Swe-Clarin: Language Resources and Technology for Digital Humanities

Lars Borin, Nina Tahmasebi, Elena Volodina
Språkbanken, Department of Swedish, University of Gothenburg
firstname.lastname@svenska.gu.se

Stefan Ekman, Caspar Jordan
Swedish National Data Service, University of Gothenburg
firstname.lastname@snd.gu.se

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

Beáta Megyesi, Jesper Näsman
Department of Linguistics and Philology, Uppsala University
firstname.lastname@lingfil.uu.se

Anne Palmér
Department of Scandinavian Languages, Uppsala University
anne.palmer@nordiska.uu.se

Mats Wirén, Kristina N Björkenstam, Gintarė Grigonytė, Sofia Gustafson Capková
Department of Linguistics, Stockholm University
firstname.lastname@ling.su.se

Tomasz Kosiński
Department of Applied IT, Chalmers University of Technology
tomasz.kosinski@chalmers.se

Keywords: Swe-Clarin, CLARIN, digital humanities, language technology

Abstract

CLARIN is a European Research Infrastructure Consortium (ERIC), which aims at (a) making extensive language-based materials available as primary research data to the humanities and social sciences (HSS); and (b) offering state-of-the-art language technology (LT) as an e-research tool for this purpose, positioning CLARIN centrally in what is often referred to as the digital humanities (DH). The Swedish CLARIN node Swe-Clarin was established in 2015 with funding from the Swedish Research Council.

In this paper, we describe the composition and activities of Swe-Clarin, aiming at meeting the requirements of all HSS and other researchers whose research involves using text and speech as primary research data, and spreading the awareness of what Swe-Clarin can offer these research communities. We focus on one of the central means for doing this: pilot projects conducted in collaboration between HSS researchers and Swe-Clarin, together formulating a research question, the addressing of which requires working with large language-based materials. Four such pilot projects are described in more detail, illustrating research on rhetorical history, second-language acquisition, literature, and political science. A common thread to these projects is an aspiration to meet the challenge of conducting research on the basis of very large amounts of textual data in a consistent way without losing sight of the individual cases making up the mass of data, i.e., to be able to move between Moretti's "distant" and "close reading" modes.

While the pilot projects clearly make substantial contributions to DH, they also reveal some needs for more development, and in particular a need for document-level access to the text materials. As a consequence of this, work has now been initiated in Swe-Clarin to meet this need, so that Swe-Clarin together with HSS scholars investigating intricate research questions can take on the methodological challenges of big-data language-based digital humanities.

Introduction: CLARIN and Swe-Clarin

CLARIN (Common Language Resources and Technology Infrastructure) is a European Research Infrastructure Consortium (ERIC), an ESFRI (European Strategy Forum on Research Infrastructures) initiative which aims at (a) making extensive language-based materials available as primary research data to the humanities and social sciences (HSS) research communities; and (b) offering state-of-the-art language technology (LT) as an e-research tool for this purpose, positioning CLARIN centrally in what is often referred to as the digital humanities (DH).

Swe-Clarin as the Swedish CLARIN node was established in 2015 with funding from the Swedish Research Council by a consortium consisting of 9 members – so-called Swe-Clarin centers – representing the Swedish academic community as well as public memory institutions. There is a good balance of academic members from across the LT field, covering existing and possible research areas and user groups, and the memory institutions provide access to many of the language-based materials of interest to the users. Swe-Clarin is coordinated by Språkbanken (the Swedish Language Bank) at the University of Gothenburg.

From the start, Swe-Clarin has aimed to establish good relations with the HSS fields and open the door to all researchers who wish to work with DH research using text and speech as primary research data. Swe-Clarin – and also CLARIN ERIC as a whole – is an organization originally established by language technologists in close collaboration with linguists, who are among the most established users of digital research e-infrastructure; machine translation experiments are almost as old as the first computers; and corpus linguistics traces its roots to about the same period. To avoid being a project by language technologists for linguists, we strive to include the HSS researchers in the process as early as possible. Our preferred way of doing this has been to establish pilot projects with at least one member from the HSS field and at least one Swe-Clarin consortium member, together formulating a research question the addressing of which requires working with large language-based materials. Ideally, the collaboration should additionally always include a data owner, a person or persons representing the institution where the text or speech data is kept – typically a memory institution, such as a library, archive or museum. In Sweden, the most important being the National Library, the National Archives, the Nordic Museum, and the Institute for Language and Folklore.

In addition to the pilot projects, we have arranged workshops and user days and we have published newsletters and a blog. The workshops held so far have been on topics such as: a general introduction to Swe-Clarin, our tools and resources; historical resources and tools; making cultural heritage textual data available for research; how to deal with imperfect OCR; and HSS research on digitized speech data, such as those of the Swedish Media Archive. We have started a series of national, hands-on training events called *Swe-Clarin on tour* using Språkbanken's widely used Korp corpus infrastructure (Borin et al., 2012) to explore previously unexplored materials in a hands-on way, giving researchers of LT and HSS the opportunity to meet and discuss research questions and the potential of using LT for DH. The experience of working with HSS researchers will help reveal the limitations of existing tools and hopefully also stimulate our target communities to enter into broader methodological discussions, thus setting the stage for future development of tools more appropriate for DH research. The first such event was held at Stockholm University in the spring of 2016. It featured the ethnographic questionnaires collected by the Nordic Museum from the late 1920s and now digitized by them, and was attended mainly by ethnologists. The tour subsequently visited Umeå in conjunction with the *Swedish Language Technology Conference* in November 2016. On this occasion, the resources under discussion were the Swedish Government Official Reports (*Statens offentliga utredningar*, SOU) in the version digitized by the National Library of Sweden, comprising more than 400 million words covering the years 1922–1998, and the event was attended mainly by political scientists. The third event in this series took place in the fall of 2017, at Södertörn

University College, in the southern part of the greater Stockholm area: the subject investigated at this event was online social media.

Swe-Clarin pilot projects: Background and motivation

The pilot projects aim to spread the word about Swe-Clarin, show the potential of using language technology in DH research, create a user base for the tools and resources developed and maintained by Swe-Clarin and, last but not least, ensure that this development is informed by input from users in the earliest possible stages of the project. We know that HSS research relies to a large extent on primary data in the form of information conveyed by language (in the form of text or speech), and our infrastructure is based on the premise that the only reasonable way that such research can scale up to meet the challenges of the massive availability of digital textual resources from all relevant periods is through the use of state-of-the-art language technology. Nevertheless, this infrastructure is still in its infancy, and a more general picture of its contribution to HSS research can only emerge after a good deal of trial-and-error. Thus, one of the main purposes of the Swe-Clarin pilot projects is to be exploratory.

