijs Aangenendt
Uppsala University

Maria Skeppstedt
Stockholm University

Karl Berglund
Uppsala University

Natural language processing (NLP) has become a field of interest for many researchers within the humanities. However, framing humanities research questions as NLP problems and identifying suitable methods can be a difficult task. Taking previous and ongoing projects from the Centre for Digital Humanities and Social Sciences at Uppsala University (CDHU) as a point of departure, this chapter presents concrete use cases of how humanities research questions can be approached using various NLP methods and tools, from ready-to-use text analysis tools to programming libraries that require basic familiarity with Python. Two case studies from the field of history and literature will be introduced to illuminate how texts can be processed for humanities research purposes. With this chapter, we hope to give the reader the means to directly explore NLP methods for their research as well as encourage further learning.

1 *Introduction*

With the increasing availability of texts in digital form, natural language processing (NLP) has become a field of interest for many researchers within the humanities. Researchers turn to NLP to extract information from collections of texts that are too large to manually analyze. This promise of "distant reading", to use the term coined by Franco Moretti, is one of the main attractions of NLP for the humanities (Moretti 2000).

Despite its potential, utilizing concepts and methods from the technical-linguistic field of NLP can be challenging. Framing humanities research questions as NLP problems, identifying suitable methods, evaluating their performance, and finally building arguments based on the output is not straightforward due to fundamental differences between the domains (Mcgillivray et al. 2020). When starting to engage with NLP, scholars often need to learn

HUM INFRA Gijs Aangenendt, Maria Skeppstedt & Karl Berglund. 2025. Applied NLP for humanities research. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 77–111. University of Tartu Library. DOI: 10.58009/aere-perennius0173

several new things at once, including technical skills (e.g., basic programming and data wrangling), statistical and mathematical concepts (e.g, "what is standard deviation?", "what is a vector?"), and – not least, but often forgotten – linguistic concepts (e.g., "what is a lemma and why does it matter?"). Resources for learning and applying NLP, however, are often technical or linguistic in nature and may not depart from a humanities point of view or be tailored to the kinds of sources commonly used in the humanities. Published research from the humanities that utilizes NLP methods may have to prioritize presenting arguments over showcasing the technical process in detail due to space constraints. As a result, learning from these studies and gaining insight into how the texts were processed becomes more difficult.

This chapter offers humanities scholars curious about the possibilities of NLP in their research a first introduction to the field. It begins with a broad overview of NLP in the humanities, including commonly used NLP methods as well as available resources and tools. Following this, two case studies present concrete examples of how research questions, from the fields of history and literature specifically, can be approached with the help of NLP. The history case study employs a ready-to-use text analysis tool, while the literature case study utilizes programming libraries that require basic familiarity with Python. With this chapter, we hope to give readers with varying levels of technical expertise the ability to experiment with NLP in their own research. Notebooks and code examples connected to this chapter are available in a GitHub repository.[1]

2 *NLP and the humanities*

Within NLP, language is analyzed in various ways, from how it sounds and is spoken (phonetics and phonology), how words and sentences are structured (morphology and syntax) to the meaning of words, word combinations and expressions (semantics) (McGillivray 2022). These approaches to language have been adopted to support humanities-based research in a variety of tasks, ranging from tracking semantic change of words to extracting topics from a collection of texts. In this chapter, we adopt a broad definition of NLP that encompasses any attempt to analyze natural language with the help of computers (Schofield 2022).

When humanities researchers use NLP, they do not always directly refer to it as such. Instead they may use terms like "distant reading" (Moretti 2013, Jänicke et al. 2015, Drucker 2017) or "text mining" (Jockers & Underwood 2015, Thompson et al. 2016), which are more closely tied to the humanities

1 https://github.com/CDHUppsala/Applied-NLP-for-humanities-research

field and its practices. There are also subfields such as "digital history" (Fridlund et al. 2020) or "computational literary studies" (Bode 2023) in which NLP methodologies are adapted to support humanities research tasks. This indicates that NLP is often not a goal in itself, but rather a means to an end. Humanities scholars use computational methods from the field of NLP in the hope of gaining new insights into a body of texts, supplementing knowledge gathered by means of traditional humanistic close reading. NLP analyses can provide meaning for the humanities in multiple ways: by detecting patterns and trends in large collections of text, but also identifying outliers that break the pattern. Humanities NLP work often combines quantitative analysis of trends with qualitative analysis of outliers, which may help explain the trends better (So 2017, Underwood 2019).

There is a wide range of NLP methods that can be used for humanities research purposes, making it difficult to choose which method is most suitable. Selecting methods can depend on the research question, the phase of the research process, and the material at hand. Choosing the method also depends on the level of transparency and explainability needed, which is especially relevant when it comes to communicating research results to peers. Results from word frequency-based methods are easier to interpret, validate, and explain than results derived from complex machine learning algorithms. A guiding principle – and here we follow de Bolla et al. (2019) – could be to keep the method *as simple as possible*, not least for the sake of transparency. If no one in your specific humanities field understands what you have done, the research will likely have less impact than if the methods deployed are more intuitively graspable.

When looking for relevant methods and tools from the field of NLP, or any other field for that matter, it is important to be mindful of potential differences compared to the humanities domain. Within NLP, researchers often aim at improving the state-of-the-art and advancing the field, trying to develop models that perform better than its predecessors on gold standard corpora. These corpora have been (at least to some degree) consistently annotated by two or more annotators to ensure reliability in evaluating model perfomances (Artstein & Poesio 2008). The performance of models is typically measured by precision, recall, and the F1-score, the harmonic mean of precision and recall. These evaluation metrics make it easier to compare different models.

Humanities researchers, on the other hand, are interested in whether the NLP method contributes to answering their research questions. The choice of method might therefore not primarily depend on the state-of-the-art and reported F1-scores. Evaluating the performance of the method may be done by examining the output qualitatively. Not in the least considering that

incorrect predictions can still be valuable and provide new directions for further qualitative analysis (Rettberg 2022). Additional factors for deciding which method to use can be 1) the explainability of the method - how well is the model able to show what features were used to make predictions, 2) how well the model is able to state the confidence with which it makes its predictions, and 3) how well the method performs on other corpora (or how easily it can be adapted to other corpora) than the one used for developing the method.

Whether methods can be applied to other types of corpora is especially relevant as humanities researchers tend to make use of project-specific corpora. Most NLP models are trained on specific datasets from a certain language – often English – as well as a particular period and domain, for instance Wikipedia pages or contemporary news articles. Consequently, these models do not necessarily perform well on other corpora (Plank 2016, Bamman 2017), including those typically used within the humanities with messy, historical, domain-specific, or odd and experimental language. Tasks considered 'solved' in NLP may therefore still be difficult to implement and require additional research to develop the best approach for the specific time period, domain, and language of the materials (Mcgillivray et al. 2020).

Finally, it is worth mentioning the importance of data acquisition. Depending on the types of sources relevant to the project, this may involve scanning documents and applying optical character recognition (OCR) or handwritten text recognition (HTR), web scraping, or transcribing audio files using speech-to-text techniques. The quality of the data gathering influences the subsequent application of NLP methods. When digitizing text materials, achieving high-quality OCR or HTR is crucial, as the performance of NLP tasks can drop when accuracy falls below 80 or even 90 percent (van Strien et al. 2020). Historical spelling, fonts, and scripts, as well as poor print or scan quality, can complicate achieving high-quality textual data. For born-digital text materials, problems often concern access, copyright, and data cleaning of various sorts.

3 *NLP methods in humanities research*

Once a corpus, or collection of texts, has been acquired, there are, broadly speaking, two main approaches for how one could study its content. In scenarios where the researcher knows in advance what they want to find or extract from the corpus, a *supervised* approach can be taken. In this case, the intended research approach must somehow be translated or transferred into an NLP solution. This translation may involve creating a list of search

terms to retrieve relevant texts or implementing a *rule-based system* – for example a classic expert system (Russell et al. 1995), where explicit logic or programming rules determine how the texts are processed. *Supervised machine learning models*, in contrast, learn by example. Here, the research approach is transferred through annotated example data, which the model then uses to learn to automatically produce similar annotations on previously unseen texts. Finally, more recent developments include *prompt engineering* (Hu et al. 2024). Similar to rule-based systems, the research approach is transferred through explicit instructions. However, instead of rules, these instructions are formulated in natural language to be interpreted by a large language model (and in practice annotated examples can also be provided as part of the instructions).

When a researcher does not have a predefined approach for how to study the texts, a more exploratory *unsupervised* approach may be taken. The aim is then to extract interesting information based on the content of the texts themselves. This could, for instance, be information retrieved with the help of corpus linguistics methods, such as identifying the most frequent words or examining differences in word usage between texts. Another option is to use NLP methods that, in an unsupervised fashion, extract categories from the text itself, e.g., based on word co-occurrence patterns. Unsupervised NLP methods require no rules or annotated data as input from the researcher. Instead, the bulk of the work consists of preprocessing the texts, with the aim of preserving the meaningful content, followed by interpreting the output. In practice, supervised and unsupervised approaches are often combined.

In this section, we present a number of NLP methods that are commonly utilized for humanities research purposes. As examples, we draw from a pool of past and ongoing projects conducted in collaboration with the Centre for Digital Humanities and Social Sciences at Uppsala University (CDHU). These projects form a set of concrete and realistic example cases of how NLP and corpus linguistics are used by researchers in practice. The examples are either drawn from longer research projects or from a pool of shorter pilots, which have been part of the centre's pilot project program. By briefly describing these examples, we aim to introduce and exemplify some of the key concepts required to understand the field of NLP, and also to provide a practical perspective on how NLP can be used in humanities research.

3.1 *Word frequencies*

Many projects in our example pool employ methods from corpus linguistics to count word frequencies in the studied text collections. As mentioned above, two main approaches can be identified: 1) a supervised approach,

if the words of interest are known beforehand, or 2) an unsupervised exploratory approach for the automatic extraction of potentially interesting words. It could, of course, also be relevant to apply both of these approaches to the same text collection. The three projects discussed below all study word frequency change in text collections that span longer periods of time. It is, however, also common to study word frequencies on synchronic text collections or in cases where the temporal dimension is not relevant for the aims of the study, for example, when comparing texts across genres or authors.

