How Can Big Data Help Us Study Rhetorical History?

Jon Viklund
Department of Literature
Uppsala University, Sweden
jon.viklund@littvet.uu.se

Lars Borin
Språkbanken/Dept. of Swedish
University of Gothenburg, Sweden
lars.borin@svenska.gu.se

Abstract

Rhetorical history is traditionally studied through rhetorical treatises or selected rhetorical practices, for example the speeches of major orators. Although valuable sources, these do not give us the answers to all our questions. Indeed, focus on a few canonical works or the major historical key figures might even lead us to reproduce cultural self-identifications and false generalizations. However, thanks to increasing availability of relevant digitized texts, we are now at a point where it is possible to see how new research questions can be formulated – and how old research questions can be addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small samples, but where a methodology has not yet established itself. The aim of this paper is twofold: (1) We wish to demonstrate the usefulness of large-scale corpus studies ("text mining") in the field of rhetorical history, and hopefully point to some interesting research problems and how they can be analyzed using "big-data" methods. (2) In doing this, we also aim to make a contribution to method development in e-science for the humanities and social sciences, and in particular in the framework of CLARIN.

1 Introduction

1.1 Background

Like many other scientific disciplines, the humanities and social sciences (HSS) are now entering the age of big data. The modern digital world and the mass digitization of historical documents together provide unprecedented opportunities to all research fields relying on text as primary research data. This includes almost all HSS disciplines. In fact, we are now at a point where it is possible to see how new HSS research questions can be formulated – and how old research questions can be addressed from a new angle or established results verified – on the basis of exhaustive collections of data, rather than small samples, but where a methodology has not yet established itself. As larger amounts of texts are digitized and made searchable, we are able to see and investigate abstract patterns in large text masses that produce, partly, a new kind of knowledge.

In 2010, a team of American researchers published a study based on the enormous amount of texts made public by Google, a corpus of over 5 million digitized books. They named the new field of study *Culturomics* (Michel et al., 2011; Aiden and Michel, 2013), since it purported to uncover cultural and linguistic developments over time by large-scale computational processing of words, the basic building blocks of texts, in a way analogous to the methodology developed in the biomedical-informatics subfields of *genomics* and *proteomics*.

However, this development comes with considerable methodological challenges. The studies published so far are both tantalizing and disappointing. One can hardly deny their potential, but most often they are arguably more in the way of proof-of-concept showcases than seriously intended HSS research efforts, and have in fact been conducted primarily by non-HSS researchers (e.g., physicists and computer scientists). The field has been characterized by a conspicuous lack of "deep" research questions. The results of these studies affirm what we already knew. They argue for the advantages of different kind of

"text mining" in large collections of data, but so far they have not produced significant results that answer important research questions in the different fields of the humanities.

Culturomics and similar high-profile big-data initiatives drawing on vast amounts of digitized text have been criticized – justifiedly so, in our opinion – both for the low level of linguistic sophistication of the basic text processing underlying them (e.g., Zimmer 2013; see also Tahmasebi et al. 2015), and for the questionable suitability, in terms of representativity, of the investigated data sets for the particular research questions asked and claims made in this body of work (e.g., Pechenick et al. 2015). Thus, critics have pointed out that the bulk of this work reveals a quite simplistic "folk-linguistic" notion of language on the part of its authors, quite far removed from how the various branches of modern linguistics construe their object of study. The practitioners of culturomics have also so far shown little awareness of the issue of validity of their big data, one which humanists and social scientists have been contending with for a long time (Franzosi, 1987).

Having said this, however, we hasten to add that the basic premise of culturomics and similar big-data HSS initiatives appears to us both sound and exciting, and capable of bringing a new dimension to HSS research. The way to overcome the methodological obstacles mentioned above may simply be a matter of arranging for – in the words of Zimmer (2013) – "better communication between disciplines that previously had little to do with each other". This is exactly where CLARIN fits into the picture, as a provider of an e-science infrastructure allowing for linguistically sophisticated large-scale text processing, explicitly directed towards addressing HSS research questions, an infrastructure which crucially includes "experts", language-technology researchers who in dialogue with HSS scholars support the latter in their use of the infrastructure..

1.2 Towards HSS e-science – a concrete case study utilizing big textual data

The aim of the work presented here is twofold.

Firstly, in this paper we will try to demonstrate the usefulness of large-scale computational methods in the field of rhetorical history, and hopefully point to some interesting research problems and how they can be analyzed. In rhetorical studies, and more so in the field of rhetorical history, these quantitative, statistical methods have barely been tested. Our main historical interest lies in Swedish 19th century rhetorical culture, and we will demonstrate how large-scale quantitative analysis of a suitable text material can be used to learn more about ideas of and attitudes towards eloquence in this period.

These issues are not easily investigated through traditional research perspectives used by historians of rhetoric. Usually the materials under investigation in such research are either rhetorical treatises, or selected rhetorical practices, for example the speeches of major orators. These are valuable sources, but they represent only a minor part of all relevant historical documents, and, consequently, they give us but a limited view of the historical period at hand. In addition since these texts conform to the traditional material for rhetorical historiography, we also tend to pose traditional research questions, which often render predictable answers. In the worst case focus on a few canonical works or the major historical key figures might even lead us to reproduce cultural self-identifications and false generalizations (Malm, 2014).

Take the idea of "the death of rhetoric", repeated in handbooks for many years. This narrative is easily confirmed following the canonical texts of, e.g., enlightenment philosophers and authors from the romantic period. It has led many rhetorical scholars to conclude that the discipline of rhetoric in some sense really was dead in the 18th and 19th century, which of course is untrue (Fischer, 2013). With a large-scale textual base, we can see beyond this old master narrative, and find historical trends pointing to other narrative paths in history.

Secondly, we aim to make a methodological contribution to HSS e-science, and concretely to the research infrastructure for HSS being developed within CLARIN ERIC.

Following Tangherlini (2013, 8), the four logical stages of HSS e-research can be characterized as: (1) collection and archiving; (2) indexing and classification; (3) visualization and navigation; and (4) analysis. In our ongoing work, we rely on existing digitized text collections for the first stage, and stage 4 is the realm of the HSS scholar, of course. Stage 2 is largely the domain of text-mining and

computational-linguistic analysis and annotation – a central concern for CLARIN, and fortunately also a very lively research area in its own right.

In the present phase of our work, the focus is on the third stage which is often underdeveloped in digital humanities projects (Warwick et al., 2008, 99f), despite the fact that humanities researchers "need to be able to orient themselves digitally as well as they can in a physical library (including being able to estimate the total size of a digital resource)" (Burrows, 2013, 578). The development of methodologies and accompanying interactive tools for visualization and navigation which expand the "exploratory space" between stage 2 and 3 is where we aim to make a concrete methodological contribution. There are ample indications that data visualization and visual analytics have an extremely important role to play here (e.g., Havre et al. 2000; Smith 2002; Allen et al. 2007; Lee 2007; Schilit and Kolak 2008; Keim et al. 2010; Chuang et al. 2012b; Chuang et al. 2012a; Hirschmann et al. 2012; Broadwell and Tangherlini 2012; Chen et al. 2012; Krstajić et al. 2012; Oelke et al. 2012; Oelke et al. 2013), but also that the involvement of the end-users – HSS researchers – in the design process is crucial for acceptance of the final solution (e.g., Ramage et al. 2009; Chuang et al. 2012b; Rohrdantz et al. 2012). The present study is part of such an initiative, where a rhetorical scholar (Viklund) is working together with a researcher in natural language processing (Borin) and an e-science infrastructure unit (SWE-CLARIN/Språkbanken) on designing, developing and evaluating language-technology based e-science tools for HSS.

