Dialect comparison in the Anselm Corpus

Stefanie Dipper

Institute of Linguistics
Ruhr-University Bochum

Historical Corpus Linguistics: Methods and Applications
SFB 1102
June 16/17 2016
Overview

- German reference corpora ("DDD")
  - annotations
  - searching the corpora in ANNIS
- Dialect comparison
  - Anselm corpus
  - comparing text variants
German reference corpora

1. **Reference Corpus Old German** (750–1050)
   - Karin Donhauser, Jost Gippert, Rosemarie Lühr
   - already available
   - [https://korpling.german.hu-berlin.de/annis3/ddd](https://korpling.german.hu-berlin.de/annis3/ddd)

2. **Reference Corpus Middle High German** (1050–1350)
   - Klaus-Peter Wegera, Claudia Wich-Reif, Stefanie Dipper, Thomas Klein
   - almost ready (todo: documentation)
   - [http://smokehead.linguistics.rub.de/annis3/rem](http://smokehead.linguistics.rub.de/annis3/rem)

3. **Reference Corpus Early New High German** (1350–1650)
   - Hans-Joachim Solms, Ulrike Demske, Stefanie Dipper, Klaus-Peter Wegera
   - work in progress

4. **Reference Corpus Middle Low German and Low Rhenish** (1200–1650)
   - Ingrid Schröder, Robert Peters
   - work in progress
Selection of texts

- Old German: everything which is there
- Middle High German:
  - Subcorpus (1 mio tokens): balanced and structured: composed of roughly equally-sized texts from different dialect areas, time periods and text sorts (everyday life, literature, poetry, science, religion)
  - Rest (1 mio tokens): more diverse selection, e.g. including texts written by different authors in different dialects
- Early New High German: balanced and structured
- Middle Low German and Low Rhenish: balanced and structured
Annotation

- **Layout**: page, line, column, etc.

- **Transcription and tokenization**:
  dipl:  *er uorfcon vorregote*  ‘find out’, ‘before god’
  mod:  *ervorscon uore gote*

- **Linguistic information**:
  - **part of speech** (HiTS): token- and type-specific
  - **lemma**
  - **inflection**

- **Punctuation**: original punctuation marks and modern sentence boundaries
HiTS: Historical tagset for part of speech
(Dipper et al. 2013)

- **Token-specific** vs. **type-specific** annotation
  - token-specific: POS of actual use
  - type-specific: POS in general / of the lemma

- Double annotation allows for monitoring language change resulting in a change of POS

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVD &gt; PTKVZ</td>
</tr>
<tr>
<td><em>Uz</em> [<em>...</em>] duont sie iuuuih</td>
</tr>
<tr>
<td>out do they you</td>
</tr>
<tr>
<td>‘they expel you’</td>
</tr>
</tbody>
</table>
ANNIS: Searching and visualization

General tool for searching and visualizing richly annotated corpora

- developed by Anke Lüdeling’s group (Krause and Zeldes 2014)
- allows for complex queries, relating to all levels of texts, annotations and metadata
- visualizes full texts, grid annotations, pointers, graphs, ...
- supports simple statistic analyses

REM’s simplified interface: searching for diplomatic and modernized forms of various regions, time periods, and text sorts
REM’s Simple Search

http://smokehead.linguistics.rub.de/simplesearch/
Interrogatio Sancti Anselmi de Passione Domini

Joint work with M. Bollmann, J. Krasselt, F. Petran, S. Schultz-Balluff, K.-P. Wegera

‘Questions by Saint Anselm about Lord’s Passion’

- Content: questions by Saint Anselm of Canterbury about Lord’s Passion, and Mary’s answers
- Highly popular text at the end of the Middle Ages
  - 230 manuscripts and prints in total (Latin, German, Dutch)
- The **Anselm corpus**: 51 manuscripts, 10 prints (all complete versions in German)
  - long and short versions (av. length: 8,000)
  - prose and verse versions
  - from 14th–16th centuries
  - from **Upper, Central and Low German**
German dialects: Upper, Central and Low German

