Swe-Clarin: Language Resources and Technology for Digital Humanities

Lars Borin[1] **Nina Tahmasebi**[1] **Elena Volodina**[1] **Stefan Ekman**[2] **Caspar Jordan**[2]

[1]Språkbanken, Department of Swedish
University of Gothenburg, Sweden
firstname.lastname@svenska.gu.se
sam@sweclarin.se, info@sweclarin.se

[2]Swedish National Data Service
University of Gothenburg, Sweden
firstname.lastname@snd.gu.se

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. The academic members are well balanced over 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, University of Gothenburg.

From the start, Swe-Clarin has aimed to establish good relations to the HSS fields and open the door for all researchers who wish to work with DH research using text and speech as primary research data. 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 small 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.

The pilot projects aim to spread the word of 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, having this development being informed by input from users in the earliest possible stages of the project. Some pilot projects are already underway (see below).

In addition to the pilot projects, we have arranged workshops and user days and published newsletters and a blog. The workshops held so far have been on topics such as: general introduction to Swe-Clarin, our tools and resources; historical resources and tools; making cultural heritage text data available for research; and HSS research on digitized speech data, such as those of the Swedish Media Archive. We have started a series of workshops called *Swe-Clarin on tour* where Språkbanken's widely used Korp corpus infrastructure (Borin et al. 2012) is used to explore previously unexplored materials in a hands-on manner, giving researchers of LT and HSS the opportunity to meet and discuss research questions and the potentials of using LT for DH. The experience from working with HSS researchers will help reveal the limitations of existing tools and hopefully also engender general methodological discussion, thus setting the stage for future development of tools more appropriate for DH research. The first such workshop was held at Stockholm University in the spring of 2016. It featured the ethnographic questionnaires collected by the Nordic Museum since the late 1920s and now digitized by them, and it was attended mainly by ethnologists. The next workshop in the series will be held in Umeå in conjunction with the Swedish Language Technology Conference in November 2016. There the material in focus will be 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.

SOME SWE-CLARIN PILOT PROJECTS

Attitudes Toward Rhetoric Over Time

In this pilot project, a historian of rhetoric at Uppsala University together with the Swe-Clarin center Språkbanken explored how Språkbanken's Korp infrastructure could be applied to the research question of how the attitudes to rhetoric expressed in Swedish public discourse have changed over the last 200 years. The focus in the pilot project was on a large (almost 1 billion words) digitized historical newspaper material provided by the National Library, but some preliminary studies of modern social media were also included for comparison. (Viklund and Borin 2016)

A Text Analysis Toolbox for Learner Language

The Swe-Clarin center at Uppsala University has developed SWEGRAM, a web service that provides automatic linguistic annotation 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. In a 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 for the national exam in Swedish and Swedish as a second language for different grades (3rd, 6th, 9th grade). (Megyesi et al. 2016)

The Annotated Strindberg Corpus

The Swe-Clarin center at Stockholm University in collaboration with the Swedish Literature Bank (Litteraturbanken) and the editorial team of the National Edition of August Strindberg's Collected Works aim to construct a linguistically annotated corpus of Strindberg's collected works. The National Edition consists of 72 volumes with about 6 million words published between 1981 and 2012. The annotated version of the corpus will enable new kinds of research to be conducted on this material, as well as pave the way for even deeper annotation in the future. (Nilsson Björkenstam et al. 2014)

LAST BUT NOT LEAST

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

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

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.

Nilsson Björkenstam, K., 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*. Cambridge: Cambridge Scholars Publishing.

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.