The traditional methodology in HSS is qualitative, corresponding to what literary scholars sometimes refer to as "close reading", and most HSS communities are still uncomfortable with large-scale quantitative approaches imported from language technology (LT) and text mining (Gooding, 2013). In this connection it is highly relevant that initiatives like culturomics and, in literary studies, "distant reading"/"macroanalysis" (Moretti, 2005; Moretti, 2013; Jockers, 2013) have so far generally not heeded Shneiderman's (1998, 523) well-known "visual information-seeking mantra": "Overview first, zoom and filter, then details on demand". These initiatives have tended to emphasize the bird's-eye aspect and have generally not provided the means of referring back to and studying individual text passages at close range. We believe that real progress in big-data HSS will only be forthcoming using methodology which combines distant and close reading, quantitative and qualitative research, and allows the researcher to move effortlessly between the two modes of enquiry (e.g., Schöch 2013).

Importantly, the actual mechanisms realizing this methodological point must work both ways. The observed broader statistical regularities must be translated into concrete and detailed research questions, but results of investigating the latter must also be made to inform the basis for future quantitative analysis. We consider it a major advancement in HSS research methodology, if qualitative close-reading methods can be applied to texts and text passages selected not on the basis of convenience, chance or tradition, but by mining very large text collections using explicit and reproducible operationalizations of principled criteria, utilizing information contributed by statistical and linguistic analysis tools and text mining tools, and where the tools are successively attuned to and informed by the research results.

The present work represents a first step in this direction, whereby a corpus infrastructure designed according to the principles described above, but aimed specifically at linguistic research, is pressed into service as a tool for introducing big-data methodology to the study of the history of rhetoric, in the hope that we will learn something new about the history of rhetoric, as well as learn more about which kind of digital tools could be useful for studying it effectively in very large volumes of text.

2 The research question: *doxa* in everyday discourse

Obviously, some questions are better suited than others for text mining. For example, we believe that it might be productive to raise issues that concern expressions of *doxa* (roughly: belief and opinion) in everyday discourse: What did people talk about, and believe was true, good, or beautiful? For instance, what were people's attitudes towards *eloquence* as a cultural phenomenon? How did people talk about eloquence, what ideas where brought forwards in relation to rhetoric, what kind of stereotypes were used, and how were they transformed?

Previous work has showed the cultural importance of of eloquence in Sweden up until the 18th century, as well as the centrality of rhetorical theory in schooling, aesthetics, sermons, political discourse etc.

(e.g., Johannesson 2005). There is a general conception that the importance of rhetoric as a discipline diminishes during the 19th century, only to increase again in our days. However, we have no significant studies that confirm the image of a general demise of rhetoric or eloquence in terms of people's attitudes towards these issues, and probably we need to revise our view also in relation to more specific materials, such as rhetorical manuals (Viklund, 2013). As demonstrated by Fischer (2013) in his study of the status of rhetoric and eloquence in the 18th century debate, the talk of rhetoric's "demise" or "death" in the period is based on an anachronistic view; one didn't conceive of rhetoric in those terms. The status of rhetoric can therefore best be evaluated through studies of the manner in which one talked about eloquence and rhetorical issues in general.

These issues are better investigated taking as the point of departure the large mass of everyday public discourse than through these singular and exceptional books and speeches. The literary scholar Franco Moretti (2013) calls these computational analyses of large masses of texts "distant reading", as opposed to "close reading", generally used in the humanities (see also Moretti 2005; Jockers 2013). Through abstraction and reduction, these quantitative analyses reveal patterns that only emerge from a distance. Of course, the results are not meaningful in themselves. The data are not explanations; these you need to supply yourself. However, what the big-data infrastructure brings to the research by these statistical readings of predominantly yet unread texts, is a tool for understanding historical transformations in a new way. What we don't know about the transformation of rhetoric is precisely the *gradual changes* of opinions in everyday discourse, the mindset of people toward an issue and how it changed.

From about 1840 we see the start of the democratization of politics and public life in Sweden, a process that culminates 1921 in the first election with universal suffrage. During these decades we have the enactment of a number of parliamentary reforms, social movements are changing the public agenda, and new groups of people are entering the public stage, moving the constraints of public debate. The long-term goal of the work in which the present study forms a part is to study the rhetorical formation of public debate on politics and social issues in this time of change. How were political issues talked about, and how did this discourse mutate over time? How did the key concepts of the debates change, and how did the rhetorical framing of these concepts change?

In the history of rhetoric studies – whether they concern rhetorical practice or theory – arguments are generally based on specific examples, which are set in relation to general notions of the "rhetorical tradition". With the large amounts of texts now at our disposal, we can now begin at the other end: to induce patterns of discourses from a vast material that can serve as starting points for much more systematic descriptions of rhetorical practices, as well as analyses of attitudes displayed in the rhetoric of the debate.

From a historical point of view, in order to understand rhetorical practices of a specific period, we need to know about doxa, about opinions and values, social cognitions formulated in language and practiced by groups of individuals in society (Amossy, 2002; Rosengren, 2002). From this angle, argumentation in public debate can be studied in terms of *topoi*, commonplaces or argumentative themes that reflect a system of public knowledge and thereby support the argument (Anscombre, 1995; Angenot, 1982). These topoi, based on values and general opinions, change over time and are difficult to describe systematically. One of our aims is to develop methods for such a systematic diachronic study.

A revision of 19th century rhetorical historiography is long overdue. Scholars have studied rhetorical performance of the period in relation to major authors (e.g., Johannesson et al. 1987; Viklund 2004) or to the specific rhetorical practices of the social movements of late 19th and early 20th century (e.g., Josephson 1991; Mral 1993).

As for the 18th century, scholars have studied public debate and political rhetoric (e.g., Skuncke 1999; Skuncke 2004; Öhrberg 2001; Öhrberg 2010). In addition, a newly finished research project on the attitudes to rhetoric in 18th century newspapers and periodicals (Fischer, 2013; Öhrberg, 2014), and rhetorical practices in the so called cultures of politeness (Öhrberg, 2011), has demonstrated the fruitfulness of meta-rhetorical studies. The status of rhetoric between late 18th and mid 20th century has been considered low, but few empirical studies have been made to support this claim. Lately Viklund (2013)

has begun to revise this simplified historiography, in a study that proposes that there is a renaissance of elocutionary rhetoric parallel to the process of democratization in Sweden.

