Lars Borin, Språkbanken Text, University of Gothenburg: lars.borin@svenska.gu.se
Eva Pettersson, Department of Linguistics and Philology, Uppsala University: eva.pettersson@lingfil.uu.se

Towards a Swedish Diachronic Corpus

Aim

To create a Swedish diachronic corpus, comparable to diachronic corpora for other languages, such as for example the Corpus of Historical American English (COHA), the Icelandic Parsed Historical Corpus (IcePaHC), the IMPACT-es corpus for Spanish, and the Diakorp corpus for Czech.

Time Plan

2019

1) Survey: existing resources and steps needed for corpus creation
2) Detailed plan of how to build the corpus

2020

1) Data collection
2) Formatting
3) Annotation
4) Corpus release (first version)

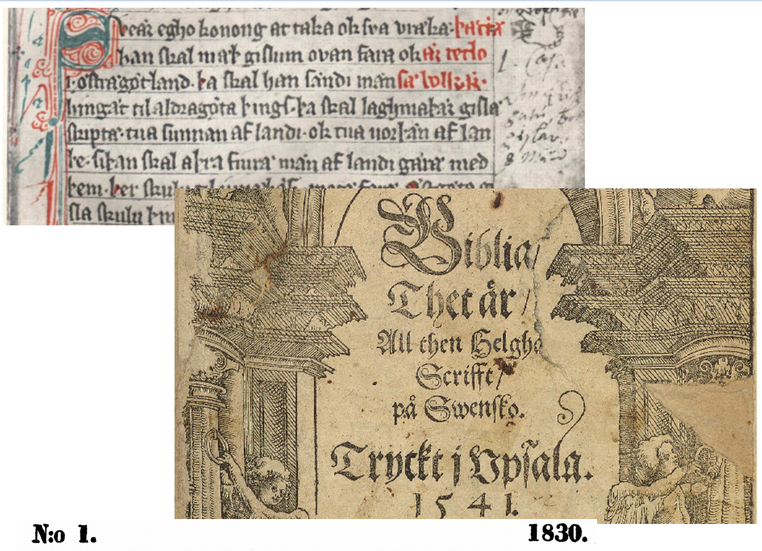
As a long-term goal, we also aim at providing NLP tools for processing text from different time periods, such as spelling modernisation tools, taggers, parsers etc.

(Some) Aspects to be considered

Time period to be covered
•From Old Swedish (~13th century) to present-day Swedish? Including Runic Swedish?
•Static corpus or monitor corpus?

Languages to be covered
•Swedish only, or languages written in Sweden during the time (Latin, German, French, Finnish, ...)?
•Texts produced in present-day Sweden, or only texts produced in regions belonging to Sweden at the time?

Corpus balance
•How do we get a balanced corpus?
•What text genres and amounts of text are available for different time periods?
•How much effort should and could we put into digitization of material for “missing” time periods?



Granularity
How fine-grained should the corpus be in terms of time periods? Subcorpora for:

- decades?
- 50-year periods?
- centuries?
- varying (dependent on period)?

Corpus size
How large a corpus should we aim for?

Format
What format should the corpus be stored in?

Metadata
•What metadata information should be included?
•What metadata standard should we use?

Annotation
•What kinds and levels of annotation are reasonable?
•What annotation standards should we use?

Target audiences
•Who do we see as our main target audiences?
  - (Historical) linguists
  - Other researchers in the humanities and social sciences
  - Computational linguists
•What needs and requirements are there for users in the targeted groups?

Accessibility
•We aim for a freely available corpus, for example via some form of Creative Commons licence
•Search interface for easy access to the contents
•Possibility to download corpora files