Ideally, a pilot project should produce results, preferably including one or more publications. It should also generate funding proposals to be submitted to one of the national or international (Nordic or European) funding agencies.

Several pilot projects are now underway. Among these we find, for example, studies of the occupational roles of women and men in early modern Sweden (appr. 1550–1800 CE); an investigation of ethnic bias in the grading of high-school tests; a study of translations of works by Swedish female writers of the early 20th century; development of information extraction methods for specialized text types such as medical texts and descriptive grammars; investigations of character interactions in works of fiction; and so on. Below, we describe four of the earliest pilot projects started with the involvement of Swe-Clarin, which have produced some tangible outputs (notably publications), and also in some cases have led to further funding proposals. These projects illustrate various ways in which language tools of the type developed out of LT research can contribute to HSS research. A common thread in these projects is the aspiration to meet the challenge of conducting research on the basis of very large amounts of textual data in a consistent way without losing sight of the individual instances making up the mass of data, i.e., to be able to move between Moretti's (2013) *distant* and *close reading* modes.

Attitudes toward rhetoric over time

Background and purpose

In this pilot project, a historian of rhetoric at Uppsala University, together with the Swe-Clarin center Språkbanken at the University of Gothenburg, explored how Språkbanken's corpus infrastructure Korp (<https://spraakbanken.gu.se/korp/>) could be applied to the following research question: How have the attitudes to rhetoric expressed in Swedish public discourse changed over the last 200 years? The focus of the pilot project was on a large corpus (almost 1 billion words) of digitized historical newspapers provided by the National Library, but some preliminary studies of modern social media were also included for comparison. Below, we provide a brief summary of the project and its results. More details are given in Viklund & Borin (2016).

In the field of rhetorical history, questions about common opinions and everyday beliefs (doxa) are paramount but difficult to investigate, especially over long periods of time. It has been a general contention that the concept of rhetoric lost its central position in cultural history during the 18th century (Vickers, 1988; Johannesson, 2005), only to emerge again in later times, e.g. as a philosophical concept in the mid 20th century (Bender & Wellbery, 1990) and as a (frequently pejorative) buzzword in contemporary public discourse. Research has normally focused on selected theoretical treatises and examples of rhetorical practices; but, by employing mass digitization of relevant sources and big data methodologies, we are now able to describe some of those transformations with greater accuracy and on the basis of much larger corpora. The aim of the project has been to initiate a study of the conceptual history of rhetoric in public discourse, focused not on major treatises but on how people have generally used the concept of rhetoric from 1800 until the present day.

The principal research question concerned the transformation of attitudes towards rhetoric expressed in Swedish public discourse over the last 200 years. Which words have been used when referring to the concept of rhetoric, and which conceptual frames have been evoked over time? What kind of topical transformations are related to the concept of rhetoric? Alongside these questions we also wanted to investigate the usefulness and accuracy of the Korp search tool for this and other similar studies of rhetoric.

Methods and procedures

The pilot project used the Språkbanken corpus infrastructure Korp, in particular the collection of historical newspapers (late 18th to early 20th century) but also collections of present-day online

discourse. The chosen methodological approach was firstly to chart the historical contours in the material synchronically and diachronically, and then to study the gradual changes more closely through representative examples. As demonstrated below, Korp's "keyword in context" (KWIC; see Figure 6) feature was used, as well as the "trend graph" (Figures 1 and 2) and a "word picture" function (Figure 3). Although originally devised for the purposes of linguistic analysis of texts, the "word picture" can be used as a kind of abstract topical map that guides you to closer readings of the corpus.

Preliminary findings

One type of results concerned word usage over time and how those trends suggested historical change. One part of the investigation focused on terms related to rhetorical performance used in the newspaper material. Figure 1 shows a visualization of the usage of the words *talare* 'speaker' and *tal* 'speech', of which there are two conspicuous trends in usage: the nine peaks between the 1830s and 1860s; and the steady increase over the last decades of the period.

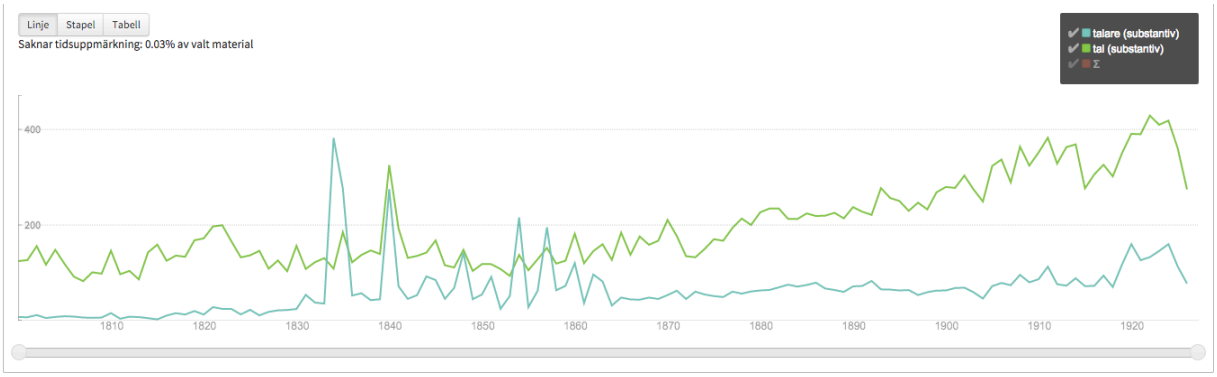

Figure 1. *Talare* 'speaker' and *tal* 'speech' 1800–1920s

The increase in usage confirms what we know about the growing interest in agitation and public debate at this time in history (e.g., Josephson, 1991; Mral, 1993). The nine peaks are interesting since they so clearly coincide with the triennial sessions of the old Riksdag (parliament) of the Estate, and the pattern disappears after 1866 when the old Riksdag was dissolved.

