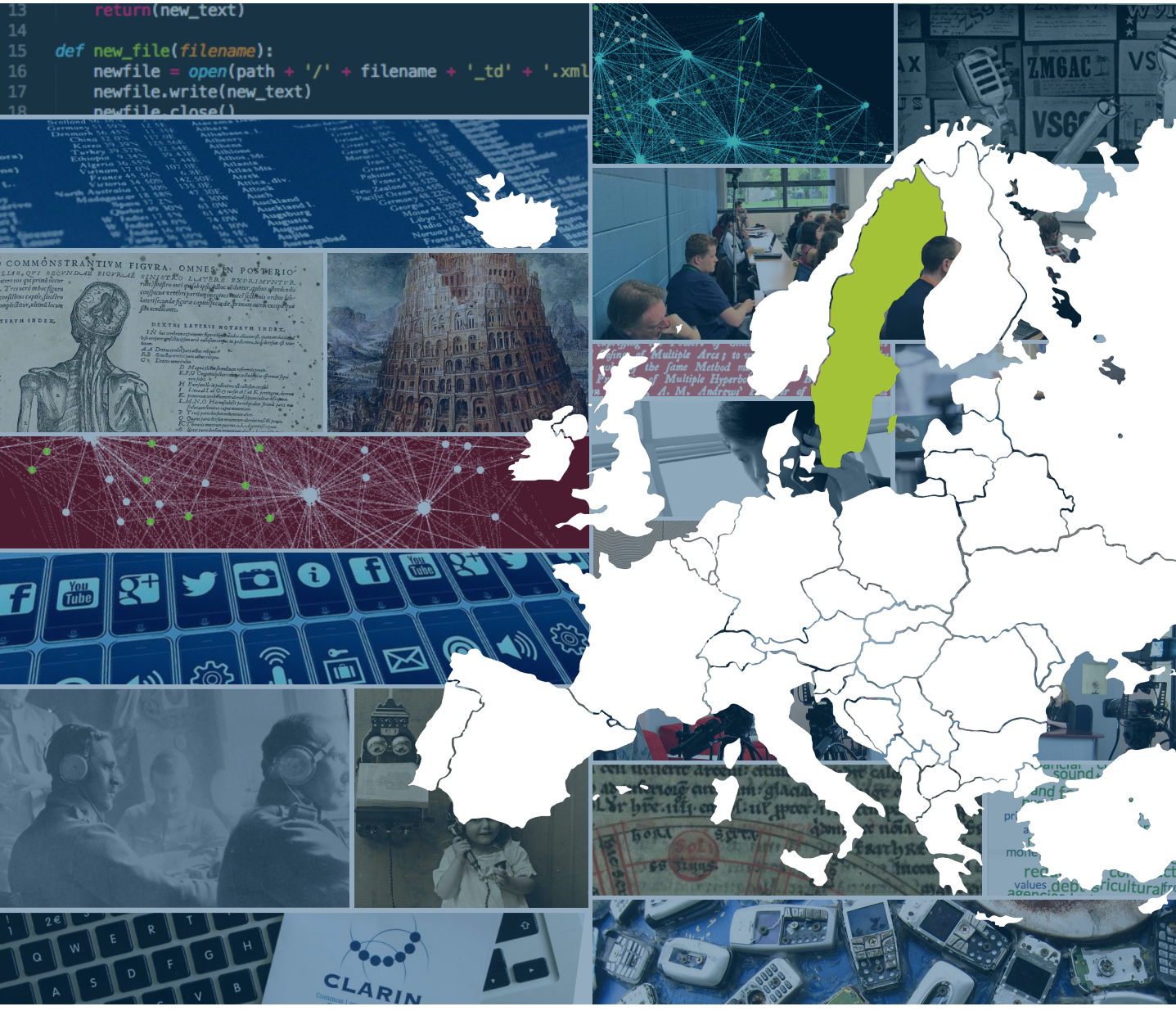
CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

Sweden



Written and edited by Darja Fišer and Jakob Lenardič

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN national consortia with the aim to increase the visibility of CLARIN consortia, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

This brochure presents Sweden and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports on a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium's infrastructure in their research

Stockholm, Sweden | photo by Víctor Vázquez | Unsplash

Sweden

Written by Darja Fišer and Jakob Lenardič

The Swedish consortium Swe-Clarin,[1] which has been a member of CLARIN ERIC since 2014, is a collaboration between the national archive Riksarkivet, the Swedish National Data Service, the Swedish language council Språkrådet, the KTH School of Computer Science and Communication and a variety of language technology research units at five universities – the Department of Computer and Information Science at Linköping University, the Humanities Lab at Lund University, Språkbanken (the Swedish Language Bank) at the University of Gothenburg, the Department of Linguistics at Stockholm University, and the Computational Linguistics Group at Uppsala University. The national coordinator of Swe-Clarinis Lars Borin, professor of natural language processing at the University of Gothenburg and co-director of Språkbanken.