The first example concerns a pilot project on the transformation of language in art criticism. Word frequency methods were applied on a collection of texts on Beethoven's piano sonatas from English newspapers and periodicals dating from the 1850s onward. The aim was to analyze the use of descriptive and evaluative terms over time (Dubremetz 2023). Since the researchers already knew which words they were interested in, a predefined set of keywords was used, containing terms like ”passionate”, ”romantic”, ”beautiful” and ”expressive”. A timeline based on the frequencies (normalized for the size of the text collection also changing over time) visualized the variation and patterns in the usage of these terms over time.

A second example comes from a pilot project that studied the career of the term 'Alt-Right' in Twitter posts from 2008 to 2020 (Ekeman 2023, Piqueras 2023). Using word frequency methods, this pilot aimed to determine when the term began to be used frequently on Twitter, how its usages varied over time, and how the content of the tweets containing the term evolved over time. Similar to the previous example, the content of the Twitter posts was analyzed through the temporal frequency changes of a set of predefined keywords, referring to names, ideologies, movements, events, and activities.

In addition, the most important words in tweets containing the term “Alt-Right”, after lemmatization and stop word removal, were automatically extracted with the word importance measure *TF-IDF* to find out how the content of the tweets varied. TF-IDF stands for “term frequency, inverse document frequency”, and it normalizes the term frequency by the number of documents in which the term appears (where document can mean, e.g., a text or time span, depending on the relevant unit for comparison). The intuition behind TF-IDF is that words appearing in many documents or time spans are often less informative (Liu 2011: 189). These words are therefore down-weighted, while the more informative words that appear in fewer documents receive a higher TF-IDF score.

The third example examined relative word frequency differences across four decades (1950s to 1980s) in the journal Allergia, published by the Swedish Asthma and Allergy Association (Söderfeldt et al. 2019). By study-

ing changes in word frequency over time, the researchers aimed to identify vocabulary shifts in the journal that could reflect transformations within the field of allergy. In the first part of the study, the relative frequency of a set of predefined keywords related to contact allergy (e.g., laundry/dish-washing soap, shampoo), including compound words (machine dishwasher soap), was measured across the four decades studied. After comparing the results with a close reading of the periodical, the researchers discovered that the predefined keywords did not capture the discourse on contact allergies, missing terms related to specific materials and substances (e.g., glue, rubber, turpentine, sodium lauryl). Recognizing this limitation, a second approach was devised where the statistical measure *log-likelihood* was used to extract over-represented words for each decade relative to their frequency in the entire corpus. Instead of relying on a predefined list of keywords, this method allowed interesting words to be extracted from the corpus itself.

3.2 *Word contexts*

A possible next step in investigating the content of a collection of texts, is to not limit the study to the words themselves, but to also investigate the linguistic or textual contexts in which they appear. The idea behind this approach is that the meaning of words can be inferred from their contexts.

An extensive study from our example pool explored the conceptual career of the word "climate" over the period 1800–2010, using two methods to study word contexts (Boyden et al. 2022). The aim of the study was to show that the concept of climate as a global entity – that is, in the meaning the concept has today, as "the climate" – only emerged in the later decades of the twentieth century. In terms of text materials, the study used the Corpus of American Historical English (COHA), a corpus covering roughly 400 millions words from a variety of historical sources: fiction books, magazines, and newspapers.[2]

The first method for studying word contexts concerns collocational analysis from the field of corpus linguistics. *Collocates* are words that have a syntagmatic relationship with the keyword, occuring within a specified window of *n* number of words to the left and right of the keyword. These words co-occur more frequently together with the keyword and can have different part of speech (e.g., noun, verb, adjective) as the keyword (Rapp 2002). In the study, the researchers tracked how collocates of "climate", i.e., words frequently used together with "climate" in a sentence, varied over time, with a particular focus on qualifying adjectives. This showed how in the first period studied (1810-1850) "tropical" and "temperate" were the

2 https://www.english-corpora.org/coha/

most frequent qualifiers, while in the latter half of the twentieth century (1960-2000), "political" and "global" become more prominent.

The second method applied in the study concerns *word embeddings*. Word embeddings capture semantic similarity between words, defined as the extent to which they occur in similar contexts (i.e., paradigmatic similarity). Words with high semantic similarity tend to co-occur with similar words, and generally belong to the same part of speech. They can often be switched out for one another without significantly altering the grammatical structure of a sentence (Rapp 2002). For an introductory non-technical discussion of word embeddings, see McGillivray (2022). In the climate study, word embeddings are used to calculate similarity distances between "climate" and other words. For instance, the study shows how the two words "climate" and "environment" appeared in very different contexts in the beginning of the nineteenth century, but that the similarity of these contexts increased over time, and today the two words are used in very similar contexts.

Word embeddings were also used to study word frequency changes in patient organization periodicals and medical professional journals in the project Acting Out Disease: How Patient Organizations Shaped Medicine (ActDisease).[3] Unlike the climate study, which focused on distances between and clusters surrounding specific words, the ActDisease project uses word embeddings to create multiple *clusters of semantically similar words* and to analyze how the content of these clusters changes over time. For instance, the clusters show how prominent names, medical professions important to the field, or frequently mentioned medical problems varied over time within the corpus. It is also possible to compare different time periods on a semantic level. For example, instead of just observing that the word "driver" was frequently used during a certain period, one can observe that an entire cluster of trafic-related words was prominent during that time (Skeppstedt et al. 2025; Chapter 6 in this handbook).

3.3 *Text categorization*

Text categorization, or classification, can be used to automatically assign texts to organized categories, which is useful when you have acquired a collection of texts that is too large to manually organize into categories. Text classifiers can help to investigate the distribution of categories in a collection of texts or extract texts belonging to specific categories for further (manual or automatic) analysis. Again, two different approaches can be taken for the task of text categorization. If the relevant categories are known beforehand, a supervised approach should be used (Marsland 2009: 7–11). Instead, if

3 https://www.actdisease.org/

the aim of the study is to discover categories based on the content of the texts themselves, an *unsupervised* approach is more appropriate (Marsland 2009: 195–200). For supervised text categorization, the predefined categories may be standard categories, as in the case of sentiment analysis ("negative", "neutral", "positive"), or specific categories tailored to the research question.

A concrete example comes from a pilot project on the commodification of wild berries. In this project, a transformer-based Finnish BERT model was fine-tuned to carry out a binary classification of nineteenth century Finnish newspaper articles about berries and berry picking. The two categories distinguished whether the articles discussed the use of the wild berries as an economic resource or not (Vats 2022, La Mela & Vats 2023). To fine-tune the BERT model, trained on general Finnish language data, and teach it to perform this particular classification task, 415 articles were manually annotated according to these predefined categories.

Another example from the ActDisease project involves classifying texts from patient organization periodicals into different genres, such as academic, news, advertisement, and guidance (Danilova & Söderfeldt 2025). Automatic genre classification makes it possible to perform more focused studies on parts of the text collection, in this case the patient organization periodical. For instance, it could be used to filter out genres not central to the study – such as advertisements – or to extract texts belonging to a specific genre of interest – e.g., the advertisements if these form the focus of the study. In order for the model to learn the features of each of the genres, a set of manually classified examples was required.

In contrast, when relevant categories are not known in advance, unsupervised text categorization offers yet another way to gain insight into large corpora without a predefined approach. A commonly used method for unsupervised text categorization, in particular within humanities research, is topic modeling (Du 2019). Topic modeling is used to discover hidden themes or topics in large corpora (Blei 2012). Automatically extracted topics are typically represented by a list of associated words that form a topic, as well as a list of texts where this topic occurs. There are various topic modeling techniques. Some rely on co-occurrence patterns within the texts (Boyd-Graber et al. 2017), while others, such as BERTopic, use neural network-based representations of text units to generate topics (Grootendorst 2022).

An (unpublished) example of topic modeling concerns a pilot project on survey data related to eHealth. Open-ended survey questions form a prototypical case where unsupervised classification is suitable; the interesting categories are not known beforehand, but it is reasonable to assume that topics can be found in the data, since some respondents will likely have provided similar answers (Baumer et al. 2017). There are also cases in our

example pool where topic modeling has been applied to corpora spanning multiple years or decades. In such cases, it is possible not only to identify potentially relevant text categories in the corpus, but also to create a timeline visualization showing how the occurrence of these topics changes over time (Söderfeldt et al. 2025, Skeppstedt et al. 2024).

3.4 *Extracting entities from text*

The final NLP task discussed here is *named entity recognition and classification* (NER). The aim of NER is to automatically detect spans of text that refer to names or other specific semantic types (Jurafsky & Martin 2008: 759–768). Commonly extracted entity types include personal names, locations, geopolitical entities, dates, times, and organizations. These extracted entities can be used as a basis for network analysis (Tamper et al. 2019), spatial and geographic text analysis (McDonough et al. 2019), or for improving information retrieval in databases (Brandsen et al. 2022).

In addition to the above mentioned standard entity types, study-specific entities are often defined based on the particular research needs. This was the case in a pilot project on Swedish excavation reports. To improve access to information on houses in the reports, five custom entity categories were created related to house type and construction details. A general Swedish BERT language model, pretrained by the National Library of Sweden, was fine-tuned to detect these entities in the reports. Since this project relied on supervised machine learning, a manually annotated dataset was created to use as training data, consisting of 1 000 examples of entities that were to be extracted (Maen 2024).

Another approach to extracting entities was used in a pilot project on digitized eighteenth century handwritten bills of sale. In this case, prompt engineering with the large language model Mixtral was employed to identify and extract entities such as buyer, seller, sum, village, and region (Badri 2024). These entities were then stored as structured data for further analysis. Unlike the project on excavation reports, this approach did not require a manually annotated dataset for training, although annotated examples of bills of sale were included as examples when instructing the model.

4 *NLP resources and tools for the humanities*

Getting into NLP does not mean you need to start from scratch. A wide range of existing resources and tools are available, offering suitable starting points for exploring NLP in the context of humanities research. These resources vary in terms of applicability and technical skills required to use them.

The availability of NLP resources depends on the language in question. Some resources are language-independent, meaning they can be applied to any, or at least many, languages, while others are specifically designed for one or a limited number of languages. For low-resource languages, there may not be any pretrained models, annotated corpora, or larger textual datasets available, which raises the bar for starting to experiment with NLP. Furthermore, tools and models developed for high-resource languages are not always easily adapted to low-resource languages, especially when there are substantial differences in writing system (Tasovac et al. 2023, Dombrowski & Burns 2023).

