Parsing the Past – Identification of Verb Constructions in Historical Text

Eva Pettersson, Beáta Megyesi and Joakim Nivre
Department of Linguistics and Philology
Uppsala University
• Historical texts constitute a rich source of data
• Lack of NLP tools for handling historical text
  – Time-consuming to manually search large amounts of text
  – Likely that language technology could substantially reduce the manual effort
• General aim: NLP of historical text
• Specific aim: Extraction of verbs and their complements from historical Swedish text, using contemporary NLP tools
Gender and Work Project
- What men and women did for a living in the Early Modern Swedish Society (1550-1800)
- Most often expressed by a verb and its complements

Now requested Anderss permission to deliver him Grain again

Olof brought wood to Stockholm to be sold
Outline

• Background
• System Overview
• Experiments and Results
• Conclusion and Future Work
Difficulties Related to Historical Text

- Different time periods
- Orthographic differences
- Lack of spelling conventions
- Morphological differences (richer verb paradigms)
- Syntactic differences (word order)
Using NLP tools for analysing historical text:

1. Using NLP tools as they are
   - Pennachiotti and Zanzotto (2008), Italian

2. Training new NLP tools on annotated historical corpora
   - Sánchez-Marco et al (2011), Spanish

3. Using NLP tools combined with normalisation and/or dictionaries covering historical language variation
   - Rocio et al (1999), Portuguese
   - Oravecz et al (2010), Hungarian
   - Pettersson and Nivre (2011), Swedish
Complement Extraction
– System Overview
Normalisation

- **Input**: Tokenised historical text
- **Two sets of handwritten rules**:
  - Rules inspired by the reformed Swedish spelling (1906)
    - /j?f/v/  
      fördärfvat, blijfva → fördärvat, bliva
    - /($vowel)$1/$1/  
      sööka → söka
  - Rules based on 17th century court records
    - /u($vowel)/v$1/  
      tuinga → tvinga
    - /sch/sk/  
      schall → skall
- **Output**: Text normalised to a more modern spelling
PoS Tagging and Verb Extraction

- **Input**: Normalised text
- PoS tagging using HunPOS trained on the Stockholm-Umeå Corpus (SUC) of contemporary Swedish
- Extraction of word forms analysed as verbs by the tagger
- **Output**: List of all verbs in the text
Parsing and Complement Extraction

- **Input:** PoS-tagged text
- Parsing with dependency parser MaltParser, trained on the Talbanken section of the Swedish Treebank
- Extraction of phrases analysed as dependents of a verb
- Valency dictionaries for discarding unlikely complements
- **Output:** List of all verbs in the text, with their complements
Dependency Types

- direct object
- indirect object
- predicative complement
- prepositional complement
- infinitive complement of object
- verb particle
- subject (if passive verb)
Valency Dictionaries

• Initial interpretation:
  \( J \text{ midler tijd } \text{kom } \text{[obj greffuinnans gotze]} \text{ fougte thijtt} \)
  However, \( \text{[obj the Countess’ estate]} \text{ bailiff came there} \)

• Valency dictionary entry for \textit{komma} ('come'): Intransitive

• Final interpretation:
  \( J \text{ midler tijd } \text{kom greffuinnans gotze fougte thijtt} \)
  However, the Countess’ estate bailiff \textit{came} there
Experimental Setup

• 10 texts ranging from 1527 to 1737
  – 6 court records, 4 documents related to the Church
  – 444,075 tokens

• Gold standard
  – 400 sentences (40 randomly selected sentences from each text)
  – 21,660 tokens, of which 3,105 verbs
  – Manually validated annotation of verbs and complements

• Comparison with contemporary text
  – 20,000 tokens from the Stockholm-Umeå Corpus
    (subset not included when training the tagger)
<table>
<thead>
<tr>
<th>Swedish</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nu begäradhe/VB₁ Anderss</td>
<td>Now requested/VB₁ Anderss</td>
</tr>
<tr>
<td>[OO₉₉₁ tillståndh OO₉₉₁] att</td>
<td>[OO₉₉₁ permission OO₉₉₁] to</td>
</tr>
<tr>
<td>lefverere/VB₂ [IO₉₉₂ honom IO₉₉₂] [OO₉₉₂ Sädh OO₉₉₂] igen</td>
<td>deliver/VB₂ [IO₉₉₂ him IO₉₉₂] [OO₉₉₂ Grain OO₉₉₂] again</td>
</tr>
</tbody>
</table>
Evaluation Metrics