As has already been mentioned, simple word searches can point to cultural changes that have not been fully explored. Part of the pilot project was to investigate which kinds of rhetorical interest have characterized different periods. One hypothesis (Viklund, 2013) was that during the period 1880–1920 there was a notable rise not only in interest in rhetorical performance in general but also of declamatory practices in particular, which seemed to be supported by the data (see Figure 2).

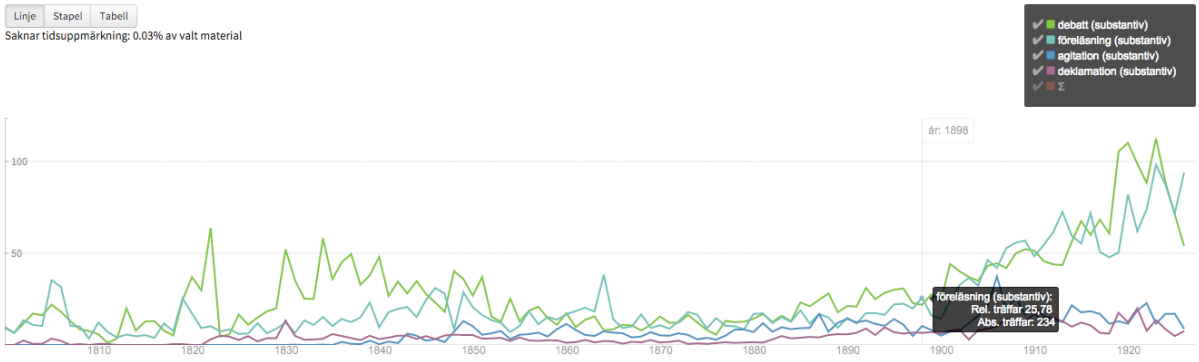

Figure 2. Terms associated with rhetorical performance (*debatt* 'debate', *föreläsning* 'lecture', *agitation* 'agitation', *deklamation* 'declamation'), 1800–1920s

Topical transformations and metaphorical conceptualizations

The "word picture" function was used in order to map different segments of the period 1800–1920s, and focus was placed on the most representative modifiers, generally adjectives (see Figure 3). In this way the most common topics were found, and so a preliminary view of some important features of contemporary attitudes towards rhetoric:

1. One common reference was **genre**: e.g. 'parliamentary', 'spiritual', 'political', 'Roman', 'academic', 'marital'
2. One described the ideal character of being **truthful** or **natural**: 'real', 'true', 'right', etc.; 'manly', 'natural', 'artless', 'unpretentious', 'simple', etc.
3. One category of modifiers contained two different **conceptual metaphors**: 'fiery', 'glowing', 'burning' (*eloquence is fire*); 'flowing', 'fluent', 'pouring' (*eloquence is water*)

The finding that two conceptual metaphors seemed to dominate the material prompted a more systematic search for similar expressions in the context of the word *vältalighet* 'eloquence'. In this way, a larger number of metaphorical expressions in the two categories was found and studied more closely in order to understand the dominant attitudes towards eloquence.

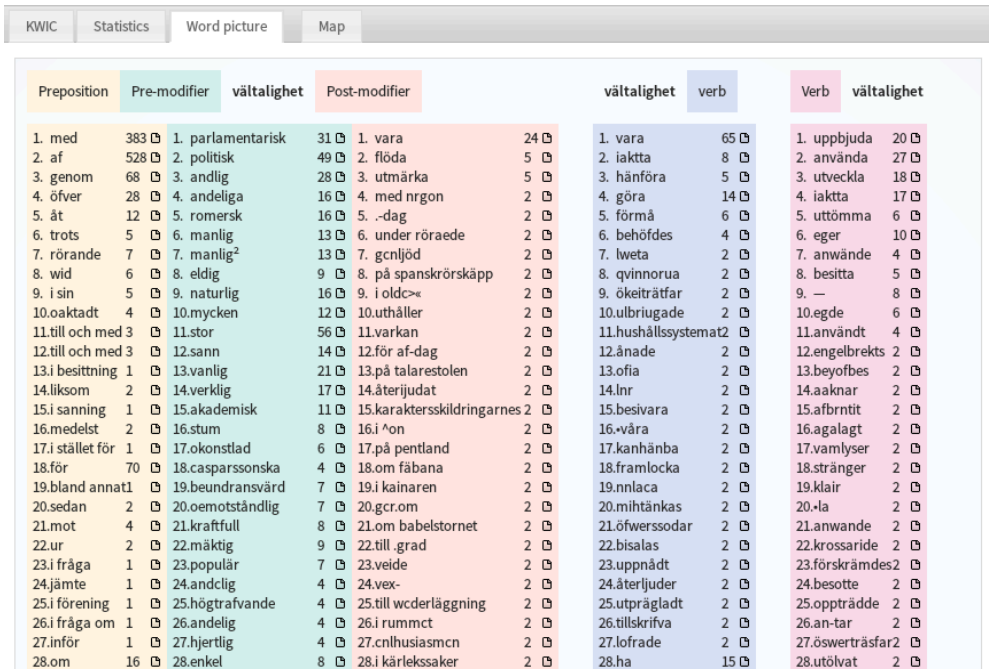

Figure 3. Korp word picture: significant syntactic neighbors of *vältalighet* ‘eloquence’, 1800–1920s

The concept *eloquence is fire* generally expresses positive values associated with rhetorical performances such as great pathos and energy. Often the *eloquence is fire* concept expresses the genius of the speaker, and so becomes a trope of transcendence. Consequently metaphors build on *eloquence as a force of nature*: they are not only natural but also overpowering, overwhelming the senses.

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

One frequently expressed idea was that the skill of the speaker must seem natural and fluent, and, on the other hand, the art of rhetoric must not be made apparent. Another image schema was concerned with the pattern *from one container to another*. The stream of eloquence flows ideally from one heart to another, and likewise the effects of rhetoric – a fire – can set the listener’s mind on fire, e.g.:

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'

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.'

A preliminary comparison with present-day debates in newspapers and digital media suggests that the use of *retorik* 'rhetoric' and *vältalighet* 'eloquence' has changed both in conceptual framing and quantity. In the last decades, *retorik* is now much more frequent than *vältalighet*. *Vältalighet* is not often used metaphorically and most often refers to historical contexts. *Retorik* is used with mainly abstract attributes, which, moreover, are negatively framed. So the rhetoric in question is usually negative e.g. 'bad', 'unfair' and 'bigoted', or else it refers to political groups in a negatively framed context: socialist, nazi, right wing, feminist, etc.