3 Methodology: Towards a big-data infrastructure for HSS

The digital infrastructure used for this investigation is an advanced corpus search tool called *Korp* (Borin et al., 2012).[1] It is developed and maintained by Språkbanken (the Swedish Language Bank) at the University of Gothenburg. Språkbanken is a national language technology research and infrastructure development center, and the coordinating node of SWE-CLARIN, the national Swedish CLARIN ERIC organization,[2] and Korp is a central component of the Swedish CLARIN infrastructure. Its development started in 2010, drawing on several decades of experience in collecting and processing Swedish text corpora, and making them available for researchers and the public. Korp is a mature corpus infrastructure with modular design and an attractive and flexible web user interface, which is also used by other national CLARIN consortia.[3] Notably, a guiding principle for its design has been exactly the requirement that the user be able to move at will between high-level and abstract overview visualizations and individual data points.

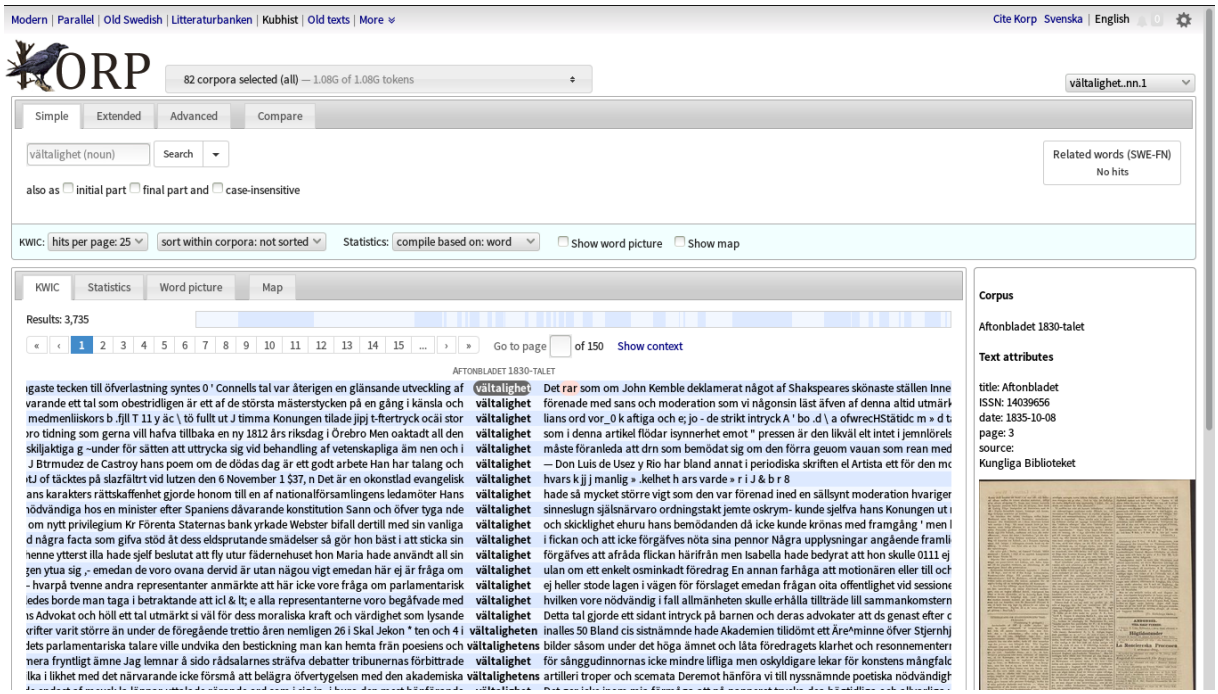

Figure 1: Korp: KWIC view for the lemma *vältalighet* 'eloquence' in Språkbanken's 19th century newspaper corpus (∼1 billion words)

Språkbanken offers access to a large amount of annotated corpora[4] and Korp offers the opportunity to make simple word searches as well as more complicated combined searches utilizing the automatic linguistic annotations present in the corpora.[5]

The results are presented in three different result views: as a list of hits with context (***k**eyword **i**n **c**ontext*: KWIC; see Figure 1); as statistical data with relative and absolute occurrence frequencies in

[1] See <http://spraakbanken.gu.se/korp/#?lang=en>.

[2] See <http://sweclarin.se>.

[3] Korp is used at least in Finland <https://korp.csc.fi>, in Estonia <https://korp.keeleressursid.ee>, and in Norway <http://gtweb.uit.no/korp/> (for the Sami languages).

[4] At the time of writing, the corpora searchable through Korp amount to over 10 billion words, out of which about 1 billion words are historical texts.

[5] Most of the corpora have annotations for part of speech, lemma, and dependency syntax. There is actively ongoing research at Språkbanken on extending and improving the annotations. For the experiments presented here, the lemma annotations have been used, and the dependency-syntax annotations are the basis for Korp's "word picture".

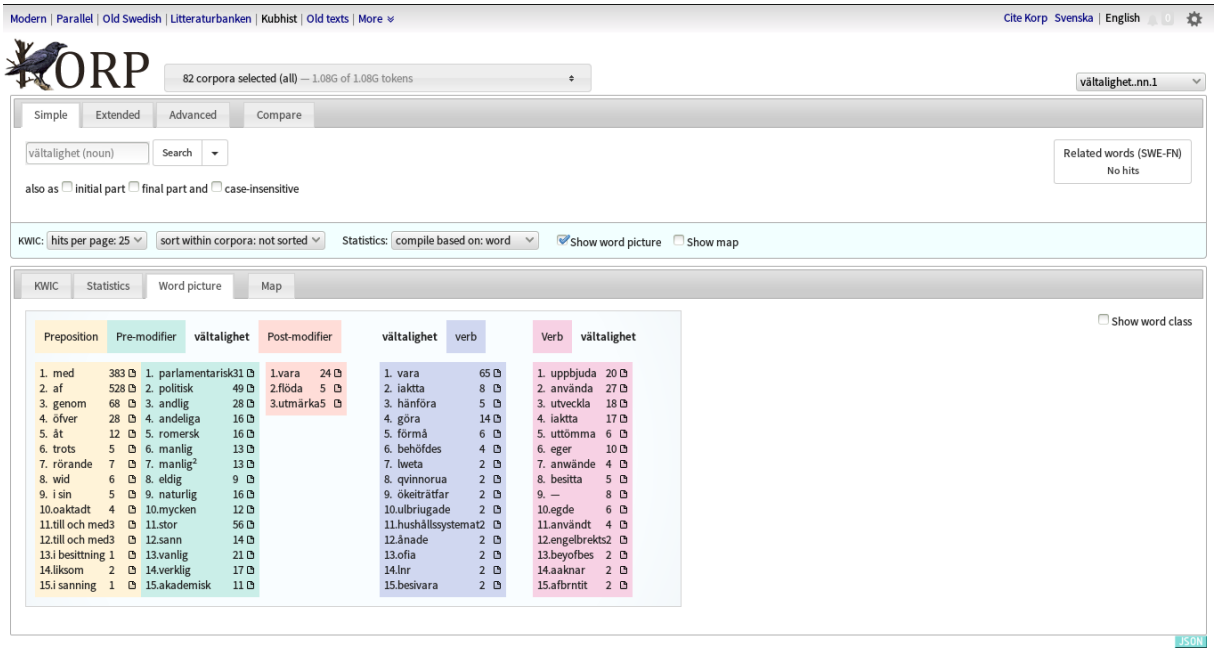

Figure 2: Korp: Word picture for the lemma *vältalighet* ‘eloquence’ in Språkbanken’s 19th century newspaper corpus (∼1 billion words)