The coordinating centre of Swe-Clarin is Språkbanken (the Swedish Language Bank), a major national and international research centre on computational approaches to language in Sweden, established already in the 1970s, which provides researchers with access to language resources, including an extremely wide range of Swedish texts, as well as state-of-the-art computational tools for the processing, compilation and linguistic analysis of corpora. The rapidly-increasing number of corpora, which are in the majority of cases available for download in standard formats, not only comprise a comprehensive collection of contemporary Swedish texts representing a wide variety of formal and informal discourse produced both in Sweden and Finland, where Swedish is an official language, but also include historical texts from most periods of written Swedish.

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

Related to the corpora are the tools of Språkbanken; for instance, the corpus infrastructure Korp, used for accessing both the above mentioned corpora, the Finnish corpora made available by FIN-CLARIN in the Language Bank of Finland, the Saami corpora provided by the Norwegian CLARIN Giellatekno node in Tromsø, and Estonian corpora available through the Estonian CLARIN ERIC centre, or the annotation tool Sparv, presenting a web-based interface to the Korp annotation toolchain, offering part-of-speech tagging, compound analysis, word sense disambiguation, named entity recognition and dependency parsing of Swedish text. Additionally, Språkbanken researchers are working in a great number of research projects – one such endeavour is the research program "Towards a knowledge-based culturomics", among whose goals is "to advance the state of the art in language technology resources and methods for semantic processing of Swedish text, in order to provide researchers and others with more sophisticated tools for working with the information contained in large volumes of digitised text, by, for instance, being able to correlate and compare the content of texts and text passages on a large scale".

Swe-Clarin has also organised successful user involvement events. One such event was the Second National Swe-Clarin workshop held in connection with the Swedish Language Technology Conference in November 2016, where two invited presentations were given, one on the text mining project BiographyNet and the other on the impact language technology has on scholarship in the humanities, followed by a poster session featuring Swedish CLARIN-supported research.

Lars Borin, National Coordinator of Swe-Clarin.

Korp

Written by Darja Fišer and Jakob Lenardič

A concordancer is one of the key tools of a language resource provider, as it serves as the main entry point to language in context. One of the best known and widely used concordancers is that provided by Swe-Clarin's Korp.[2] A versatile and user-friendly tool, it is the main corpus infrastructure of Språkbanken and is used extensively by the Swedish and Finnish consortia, as well as in an Estonian and a Norwegian CLARIN centre. Through Korp, researchers can access some of the consortia's most important language resources, such as Swe-Clarin's Riksdagsen öppna data corpus (see page 8) and FIN-CLARIN's Suomi24 corpus.

Korp has been developed by a team of about eight people at Språkbanken at the University of Gothenburg and consists of three components:

- the Korp corpus pipeline, which is used for the import, annotation and export of corpora;
- the Korp backend, which consists of a series of web services used for searching and retrieving both the corpora and their associated annotations and metadata; and
- the Korp frontend, which is the graphical user interface communicating with the backend.

The exhaustive corpus collection of Språkbanken, which is accessed through Korp, consists of over 400 corpora with more than 13 billion tokens and almost one billion sentences representing mainly modern written Swedish, but also the older language, going back all the way to the Old Swedish of the Middle Ages.

Through the Korp corpus pipeline, researchers can import and annotate their own data. A pivotal characteristic of the pipeline is its dynamic nature, which allows researchers to integrate their existing annotations into the Korp infrastructure and use it as the basis for other types of annotation. The pipeline also provides researchers with a series of automatic annotation options – tokenisation, sentence splitting, links to the lexical persistent identifiers, lemmatisation, compound analysis, PoS/MSD tagging, and syntactic dependency parsing.