Conclusion

The pilot project outcomes should only be considered preliminary results, and there is still work to be done before the research questions can be answered satisfactorily. However, the study demonstrated the usefulness of Korp as a tool for getting an overview of large textual collections and diverse periods, and it was evident that the statistical analyses, the visualizations and the "word picture" were helpful for the understanding of how certain topics and attitudes changed over time. From the perspective of research in the Humanities, there is certainly potential for developing more advanced search methods in Korp. For example, one improvement would be to add more 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.

A text analysis toolbox for learner language

Background

The Swe-Clarin center at Uppsala University has developed SWEGRAM, a web service (<http://stp.lingfil.uu.se/swegram/>) that provides automatic linguistic annotation of Swedish texts at word and sentence level, which can subsequently be used to derive statistics on different linguistic characteristics of the texts: for example, the number of words and sentences in a text, the average length of a word, the distribution of word classes or different measures of

readability. SWEGRAM can be used for the annotation and analysis of any text(s) in Swedish (Näsman et al., 2017). SWEGRAM is intended to be easy to use by anyone who is interested in text analysis and to enable researchers in the Humanities and social sciences to conduct large quantitative studies on text-related features. SWEGRAM is freely available and can be used on-line. The user can upload one or several texts, annotate them and send them for further statistical analysis. The results are made available to users in the form of a downloadable plain text file or shown directly on the web page. Below, we describe the tools involved and give an example of a corpus on learner language, which we created by using SWEGRAM.

Automatic linguistic annotation

The linguistic annotation of SWEGRAM consists of standard tools for the automatic processing of Swedish texts. The user can upload one or several texts in various formats (doc, docx, odt, rtf or txt) and a preprocessing tool transforms them into a text file encoded in unicode (utf-8). The texts are tokenized to separate the words from punctuation marks and segment the sentences. Misspelled words are automatically corrected by a normalizer. The corrected text is then annotated with a part-of-speech (PoS) tagger to analyze the words with their main PoS and morphological features along with lemmatization to find the base form of each word. We use two types of PoS annotation. One is based on the universal PoS tagset (Nivre et al., 2016), which consists of 17 main part-of-speech categories, and the other tagset consists of the Stockholm-Umeå Corpus tagset (Capková & Hartmann, 2016), which contains 23 main PoS categories with their morphological features. Lastly, syntactic parsing is applied to describe the syntactic structure of the sentences using universal dependencies (Nivre et al., 2016). The tools are connected in a pipeline for easy and fast processing. For the annotation, we use automatic, freely available tools developed for Swedish within the computational linguistic community. The pipeline of the annotation is shown in Figure 4. Any module can be deactivated, which enables users to exclude some part of the annotation if they wish and use their own annotation instead.

Figure 4. Annotation pipeline.

The annotation format is designed so the data can be easily examined and manually corrected. We use the CoNLL-U format (Nivre et al, 2016), which is tab-separated with one token per line and has various linguistic fields represented in various columns. An example of the annotation of the sentence *Den kalla vinden slåg mot mina kinder.* 'The cold wind hit my cheeks.' is shown in

Figure 5 where the first two columns represent the sentence (TEXT ID) and token ID, and the original tokens as written by the author are given in column three (FORM). Misspelled and corrected words are listed in column four (NORM) followed by the lemma in column five. The universal PoS tag and SUC tag are given in column six and seven, respectively with morphological information as given by the SUC tagset (C-FEATS) and universal PoS tagset (U-FEATS). The last two columns represent the syntactic structure in terms of universal dependencies.

TEXT ID	ID	FORM	NORM	LEMMA	U-POS	CPOS	C-FEATS	U-FEATS	HEAD	DEPREL	Translation
2.2	1	Den		den	DET	DT	UTR|SIN|DEF	Definite=Def|Gender=Com|Number=Sing	3	det	The
2.2	2	kalla		kall	ADJ	JJ	POS|UTR/NEU|SIN|DEF|NOM	Case=Nom|Definite=Def|Degree=Pos|Number=Sing	3	amod	cold
2.2	3	vinden		vind	NOUN	NN	UTR|SIN|DEF|NOM	Case=Nom|Definite=Def|Gender=Com|Number=Sing	4	nsubj	wind
2.2	4	slåg	slog	slå	VERB	VB	PRT|AKT	Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act	0	root	hit
2.2	5	mot		mot	ADP	PP	_	_	7	case	[against]
2.2	6	mina		min	DET	PS	UTR/NEU|PLU|DEF	Definite=Def|Number=Plur|Poss=Yes	7	nmod:poss	my
2.2	7	kinder		kind	NOUN	NN	UTR|PLU|IND|NOM	Case=Nom|Definite=Ind|Gender=Com|Number=Plur	4	nmod	cheeks
2.2	8	.		.	PUNCT	MAD	_	_	4	punct	.

Figure 5. Example of a sentence annotation: ‘The cold wind hit my cheeks.’

Automatic linguistic analysis

Users can upload one or several annotated texts described above for further quantitative analysis of various textual and linguistic features. The statistical calculations include general statistics about the texts and their linguistic characteristics, including the number and the average length of texts, sentences, words and tokens in all uploaded files and separately for each file. Statistics are also calculated for all PoS or for specific ones decided by the user. Readability measures such as LIX, OVIX and the nominal ratio are also calculated. The tool provides the user with frequency lists for all texts and for individual texts based on lemmas or tokens, with or without delimiters. The frequency lists can be sorted based on frequencies or words (lemmas or tokens) in alphabetical order. The frequency lists can also be limited to specific parts of speech. The user can also search for words with a certain length above, below or at a specific threshold value. In addition, spelling errors can be listed and sorted by frequency, for all uploaded texts and for individual texts.

Users can also view the uploaded texts and perform different types of searches in the annotated text. This includes searching for words, lemmas and PoS tags that either start with, end with, contain or exactly match a user-defined search query. The results are then printed and sorted according to what texts they appear in.

Finally, users can specify whether the output should be delivered as a downloadable tab-separated data file, which can be imported into other programs such as Excel for further analysis, or shown directly in the browser.