subcorpora, which for example give you the opportunity to create a trend graph – relative frequency plotted over time – for one or several words or lemmas (see Figures 3–5 below); and, thirdly, as a “word picture”, which shows the most typical fillers of selected syntactic dependency relations of a word (most typical subjects and objects of a verb, most typical adjectival modifiers of a noun, etc.; see Figure 2). In line with the general interface design principles referred to earlier, the Korp user can move freely between the more comprehensive views and the KWIC view. Thus, clicking a data point on the trend graph, or the document symbol in one of the word picture items, will open a new KWIC view showing the corresponding search hits and their contexts. In the case of non-copyrighted material – e.g., the historical press texts used here – the KWIC view context can be expanded to a longer text passage. For this corpus and some other historical corpora, there is also a link to the digitized page image (see Figure 1).

Although originally devised for the purposes of linguistic analysis of texts, the word picture can be used as a kind of abstract topical maps that guide you to closer readings of the corpus. The corpus used in this study is a collection of historical newspapers from the late 18th to early 20th century, digitized by the Swedish Royal Library. The total corpus contains about one billion words, or almost 70 million sentences. On the one hand this is small in comparison with the Google Books dataset, but, as already mentioned, our corpus is annotated with linguistic information, including lemmatization made using high-quality Swedish lexical resources (modern and historical), which goes a long way towards compensating for the smaller size of the corpus by providing much greater accuracy (Borin and Johansson, 2014; Tahmasebi et al., 2015). Notably, however, and importantly, it is still far larger – by orders of magnitude – than any material previously used for studying the development of Swedish public discourse during this period.[6]

[6]For example, studying the historical linguistic development of the texts in the early Swedish Social Democratic press, Ledin (1995) works with a sample comprising less than 0.5% of the issues published during the 21-year period he examines, and Byrman (1998; 2001) samples short news items from less than 0.03% (issues) of the investigated newspapers for her study of the diachronic development of the language of short news items and public notices. Similarly, the corpus used by Lagerholm (1999) for his investigation of “orality in writing”, is made up of 16,000-word samples collected at 50-year intervals over a period of 200 years, 64,000 words in total.

4 Preliminary findings

So how can this search tool help us to learn more about the history of rhetoric? As previously noted, big-data methods facilitate studies of historical transformations. When did certain words come into use, and when did they disappear? How does the interest in certain topics change over time? More particularly, the Korp infrastructure helps us to visualize certain trends in what people talked about in daily newspapers, which is of great value when studying the topics of rhetoric, eloquence and debate.

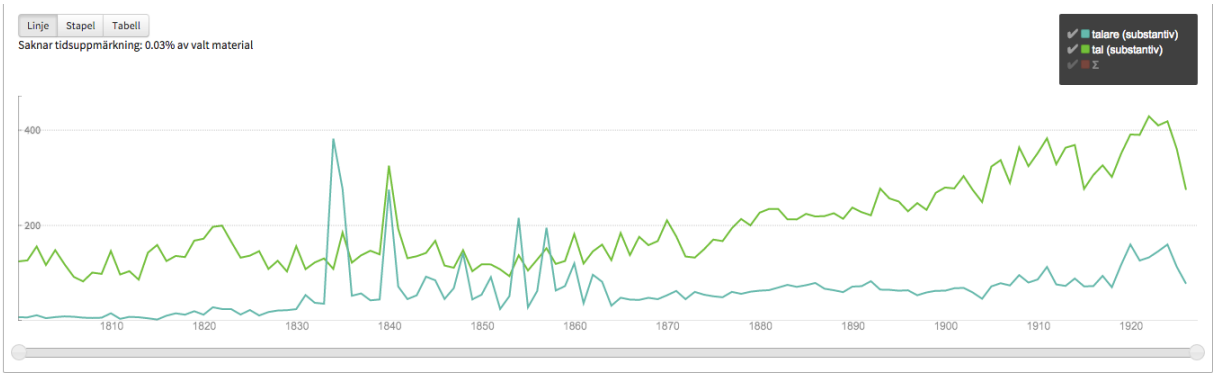

Figure 3: *Talare* 'speaker' and *tal* 'speech', 1800–1920s

The trend graphs in Figure 3 show the use of the words *talare* 'speaker' and *tal* 'speech' between 1800 and the 1920s in the Swedish press corpus. The nine peaks between the 1830s and 1860s coincide with the parliament sessions held every third year. In 1866 the old parliament where only the upper strata of society were represented, was dissolved, in favor of a new, more egalitarian parliamentary system. Apparently the newspapers were very keen on reporting the discussions in the old parliament, but less so thereafter, when the parliament met annually. The graph prompts a number of interesting questions: Why the sudden rise of interest in speakers around 1835, and why did the interest in the debates diminish? And still one notices a slow but steady increase of interest in the 'speaker' and in 'speech' in newspapers from the 1860s to the 1920s. Whereas the first peaks point to institutional changes and how they were reflected in press reports, it seems probable that the subsequent rise might suggest a more general interest in public debate and public opinion.

The testing of another hypothesis might support this assumption, one that seems plausible in light of what we know of the emergence of social movements in this period. We ought to be able to see a correlation between on the one hand democratization and the rising interest in politics, and, on the other, an increasing interest in rhetorical practices: oral performances and debate. As a simple way of testing this one might see to what degree people talked about 'politics' and 'democracy'. That is, through these search queries one can get an idea of the topical changes in the newspapers. The graphs in Figure 4 seem to confirm that there is an increasing interest in politics towards the end of the investigated period.

The next step would be to investigate some terms that we associate with rhetorical performance: *föreläsning*, *föredrag* 'lecture'; *tal* 'speech'; *declamation* 'declamation'; *debatt* 'debate'; *framförande* 'delivery, performance'; *agitation* 'agitation'; *anförande* 'speech'. The distribution in our corpus material of some of these is shown in Figure 5.