1. Identification of verbs
   - Precision and recall of verb analyses

2. Identification of complements
   - Precision and recall of requested dependency types
   - Exact match required
   - Labels are not considered

3. Identification of holistic verb constructions
   a) Number of fully correct complement sets
   b) Number of partially correct complement sets
   c) Number of incorrect complement sets
   d) Number of missing complement sets
Verb Extraction Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data</td>
<td>75.4</td>
<td>60.0</td>
<td>66.9</td>
</tr>
<tr>
<td>Normalised data</td>
<td>78.4</td>
<td>76.0</td>
<td>77.2</td>
</tr>
<tr>
<td>Contemporary text (SUC)</td>
<td>99.1</td>
<td>99.1</td>
<td>99.1</td>
</tr>
</tbody>
</table>

- Promising results for historical text
- Recall increases from 60% to 76% with normalisation
- Tokens erroneously analysed by the tagger:
  - Old spelling is identical to an existing, but different, word form in contemporary language
    - e.g. *skal* (=noun 'shell'), but should be *skall* (=verb 'shall/should')
  - Ambiguous words
  - Tokens that need further normalisation
## Complement Extraction Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data</td>
<td>24.8</td>
<td>27.5</td>
<td>26.1</td>
</tr>
<tr>
<td>Normalised data</td>
<td>28.3</td>
<td>28.2</td>
<td>28.2</td>
</tr>
<tr>
<td>+Valency Dictionaries</td>
<td>33.1</td>
<td>25.5</td>
<td>28.8</td>
</tr>
<tr>
<td>Contemporary text (SUC)</td>
<td>68.2</td>
<td>70.7</td>
<td>69.5</td>
</tr>
<tr>
<td>+Valency Dictionaries</td>
<td>71.8</td>
<td>56.2</td>
<td>63.0</td>
</tr>
</tbody>
</table>

- Normalisation has a smaller effect than for verb extraction
- The use of valency dictionaries improves precision for historical text, but has a large negative impact on recall for contemporary text
## Extraction of Holistic Verb Constructions

<table>
<thead>
<tr>
<th></th>
<th>Fully Correct</th>
<th>Partially Correct</th>
<th>Incorrect</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data</td>
<td>32.6</td>
<td>20.3</td>
<td>29.3</td>
<td>17.8</td>
</tr>
<tr>
<td>Normalised data</td>
<td>34.5</td>
<td>19.5</td>
<td>25.2</td>
<td>20.8</td>
</tr>
<tr>
<td>+Valency Dictionaries</td>
<td>38.7</td>
<td>16.9</td>
<td>18.9</td>
<td>25.5</td>
</tr>
<tr>
<td>Contemporary text (SUC)</td>
<td>30.3</td>
<td>54.2</td>
<td>9.1</td>
<td>6.4</td>
</tr>
<tr>
<td>+Valency Dictionaries</td>
<td>30.8</td>
<td>47.9</td>
<td>6.8</td>
<td>14.6</td>
</tr>
</tbody>
</table>

- 55.6% of the verbs in the historical texts are assigned a partially or completely correct set of complements.
- The percentage of fully correct complement sets is higher for the historical texts than for the contemporary text.
Conclusion

• Contemporary NLP tools are useful for extracting verbs and their complements from historical text, provided that normalisation is performed.

• Verbs are identified with an F-score of 77.2%.

• 55.6% of the verbs are assigned a partially or completely correct complement set.

• This kind of language technology application has the potential to facilitate the time-consuming, manual search for information in historical documents.
Future Work

- Exploring alternative normalisation techniques
- Exploring system performance over different time periods and text types, and what adaptations would possibly be needed to handle certain text types
- Identifying other categories than verbs and complements
- Tests performed by historians in the Gender and Work project