Creation of a learner language corpus

SWEGRAM can be used for the automatic creation of linguistically annotated corpora. In collaboration with researchers at the Department of Scandinavian Languages at Uppsala University, SWEGRAM has been made the basis for a web-based tool for annotation and quantitative analysis of student essays. Using SWEGRAM, we created the Uppsala Corpus of Student Writings (Megyesi et al., 2016), consisting of 2,500 texts with over 1.5 million tokens, produced as part of a national test in Swedish and Swedish as a second language. The students whose writings have been included in the corpus range in age from nine (in year three of primary school) to nineteen (the last year of upper secondary school). Parts of the texts have been annotated on several linguistic levels using SWEGRAM. The corpus is intended to be a monitor corpus and is currently under development.

The development of SWEGRAM goes hand in hand with the development of learner language corpora, which form an important empirical basis for research aiming to build LT applications capable of improving the normalization of erroneous text sequences not only on the word level but on the sentence level, in order to correct errors such as morphological disagreement and incorrect word order.

The annotated Strindberg corpus

Background

The Swe-Clarin center at the Department of Linguistics at Stockholm University is constructing a linguistically annotated corpus of the *National Edition of August Strindberg's Collected Works*, which were published as 72 printed volumes 1981–2012. Strindberg's texts comprise about 9.6 million tokens or 20,000 printed pages; in addition, the volumes include a large amount of editorial commentaries and word explanations, as well as an index. The work described here is a collaboration with the Swedish Literature Bank (Litteraturbanken) at the University of Gothenburg and the editorial team of the National Edition of Strindberg's Collected Works at Stockholm University. The aim of this infrastructural work is to increase the application of corpus-linguistic methods to literary science in general and to Strindberg's texts in particular. Literary science is a field where such methods have been relatively rare compared to other disciplines with primary data in the form of natural language such as linguistics, history or social sciences (Balossi, 2014, p. xiii). In contrast, the related area of stylometric studies (capturing the style of a particular

author based on quantitative criteria) has long been using computational methods (Oakes, 2009). This work is a continuation of a previous project on a smaller scale which generated a corpus of Strindberg's autobiographical works, the Stockholm University Strindberg Corpus (SUSC; Björkenstam, Gustafson Capcová & Wirén, 2014).

The development of the corpus is planned in three versions (one unannotated and two annotated), corresponding to the needs of different types of researchers:

1. A raw-text version without annotation, with a simple structure for representing basic textual units such as chapters, paragraphs and headings using interspersed blank lines. This version is intended for researchers who want to work with the raw text only, for example, by using their own scripts or an on-line corpus annotation and lexical analysis tool such as Sketch Engine (Kilgarriff et al., 2014).
2. A CoNLL version with one word per line and linguistic annotation distributed across columns. This version is intended for researchers who want to work with the annotated corpus without going through a search interface (as in item 3 below). Furthermore, it is straightforward to import this format to a spreadsheet format such as Excel, for researchers who want to compute statistics or add their own annotation in that way.
3. An XML version with an XML schema that encodes the structure and annotation of the text. This version is meant to be bundled with an independent search engine such as the IMS Open Corpus Workbench 5 (CWB) or to be integrated with the Korp and/or Strix infrastructures at Språkbanken. It is primarily intended for literary researchers who want to work with the annotated corpus through a search interface or concordancer.

A crucial decision is how to optimize the quality of the automatically produced linguistic annotation of the corpus, whose major components will be parts of speech and syntactic structure. On the one hand, this task is facilitated by the language of the National Edition having been modernized in several respects. A detailed account of the principles behind this modernization is available in the *Presentation of the National Edition* (1990), and will only be briefly described here. Thus, to the extent that Strindberg used older spellings that were standard before the spelling reform in 1906, these are modernized in accordance with *The Swedish Academy Glossary* (Svenska Akademiens ordlista, SAOL) for the benefit of present-day readers. For example, the old form *af* is spelled *av* 'of' in the modernized form, and similarly *hvilka – vilka* 'which', *qvinna – kvinna* 'woman', *jern – järn* 'iron', etc.

On the other hand, there are many exceptions to the modernization of the text: First, Strindberg's spelling is retained whenever it is considered to reflect a spoken register or to characterize the speakers, and in general when it is judged to have a stylistic function. Examples of this are *intressangt* 'interesting', *dalkarar* 'men from Dalecarlia', *dronning* 'queen', *körka* 'church' and *reståranger* 'the restaurant'. Secondly, Strindberg's inflections are kept unchanged. Thus, plural endings of verbs, like *voro* 'were', *gingo* 'walked' and *kommo* 'came', which generally did not go out of fashion until the 1940s, are retained. Similarly, rare forms carrying stylistic function, like *huven* 'heads', *soplårn* 'the garbage bin', *instrumenter* 'instruments', *fästmänner* 'fiancés' and *lyftade* 'lifted' are kept.

Beyond the word level, the main problem for automatic annotation is Strindberg's unorthodox use of punctuation and abbreviations, and the resulting difficulties in segmenting the texts into sentences. Another consideration for segmentation is the different conventions exhibited in the different genres of Strindberg's work, such as novels, plays and poetry. Yet another challenge is his use of passages in foreign languages, e.g. Danish and French.

The word-level idiosyncrasies of Strindberg's language, mapping archaic or stylistically motivated forms to their standard, present-day counterparts, could be dealt with using a preprocessing step called normalization, that is, mapping the word forms to their present-day correspondences. Pettersson (2016) has developed and compared several such methods for Swedish historical texts. Preliminary experiments with Pettersson's tools for normalization applied to a subset of the National Edition have given promising results in the sense that annotation of the output texts becomes more accurate than that of the original text. Above the word level, we shall specify methods for proper segmentation of the texts into sentences, and possibly apply language identification to recognize passages in foreign languages.

When the text is suitably normalized, we shall apply a linguistic analysis chain such as efselab (Östling, 2016) for the purpose of producing an annotated version of the corpus, including part-of-speech tagging, named entity recognition and dependency parsing. An additional step is then needed to map the annotation back to the original text. Since the original and normalized texts can be regarded as a parallel corpus, we can use alignment to keep track of the links between tokens (Tiedemann, 2011). To the extent that normalization only changes word forms and punctuation, however, this problem is much simpler than the general alignment problem, which typically involves material in different languages.

Towards interactive visualization of public discourse in time and space

Background, research question and data set

The Swe-Clarin center Språkbanken at the University of Gothenburg has worked together with political scientists at their own university to develop digital research tools for investigating political discourse in social media.