The Korp frontend is a graphical search interface, and thus the aspect of the corpus that researchers usually first come in contact with. The Korp frontend gives users the flexibility to search through the corpora by giving them the option to use simple queries or the CQP-corpus query language. After performing a search, users can then find the concordances under the KWIC tab (Figure 1), which also brings up a sidebar on the right-hand side that shows how the relevant token is annotated. Other functions of Korp include the ordbild (the word picture) tab (Figure 2), which shows the most relevant syntactic collocates of a lemma or text word; the related words tab, where a list of semantically-related lemmas is given; and the statistics tab, which provides users with a statistical overview of the token, as a table with a row for every unique hit and a column for every selected corpus, or in the form of a graph showing frequency of one or more linguistic phenomena over time (Figure 3). The Korp backend, which provides access to corpora, their annotations and their metadata, can be downloaded here, while most of the corpora that can be searched through Korp are available for download in Språkbanken.

[2] https://spraakbanken.gu.se/swe/node/1535

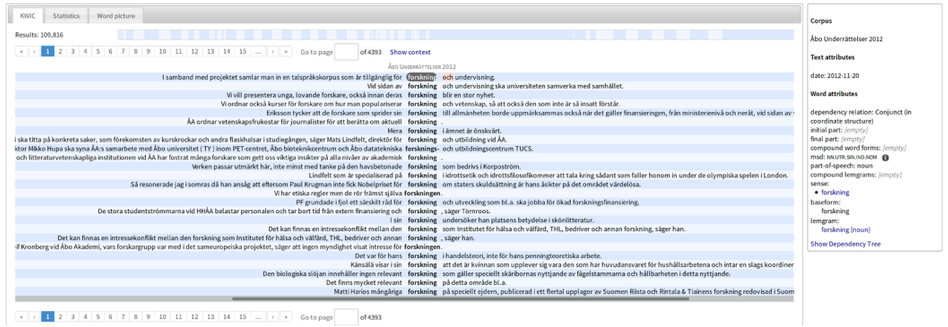


Figure 1: Concordances for the lemma "forskning" (“research”).

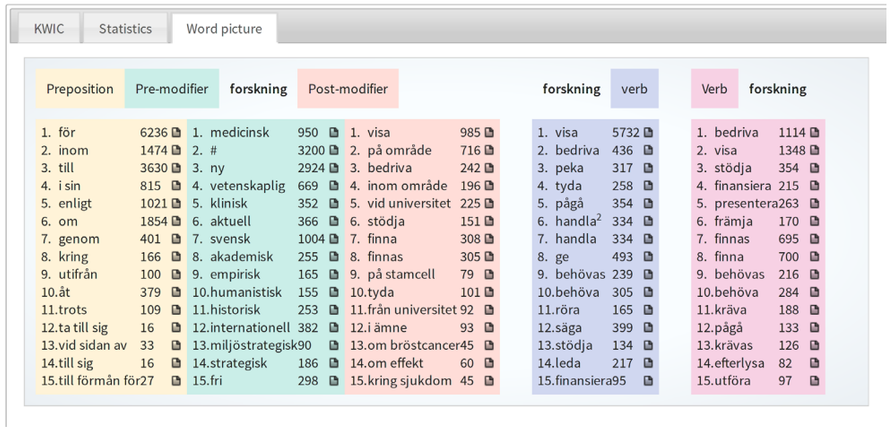


Figure 2: The word image for the lemma "forskning" (“research”).

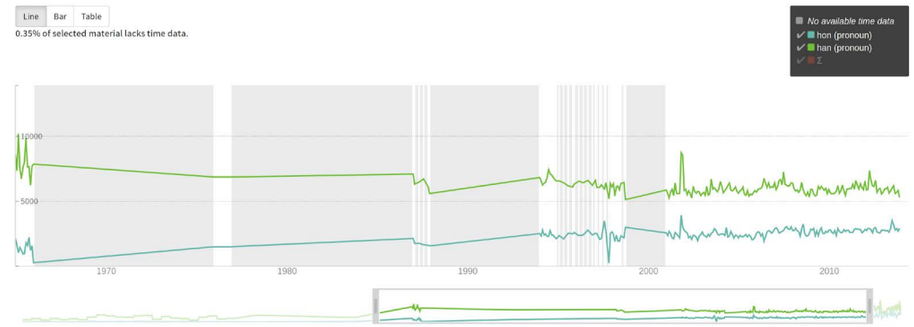


Figure 3: The trend diagram for the personal pronouns "hon" (“she”) and "han" (“he”) in the modern newspaper corpus.