Tools exist that do not require any programming knowledge. Voyant Tools[4], a web-based application for text analysis, and AntConc[5], a text-analysis toolkit, are examples of programs that allow users to perform computational text analysis through a graphical user interface. While these tools provide suitable starting points for experimenting with NLP and corpus linguistics, they do have certain limitations (Arnold & Tilton 2019). As a user you are limited to the functionalities the developers have decided to include in the tool, which may not be sufficient for answering particular research questions or working with certain types of text materials (Jockers & Underwood 2015). In the case of Voyant Tools, for instance, texts need to be uploaded to a server for analysis, which cannot always be done with sensitive or copyrighted data.[6] Additionally, in terms of transparency, it can be unclear how exactly the texts have been processed behind the interface. Lastly, these types of tools also tend to work better with small or medium-sized corpora rather than large text collections.

Resources are also available for those with knowledge of programming. Python libraries such as the Natural Language Toolkit (NLTK)[7], spaCy[8], and Gensim[9] are commonly used to perform NLP tasks such as those discussed in the previous section. Compared to ready-made tools, these libraries offer a higher degree of flexibility and customization, more transparency and control over the processing steps, and the ability to work with larger text collections. However, they require a higher level of technical expertise, including programming skills and the ability to troubleshoot when packages do not work straightaway. Interactive coding notebooks, such as Jupyter

4 https://voyant-tools.org/
5 https://www.laurenceanthony.net/software/antconc/
6 Voyant Server is a standalone version of Voyant Tools which can be installed and run locally on a computer https://github.com/voyanttools/VoyantServer
7 https://www.nltk.org/
8 https://spacy.io/
9 https://radimrehurek.com/gensim/

Notebook[10], Google Colab[11] and Kaggle[12] take away some of these challenges. In these notebooks, code is presented in executable blocks often alongside explanations of what each block does. Using and modifying existing notebooks reduces the need to code from scratch, but still requires familiarity with programming, to load data correctly, adjust parameters, and install the necessary libraries.

There are also programming resources designed specifically for humanities researchers and their research questions, often in the form of interactive notebooks. The website Programming Historian offers beginner-friendly tutorials for humanists interested in learning about digital tools and techniques.[13] The website includes a variety of tutorials with step-by-step instructions on NLP topics such as word embeddings, topic modeling, and sentiment analysis. Another resource in this category is the course Introduction to Cultural Analytics and Python designed by Melanie Walsh.[14] The course has a broader scope than just NLP. Alongside a chapter on text analysis, it includes introductory chapters on the basics of Python, data collection and analysis, and network analysis.

In the Swedish context, the Huminfra nodes provide access to many useful resources on NLP and other topics relevant to the digital humanities. The National Library of Sweden publishes language models trained on Swedish data that can be used to perform NLP tasks, for example models that can be fine-tuned for entity extraction as exemplified in the pilot study on Swedish excavation reports.[15] Språkbanken offers digital tools for textual analysis and access to Swedish language resources.[16] For example word embedding models trained on Swedish historical newspaper material (Hengchen & Tahmasebi 2021) and a variety of research tools such as Mink, with which texts can be lemmatized and tagged, e.g., with part-of-speech.[17] Another resource is Efselab (Efficient Sequence Labeling, Östling 2018)[18], which we have used in the case studies for lemmatization and part-of-speech tagging.

10 https://jupyter.org/
11 https://colab.google/notebooks/
12 https://www.kaggle.com/
13 https://programminghistorian.org/en/
14 https://melaniewalsh.github.io/Intro-Cultural-Analytics/welcome.html
15 https://huggingface.co/KBLab
16 https://spraakbanken.gu.se/
17 https://spraakbanken.gu.se/mink/
18 https://github.com/robertostling/efselab

5 *Case studies*

In this section, two case studies are discussed to present concrete examples of how methods from NLP and corpus linguistics can be used to answer research questions in the field of history and literature. The case studies illustrate how textual data can be collected, processed, and analyzed in various ways, using both a ready-to-use text analysis tool and programming libraries. Considering the abundance of available resources and tutorials for working with English texts, we will provide examples using Swedish corpora. The basic principles discussed in this section can, however, be applied to other languages of interest. The history case study uses parliamentary motions retrieved from the open data portal of the Swedish parliament (*Riksdagens öppna data*), while the literature case study draws on novels retrieved from The Swedish Literature Bank (*Litteraturbanken*).

5.1 *Case study 1: History*

Parliamentary records are often used by historians interested in applying NLP methods. From a practical perspective, their advantages for computational exploration include accessibility, a relatively uniform structure and format, and coverage across long time periods. From a research standpoint, parliamentary records are an interesting resource to explore as they reflect the political, economic and societal developments of a given era (Ohlsson et al. 2022). This case study explores parliamentary motions. Motions are proposals for decisions submitted by one or more members of parliament. They can be introduced either in response to a specific proposal from the government or during the general private member's motion period when motions can be submitted on virtually any topic.[19]

To explore and analyze the motions, we use the text analysis toolkit AntConc (Anthony 2024). AntConc offers various tools for text analysis and corpus exploration. The program is user-friendly and requires no programming experience, although preprocessing the texts in advance can help get more meaningful results out of some of the tools. Laurence Anthony, the developer, has created a series of video tutorials on YouTube that show how to get started with AntConc and use the tools included in the program.[20] For the purpose of this case study, we will look more closely into the following tools: the Word list, Key Word in Context (KWIC), File, Collocate, and Keyword list tool.

19 https://www.riksdagen.se/en/news/the-general-private-members-motions-period/

20 Search for 'AntConc (Version 4.2) Tutorials' https://www.youtube.com/playlist?list=PLiRIDpYmiC0SjJeT2FuysOkLa45HG_tIu

For historians, it is useful to be able to zoom in and out of a corpus, looking at one text at a time as well as getting a broader overview of the corpus content, moving between so-called 'distant reading' and 'close reading'. AntConc supports this approach. By dividing the text collection into subcorpora it is also possible to study temporal change. For example, Jarlbrink & Norén (2023) use AntConc to study the evolving concept of propaganda in parliamentary debates from 1871 to 2019, combining close reading with the analysis of n-grams, co-occurrences, and keywords. Moving between distant and close reading helps overcome some of the limitations of large-scale text analysis. Calculated word frequencies and word co-occurrences do not automatically return significant or relevant terms that indicate change. Domain knowledge of the materials and their historical context, as well as returning to the original texts, is essential for interpreting the results and drawing meaningful conclusions (Guldi 2023).

The texts used for the AntConc experiments were retrieved from the open data portal of the Swedish parliament (Riksdagens öppna data).[21] Motions put forward by members of parliament between 1971 and 1979 were downloaded as plain text files along with a metadata file containing the date of the motion, its title, and initiator(s). Party affiliation is not included in the metadata and can therefore not easily be utilized for analysis. In total, the corpus consists of 20 972 motions.

Besides the data portal of Swedish parliament, motions can also be accessed and explored in a web interface developed by Gothenburg Research Infrastructure in Digital Humanities (GRIDH) in collaboration with Chalmers University of Technology.[22] Additionally, the project Swedish Riksdag 1867–2022: An Ecosystem of Linked Open Data (SWERIK) is in the process of making parliamentary records, including motions, more accessible for quantitative and qualitative research.[23] Proceedings from parliamentary debates curated by SWERIK can be explored through an interface that supports keyword searches, filtering by year and party affiliation, and the study of collocates and n-grams.[24] The curated corpus of motions are at the time of writing this chapter only available on their GitHub repository as XML files.[25] Users with programming experience may prefer to retrieve motions

21 In the data portal, a wide range of parliamentary documents are available, including proceedings from debates, motions, bills, and public investigations (SOU). https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/

22 https://riksdagsmotioner.dh.gu.se/

23 https://swerik-project.github.io/

24 https://riksdagsdebatter.se/

25 https://github.com/swerik-project/riksdagen-motions

from this repository, as the SWERIK files have improved OCR and party affiliation included in the metadata.

This case study demonstrate the use of AntConc for historical research in two scenarios: one where we already know what we want to explore in the corpus and one where we adopt an exploratory approach and aim to discover patterns within the corpus itself. For the first scenario, we are interested in examining motions discussing the topic of nuclear power. Nuclear power as an alternative source of energy became an item on the Swedish political agenda during the 1970s as a result of the global oil crisis.

Loading files into AntConc can be done in the corpus manager (navigate to "File" and click on "Open Corpus Manager"). To load the motions, choose "Raw File(s)" as corpus source, give the corpus an appropriate name, and add the files or directory containing the motions that you would like to include in the corpus. For this case study, we will use the standard settings. Press create. Once the process is finished, a database file (.db) is created, which can be reopened at any time from the corpus manager. AntConc supports common file formats, including text (.txt), word (.doc and .docx), and pdf files (.pdf). In the corpus manager, you can see how many files the corpus contains, the number of tokens (in this case the number of words), and the number of token types. In our corpus, there are 20 972 motions, 15 433 308 tokens, and 299 311 token types.

With the corpus loaded into AntConc, we can start exploring its content. The **Word list tool** lets you count the frequencies of words in the corpus. This can be done in two ways. By pressing start without entering any words in the search bar, you get an overview of the frequency of all the words in the corpus. It is also possible to count frequencies of a specific term or multiple terms. The asterisk symbol (*) before or after the word can be used to find variations of that word. Two vertical bars (||) between the words function as OR, allowing you to search for two or more words. Other available wildcard symbols include the question mark (?) for any single character or the plus symbol (+) for one or more characters. These wildcards can also be used in some of the other tools described below.

In addition to frequency, the range of the words in the corpus is also indicated in the Word list tool, that is the number of documents in which the word appears. Range helps determine whether a word is used occasionally across many texts or frequently in only a few. This is important when making arguments about the content of a corpus: if a word occurs a hundred times in one but very rarely in other texts, it says a lot about the content of that particular text but it may not be representative of the corpus as a whole.

