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

<table>
<thead>
<tr>
<th>2019</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Survey: existing resources and steps needed for corpus creation</td>
<td>1) Data collection</td>
</tr>
<tr>
<td>2) Detailed plan of how to build the corpus</td>
<td>2) Formatting</td>
</tr>
<tr>
<td></td>
<td>3) Annotation</td>
</tr>
<tr>
<td></td>
<td>4) Corpus release (first version)</td>
</tr>
</tbody>
</table>

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?

- **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?

- **Metadata**
  - What metadata information should be included?
  - What metadata standard 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