Public discourse has been characterized as being "among the most remarkable inventions of the early 19th century" (Nordmark, 2001, p. 42 [our translation]). It has been repeatedly transformed over its long history: technologies have evolved, new media have appeared, and participation has become increasingly inclusive. The most recent manifestations of public discourse are the various social media that have emerged only over the last decade or so, complementing or perhaps even supplanting traditional print and broadcast media as the main arena of public discourse and opinion formation, involving many more citizens in a much more interactive mode than ever before.

However, there are many questions about public discourse as conducted in social media, questions about the demography and representativeness of participation, whether the issues are the same as in traditional media, and whether public opinion formation processes have become fundamentally different as a result. The design of e-science tools for HSS research was investigated in order to enable HSS researchers to work with massive amounts of richly annotated textual data, e.g. those resulting from mining and processing social media such as Twitter.

The data used for the work summarized in this paper consist of Swedish tweets collected from Twitter's public streaming API during a narrow time window around two televised Swedish party leader debates in October 2013 and May 2014, before the national elections in September 2014, a total of almost 340,000 tweets. The study focused on six topics, which had been preselected for the televised debates, and their respective share of the total programming time was manually estimated by simply counting the number of minutes of broadcast time of each topic. The topics were considered to stand for the *left – right (LR)* and *green/ alternative/ libertarian – traditional/ authoritarian/ nationalist (GAL-TAN)* dimensions of political issues. The topics identified within these dimensions were *labor market, healthcare, education* (LR) and *climate, refugees/immigration, crime* (GAL-TAN).

The study of political discourse analysis in social media performed by the team of political scientists, in collaboration with Swe-Clarin experts, was conducted with the use of a mix of

manual and automatic methods. The tweets were classified into the six topics using basic information retrieval techniques, viz. index terms and tf–idf (term frequency – inverse document frequency). Tweets could be assigned to more than one topic. This was used to determine the distribution of topics over tweets with the ultimate goal of testing the hypothesis that GAL-TAN issues should be more salient in social-media discourse than in traditional television, thought to flow from the different demographics of these media. The resulting publication is under review for a journal and cannot be referenced here in the interest of preserving anonymity.

Data preparation

The work presented here constitutes a further development of the methodology of the earlier study, making deeper use of language technology and also a sophisticated spatiotemporal visualization platform. For a more detailed description of the work discussed here, see Borin & Kosiński (2016).

Our point of departure has been the index term list prepared for the earlier study. An automatic morphological analyser was first used for processing this list, in order to reduce multiple inflected forms of the same lexical entry present in the list to the entry itself. A substantial 26% decrease of the average number of index terms per topic was achieved by this measure, together with a manual reduction of compounds missing from the lexicon but analysed by the morphological analyser to a common prefix or suffix. e.g., *flyktingstatus* ‘refugee status‘, *flyktingsmuggling* ‘refugee smuggling’ and *flyktingorganisation* ‘refugee organization’ could all be subsumed under a classification criterion requiring simply that the term exhibit the compound first member *flykting..nn.1* ‘refugee (n)’ (see Figure 6). The reduced index list covers many more text word types than the original list, of course, for the reason described below.

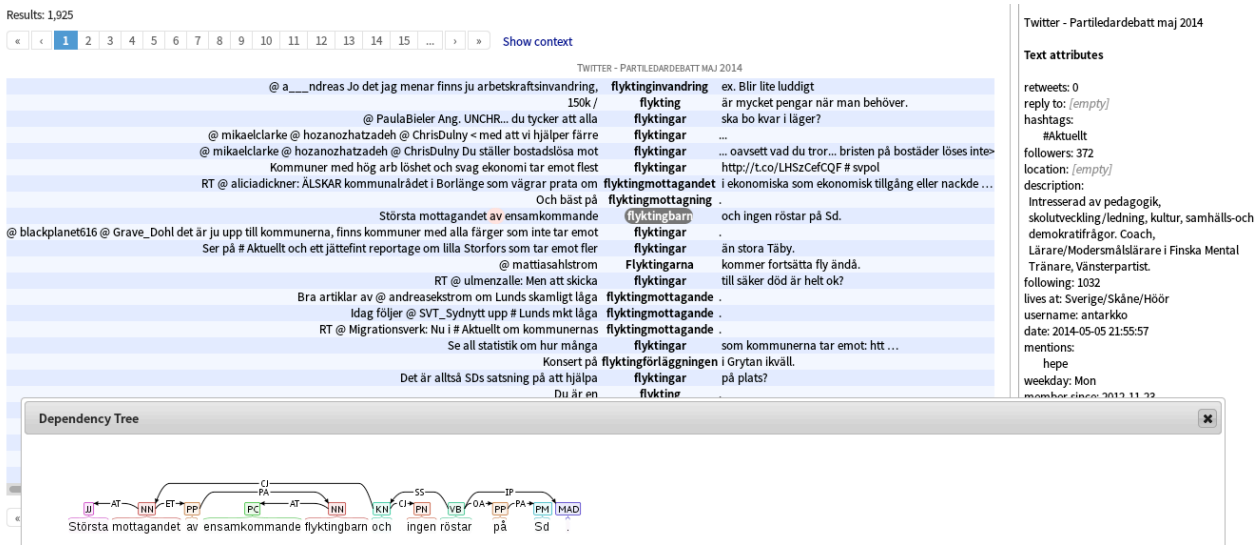

Figure 6. Korp KWIC view and dependency tree

The text of the tweets was provided with linguistic annotation layers, using the standard Korp import pipeline, and these annotations could then be used for matching against the index terms. The following criteria were employed: (c1) precise text-to-word match (the single criterion used in the previous study); (c2) a lexical entry match; (c3) a compound prefix and suffix match; (c4) a compound prefix only match; and (c5) a compound suffix only match. The criteria have been prioritized in the order described. Multiple topic assignment was allowed, as in the original study. These criteria ensure that topic matching will capture all relevant lexical units (modulo homography/polysemy), even when they are embedded in a compound. This arguably improves classification. In fact, our classification results are slightly different from those of the earlier study. Notably, the two most common topics – *labor market* and *education* – switch places. This deserves further study, which however falls outside the scope of this presentation.