5 'Eloquence' in 19th century Swedish public discourse

All these results point to an increase in talk about rhetorical practices. As opposed to the words mentioned earlier that become markedly more frequent during the 19th century, for example *talare* 'speaker', *debatt* 'debate', and *deklamation* 'declamation', the word *vältalighet*, 'eloquence', has a relatively stable trend curve. One can simply surmise that it is a constant cultural concern all through the period. Navigating through a large part of the examples generated by the word picture function of Korp (see Figure 2) it was obvious that although the context in which the word was used did change to some extent, it was not worth investigating at this point. For example, the use of the word in a political context increased, but that

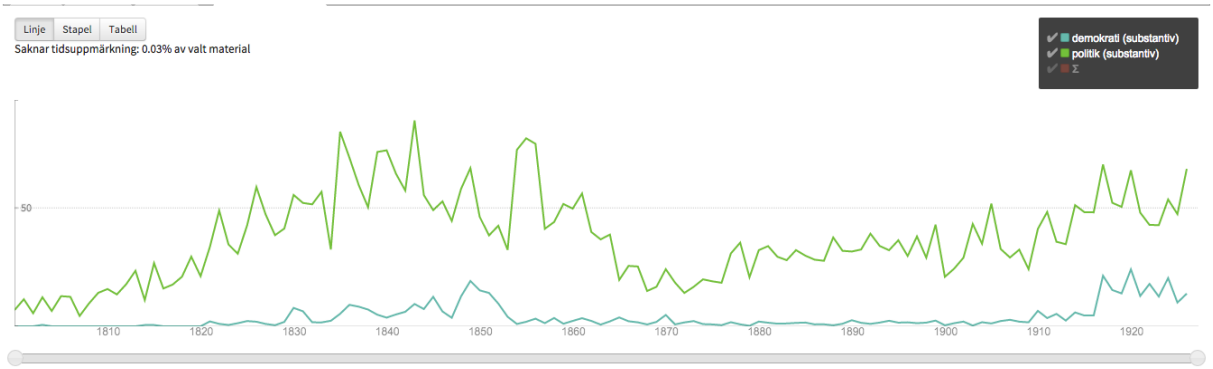

Figure 4: *Demokrati* ‘democracy’ and *politik* ‘politics’, 1800–1920s

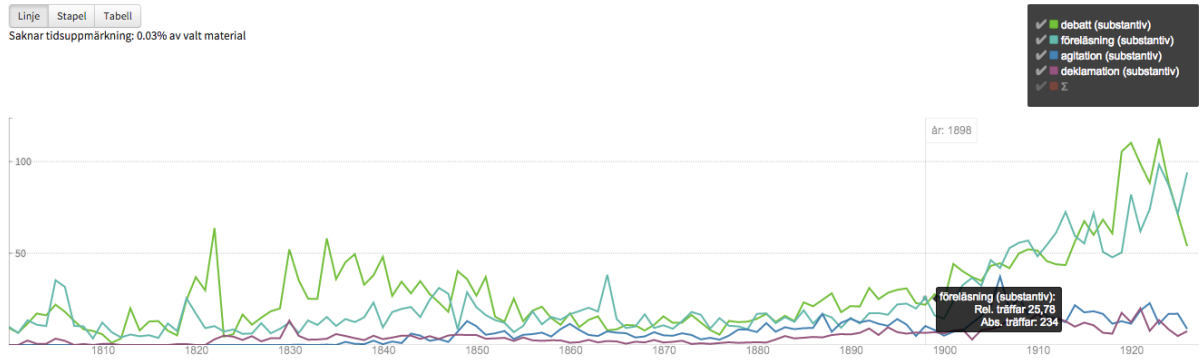

Figure 5: Terms associated with rhetorical performance, 1800–1920s

was only predicable, due to the process of democratization mentioned above. So initially, it seemed most productive to see what kind of generalizations one might make about the use of the word over the whole period, generalizations that could be the starting point for a study of the transformation of the concept of rhetoric in the 20th century. What kinds of attitudes toward *vältalighet* ‘eloquence’ are revealed in the newspaper texts?

The term “attitude” is used in a broad sense, as a manner of thinking about an object or phnomenon, often in terms of a positive or negative evaluation. In modern rhetorical criticism attitudes are often analyzed in relation to style and word choices (e.g., Burke 1969; Billig 1996): the way we use language reveals our attitudes toward a thing or person. Attitudes in this sense are not necessarily connected to a person’s stated intentions. In the discussion below, looking for attitudes is, on the one hand, a perspective that makes it easier to sort out central meanings in the use of the word ‘eloquence’, and, on the other, one that highlights a certain framing of the word.

So what kinds of attitudes toward eloquence and rhetoric are displayed in the material? The methodological gain with this broad starting point is that it does not exclude any new information the searches in Korp might produce. As indicated earlier the word picture function in Korp can be used as a kind of topical search tool: what words go together with this noun, *vältalighet* ‘eloquence’, and what do they reveal about the uses of the word? The searches produce for example the most common, or the most representative, modifiers, generally adjectives, and when studied in greater detail, one could observe a pattern of notions that describe certain qualities of ‘eloquence’. They can be divided into three main groups, plus an “other” category:

- **Genre:** ‘parliamentary’, ‘spiritual’, ‘political’, ‘Roman’, ‘academic’, ‘marital’, etc.
- **Truthfulness and naturalness:** ‘real’, ‘true’, ‘right’, etc.; ’manly’, ’natural’, ‘artless’, ‘unpretentious’, ‘simple’, etc.

- **Conceptual metaphors:** 'fiery', 'glowing', 'burning' (eloquence is fire); 'flowing', 'fluent', 'pouring' (eloquence is water)
- **Other:** 'mute', 'normal', 'mighty', 'great', 'irresistible', 'bold'

The categories retrieved from the word picture help us to better understand the linguistic contexts in which one talked about eloquence, as well as the attitude towards rhetoric evidenced in the corpus. The words prompt a number of interesting explorations. One might for example investigate positive vs. negative connotations of the concept of eloquence, or one might look into gender differences in relation to the various examples and the categories produced by the computational reading.

We found the conceptual metaphors interesting. It was surprising to see that the majority of metaphors connected to the notion of 'eloquence' so clearly could be divided in two main categories: either *eloquence is fire* or *eloquence is water*. In these kind of metaphors, described first by cognitive linguists, one conceptual domain – the target domain, here *eloquence* – is mapped onto another – the source domain, here *fire* and *water*. These concepts are often analyzed in terms of image schemas, which, in simplified terms can be said to be mental patterns that structure the metaphorical expression. Talking about eloquence as warmth, and as sparks, for example, demonstrates how the metaphor can be based on motor or sensory experience. And water, to add another example, is in these expressions often *flowing from one container to another*; so the expressions are orientational, and often they are emphasizing the quality of *fluidity* or *fluency*. (Lakoff and Johnson, 1980; Geeraerts, 2006) These image schemas can help us to understand what the metaphors are saying about the attitudes expressed in the examples.

Methodologically we here used all the context material retrieved in connection to the word *vältalighet* 'eloquence' and searched for the most common words in the conceptual metaphor clusters, and in that way ended up with many examples of these metaphorical expressions that were not necessarily associated with the specific word 'eloquence', but with the more general concept of 'rhetoric'.

So what can be learnt about how 'eloquence' was perceived from these clusters of words? After studying the many instances of these metaphors we found that they generated a deeper understanding of the attitudes toward 'eloquence' during the period under study, and we have summarized these insights into four points.

1. Positive values: A general knowledge retrieved from these expressions concerns which positive values are emphasized. The fire metaphors generally express images of force: pathos, burning hearts, passion, and energy.