To get an overview of the frequency of the word nuclear power and its variations, we can search for "kärnkraft*". In this case, the definite form of

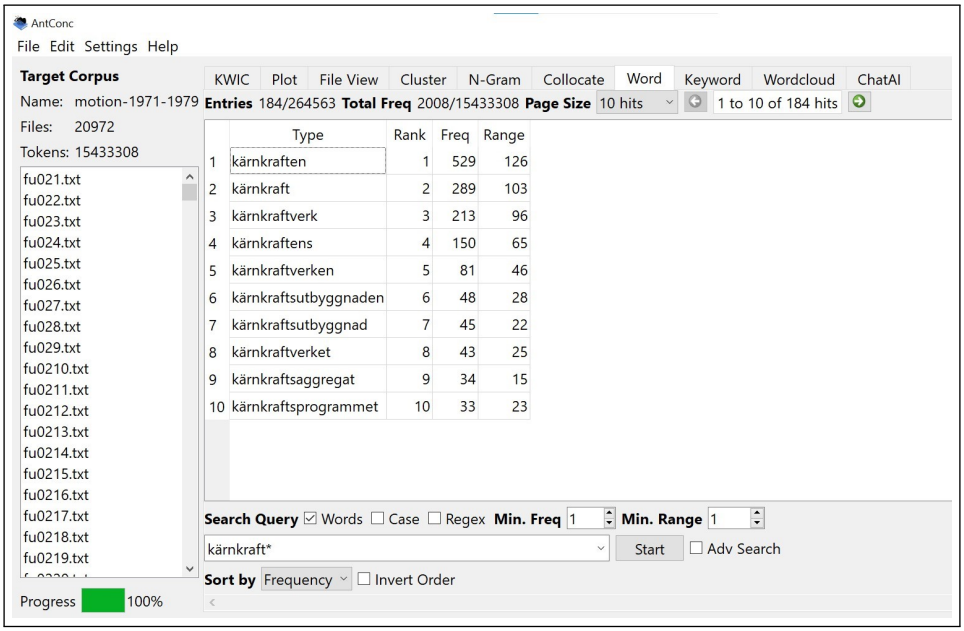


Figure 1: Ten most frequent words of nuclear power and its variations.

nuclear power (kärnkraften) is the most common in the corpus, appearing 529 times (Freq) in 126 motions (Range), which corresponds to 0.6% of the total number of motions in the corpus (Figure 1).

If you find an interesting word that you would like to inspect more closely, you can click on it, and the program redirects you to the **KWIC tool**, also known as the concordance tool. Here, all occurences of that word are listed with a context window of *n* number of tokens on each side of the word. The hits can be displayed in different orderings (see Sort Options), for example, based on the context to the right or the left of the word. This way patterns in the language use can be identified. It is also possible to export the results as a tab separated text file (File > Save Current Tab Results) or csv file (File > Save Current Tab Database Tables) for further analysis in another program.

If we want to learn more about how the motions discuss the construction or expansion of nuclear power, we can, for example, click on "kärnkraftutbyggnaden". Now, the 48 hits for this word are displayed in the KWIC tool, sorted by the right-hand context of the word (Figure 2).

Again, if you find an interesting instance of a word in the KWIC tool that you would like to examine more closely, it is possible to navigate to the content of a single file. Clicking on a hit will redirect you to the **File tool**. Here, the full-text of the motion will be displayed with the hit(s) highlighted.

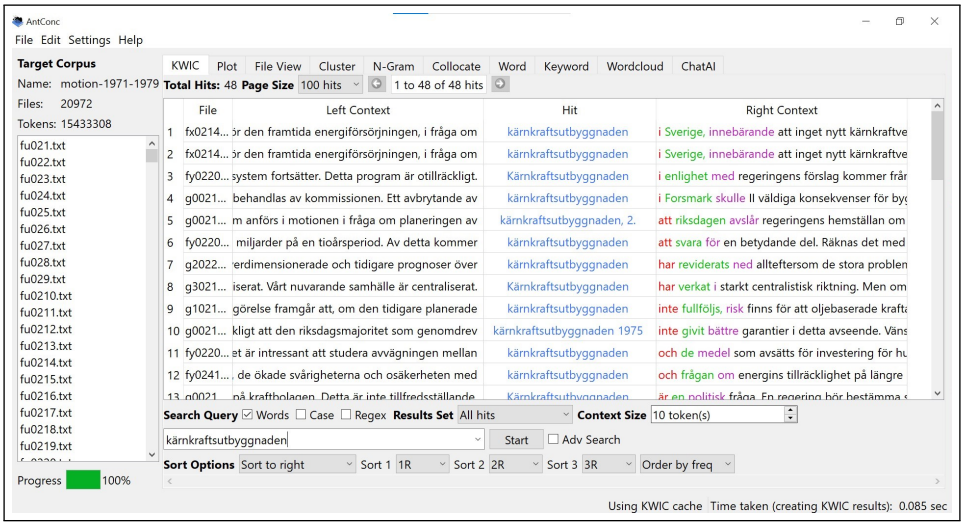


Figure 2: Hits for kärnkraftutbygganden in KWIC tool.

It is also possible to access the content of any of the motions directly by clicking on the name on the left side of the interface. If we click on the first hit for "kärnkraftuppbyggnaden" we can see the entire text of the motion, with each hit highlighted in blue (Figure 3).

So far we have used the texts directly as they were retrieved from the data portal. However, it is also an option to preprocess the text to increase the usefulness of some of the tools in AntConc. For example, when counting word frequencies, it can be helpful if the definite and indefinite forms of a word or various conjugations of a verb are counted as the same. Preprocessing steps are done to make the texts more understandable for machines, keep words that carry the most information, and in some cases also to reduce the size of the dataset. It often takes multiple iterations to get the preprocessing 'right' and how it exactly should be done is determined by the research questions that are to be explored. During the preprocessing a lot of decisions are made about what words to keep and what words to consider as stop words or noise. The choices made during this stage can greatly impact the final output, making preprocessing an integral part of the research process (Schofield 2022). Tahmasebi & Hengchen (2019) provide a detailed overview of the various steps a preprocessing pipeline may contain for this kind of study.

The corpus of motions was preprocessed in the following way. After collecting the motions as text files, the first preprocessing step was to clean the texts, in this case removing metadata text chunks, not part of the actual motions. We identified six such repetitive text chunks, which contained

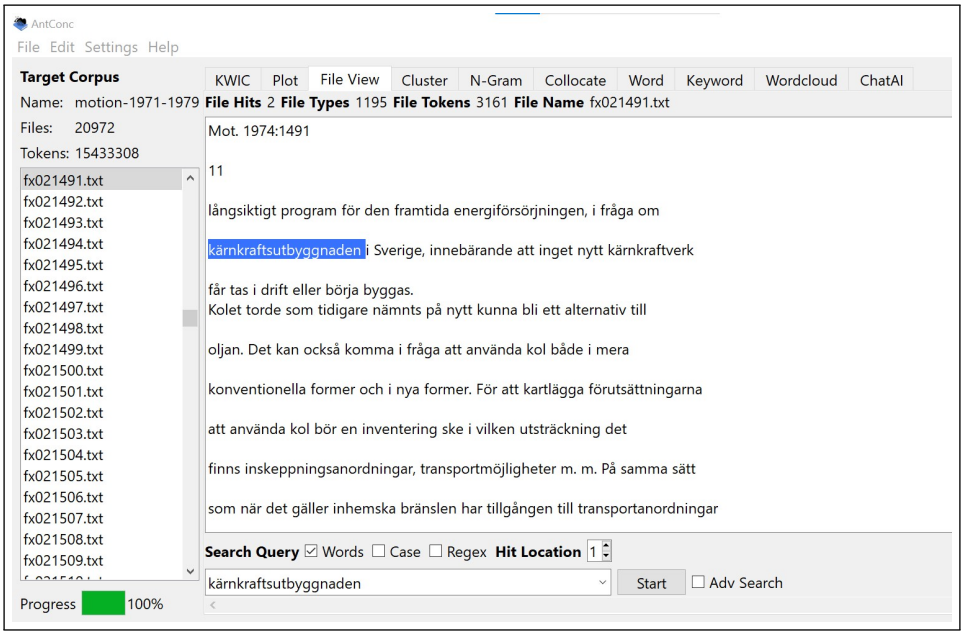


Figure 3: First hit for kärnkraftutbygganden in File tool.

information regarding the quality of the OCR as well as the different printing houses that had been responsible for printing the motions. The second step consisted of tokenizing and lemmatizing the texts. Tokenization refers to the process of splitting up the texts into tokens, smaller linguistic units, while lemmatization is the process of transforming each word into its base or dictionary form (Tahmasebi & Hengchen 2019). For this, we used the previously mentioned Swedish tokenizer/lemmatizer Efselab (Östling 2018). As third and final preprocessing step, stopwords were removed from the texts. We used the Swedish stopword list included in NLTK, a Python programming library for NLP, and expanded it with a list of manually compiled stop words specific to the corpus in question. Whether to remove stop words or not depends on the research question. When focusing on content, words such as "the", "a", "we", "your", are often regarded as uninteresting, but for other explorations there can be reasons to keep them. The code used to preprocess the corpus is available in the GitHub repository.[26]

For newcomers to NLP and programming, using libraries such as Efselab or spaCy to preprocess the texts might not be a possibility in the beginning. There are tools available that can perform lemmatization without the need

26 https://github.com/CDHUppsala/Applied-NLP-for-humanities-research

for programming, for example TagAnt[27] and the previously mentioned Mink. These tools have, however, not been explored and evaluated for this case study. AntConc also includes a functionality to filter the words displayed in some of its tools. This can be done under Settings, Global Settings, Tool Filters. The tool filters can be used in two ways. First, it can be used as a curated keyword list to only display words that you are interested in.[28] Secondly, it is possible to use the tool filter functionality as a stop word list. Instead of removing the words from the actual motions, the words are filtered out from the results displayed in the tools. The list can be iteratively extended.

For this case study we have worked with two corpora. One corpus consists of the 'raw' text files without any preprocessing, adding the files as they were retrieved from the open data portal. This corpus is useful for close reading the texts in the KWIC and file tool. The second corpus contained all words in lemmatized form and had stop words removed, which can be used for looking for co-occurrences and finding keywords that characterize a subsection of the corpus as will be shown below.