Interactive visualization of big textual data

Human cognitive capacity has difficulty in handling large amounts of quickly changing data. This is one of the reasons why information visualization is a growing and vibrant field. Also, traditional methods of text analysis fail when confronted with big, textual data. Very large textual datasets need to be pre-processed in order to be presented to the user in a meaningful way. In our case, LT tools have been employed to provide greater abstraction, forming the basis for a meaningful overview of the data. Interactive exploration of the pre-processed data may require another kind of pre-processing so that response times are kept to a minimum. For our purposes, we required a visual analytics tool fulfilling the following desiderata:

(f1) Open-source licensing (to allow for unconstrained publication of the results);

(f2) high-dimensional data, real-time and interactive visualization support;

(f3) pixel-oriented technique support (partly as a consequence of (f2)) (Keim, 2000);

(f4) temporal dimension support (to allow for, for example, historical data analysis);

(f5) spatial, customizable dimension support (to allow for spatial data analysis);

(f6) support for additional, customizable dimensions.

A mature platform offering the specified characteristics along with the required extensibility is *Nanocubes* (Lins et al., 2013). It was used in the project for providing an interactive visualization of the Twitter data in temporal, spatial and 8 separate customized dimensions. It was also extended in order to provide the capability to inspect the original data, underlying the presented, aggregated results of pre-processed datasets. See Figure 7.

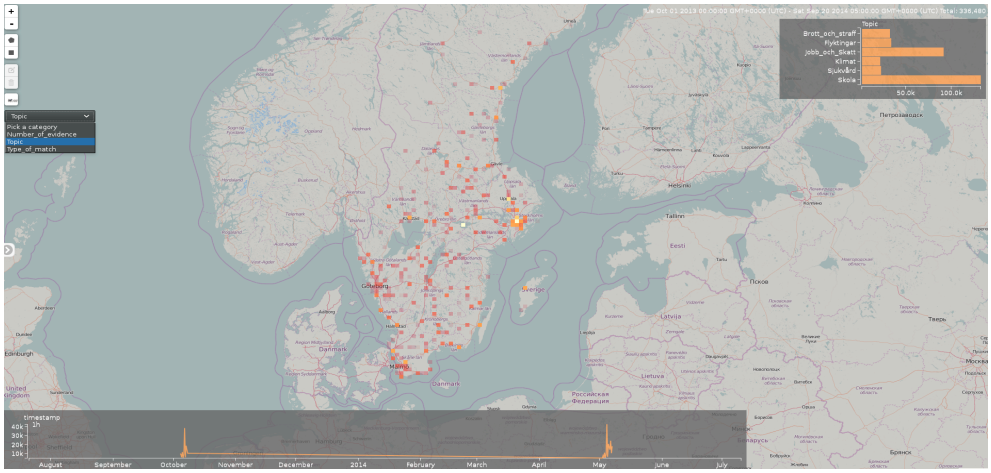

Figure 7. Swedish political topics on Twitter visualized with Nanocubes

Visual browsing history is tracked by the engine, i.e. selected regions and category ranges or sections are saved for the user locally. Once the user selects the “dive in” option, all of the above are used to specify a query and present the subset of the source tweets corresponding to the selections performed. For instance, it is easy to investigate if there are differences in the political discussion topic profiles between arbitrary geographical regions such as the north and south of the country (see Figure 8). For the purposes of the current study, three custom dimensions of the data were added to the spatial and temporal dimensions: *Topic*, *Type of match* (corresponding to the LT analysis criteria) and *Strength of evidence* (a score of matching words).

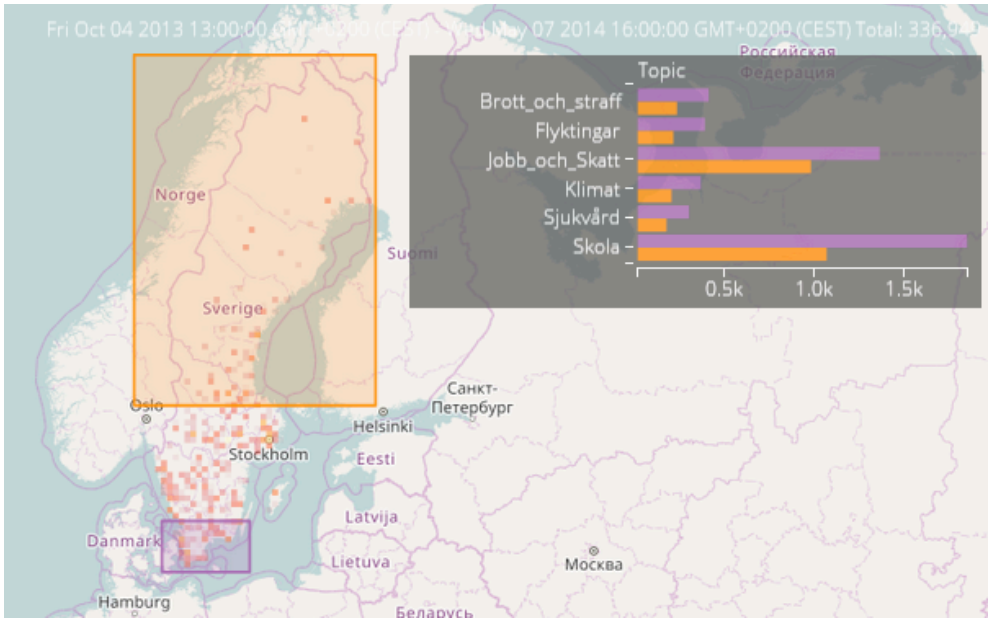

Figure 8. Swedish North–South differences in Twitter discussion topic distribution

Discussion and future work

What this work did not address was source data deficiencies (from the point of view of the LT tools used to process it). One aspect of this is the language of social media in general, and perhaps tweets in particular. A full-sized modern Swedish lexical resource is used for the lexical analysis of the tweet text, viz. SALDO (Borin et al., 2013), with approximately 140,000 entries, covering Modern Swedish inflected forms to the order of two million, but almost exclusively correctly spelled standard forms. We can feel confident that we find all or most of the relevant *correct* words for making the topic classification, especially with the compound analysis, but we do not know how many relevant words we miss because they are misspelled or otherwise deviate from the standard. Another aspect of data inaccuracy is that the location precision related to the tweets was insufficient, where only 17% of tweets offered geographical coordinates in an explicit way and 18% more was achieved by matching the location field metadata against a Swedish place name gazetteer, for a total of 35% of the tweets being provided with a geographical location. While this comprises some 120,000 tweets, it still constitutes a smaller share of the total, and it would be desirable to increase this share somehow. The matching of location field items against the gazetteer could be improved, and we also plan to move in the direction of exploring textual clues for geolocation (see, e.g., Berggren et al., 2015).