> *Hans herravälde öfver sinnena måste man, såvida man icke är alltför intagen af fördomar, uteslutande tillskrifva hans lågande vältalighet, hans alltid slagfärdiga dialektik och särskildt det glansfulla sätt på hvilket han vet att försvara de demokratiska grundsatserna.*
>
> 'His mastery over the senses must, unless one is too captivated by prejudice, be ascribed exclusively to his fiery eloquence, his invariably witty dialectics, and especially the glittering manner in which he is able to defend the democratic tenets.'

The metaphors build on eloquence as a force of nature. Fire, heat, glow and lightning are not only natural phenomenon, they are also overpowering, i.e. they have the power to overwhelm the senses of the listener. Another positive quality is the association to genius:

> *[. . .] parlamentarisk vältalighet. Såsom lyrisk skald lyser Béranger i synnerhet genom ingifvelsens eld och äkta originalitet [. . .]*
>
> '[. . .] parliamentary eloquence. As lyric poet Béranger shines especially through the fire of inspiration and genuine originality [. . .]'

One should note that the conceptual metaphor *eloquence is fire* affirms the distinction between the two main modes of persuasion: with reason, *logos*, or with emotion, *pathos*, since it always expresses the force of the latter. One could say that these expressions reflect one mode of rhetorical proof while deflecting another. That dichotomy is, of course, generally only implicitly present in the examples, but

once in a while you can find it thematized in the texts, as in the example where someone is contrasting eloquence in Norway and Sweden: the heat of passion is contrasted with the calmer and colder nature of reason:

> *Norska tribunens vältalighet är af en lugnare natur i det man mer lägger an på att verka genom skäl än granna ord mera söker verka på förståndet än känslan hvars villfarelse man fruktar; denna tribun är icke derföre mindre mäktig och verksam. Om den ock ej disponerar tordönet och ljungelden så har den likväl grunder kallblodighet ståndaktighet och mod. Debatterna gå sällan utom den lugna diskussionens gränser men föredragen som nästan alltid äro muntliga och improviserade ersätta som oftast genom grundlighet och öfvertygande kraft hvad dem brister i liflighet och värma.*
>
> 'The eloquence of the Norwegian tribune is of a quieter nature in that one makes a point of rather acting by reason than gaudy words, more seeking to act on the mind than the feeling, the delusion of which is feared; this tribune is not therefore less powerful and effective. Even though it may not possess thunder and lightning, it is nonetheless grounded in coolness, steadfastness, and courage. The debates rarely exceed the limits of calm discussion, but the talks that almost always are oral and improvised, usually substitute what they lack in vividness and heat with thoroughness and persuasive force.'

2. Attitudes toward gender: The concept *eloquence is fire* also highlights attitudes toward gender. The fire metaphors are clearly coded as a male feature – there are no women described or speaking in this category. This is not surprising; force and genius are qualities that traditionally have been seen as male. But an awareness of the consistency is important; it would be interesting to see at what time in history this trend is broken. In the other metaphorical concept – *eloquence is water* – we have examples that refer to both men and women.

3. Attitudes toward eloquence as an art: The *eloquence is water* concept frames the attitudes to rhetoric in a way that it can be used either positively or negatively (ironically), as opposed to the *eloquence is fire* concept that almost always is used in an unambiguously positive sense. When eloquence is described as flowing, streaming etc. the semantic orientation has more to do with rhetorical ability than degree of pathos.

> *Hans klara och lätta diction flöt rikt ex tempore och blef jemväl full af eld när det behöfdes*
>
> 'His clear and easy diction flowed abundantly ex tempore and yet became full of fire when needed'
>
> *Alla dessa enskildheter flödade från timmermannens läppar i en ström af enkel vältalighet, hvilken inga lektioner i "uttalslära" hade kunnat föröka med ett grand af ytterligare effekt*
>
> 'All these particularities flowed from the carpenter's lips in a stream of simple eloquence, which no lessons in "pronunciation" could multiply with a mote of additional power'

Here the image schema describing the technical ability of delivery – diction – has to do with fluency. A person characterized with flowing eloquence has the rhetorical skill of speaking naturally without displaying too much art. Both thoughts and feelings come "streaming from the speaker", and this fluency can be a sign of natural ability – opposed to the art of rhetoric as in the latter example – or just a sign that one masters the art. For this reason, the metaphorical concept is often used negatively when there is a conflict between art and genuine thoughts and feelings:

> *Med glödande öfvertygelse, med prålande ord och flytande vältalighet framhålla de, nationernas representanter, var och en på sitt håll, sanningen sådan den bäst passer sig för dem*
>
> 'With glowing conviction, with pompous words and fluent eloquence, those representatives of the nation emphasize, each in their own way, the truth as they see fit'

Han ägde en obeskriflig vältalighet, men jag tyckte nästan han talade med alltför stor lätthet – det flödade öfver som en flod

'He possessed a tremendous eloquence, but I almost thought he was speaking with too much ease – it flowed over like a river'

Men inte ens den mest ifriga anhängare af »saken», kunde i hettan uthärda 2 timmars flödande vältalighet från 20 talarstolar på en gång

'But not even the most ardent supporter of the "cause", could in the heat endure two hours of flowing eloquence from 20 podiums at once'

4. Attitudes toward eloquence: from one heart to another: The most characteristic image schema of the two metaphorical concepts concerns the pattern *from one container to another*. A fire is burning *from* the soul, and a flame, a bolt of lightning or an electric spark is coming *from* the speaker to the audience and sets the listener's mind on fire. Likewise, the stream of eloquence is sometimes described as coming, with a cliché often used, from the heart of a speaker.

Han älskade alltid att dröja vid denna stora tanke; och äfven nu sökte han med all sin brinnande vältalighet att inskrifva den outplånligt i sina åhörares hjertan.

'He always loved to dwell on this great idea; and also this time he sought with all his fiery eloquence to write it indelibly into the hearts of his listeners.'

Engelbrekts flammande vältalighet hade bland det mäktiga Söderköpings talrika borgerskap upptändt en eld som ännu glödde under askan [. . .]

'Engelbrecht's flaming eloquence had among the mighty and numerous burghers of Söderköping kindled a fire which still glowed under the ashes . . . '

Vi voro dock nog djerfve att trotsa denna ordförandens uppmaning och stannade alltså , afvaktande om något af den visdom , som flödade öfver hans läppar, möjligen kunde tränga in i våra dumma hufvuden och vi derigenom sättas i stånd att begripa hvad en utgift på 2,000 kr [. . .]