The **Collocate tool** gives additional options to study a keyword and its frequently surrounding words. The tool shows how a keyword co-occurs with other words in the corpus within a certain distance. The window span can be adjusted on either side of the word. Random distribution is used as a base line to see if words appear more often or more seldom together. This method is sensitive to statistical outliers. It may therefore be a good idea to set a minimum threshold for terms to be included in the calculations (freq) and in how many of the documents they should appear (range). From the collocate tool, it is possible to navigate to the KWIC tool by clicking on one co-occurring word. Depending on the size of the corpus, it might take a while before the results are generated.

In Figure 4 we can see what words tend to co-occur with the lemma nuclear power ("kärnkraft") within a window span of five tokens to the left and the right side of the search term. From the results, we get a sense of the different contexts in wich nuclear power is discussed, from investing (uppbyggnad, satsning) to decommissioning (avveckla) nuclear power, and the possibility of having a referendum on the topic (folkomröstning). Dividing the corpus into multiple time periods and rerunning the collocate analysis might reveal that in the beginning of the period more motions dicussed the building up of nuclear power while at the end of the period the decommissioning of nuclear power was discussed more. From the collocates

27 https://laurenceanthony.net/software/tagant/

28 The rank of the keywords does not change so it might be necessary to adjust the page size or use the arrows to navigate to the next results to see all the words included in the list.

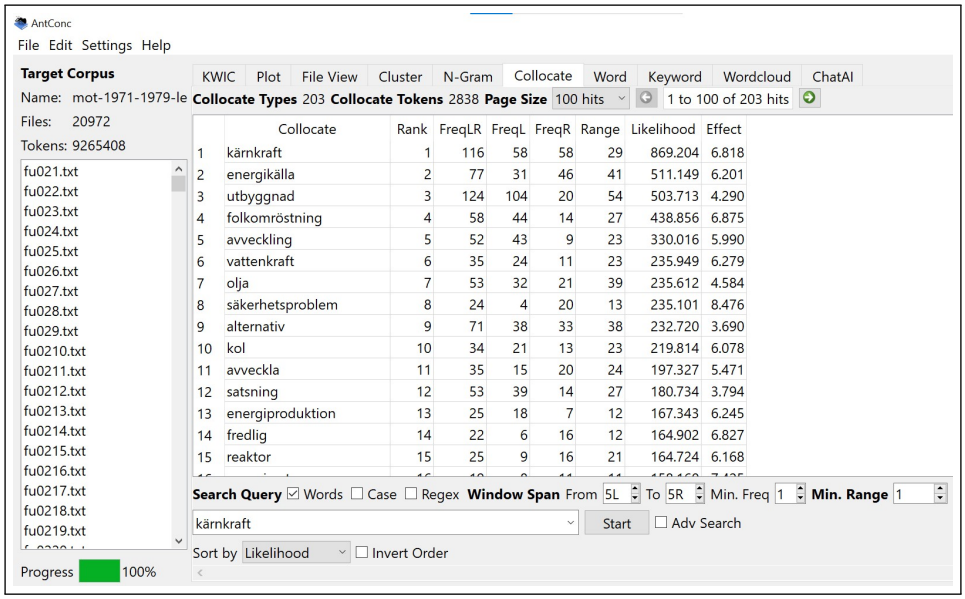
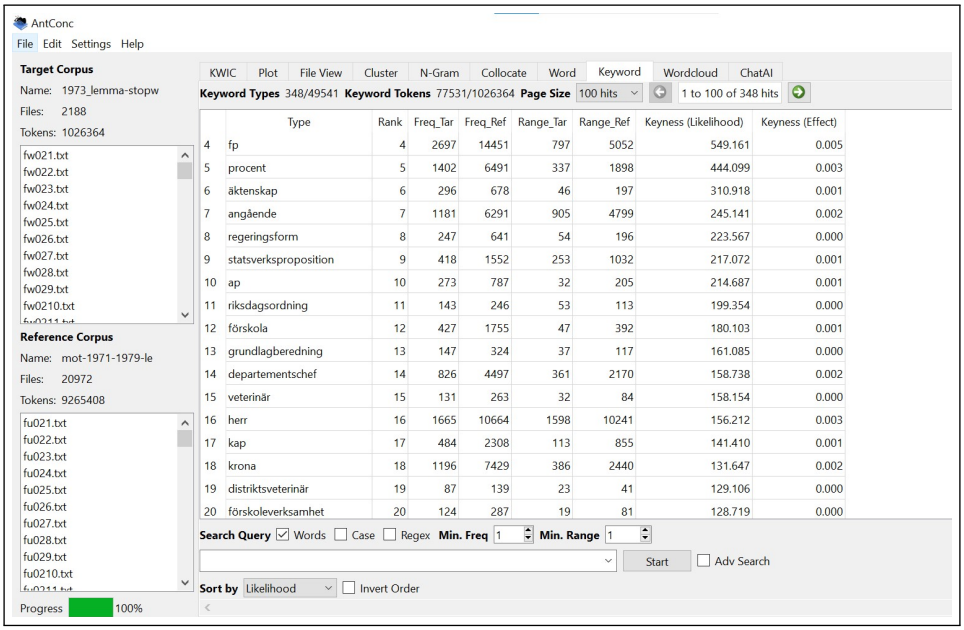


Figure 4: Collocate hits for kärnkraft.

alone it is not possible to deduce the exact context in which the words were used. Therefore, it is important to inspect the actual sentences and texts in which they appear, which we can reach through the KWIC and File tool, which have been introduced above.

Finally, studying the content of the corpus in a more exploratory fashion is also possible with AntConc. As previously mentioned, the Word list tool can be used to get an overview of the most frequent words in the corpus. This can be done by pressing start without entering any search term. Additionally, the **Keyword list tool** can be used to compare a subsection of the corpus (target corpus) against the entire corpus (reference corpus), as was done in the study on the patient organization periodical Allergia discussed previously (Söderfeldt et al. 2019). To use this tool, two corpora need to be loaded into AntConc: the subset of the corpus you would like to analyze and the entire original corpus. Once both corpora are loaded, open the subset as target corpus and the entire original corpus as reference corpus in the corpus manager. Then navigate to the Keyword tool and press Start. Based on a likelihood measure, this tool can show which words appear unusually frequently or infrequently in the subset in comparison to the entire corpus. It shows what words are characteristic for the subset. To inspect the negative keywords, words that appear unusually infrequently, navigate to Settings, Tool Settings, Keyword and select Show negative keywords.



	Type	Rank	Freq_Tar	Freq_Ref	Range_Tar	Range_Ref	Keyness (Likelihood)	Keyness (Effect)
4	fp	4	2697	14451	797	5052	549.161	0.005
5	procent	5	1402	6491	337	1898	444.099	0.003
6	äktenskap	6	296	678	46	197	310.918	0.001
7	angående	7	1181	6291	905	4799	245.141	0.002
8	regeringsform	8	247	641	54	196	223.567	0.000
9	statsverksproposition	9	418	1552	253	1032	217.072	0.001
10	ap	10	273	787	32	205	214.687	0.001
11	riksdagsordning	11	143	246	53	113	199.354	0.000
12	förskola	12	427	1755	47	392	180.103	0.001
13	grundlagberedning	13	147	324	37	117	161.085	0.000
14	departementschef	14	826	4497	361	2170	158.738	0.002
15	veterinär	15	131	263	32	84	158.154	0.000
16	herr	16	1665	10664	1598	10241	156.212	0.003
17	kap	17	484	2308	113	855	141.410	0.001
18	krona	18	1196	7429	386	2440	131.647	0.002
19	distriktsveterinär	19	87	139	23	41	129.106	0.000
20	förskoleverksamhet	20	124	287	19	81	128.719	0.000

Figure 5: Keywords for 1973 compared to the entire corpus.

As an example, we can find the over-represented words for the year 1973 compared to the entire corpus 1971-1979. The output in Figure 5 displays the frequency and range of each word in the target corpus and the reference corpus. The output suggests that an additional preprocessing of the texts may be necessary, for example party abbreviations such as "fp" for Folkpartiet (the predecessor of the Liberal party) or honorifics (herr) could be included in the list of stop words if these are not of interest for the research question. Other words may reveal characteristics of motions submitted during the year 1973, such as words about preschools (förskola, förskoleverksamhet) or the veterinarian (veterinär) that could warrant further exploration.

Ready-to-use tools like AntConc are valuable resources for historians to start experimenting with NLP and corpus linguistics methods without the need to know or learn programming. As shown in this case study, AntConc supports the study of interesting words that are known beforehand but also exploratory approaches for finding interesting words based on the content of the corpus. Such tools do have its limitations compared to programming libraries, as the user is limited to the included functionality and, in the case of AntConc, needs to manually load different versions of the corpus separately in order to analyze them.

5.2 *Case study 2: Literary style*

Our second case study concerns literature, and more specifically literary style in Swedish 20th century prose fiction. "Literary style" is a complex notion within literary studies that can refer to a blend of things on different levels, including syntactic, grammatical, and thematic patterns of various kinds. Yet, certain more formalized aspects of literary style are well-suited to trace by computational means. In fact, the possibilities within computational literary studies has led to a new-found scholarly interest in literary style more broadly (Allison et al. 2013, Herrmann et al. 2015). It is this vein of "instrumentalized" or "operationalized" literary stylistics that the following case study belongs to.

Basically, we are interested in research questions concerning the relationship between literary style – broadly understood as described above – and author position in the literary field. A foregrounding hypothesis is that style relates to levels of public impact and literary prestige. To put it in a very straightforward way: the bigger the audience, the less experimental the writing style. But we also have a second and somewhat contradictory hypothesis that although this might be true in most cases, there are likely both exceptions and other parts of what we here label as style – how literature is written rather than its themes or narrative structures – at work.

To investigate this, we constructed a corpus of 50 novels written by 21 Swedish authors and published between 1910 and 1967. This selection emanates from a much larger study (Berglund & Svedjedal 2025), where principles for corpus selection are discussed thoroughly. For this toy example case study, it is sufficient to say that it includes a broad range of prose fiction that spans from niche and experimental modernist writing with high prestige in the literary field to popular genre fiction with little prestige. Also of importance is that all works in the corpus are written with modern Swedish spelling.

