This project will bring forth tools for computationally detecting lexical and semantic changes as well as word sense induction for Swedish. These tools give us a chance to study language changes in their own right but also overcome hurdles with research on historical text. We will bring together historical linguistics with NLP and data science, a unique collaboration needed to study SC and LR in full.

We believe the results of the project will advance the field in several research areas, offer benefits for researchers in the DHSS and lower the threshold for the public to make use of our textual resources. These areas include but are not limited to:

- Reliable methods for detecting semantic and lexical change in both synchronic and diachronic data contexts; how do language changes spread, in which media do they appear firstly and what is the change rate? We will further the studies in e.g., Vejdemo and Hörberg (2016).
- We offer researchers in the DHSS a method to gather evidence and analyze text related to their concept of interest without being experts in diachronic or synchronic language change, thus reducing the threshold to target large-scale text mining and the risk of drawing wrongful conclusions based on word sense and sentiment changes.
- We open up the vast resources of our digital archives, like the cultural treasures of Språkbanken, to the public and nonprofessional users. These users wish to use the resources without investing considerable time to (1) find all words related to their search query and then, once they have found the texts they are interested in, (2) look up words to find if their meanings have changed.
- Many downstream NLP applications, such as semantic role labeling and word sense disambiguation would benefit from robust methods to detect lexical and semantic change. The resulting tools will feed back into the Korp processing pipeline making the results directly usable. We will openly release modern and historical (sense-differentiated) word vectors, similar to HistWords (Hamilton et al., 2016) for direct inspection and use in other NLP applications.

The plan

The application areas aim to highlight all kinds of change and apply to researchers as well as the public. The use cases will be integrated into the Korp pipeline (Borin et al., 2012) where possible, and in the Strix environment, a new interface under development that offers a close-reading capacity where texts, not sentences, and their temporal comparison are in focus.

Close reading (Simplify research) In this use case, we will help users (researchers and laymen) to firstly *find*, and secondly *understand* content in digital archives by (1) implementing features that suggest extension to search queries with relevant word replacements and their validity periods, and (2) in a text, highlight words that have changed their meaning. Changes will be accompanied with the original passages of text such that users can verify the results, or get new entry points into the corpus.

Distant reading (Quantify Research hypothesis) Many researchers are moving into DHSS (as seen by the increasing number of centers for Digital Humanities across Swedish universities), drawn in by the promise of large amounts of data and automatic methods for analyzing them. In this use case, we will collaborate with three research groups to help quantify their hypotheses.

Firstly, we will collaborate with Sarah Valdez at the Institute for Analytical Sociology in Norrköping, who works with analyzing differences in meaning for concepts in politics (e.g. democracy, freedom, immigration) for different political parties, this relates primarily to word sense induction on 20th century Swedish political party programs and election manifestos.

Secondly, we will collaborate with a group of concept historians led by Henrik Björck who are investigating the rate and spread of abstraction of the *market*, that goes from a concrete time and place to an abstract concept, like job or stock markets. We will study the interplay of these concept and investigate when *the market* becomes an agent that affects people rather than the other way around. Once the first abstract market has been established as a concept, does the process move faster and faster with new abstract markets and can we quantify this rate?

Thirdly, we will work with historical linguists led by Lena Rogström to show that scientific texts in the 18th century attributed human-like features to animals and plants, see e.g., Linné and Bjerkander. By applying semantic change detection to the Royal Science Academy texts, as well as to texts from other contemporary genres, we can help quantify the hypothesis, and find spread and change rate. This collaboration will take place after the digitization of the relevant texts, pending a funding application.

L. Borin, M. Forsberg, and J. Roxendal. Korp – the corpus infrastructure of Språkbanken. LREC, 2012.
W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *ACL*, 2016.
S. Vejdemo and T. Hörberg. Semantic Factors Predict the Rate of Lexical Replacement of Content Words. *PLOS ONE*, pages 1–15, Jan. 2016.