'We were, however, bold enough to defy the chairperson's request and therefore stayed, awaiting to see if some of the wisdom that flowed from his lips, could possibly penetrate our stupid heads so that we could be able to comprehend what an expenditure of 2,000 kr . . . '

Då Champagnen började flöda , då sprungo också alla snillets källor upp och den mest eldiga vältalighet slog i hvart ögonblick sin elektriska gnista i förvånade åhörares sinnen

'When the champagne began to flow, then all the sources of genius also erupted and the most fiery eloquence struck, in every passing moment, its electric spark in the minds of the astonished listeners'

The last, more ironical instances of the metaphor seem to suggest that the writers have no problems turning the concept of the ideal orator upside down. But despite of this one can assume that the very nature of these orientational metaphors suggests a cognitive frame indicating that eloquence has a special status as a communicative tool. But for how long are these expressions used? Today this idea of communication still exists, but one would not primarily find these metaphors in connection with the words 'eloquence' or 'rhetoric'. More likely one would find them in sentimental discourses about love or friendship.

Once again, it would be interesting to investigate how these verbal conventions transform over time, because of course they do. Even though we do not yet have enough comparable data for most of the 20th century to show the details of the changes that have taken place, we have large amounts of evidence from the latest form of written public discourse, i.e., social media such as blogs and online discussion forums, represented by close to 8 billion words accessible through Korp in Språkbanken. Thus, we can compare today's attitudes towards rhetoric to those uncovered in the 19th century material.

Today, for one, *vältalighet* 'eloquence' is rarely used – the word *retorik* 'rhetoric' has taken its place – and both these words show quite different distributions in modern corpora, such as newspapers and

blogs from the last four decades, as compared to the 19th century material. Looking at the word pictures generated on basis of the modern material, we find almost exclusively two kinds of modifiers: (1) *type of rhetoric* (for example: 'political', 'feminist', 'religious', 'social democratic'); or (2) *negative qualities* (for example: 'empty', 'made-up', 'aggressive'); and of course the two categories combined (for example: 'racist', 'populist', 'scholastic').

6 Conclusions and future work

Above we have described some initial experiments where a very large historical corpus of Swedish newspaper text and the state-of-the-art corpus infrastructure of SWE-CLARIN have been brought to bear on research questions in the field of rhetorical history.

A central and important purpose of this work was to investigate the method itself: Does it produce new knowledge? Yes and no. For instance, the fact that two conceptual metaphors – *eloquence is fire* and *eloquence is water* – were so dominant in the material definitely adds to our knowledge of *doxa* during the 19th century. Most other results were expected, or at least not surprising. But even if the results only confirm old knowledge it is worthwhile: if the method reveals results that confirm old knowledge then it might be able to see also new things which until now have not been acknowledged.

The method is promising, and we see several natural directions in which this work can be continued. The material studied here covers the 19th and the beginning of the 20th century. As described above, we also did a preliminary study using 21st century social-media and news text. The two studies together indicate that there have been considerable changes in the attitudes toward eloquence and rhetoric as expressed in Swedish public discourse over the last two centuries, but the details of these changes (including their timing) remain beyond our ken for the time being. With a corpus that covered also the 20th century it would be possible to advance our knowledge on this score.

Finally, the present corpus infrastructure – intended mainly for linguistically oriented research – proved useful for the research questions that we have described above. However, the old adage about what happens when you have a hammer is certainly valid here: We did adapt our research questions so that they could be addressed using the linguistic annotations, search functions and visualizations that Korp makes available. These should be complemented by text processing and interfaces more geared towards supporting more general digital humanistic inquiry. In particular, to the form-oriented search useful to linguists we would like to add *content-oriented* search modes – based on information-retrieval or information-extraction techniques or content-classification technologies such as topic modelling, vector-space models, or word embeddings – accessible through interfaces that would still allow the user to move easily and effortlessly between various forms of macro view visualization ("distant reading") and individual instances ("close reading"). To repeat: This we believe to be a crucial – even necessary – feature of any such tool.

Acknowledgements

The research described here has been supported in part by a framework grant from the Swedish Research Council (*Towards a knowledge-based culturomics* 2012–2016; project no 2012-5738). The basic research infrastructure development involved has been made possible by the Swedish Research Council's funding of SWE-CLARIN (2014–2018), the Swedish node of the CLARIN ERIC.

References

Erez Aiden and Jean-Baptiste Michel. 2013. *Uncharted: Big data as a lens on human culture*. Riverhead Books, New York.

Robert B. Allen, Andrea Japzon, Palakorn Achananuparp, and Ki Jung Lee. 2007. A framework for text processing and supporting access to collections of digitized historical newspapers. In M. J. Smith and G. Salvendy, editors, *Human interface, Part II, HCII 2007*, number 4555 in LNCS, pages 235–244. Springer, Berlin.

Ruth Amossy. 2002. How to do things with doxa: Toward an analysis of argumentation in discourse. *Poetics Today*, 23(3):465–487.

Marc Angenot. 1982. *La parole pamphlétaire: Typologie des discours modernes*. Payot, Paris.

Jean-Claude Anscombre. 1995. *Théorie des topoï*. Kimé, Paris.

Michael Billig. 1996. *Arguing and thinking: A rhetorical approach to social psychology*. Cambridge University Press, Cambridge.

Lars Borin and Richard Johansson. 2014. Kulturomik: Att spana efter språkliga och kulturella förändringar i digitala textarkiv. In Jessica Parland-von Essen and Kenneth Nyberg, editors, *Historia i en digital värld*.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Peter M. Broadwell and Timothy R. Tangherlini. 2012. TrollFinder: Geo-semantic exploration of a very large corpus of Danish folklore. In *The Third Workshop on Computational Models of Narrative*, pages 50–57, Istanbul. ELRA.

Kenneth Burke. 1969. *A rhetoric of motives*. University of California Press, Berkeley.

Toby Burrows. 2013. A data-centred 'virtual laboratory' for the humanities: Designing the Australian Humanities Networked Infrastructure (HuNI) service. *Literary and Linguistic Computing*, 28(4):576–581.

Gunilla Byrman. 1998. Tidningsnotisen i förändring 1746–1997. Institutionen för nordiska språk, Lunds universitet. Svensk sakprosa, rapport nr 15.

Gunilla Byrman. 2001. Municipalstämma hölls igår i Tomelilla .... Svenskt notisspråk 1746–1997. In Björn Melander and Björn Olsson, editors, *Verklighetens texter. Sjutton fallstudier*, pages 443–483. Studentlitteratur, Lund.

Annie T. Chen, Ayoung Yoon, and Ryan Shaw. 2012. People, places and emotions: Visually representing historical context in oral testimonies. In *The Third Workshop on Computational Models of Narrative*, pages 45–49, Istanbul. ELRA.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012a. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*.

Jason Chuang, Daniel Ramage, Christopher D. Manning, and Jeffrey Heer. 2012b. Interpretation and trust: Designing model-driven visualizations for text analysis. In *ACM Human Factors in Computing Systems (CHI)*.

Otto Fischer. 2013. *Mynt i Ciceros sopor. Retorikens och vältalighetens status i 1700-talets svenska diskussion*, volume 1 of *Södertörn Retoriska Studier*. Södertörns högskola, Huddinge.

Roberto Franzosi. 1987. The press as a source of socio-historical data: Issues in the methodology of data collection from newspapers. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 20(1):5–16.

Dirk Geeraerts, editor. 2006. *Cognitive linguistics: Basic readings*. De Gruyter, Berlin.

