Diabase: Towards a diachronic BLARK in support of historical studies

Lars Borin, Markus Forsberg, Dimitrios Kokkinakis

Språkbanken • Centre for Language Technology
University of Gothenburg

LREC 2010, Valletta, 19th May, 2010
topics

1. text and speech as historical research data, and language technology
2. our work with LT for historical studies: two examples
3. methodological musings on BLARKs and language variation
LT in historical studies

In historical studies, text – and speech, i.e., language – are central as both primary and secondary research data sources.

In today’s world, the normal mode of access to text, speech, images and video is in digital form. Modern material is born digital and older material is being digitized on a vast scale in cultural heritage and digital library projects.

LT can help historians and other researchers make effective use of this flood of language data from all historical periods.
Språkbanken (the Swedish Language Bank)

In recognition of the groundbreaking corpus linguistic work initiated by Sture Allén at the University of Gothenburg in the 1980s (which had resulted in the creation of one of the first large electronic text corpora in another language than English, Press-66, one million words of news text), Språkbanken (the Swedish Language Bank) was instated in 1975 as a national center (and funded on the national level) with a result to collect, process and store (Swedish) text corpora, and to make linguistic data extracted from the corpora available to researchers and to the public. Today, Språkbanken possesses a unique combination of competences in the areas of Swedish text corpora, Swedish computational lexicons and language technology tools for the processing, annotation and presentation of text corpora.

Språkbanken tidbits

<table>
<thead>
<tr>
<th>word</th>
<th>description</th>
<th>translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>gränshes</td>
<td>beast</td>
<td>gras</td>
</tr>
</tbody>
</table>

corpus search:

Corpora

<table>
<thead>
<tr>
<th>Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAROLE/SUC</td>
</tr>
<tr>
<td>Two annotated corpora</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corpus collection</th>
</tr>
</thead>
</table>

Genetic interface

| ORDAT |
| Yearbook of “Svenska Dagbladet” 1929 - 1998 |

Domain specific corpus

<table>
<thead>
<tr>
<th>The Journal of the Swedish Medical Association’s Archives (1968-2009) processed and searchable</th>
</tr>
</thead>
</table>

Lexical resources

<table>
<thead>
<tr>
<th>Lexical resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALDO</td>
</tr>
<tr>
<td>Lexical resource for Swedish language technology</td>
</tr>
</tbody>
</table>

| Söderwall och Schister |
| Dictionary of Old Swedish |

| Online ordbok |
| Dictionary of 19th century Swedish |

| Svenska ordb., LEXIN |
| Basic Swedish dictionary |

| SAOCH |
| The Swedish Academy Dictionary |

more...
Segletsen inomskärs mellan Strömstad och Göteborg utmärker sig genom ett egensinne och en nyckfullhet i lynn, som mer än en gång under färden pröfvar icke blott passagerarnes, utan även de vårdige vedskutspatronernas lugna tålmod och bekanta trygghet.

Der är ett odrägligt kryssande fram och åter: än motvind och storm, än god vind, men antingen med så hastigt påkommende stiltje, att den goda vinden tjenat till intet, eller ock med så starka orkaner och sjögång, att skutan darrande brakar i sina fogningar och far lik en spån upp och ned på de vreda böljorna. Och alla dessa särskilda fall smygas så tätt och lömskt hvarandra i spåren, att hela resan närmast kan jämföras med ett smuggling-äventyr, der man endast genom tusen beräkningar och svårigheter slingrar sig fram.
<table>
<thead>
<tr>
<th>Författare</th>
<th>Titel</th>
<th>Passage</th>
</tr>
</thead>
</table>
semantic search in 19th c. fiction

CONPLISIT – the components:

- **SALDO** – a modern semantic lexicon with inflectional morphology (≈73,000 senses)
- **Dalin** – a large 19th century lexicon (≈63,000 lemmas)
- an orthographic mapping database SALDO–Dalin
- a morphology for regular 19th c. open parts of speech
- **Litteraturbanken** (≈100 19th c. novels)
- a research question: can 19th c. fiction throw light on the emergence of consumer society in Sweden?
Dalin–SALDO round trip
some issues encountered

- the 1906 spelling reform – in reality lasting some decades around 1900
- large synchronic orthographic variation before the 19th century – and again today!
- slightly different inflectional morphologies 19th–20th century – gradual abandonment of verb-subject agreement during the first half of the 20th century until it officially went out of use around 1950
- very different inflectional system (and syntax) in the Old Swedish period
- changes in vocabulary
- changes in word meanings
methodological musings: what’s in a BLARK?

- linguistically annotated text corpora
- speech databases
- tools for basic text and speech processing
- basic lexical resources
- tools for linguistic annotation of text (POS taggers, chunkers, parsers)
- text-to-speech and speech-to-text systems

... in interoperable, standardized formats
BLARKs are static, . . .

As it is normally conceived of and presented, the BLARK assumes a modern standard language variety as the object of description, at least as far as the written language part of the BLARK is concerned, which is the part that we are competent to make judgements about. Part of the reason for this is certainly historical: The BLARK has been – and continues to be – informed more than anything else by language technology work on modern stable written standard languages.
Modern linguistics increasingly recognizes variation as a fundamental and essential characteristic of human language. In this regard, the study of history through textual primary sources makes up an interesting and challenging testbed, where the robustness and the generality of existing language technology are subjected to the acid test of messy and multilingual reality, more so than in many other application areas, since we have to deal with, *inter alia*, historical, non-standardized language varieties in addition to a number of modern standard languages.
linguistic variation

Language varies (at least),

- by community (languages, dialects, sociolects)
- by subject, purpose or medium (topics, genres)
- by time (historical language stages)

The BLARK attempts to abstract away from all three; it can be thought of as reflecting a modern standard language, which is topic- and genre-neutral.
Our work described above can be seen as the first steps towards the development of Diabase, a Swedish BLARK extended along the diachronic – or time – axis. This work also raises some interesting methodological questions about the description of variation in the linguistic resources making up the BLARK:

- fundamental/norm(al) ∼ deviation (e.g., ‘correct’ vs. ‘incorrect’ verb agreement in a novel published in 1935; correct vs. incorrect spelling in an internet page in 2010)
- one system ∼ a mix of systems (e.g., pre- vs. post-1906 spelling)
thank you for listening!