The Riksdag's Open Data Corpus

Written by Darja Fišer and Jakob Lenardič

Since parliamentary speech has a great societal impact on account of its language and content, the creation and availability of large parliamentary multimodal corpora—a topic that was the subject of a CLARIN-PLUS workshop in 2017—play a pivotal role in humanitarian and social research.

The Riksdag's open data[3] is one such corpus. It is the digitised collection of Swedish parliamentary data and consists of roughly 30,000 documents pertaining to Sweden's national political decision processes. It has been made available for download on the website of the Swedish parliament. In addition, the Swedish National Library has digitised and published the public reports of inquiry for the period between 1922 and 1999 under the CC0 licence on the parliamentary website, with newer reports now being digitised from the very outset.

This parliamentary corpus is available in Korp and consists of 1.25 billion tokens. It can also be downloaded in the XML format from the resource page of Språkbanken. The annotation was performed with the Swe-Clarin's tool Sparv and consisted of tokenization, lemmatization, as well as lemgram (inflectional paradigm) and word sense identification, and compound splitting.

The resource has been successfully used by scholars working in the Social Sciences and Digital Humanities. Fredrik Norén from the Department of Culture and Media Studies at Umeå University has researched how social information in Sweden was structured in the period between 1965 and 1975, with a focus on uncovering how the government informed its citizens and communicated with them during this period. He has used Korp to search through SOU, a subset of the parliamentary corpus that contains the official reports of the government.

Norén has also collaborated with Roger Mähler from the Centre of Digital Humanities at Umeå University to analyse the changes in governmental discourse on the basis of the distribution of nouns. Using topic modelling, they were able to identify how information discourse arose in the 1960s and infiltrated governmental policies. Norén and Pelle Snickars from Umeå University have also used similar methods to analyse policies related to Swedish film in the 20th century on the basis of 4,500 reports in the SOU corpus. All in all, digitised language data like the Riksdag's open data corpus have made it possible to study the evolution of concepts like information in great detail, and by extension, they unveil historical change in a more precise and nuanced manner than ever before.

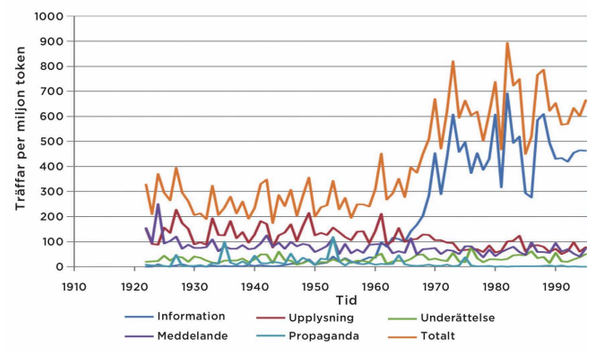


Figure 4: Frequencies of the lemmas "information", "upplysning" ("information"), "underrättelse" ("notification"), "meddelande" ("message") and "propaganda" throughout the 20th century.

[3] https://data.riksdagen.se/in-english/

Tutorial and Workshop on Automatic Sentence Selection from Corpora

Written by Darja Fišer and Jakob Lenardič

One of the most valuable aspects of an international research infrastructure such as CLARIN ERIC is the knowledge sharing that occurs among the national consortia. A successful example of this is the tutorial[4] and workshop[5] on automatic sentence selection for dictionary construction. The event, which was organised by Ildikó Pilán and Elena Volodina from Språkbanken, took place at the University of Gothenburg from 26 May to 1 June 2017 and brought together researchers from the Swedish, Estonian and Slovenian consortia.

The aim of the tutorial was to give an introduction to corpus data processing with Python and machine-learning approaches for lexicography, as well as offer opportunity for practical hands-on sessions with scikit learn and WEKA.

At the workshop, Ildikó Pilán described the HitEx extraction system, which is being developed at Språkbanken and is tailored to the automatic identification of corpus sentences for the exercises aimed at learners of Swedish as a second language. Adapted to various language-proficiency levels on the basis of the CEFR criteria, HitEx is a powerful system that allows for dynamic machine-assisted learning as it provides teaching professionals, lexicographers and students with options to set their own parameters, such as the difficulty level of the words they wish to learn. Iztok Kosem presented how the automatic extraction of corpus data has been successfully implemented in Slovene lexicography. He introduced the Collocations Dictionary of Slovene project, which has just released the first corpus-based dictionary of collocations for Slovene. Kristina Koppel presented the on-going work on compiling the Estonian Collocations Dictionary, which is in development will primarily be aimed at learners of Estonian at the B2-C1 levels.