Paul Gooding. 2013. Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing*, 28(3):425–431.

S. Havre, B. Hetzler, and L. Nowell. 2000. ThemeRiver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*, pages 115–123, Salt Lake City.

Hagen Hirschmann, Anke Lüdeling, and Amir Zeldes. 2012. Measuring and coding language change: An evolving study in a multilayer corpus architecture. *ACM Journal on Computing and Cultural Heritage*, 5(1):article 4.

Matthew L. Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, Urbana/Chicago/Springfield.

Kurt Johannesson, Eric Johannesson, Björn Meidal, and Jan Stenkvist. 1987. *Heroer på offentlighetens scen. Politiker och publicister i Sverige 1809–1914*. Tidens förlag, Stockholm.

Kurt Johannesson. 2005. *Svensk retorik. Från medeltiden till våra dagar*. Norstedts, Stockholm.

Olle Josephson. 1991. *Diskussionsskolan 1886: Språkmiljö, argumentation och stil i tidig arbetarrörelse*. Number 1 in Arbetarrörelsen och språket. Avdelningen för retorik, Uppsala universitet, Uppsala.

Daniel A. Keim, Leishi Zhang, Miloš Krstajić, and Svenja Simon. 2010. Solving problems with visual analytics: Challenges and applications. *ACM Transactions on Embedded Computing Systems*, 4(4):article 39.

Miloš Krstajić, Mohammad Najm-Araghi, Florian Mansmann, and Daniel A. Keim. 2012. Incremental visual text analytics of news story development. In *Proceedings of Conference on Visualization and Data Analysis (VDA '12)*.

Per Lagerholm. 1999. *Talspråk i skrift. Om muntlighetens utveckling i svensk sakprosa 1800–1997*. Number A 54 in Lundastudier i nordisk språkvetenskap. Lunds universitet, Institutionen för nordiska språk, Lund.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press, Chicago.

Per Ledin. 1995. *Arbetarnes är denna tidning. Textförändringar i den tidiga socialdemokratiska pressen*. Number 20 in Acta Universitatis Stockholmiensis: Stockholm Studies in Scandinavian Philology, New Series. Almqvist & Wiksell International, Stockholm.

John Lee. 2007. A computational model of text reuse in ancient literary texts. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 472–479, Prague. ACL.

Mats Malm. 2014. Digitala textarkiv och forskningsfrågor. In Jessica Parland-von Essen and Kenneth Nyberg, editors, *Historia i en digital värld*.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, (331).

Franco Moretti. 2005. *Graphs, maps, trees: Abstract models for a literary history*. Verso, London/New York.

Franco Moretti. 2013. *Distant reading*. Verso, London/New York.

Brigitte Mral. 1993. *Kommunikation och handlande i Malmö kvinnliga diskusionsklubb 1900–1904*. Number 6 in Arbetarrörelsen och språket. Avdelningen för retorik, Uppsala universitet, Uppsala.

Daniela Oelke, Dimitrios Kokkinakis, and Mats Malm. 2012. Advanced visual analytics methods for literature analysis. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 35–44, Avignon. ACL.

Daniela Oelke, Dimitrios Kokkinakis, and Daniel A. Keim. 2013. Fingerprint matrices: Uncovering the dynamics of social networks in prose literature. *Computer Graphics Forum*, 32(3):371–380.

Ann Öhrberg. 2001. *Vittra fruntimmer. Författarroll och retorik hos frihetstidens kvinnliga författare*. Gidlunds, Hedemora.

Ann Öhrberg. 2010. "Fasa för all flärd, konstlan och förställning". Den ideala retorn inom 1700-talets nya offentlighet. *Samlaren*, 131.

Ann Öhrberg. 2011. Between the civic and the polite. Classical rhetoric, eloquence and gender in late eighteenth century Sweden. In Otto Fischer and Ann Öhrberg, editors, *Metamorphoses of Rhetoric. Classical Rhetoric in the Eighteenth Century*, number 3 in Studia Rhetorica Upsaliensia. Uppsala University, Uppsala.

Ann Öhrberg. 2014. *Samtalets retorik. Belevade kulturer, offentlig kommunikation och kön i svenskt 1700-tal*. Symposions förlag, Höör.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, 10(10):e0137041, 10.

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D. Manning, and Daniel A. McFarland. 2009. Topic modeling for the social sciences. In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada.

Christian Rohrdantz, Michael Hund, Thomas Mayer, Bernhard Wälchli, and Daniel A. Keim. 2012. The world's languages explorer: Visual analysis of language features in genealogical and areal contexts. *Computer Graphic Forum*, 31(3):935–944.

Mats Rosengren. 2002. *Doxologi. En essä om kunskap*. Rhetor förlag, Åstorp.

Bill N. Schilit and Okan Kolak. 2008. Exploring a digital library through key ideas. In *Proceedings of JCDL'08*, pages 177–186, Pittsburgh. ACM.

Christof Schöch. 2013. Big? Smart? Clean? Messy? Data in the humanities. *Journal of Digital Humanities*, 2(3). <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/>.

Ben Shneiderman. 1998. *Designing the user interface*. Addison-Wesley, Reading, Mass., 3rd ed edition.

Marie-Christine Skuncke. 1999. Den svenska demokratidebatten 1766–1772. In Rut Boström Andersson, editor, *Ordets makt och tankens frihet. Om språket som maktfaktor*. Uppsala universitet, Uppsala.

Marie-Christine Skuncke. 2004. Press and political culture in Sweden at the end of the Age of liberty. Enlightenment, revolution and the periodical press. In Hans-Jürgen Lüsebrink and Jeremy D. Popkin, editors, *SVEC 2004:06*. Voltaire Foundation, Oxford.

David A. Smith. 2002. Detecting and browsing events in unstructured text. In *Proceedings of SIGIR'02*, Tampere. ACM.

Nina Tahmasebi, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik Johansson, Richard Johansson, Mikael Kågebäck, Olof Mogren, Pierre Nugues, and Thomas Risse. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2–4):169–187.

Timothy R. Tangherlini. 2013. The folklore macroscope. Challenges for a computational folkloristics. *Western Folklore*, 72(1):7–27.

Jon Viklund. 2004. *Ett vidunder i sitt sekel. Retoriska studier i C.J.L. Almqvists kritiska prosa*. Gidlund, Hedemora.

Jon Viklund. 2013. Performance in an age of democratization: The rhetorical citizen and the transformation of elocutionary manuals in Sweden ca. 1840–1920. Paper presented at ISHR [International Society for the History of Rhetoric] biannual conference in Chicago.

Claire Warwick, Melissa Terras, Paul Huntington, and Nikoleta Pappa. 2008. If you build it will they come? The LAIRAH study: Quantifying the use of online resources in the arts and humanities statistical analysis of user log data. *Literary and Linguistic Computing*, 23(1):85–102.

Ben Zimmer. 2013. When physicists do linguistics. Is English 'cooling'? A scientific paper gets the cold shoulder. *Boston Globe*, February 10.