Generally, however, the composition of the corpus is key, and forms the basis for which research questions can be asked. How representative is a certain material in relation to an era or a genre? Are you carrying out an analysis that covers everything you want to investigate (for instance: all novels by Selma Lagerlöf), or have you made a selection? How reliable is your material? For a computational literary analysis to be credible, it is important to be able to answer such questions. The goal is not to create the perfect corpus but to explain and describe each corpus along with its strengths and limitations.

As of materials, we depart from freely available digitized Swedish fiction in Litteraturbanken.[29] From the site we downloaded novels as plain text

files (.txt), and then manually deleted everything apart from the literary text (that is, colophon information about the book such as title, author, and publishing house, prefaces and afterwords that do not belong to the novel text, footnotes and other kinds of editorial comments, etc.). Since the text quality in Litteraturbanken is controlled and excellent, no further data curation of the text material was needed in this case. In total, the curated corpus contains 3.97 million tokens.

In parallel, we made a metadata file with basic information about each of the 50 novels. This includes author name, title of novel, and year of publication. To be able to track how literary prestige relates to style, we also made a literary sociological distinction between three groups of authors: 1) canonized authors with high literary prestige but small readerships in their time (a total of 20 novels in the corpus); 2) canonized authors with high literary prestige *and* large readerships in their time (26 novels); and 3) popular genre fiction authors with low literary prestige but large readerships in their time (4 novels). Literary prestige and levels of canonization were based mainly on major literary prizes and descriptions in literary historiography. Sizes of readerships were based mainly on the number of editions and copies sold (for details, see Berglund & Svedjedal 2025). The different sizes of the three groups are an effect of the focus of Litteraturbanken on canonical Swedish literature. To be able to track this more seriously, other sources would be needed as a complement.

These 50 novels can be imported into AntConc and analyzed in various ways. However, the kinds of stylistic features that we are interested in here are easier to detect by means of programming. This holds true, not least when it comes to visualizations of findings, where AntConc falls short. For all visualizations below, we use standard python libraries *seaborn* and *matplotlib*. You need to learn their specific syntax to be able to plot pretty graphs, but when you do, the possibilities for variation are endless. The Notebook scripts that we have used in this example are available in the GitHub repository.[30]

As a starting point, we ran the corpus through Efselab/Swepipe (Östling 2018). Included in this pipeline and of interest for our analyses are word segmentation, paragraph and sentence segmentation, part-of-speech tagging, and lemmatization. As an output you get a tab-separated conll-file, which includes all of the information above. For example, the first sentence in Selma Lagerlöf's *Körkarlen* (1912) reads: "Det var en stackars liten slumsyster, som höll på att dö." In the output from Efselab/Swepipe, this is transformed as shown in Table 1 (structured as a table for readability).

29 https://litteraturbanken.se/

30 https://github.com/CDHUppsala/Applied-NLP-for-humanities-research

Table 1: Example output from Efselab. (U-POS is universal part of speech.)

Token	Part of speech	U-POS	Lemma
Det	PN|NEU|SIN|DEF|SUB/OBJ	PRON	det
var	VB|PRT|AKT	AUX	vara
en	DT|UTR|SIN|IND	DET	en
stackars	JJ|POS|UTR/NEU|SIN/ PLU|IND/DEF|NOM	ADJ	stackars
liten	JJ|POS|UTR|SIN|IND|NOM	ADJ	liten
slumsyster	NN|UTR|SIN|IND|NOM	NOUN	slumsyster
,	MID	PUNCT	,
som	HP| – | – | –	PRON	som
höll	VB|PRT|AKT	VERB	hålla
på	PL	ADP	på
att	IE	PART	att
dö	VB|INF|AKT	VERB	dö
.	MAD	PUNCT	.
PARAGRAPH _BREAK			PARAGRAPH _BREAK

Although probably confusing at first sight, at least for a humanities scholar, everything that we need for further processing and analysis can be found here. The original text is in the first column, two levels of part-of-speech tagging in the two following columns, and the lemmatized token in the fourth column. End of sentence is indicated by a major delimiter ("MAD" in the second column) followed by a blank row.

So, with all foundational NLP markup at hand: what to search for in terms of style? Let's start with the most low-hanging fruit: punctuation. How authors make use of things like commas, semicolons, and question marks can actually tell us quite a bit about the kind of prose at hand, not least if authors are compared to each other. Since the length of novels vary a lot, absolute numbers are not interesting here. What we need are relative frequencies. To get this is fairly straightforward: we run a script that counts all tokens of a certain kind of punctuation, say semicolon (";"), and then we divide the number of hits with the total number of tokens for the novel in question. If we make a bar plot out of these relative frequencies for semicolon, and highlight the three sociological categories of author through a legend, we get the following graph (Figure 6).

What Figure 6 shows is that semicolons are very unevenly distributed in the corpus. While some authors use it recurrently, around half of the novels

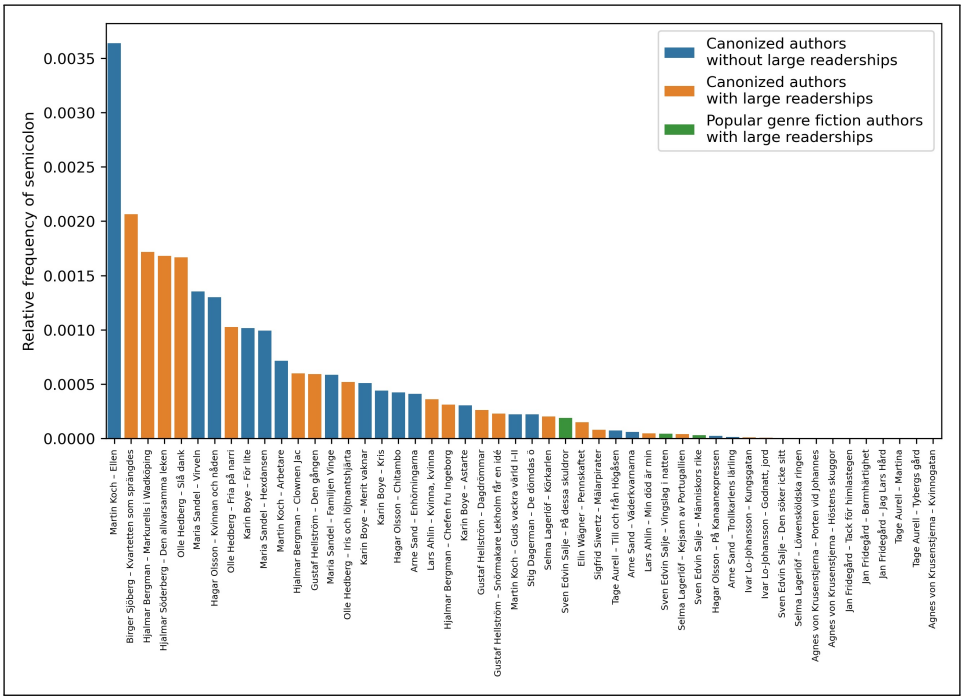


Figure 6: Bar plot of semicolon use.

in the corpus never or very rarely contain any semicolons at all. This typical stylistic feature, then, seems to be tied to novels by certain authors: Martin Kock, Birger Sjöberg, Hjalmar Bergman. Although the popular genre fiction novels in the corpus all have low levels of semicolon, it is not possible to make a link between literary prestige and semicolon use. For instance, Nobel prize laureate Selma Lagerlöf seldom uses semicolons. The same goes for the 1940s modernist Tage Aurell.

Similar calculations can be done with all kinds of punctuation, and they will indicate different aspects of style: high levels of commas indicate lots of subordinate clauses and enumerations, frequent question marks indicate dialogue, lots of exclamation marks dialogue and/or an expressive tone, etc.

But let us move on to two other metrics that are indicative of literary style: length of sentences and length of paragraphs. We use the occurrence of one or several blank lines as an indication of a paragraph break. For sentence lengths, we make use of the major delimiter markup in the Efselab/Swepipe output. For sentences, the most straightforward way is to write a script that divides the number of tokens in each novel with its number of tokens tagged as major delimiters ("MAD") in the second column. For paragraph

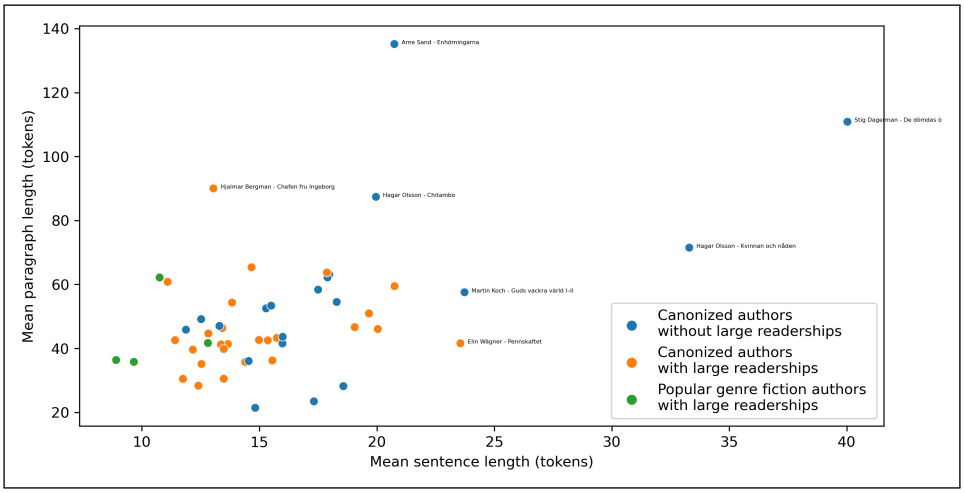


Figure 7: Scatter plot of mean sentence and paragraph length.

breaks, we have added the tag "PARAGRAPH_BREAK" to the output from Efselab/Swepipe to indicate a blank row in the original text. We can, thereby, similarly divide the number of tokens in each novel with its number of paragraph breaks.[31] If we make a scatter plot of how the 50 novels in our corpus are distributed according to these two metrics, we get the result displayed in Figure 7.