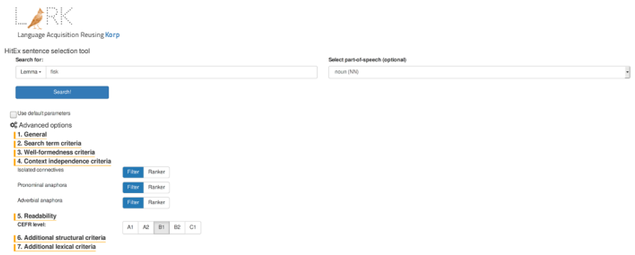


Figure 5: The HitEx user interface for sentence selection with advanced search options.

Results

Rank	Score	Sentence
No sentences matched your criteria. See below for sentences with violations.		

Results with violations

Rank	Score	Sentence
1	-1	När vi äter fet **fisk** får vi i oss fetter som är bra för kroppen .
2	-1	Han fick för sig att människor med munnen öppen såg ut som **fiskar** .
3	-1	Det finns risk att **fisken** försvinner från havet .
4	-1	Marocko tillåter att EU **fiskar** utanför Västsahara .

Different CEFR level: B2

Contains proper names: Marocko, EU, Västsahara

Typicality: 0

Figure 6: Corpus example sentence selection results for "fisk" ("fish") at B1 (intermediate) level.

[4] https://sweclarin.se/swe/sentence-selection-tutorial

[5] https://sweclarin.se/swe/mini-workshop-collocations-and-sentence-selection

Maria Ågren

Maria Ågren is a professor of history working at Uppsala University, Sweden. She leads the Gender and Work (GaW) project in which she has collaborated with Swe-Clarin researchers to create the GaW database, a collection of annotated historical language data that reveal the ways men and women supported themselves in the early modern history of Sweden. The interview was conducted by e-mail correspondence by Jakob Lenardič and edited by Darja Fišer.

1. Could you please briefly describe your background and tell us what your recent research is about?

I received my first degree in history and Swedish. I also have a diploma for teaching these two subjects in upper-secondary school; however, I never worked as a teacher because I enrolled as a graduate student in history instead. Since 2001, I have been a professor of history and my most recent research has been the Gender and Work (GaW)[6] project that I am leading.

2. How did you get involved with Swe-Clarin and what impact has this collaboration had on your research?

In the Gender and Work project, we are interested in finding snippets of information about people's jobs in historical documents, such as farm accounts, diaries, and court protocols. These snippets usually take the form: mend boat, sell eggs, take care of old people, and so on. At an early stage of the project, we told a linguist about our interest in building a database in which this type of information could be stored. She then exclaimed: "Aha! You are interested in verbs!" This comment had two far-reaching effects for our research project. First, we realised that we should call our method verb-oriented because it is a short and efficient way of explaining our approach that everyone immediately understands. Second, this linguist encouraged us to contact Professor Joakim Nivre from Swe-Clarin, which has led to a fruitful collaboration.

3. Which tools and corpora have you used and how did you integrate them into your existing research?

I did not use any existing corpora. Instead, the project has built its own corpus, the GaW database. Project members have gathered and classified thousands of fragments of information from a variety of handwritten historical sources that describe the ways people sustained and provided for themselves. The first stage of the project, which ran between 2010 and 2014 focused on the historical period from 1550 to 1800. The project now continues (from 2017 to 2021) with a focus on the period between 1720 and 1880. The GaW database is accessible to researchers, students, and the general public.

4. Have corpus data helped you reveal any interesting societal and linguistic trends of the periods you are interested in that would have been more difficult to uncover were it not for corpus-based methodology?

Yes, if one accepts my claim that the GaW database is a form of corpus then its data have been absolutely vital to the project. I would even say that most of our results could not have been achieved without it. Likewise, if one accepts that the verb-oriented method is a corpus-based methodology, then the answer is most definitely "yes". We have made many interesting discoveries about early modern society.

[6] http://gaw.hist.uu.se/what-is-gaw/research+project/

5. Could you describe the project in more detail? How did the language differ from contemporary Swedish? Are there any interesting differences from a socio-historical point of view? In what way have gendered roles/expectations changed from that time until now?