On the plus side – and this is a considerable advantage – the interactive visualization, especially when compared to the static graphs presented in the previous study, has made it possible to

analyse the source, big textual data in real time and with respect to multiple dimensions, which can be selected from the set of all dimensions offered by the dataset.

Summary and outlook

We hope to have shown with the four brief Swe-Clarin pilot project presentations above, that Swe-Clarin, and CLARIN ERIC, can make substantial contributions to digital humanities. Some interesting solutions and results have emerged from the pilot projects, but they also point to a number of areas in clear need of more development. In particular, in several of the pilot projects and also in the workshops and user events organized by Swe-Clarin, we have repeatedly seen a need for document-level access to the text materials and, as a consequence of this, work has now been initiated in Swe-Clarin to meet this need, the most prominent example being Språkbanken's recently launched Strix project (<https://spraakbanken.gu.se/strix/>). More generally, the intention is that the infrastructure itself (the LT tools and resources made available to HSS researchers) should be developed to meet the further requirements exposed as a consequence of the experiences of the pilot projects and other collaborations. We believe that such work must be conducted in the same spirit as the pilot projects, in collaboration between Swe-Clarin's LT experts, language resource owners and HSS scholars with complex research questions, and a willingness to take on the methodological challenges of big-data language-based digital humanities.

Thus, we strongly encourage you to contact us if you are interested in any of our resources, in conducting a pilot study with us, or if you have any ideas or questions regarding digital humanities research with respect to language technology and resources: <info@sweclarin.se>. See also <https://sweclarin.se>.

REFERENCES

Balossi, G. (2014). *A corpus linguistic approach to literary language and characterization: Virginia Woolf's The Waves*. Amsterdam: John Benjamins Publishing Company.

Bender, J. & Wellbery, D. E. (1990). Rhetoricality: On the modernist return of rhetoric. In J. Bender & D. E. Wellbery (Eds), *The ends of rhetoric: History, theory, practice* (pp. 3–39). Stanford CA: Stanford University Press.

Berggren, M., Karlgren, J., Östling, R., & Parkvall, M. 2015. Inferring the location of authors from words in their texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics* (pp. 211–218). Linköping: LiU EP.

Björkenstam, K. N., Gustafson Capková, S., & Wirén, M. (2014). The Stockholm University Strindberg Corpus: Content and possibilities. In R. Lysell (Ed.), *Strindberg on international stages/Strindberg in translation* (pp. 21–40). Cambridge: Cambridge Scholars Publishing.

Borin, L., Forsberg, M., & Lönngren, L. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47, 4, 1191–1211.

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012* (pp. 474–478). Istanbul: ELRA.

Gustafson-Capková, S. & Hartmann, B. (2006). *Documentation of the Stockholm - Umeå Corpus*. Stockholm University: Department of Linguistics.

Megyesi, B., Näsman, J., & Palmér, A. (2016). The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. In *Proceedings of LREC 2016* (pp. 3192–3199). Portorož: ELRA.

Johannesson, K. (2005). *Svensk retorik. Från medeltiden till våra dagar*. Stockholm: Norstedts.

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

Keim, D. A. (2000). Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6, 1, 59–78.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, 1, 7–36.

Lins, L., Klosowski, J. T., & Scheidegger, C. 2013. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19, 12, 2456–2465

Megyesi, B., Näsman, J., & Palmér, A. (2016). The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. In *Proceedings of LREC 2016* (pp. 3192–3199). Portorož: ELRA.

Moretti, F. (2013). *Distant reading*. London: Verso.

Mral, B. (1993). *Kommunikation och handlande i Malmö kvinnliga diskussionsklubb 1900–1904*. Number 6 in Arbetarrörelsen och språket. Uppsala: Avdelningen för retorik, Uppsala universitet.

Näsman, J., Megyesi, B., & Palmér, A. (2017). SWEGRAM – A web-based tool for automatic annotation and analysis of Swedish texts. In *Proceedings of the 21st Nordic Conference of Computational Linguistics* (pp. 132–141). Linköping: LiU EP.

Nivre, J., de Marneffe M-Ch., Ginter, F, Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., & Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC 2016* (pp. 1659–1666). Portorož: ELRA. See <http://universaldependencies.org>.

Nordmark, D. (2001). Liberalernas segertåg (1830–1858). In K.-E. Gustafsson & P. Rydén (Eds), *Den svenska pressens historia. II: Åren då allting hände (1830–1897)* (pp. 18–125). Stockholm: Ekerlids förlag.

Oakes, M. (2009). Corpus linguistics and stylometry. In *Corpus linguistics: An international handbook* (vol. 2, pp. 1070–1090). Berlin: De Gruyter Mouton.

Östling, R. (2016). Efficient sequence labeling. <https://github.com/robertostling/efselab>.

Pettersson, E. (2016). Spelling normalisation and linguistic analysis of historical text for information extraction. Ph.D. thesis. Uppsala: Uppsala University, Department of Linguistics and Philology.

Presentation av Nationalupplagan och redogörelse för redigeringsprinciperna (1990). In *Nationalupplagan av August Strindbergs samlade verk, volym 1, Ungdomsdramer I* (pp. 275–333). Stockholm: Norstedts. <http://www.strind.su.se/preskomp.htm>.

Tiedemann, J. (2011). *Bitext alignment*. San Rafael, CA, USA: Morgan & Claypool.

Viklund, J. (2013). Performance in an age of democratization: The rhetorical citizen and the transformation of elocutionary manuals in Sweden ca. 1840–1920. Paper presented at the ISHR (International Society for the History of Rhetoric) biennial conference in Chicago.

Viklund, J., & Borin, L. (2016). How can big data help us study rhetorical history? In *Selected Papers from the CLARIN Annual Conference 2015* (pp. 79–93). Linköping: LiU EP.

ACKNOWLEDGEMENTS

The work described here was supported by an infrastructure operation grant from the Swedish Research Council to *Swe-Clarin* (contract no. 2013-02003). The work on interactive visualization of public discourse has also received support in the form of a framework grant from the Swedish Research Council to the *Knowledge-based culturomics* project (contract no. 2012-5738). We would also like to thank the anonymous reviewers for their constructive critique of our exposition, which has hopefully allowed us to improve it considerably.