What immediately stands out in Figure 7 is that a majority of the novels in the corpus cluster in the lower left corner. Most novels investigated have a mean sentence length of between 10 and 20 tokens, and a mean paragraph length between 30 and 60 tokens. But what we also see in the graph is a smaller group of outlier novels, where both sentences and paragraphs are much longer. These outliers are experimenting with language in different ways: the modernist prose of Stig Dagerman and Hagar Olsson stands out in particular with their long and complex sentences, while the long paragraphs in Arne Sand's *Enhörningarna* (1965) can be explained by the novel's self-reflexive and diary-like internal monologue and its more or less total lack of dialogue.

The most extreme outliers are all written by the group of canonized authors without large readerships, which is at least an indication that these kinds of stylistic experiments is not something that the common reader tends to appreciate. Too bold claims should not be made out of this small selection, but this tendency shows in Figure 7. Novels by popular genre fiction authors

31 You could also use spaCy for this – its sentence tokenizer is even more easy to use than Efselab/Swepipe, see Berglund & Svedjedal (2025: 180–189).

Table 2: Mean sentence and paragraph length per author category

Author group	Novels in corpus	Sentence length (tokens)	Paragraph length (tokens)
Canonized authors without large readerships	20	18.8	57.1
Canonized authors with large readerships	26	15.0	46.1
Popular genre fiction authors with large readerships	4	10.5	44.0

with large readerships (the green dots) are clustered towards the lower left corner. Novels written by canonized authors with large readerships (the orange dots) do so as well, but slightly more in the upper right direction. Novels by canonized authors without large readership (blue dots) are the most scattered, and none of them have very short sentences on average. If calculated on a group level – by taking mean scores of the mean values for all novels respectively – the patterns stand clear, as shown in Table 2.

Also part of speech distribution has something to say about literary style. High amounts of verbs indicate plot-driven narratives with plenty of dialogue, whereas high amounts of adjectives indicate narratives where depictions and settings are important. In general, popular genre fiction is to a high extent plot-driven where canonical fiction are more diverse (Berglund et al. 2019, Berglund & Dahllöf 2021).

To get figures on parts of speech, we depart from the third column of the Efselab/Swepipe output. This information builds on the part of speech tags in Universal Dependencies (U-POS), with 17 classes in total.[32] For our analysis, we are interested in adjectives (tagged as "ADJ") and verbs ("VERB"). To calculate the share of these parts of speeches, we proceed in a similar vein as previously: we write a script that checks every token in every novel, and count all tokens that are tagged as adjectives and verbs respectively. We then divide the number of adjectives/verbs with the total number of tokens for every novel. From this, we get two new metrics: verb share and adjective share.

32 See https://universaldependencies.org/u/pos/ Efselab/Swepipe also provides a more thorough part of speech markup in the second column of its output.

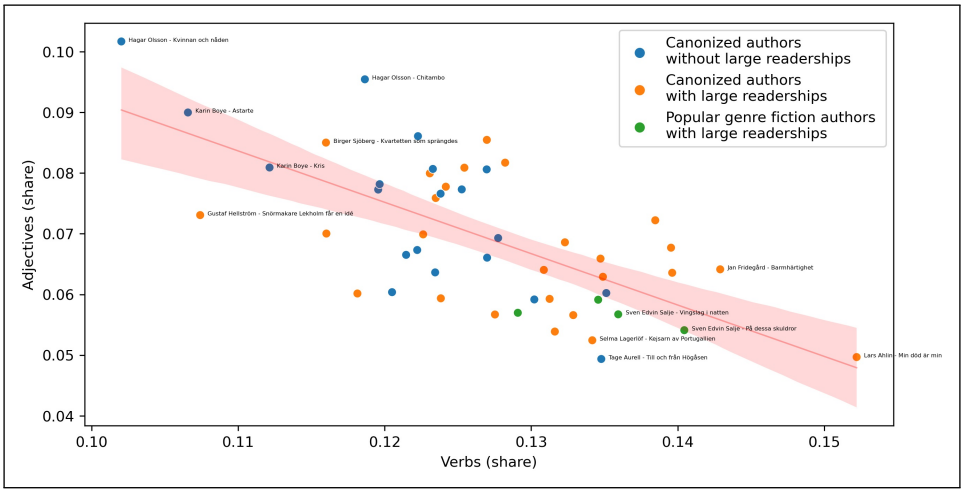


Figure 8: Scatter plot of verb and adjective share.

If we make a scatter plot of verb and adjective share for the 50 novels in our corpus (Figure 8) we see that there is a negative correlation between these two metrics: in general, the more adjectives in a novel, the fewer verbs, and vice versa. The strength of this correlation can be calculated by means of a linear regression analysis – this is what the red regression line in Figure 8 indicates. If we use *lingress* from the standard python library *scipy.stats*, we get that the correlation between share of verbs and share of adjectives for the novels in our corpus is $R = -0.668$, which in statistical terms suggest a fairly strong negative correlation.

This outcome makes sense and is no surprise. What is more interesting, though, is to take a closer look at the outliers. In the upper left corner (lots of adjectives, few verbs) of Figure 8, we find modernist writer Hagar Olsson and Karin Boye's more experimental novels. In the lower right corner (lots of verbs, few adjectives), we find the novels of the only popular genre fiction author in the corpus – Sven Edvin Salje – but also novels by canonized authors such as Selma Lagerlöf and Lars Ahlin. If we calculate the mean shares on a group level, there are apparent differences regarding part of speech use: canonized authors without large readerships are the most adjective heavy, while the popular genre fiction authors use the most verbs (see Table 3). What is important to also keep in mind is that canonized authors without large readerships are the most diverse in their literary style.

We could go on and calculate a number of other metrics concerning literary style (see further Berglund & Svedjedal (2025)), but for this toy example case study we hope to have proven our point: that basic NLP markup

Table 3: Verb and adjective share per author category

Author group	Novels in corpus	Verbs (mean perc.)	Adjectives (mean perc.)
Canonized authors without large readerships	20	12.2	7.4
Canonized authors with large readerships	26	12.9	6.8
Popular genre fiction authors with large readerships	4	13.5	5.7

(in this case derived from Efselab/Swepipe) along with easily graspable visualizations (here from matplotlib/seaborn) can be used for all kinds of insights into literary history and literary style.

6 *Conclusion*

This chapter has aimed to give humanities scholars who are curious about the possibilities of NLP a first introduction to the field. By discussing some of the challenges involved in getting started with NLP, as well as presenting practical examples of how NLP methods can be utilized within the humanities, we hope to have offered the reader an opportunity to begin experimenting with NLP in their own work.

The field of NLP offers a wide variety of computational methods for large-scale text analysis. The two case studies presented in this chapter demonstrate that these methods can be integrated in humanities research by any scholar regardless of their technical background, either through ready-to-use text analysis tools or programming libraries. More advanced NLP applications based on machine learning may fall outside the scope of many humanities scholars and require interdisciplinary collaboration with NLP experts and data scientists. However, having an understanding of the key NLP concepts, available methods, and how they differ from traditional humanities approaches – which this chapter has aimed to provide – can help facilitate such collaborations in future projects.

Notebooks

Code and notebooks connected to the case studies can be found on: https://github.com/CDHUppsala/Applied-NLP-for-humanities-research.

Acknowledgments

Part of the work here was conducted within the ActDisease project. ActDisease is funded by the European Union (ERC ActDisease, ERC-2021-STG 101040999). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Allison, Sarah, Marissa Gemma, Ryan Heuser, Franco Moretti, Amir Tevel & Irena Yamboliev. 2013. Style at the scale of the sentence. *Pamphlets of the Stanford Literary Lab* 5. https://litlab.stanford.edu/LiteraryLabPamphlet5.pdf.

Anthony, Laurence. 2024. *AntConc* (*Version 4.3.1*) [*Computer Software*]. Tokyo, Japan: Waseda University. https://laurenceanthony.net/software/antconc.

Arnold, Taylor & Lauren Tilton. 2019. New data? The role of statistics in DH. In Matthew Gold & Lauren Klein (eds.), *Debates in the digital humanities 2019*. Minneapolis: University of Minnesota Press. DOI: 10.5749/9781452963785.

Artstein, Ron & Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596. DOI: 10.1162/coli.07-034-R2.

Badri, Sushruth. 2024. *Digitising deeds*. https://github.com/CDHUppsala/digitising-deeds.

Bamman, David. 2017. Natural language processing for the long tail. In *Digital Humanities 2017 conference abstracts*, 382–384. Montreal, Canada. https://dh2017.adho.org/abstracts/408/408.pdf.

Baumer, Eric P. S., David Mimno, Shion Guha, Emily Quan & Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68(6). 1397–1410. DOI: 10.1002/asi.23786.

Berglund, Karl & Mats Dahllöf. 2021. Audiobook stylistics: Comparing print and audio in the bestselling segment. *Journal of Cultural Analytics* 6(3). 1–30. DOI: 10.22148/001c.29802.

Berglund, Karl, Mats Dahllöf & Jerry Määttä. 2019. Apples and oranges? Large-scale thematic comparisons of contemporary Swedish popular and literary fiction. *Samlaren: Tidskrift för forskning om svensk och annan nordisk litteratur* 140. 228–260.

Berglund, Karl & Johan Svedjedal. 2025. *Det stora lästa. Fjärrläsningar av klassiska svenska romaners stilvärldar, 1910–1999*. Uppsala: Uppsala University. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-557694.

Blei, David M. 2012. Topic modeling and digital humanities. *Journal of Digital Humanities* 2(1). 8–11.

Bode, Katherine. 2023. What's the matter with computational literary studies? *Critical Inquiry* 49(4). 507–529. DOI: 10.1086/724943.

Boyd-Graber, Jordan, Yuening Hu & David Mimno. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval* 11(2-3). 143–296. DOI: 10.1561/1500000030.

Boyden, Michael, Ali Basirat & Karl Berglund. 2022. Digital conceptual history and the emergence of a globalized climate imaginary. *Contributions to the History of Concepts* 17(2). 95–122. DOI: 10.3167/choc.2022.170205.

Brandsen, Alex, Suzan Verberne, Karsten Lambers & Milco Wansleeben. 2022. Can BERT dig it? Named entity recognition for information retrieval in the archaeology domain. *Journal on Computing and Cultural Heritage (JOCCH)* 15(3). 1–18. DOI: 10.1145/3497842.