Gender and Work is a combined research and digitisation project at the Department of History at Uppsala University. The aim of the project is to acquire knowledge about the work of both men and women in the past. With the project we have been able to show the importance of the two-supporter model in early modern society; that is, that there was an expectation and practical reality of both men and women contributing to the household's survival. The project has also shown that what people did for a living in the past had more to do with marital status than with gender. The difference between what married and unmarried people did for a living was larger than that between what men and women did for a living.

One could say that early modern gender roles were more similar to the ones we have today – that is, both spouses worked, both spouses were expected to take care of children, even if the mother was thought to have a somewhat larger responsibility in this respect, people worked long days and could have to travel far to earn a living – than the ones that developed within the nineteenth-century bourgeoisie.

The Swedish language at the time was of course quite different from modern Swedish. There were no spelling rules, for instance, which makes for varied and, one might say, unorthodox spelling practices. It happened that German words were used in Swedish sentences. For researchers who use the corpus, the language itself is not the largest problem, since all scholars involved in the project are historians specialising in the early modern period and they are therefore all accustomed to reading early modern Swedish. The handwriting, on the other hand, is more of a problem; sometimes, the handwriting is so bad that you simply cannot make out what the text is about.

6. Has your field in general embraced the available digital text collections and language technologies? Do Swedish historians make use of language technology or collaborate with research infrastructures such as Swe-Clarin?

I think the answer to this question must be "no". In my opinion the Gender and Work project has been a pioneer within the historical disciplines in Sweden, especially because the collaboration with researchers working with language technology has allowed us to overcome a variety of technical difficulties we faced when dealing with the historical documents.

7. Could you elaborate on these methodological and technical challenges that a researcher working with historical text collections faces with respect to the available infrastructure?

There are two large problems: (1) the inconsistent spelling and (2) the fact that a majority of documents are only available in the original, handwritten form. The former problem is less daunting. In fact, there has been a highly successful collaboration with Professor Joakim Nivre and his now former PhD student Eva Pettersson from Swe-Clarin, which has led to substantial progress in overcoming the inconsistencies in spelling. For more in-depth information regarding this, see Eva Pettersson's doctoral dissertation "Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction".

The latter problem is much more difficult to overcome. In the early modern period, state bureaucracies swelled and this led to a big increase in the production of documents, all of which are valuable to historians. But most of these sources are only available in handwritten form; rarely have they been printed (in which case they can be OCR-read) or digitised directly. If they are not available in digitised form, they cannot be processed automatically. If there were an easy way of transforming these handwritten documents to digital texts, then the corpus of early modern text material would grow enormously.

Since this is not the case, collaborative interdisciplinary projects like the one between Nivre and Pettersson on the one hand, and GaW on the other hand, are very rare. In our case, the historians read and annotated the texts manually, but at the same time also digitised them. This provided Pettersson with the language material on which she could train her normalisation tool. This tool identifies verbs, and particularly verbs describing work activity. The tool is not yet developed to perfection, but hopefully it will one day be possible to run it on digitised texts from the early modern period, and in this way speed up the processing of historical texts.

If you are interested in our approach to the extraction of information from historical texts, I suggest that you check the paper "HistSearch – Implementation and evaluation of a web-based tool for automatic information extraction from historical text" by Eva Pettersson, Jonas Lindström, Benny Jacobsson and Rosemarie Fiebranz.

8. What's your vision for CLARIN 10 years from now? What in your opinion should CLARIN focus on providing?

That it will be a permanent collection of resources and will contain more text corpora from the period between the Middle Ages and ca. 1800. This is the period during which many more documents were produced than in the Middle Ages, and most of them were not printed. After around 1800, handwriting became more similar to that we see today, and more documents were written on typewriters, making them easier to process automatically. The period from 1500 to 1800, on the other hand, is the period that is still largely unsupported in terms of corpora and text processing tools.

Stockholm, Sweden | photo by Davids Kokainis | Unsplash

COLOPHON

This brochure is part of the 'Tour de CLARIN' volume I (publication number: CLARIN-CE-2018-1341, November 2018).

Coordinated by
Darja Fišer, Jakob Lenardič and Karolina Badzmierowska

Written by
Darja Fišer and Jakob Lenardič

Edited by
Darja Fišer and Jakob Lenardič

Proofread by
Paul Steed

Designed by
Karolina Badzmierowska

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2018-1341
November 2018

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands
www.clarin.eu

CLARIN
Common Language Resources and Technology Infrastructure

CLARIN 2018