Danilova, Vera & Ylva Söderfeldt. 2025. Classifying textual genre in historical magazines (1875-1990). In Anna Kazantseva, Stan Szpakowicz, Stefania Degaetano-Ortlieb, Yuri Bizzoni & Janis Pagel (eds.), *Proceedings of the 9th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (latech-clfl 2025)*, 160–171. Albuquerque, New Mexico. DOI: 10.18653/v1/2025.latechclfl-1.15.

de Bolla, Peter, Ewan Jones, Paul Nulty, Gabriel Recchia & John Regan. 2019. Distributional concept analysis: A computational model for history of concepts. *Contributions to the History of Concepts* 14(1). 66–92. DOI: 10.3167/choc.2019.140104.

Dombrowski, Quinn & Patrick J. Burns. 2023. Language is not a default setting: Countering DH's English problem. In Matthew K. Gold & Lauren F. Klein (eds.), *Debates in the digital humanities 2023*. Minneapolis: University of Minnesota Press. DOI: 10.5749/9781452969565.

Drucker, Johanna. 2017. Why distant reading isn't. *PMLA/Publications of the Modern Language Association of America* 132(3). 628–635. DOI: 10.1632/pmla.2017.132.3.628.

Du, Keli. 2019. A survey on LDA topic modeling in digital humanities. In *Book of abstracts DH2019*. DOI: 10.34894/H9UYPI.

Dubremetz, Marie. 2023. *Philosthetic project*. https://gitlab.com/mardub/philosthetic_pilot.

Ekeman, Karl. 2023. *In want of a sovereign: Metapolitics and the populist formation of the alt-right*. Uppsala University. (Doctoral dissertation).

Fridlund, Mats, Mila Oiva & Petri Paju. 2020. *Digital histories: Emergent approaches within the new digital history*. Helsinki: Helsinki University Press. DOI: 10.2307/j.ctv1c9hpt8.

Grootendorst, Maarten. 2022. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. DOI: 10.48550/arXiv.2203.05794.

Guldi, Jo. 2023. *The dangerous art of text mining: A methodology for digital history*. Cambridge: Cambridge University Press. DOI: 10.1017/9781009263016.

Hengchen, Simon & Nina Tahmasebi. 2021. A collection of Swedish diachronic word embedding models trained on historical newspaper data. *Journal of Open Humanities Data* 7(2). DOI: 10.5334/johd.22.

Herrmann, J. Berenike, Karina van Dalen-Oskam & Christof Schöch. 2015. Revisiting style, a key concept in literary studies. *Journal of Literary Theory* 9(1). 25–52. DOI: 10.1515/jlt-2015-0003.

Hu, Yan, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts & Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. eng. *Journal of the American Medical Informatics Association* 31(9). 1812–1820. DOI: 10.1093/jamia/ocad259.

Jänicke, Stefan, Greta Franzini, Gerik Scheuermann & Muhammad Cheema. 2015. On close and distant reading in digital humanities: A survey and future challenges. A state-of-the-art (STAR) report. In *Eurographics conference on visualization (EuroVis)*, 83–103.

Jarlbrink, Johan & Fredrik Norén. 2023. The rise and fall of 'propaganda' as a positive concept: A digital reading of Swedish parliamentary records, 1867–2019. *Scandinavian Journal of History* 48(3). 379–399. DOI: 10.1080/03468755.2022.2134202.

Jockers, Matthew L. & Ted Underwood. 2015. Text-mining the humanities. In Susuan Schreibman, Ray Siemens & John Unsworth (eds.), *A new companion to digital humanities*, 291–306. Wiley Online Library. DOI: 10.1002/9781118680605.ch20.

Jurafsky, Daniel & James H. Martin. 2008. *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. 2nd edn. Saddle River, NJ, USA: Prentice Hall.

La Mela, Matti & Ekta Vats. 2023. Automatic classification of historical texts using a BERT model: News about wild berries, 1860-1910. In *Book of abstracts, DH Benelux*. DOI: 10.5281/zenodo.7990442.

Liu, Bing. 2011. *Web data mining: Exploring hyperlinks, contents, and usage data*. 2nd edn. Heidelberg: Springer. DOI: 10.1007/978-3-642-19460-3.

Maen, Adam. 2024. *Excavating text*. https://github.com/CDHUppsala/excavating-text.

Marsland, Stephen. 2009. *Machine learning: An algorithmic perspective*. Boca Raton, FL: Chapman & Hall/CRC.

McDonough, Katherine, Ludovic Moncla & Matje Van de Camp. 2019. Named entity recognition goes to old regime France: Geographic text analysis for early modern French corpora. *International Journal of Geographical Information Science* 33(12). 2498–2522. DOI: 10.1080/13658816.2019.1620235.

McGillivray, Barbara. 2022. How to use word embeddings for natural language processing. In *Sage research methods: Doing research online*. London: SAGE Publications Ltd. DOI: 10.4135/9781529609578.

Mcgillivray, Barbara, Thierry Poibeau & Pablo Ruiz. 2020. Digital humanities and natural language processing: "Je t'aime... Moi non plus". *Digital Humanities Quarterly* 14(2).

Moretti, Franco. 2000. Conjectures on world literature. *New left review* 2(1). 54–68. DOI: 10.64590/hxj.

Moretti, Franco. 2013. *Distant reading*, vol. 93. New York: Verso.

Ohlsson, Claes, Victor Wåhlstrand Skärström & Henrik Björck. 2022. The market as a concept in Swedish parliamentary records from 1867 to 1970: A mixed methods study. *Digital Humanities in the Nordic and Baltic Countries Publications* 4(2). 22–34. DOI: 10.5617/dhnbpub.11258.

Östling, Robert. 2018. Part of speech tagging: Shallow or deep learning? *Northern European Journal of Language Technology* 5. 1–15. DOI: 10.3384/nejlt.2000-1533.1851.

Piqueras, Matias. 2023. *Alt-right formations*. https://github.com/CDHUppsala/alt-right-formations.

Plank, Barbara. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of konvens 2016*, 13–20. DOI: https://doi.org/10.48550/arXiv.1608.07836.

Rapp, Reinhard. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *COLING 2002: The 19th international conference on computational linguistics*. https://aclanthology.org/C02-1007/.

Rettberg, Jill Walker. 2022. Algorithmic failure as a humanities methodology: Machine learning's mispredictions identify rich cases for qualitative analysis. *Big Data & Society* 9(2). DOI: 10.1177/20539517221131290.

Russell, Stuart J., Peter Norvig & John F. Canny. 1995. *Artificial intelligence: A modern approach*. Englewood Cliffs: Prentice Hall.

Schofield, Alexandra. 2022. The possibilities and limitations of natural language processing for the humanities. In James O'Sullivan (ed.), *The Bloomsbury handbook to the Digital Humanities*, 169–178. London: Bloomsbury.

Skeppstedt, Maria, Gijs Aangenendt, Vera Danilova & Ylva Söderfeldt. 2024. Topics in periodicals from the Swedish diabetes association 1949 – 1990: Extending the topic modelling tool Topics2Themes with a timeline visualisation. In *Selected papers from the CLARIN annual conference 2023*. DOI: 10.3384/ecp210015.

Skeppstedt, Maria, Magnus Ahltorp, Kostiantyn Kucher, Gijs Aangenendt, Matts Lindström & Ylva Söderfeldt. 2025. The Word Rain visualisation technique applied to digital history: How to visualise, explore and compare texts using semantically structured word clouds. In Elena Volodina, Gerlof Bouma, Dana Dannélls & Dimitrios Kokkinakis (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 147–182. University of Tartu Library. DOI: 10.58009/aere-perennius0175.

So, Richard Jean. 2017. All models are wrong. *PMLA/Publications of the Modern Language Association of America* 132(3). 668–673. DOI: 10.1632/pmla.2017.132.3.668.

Söderfeldt, Ylva, Karl Berglund & Matts Lindström. 2019. Towards mining the history of the active patient: A mixed-methods discourse analysis of the journal Allergia, 1957-1990. In *Uppsala papers in history of ideas* (19). Uppsala: Uppsala University. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-394915.

Söderfeldt, Ylva, Andrew Burchell, Julia Reed & Maria Skeppstedt. 2025. Topic timelines for enabling close and distant reading of discursive shifts: A pilot case using periodicals of European diabetes organizations. *Journal of Open Humanities Data* 11(1). DOI: 10.5334/johd.286.

Tahmasebi, Nina & Simon Hengchen. 2019. The strengths and pitfalls of large-scale text mining for literary studies. *Samlaren: Tidskrift för forskning om svensk och annan nordisk litteratur* 140. 198–227.

Tamper, Minna, Petri Leskinen & Eero Hyvönen. 2019. Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In *International conference on computational linguistics*

and intelligent text processing, 199–214. DOI: 10.1007/978-3-031-24337-0_15.

Tasovac, Toma, Nick Budak, Natalia Ermolaev, Andrew Janco & David Lassner. 2023. Bridging the gap between digital humanities and natural language processing: A pedagogical imperative for humanistic NLP. In Lorella Viola & Paul Spence (eds.), *Multilingual Digital Humanities*, 114–126. London: Routledge. DOI: 10.4324/9781003393696-10.

Thompson, Paul, Riza Theresa Batista-Navarro, Georgios Kontonatsios, Jacob Carter, Elizabeth Toon, John McNaught, Carsten Timmermann, Michael Worboys & Sophia Ananiadou. 2016. Text mining the history of medicine. *PLOS One* 11(1). DOI: 10.1371/journal.pone.0144717.

Underwood, Ted. 2019. *Distant horizons: Digital evidence and literary change*. Chicago: University of Chicago Press.

van Strien, Daniel, Kaspar Beelen, Mariona Ardanuy, Kasra Hosseini, Barbara McGillivray & Giovanni Colavizza. 2020. Assessing the impact of OCR quality on downstream NLP tasks. In *Proceedings of the 12th international conference on agents and artificial intelligence*, 484–496. Valletta, Malta: SCITEPRESS. DOI: 10.5220/0009169004840496.

Vats, Ekta. 2022. *BerryBERT*. https://github.com/CDHUppsala/BerryBERT.

List of abbreviations

HTR	Handwritten Text Recognition
KWIC	Key Word in Context
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
POS	Part-of-Speech
TF-IDF	Term Frequency - Inverse Document Frequency

Corresponding author

Gijs Aangenendt
Department of History of Science and Ideas
Uppsala University
gijs.aangenendt@idehist.uu.se