Source: Wikimedia

Stefanie Dipper

Dialect comparison in the Anselm Corpus
St. Anselm was very glad and said: *Tell me, dear Lady: how did the martyrdom of your dear child start?* Our Lady said: *As my child had eaten the Last Supper together with his disciples before his martyrdom, and as they had stood up from the table, Judas Iscariot went to the nobles of the Jews*
## Variance in the Anselm text

<table>
<thead>
<tr>
<th>Dialect</th>
<th>Sentence</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAV</td>
<td>Do mein chindchind híet geezzen · mít feínen Jungern · vor feíner marter daz íungíft mal mal ·</td>
<td>‘As my (dear) child had eaten the Last Supper together with his disciples (before his martyrdom)’</td>
</tr>
<tr>
<td>ALEM</td>
<td>Do mín kíntkínt hatt gefízen das lung maťzmaťz mít fíñen lungren vor fíner marter</td>
<td></td>
</tr>
<tr>
<td>WCG</td>
<td>Da myn kíntkínt hatte gefízen mit fíñen jungern daz jungíte maťzemáťze</td>
<td></td>
</tr>
<tr>
<td>ECG</td>
<td>do myn lybes kyntkynt das obent brotbrot hatte gefízen · myt fíñen iungeren</td>
<td></td>
</tr>
<tr>
<td>MOD</td>
<td>da mein liebes kindkind das abendbrotbrot hatte gegessen mit seinen jüngern</td>
<td></td>
</tr>
</tbody>
</table>

E.g. **chind – kínt – kynt ‘child’**

**mal – masz – masze – brot ‘meal’**
Normalisierung

Joint work with M. Bollmann, J. Krasselt, F. Petran

Idea:

- learn mappings from historical forms to modern forms
- based on an aligned parallel diachronic corpus

- myn, mein → mein ‘my’
- kynt, chind → Kind ‘child’
Combination of different normalization methods:

1. **Wordlist mapping**
   - e.g. vnd → und

2. **Rule-based normalization**
   - Character rewrite rules
     - e.g. v → u / # _ n

3. **Distance-based normalization**
   - (Weighted) Levenshtein distance
     - e.g. vn → un
## Evaluation

Combination of methods 1+2+3 (cascade)

- Data: manually-normalized data (Anselm, 15c)
- Results in % accuracy

<table>
<thead>
<tr>
<th>Text</th>
<th>Baseline</th>
<th>Normalizations (# tokens)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Berlin</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23.05</td>
<td>68.99</td>
</tr>
<tr>
<td>Melk</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>39.32</td>
<td>69.10</td>
</tr>
</tbody>
</table>

Note: All tokens are lower-cased and “simplified”, e.g.:

- fī → si
- z̈ → zve
## Further variation in the Anselm Corpus

<table>
<thead>
<tr>
<th></th>
<th>1. Wien (W1)</th>
<th>2. Halle (H1)</th>
<th>3. München (M4)</th>
<th>4. Karlsruhe (Ka1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>da mein lib s chind</td>
<td><strong>da mein lieber son ihefuż</strong></td>
<td><strong>daz nachtmal</strong></td>
<td><strong>mit feinē iungern̄</strong></td>
<td><strong>hatte gezen</strong></td>
</tr>
<tr>
<td>het gęffenn</td>
<td><strong>daz nachtmal</strong></td>
<td><strong>am heiligen grün dornftage</strong></td>
<td><strong>het ge effen</strong></td>
<td><strong>daz ivngefte maz</strong></td>
</tr>
<tr>
<td>mit feinen Jn-ngern</td>
<td><strong>mit fineden iungen̄</strong></td>
<td></td>
<td><strong>vor feiner marter</strong></td>
<td><strong>mit sinen ivng’n</strong></td>
</tr>
<tr>
<td>vor feiner marter</td>
<td><strong>geffen hatte</strong></td>
<td></td>
<td><strong>das iungft effen</strong></td>
<td><strong>vor sin ‘mart’</strong></td>
</tr>
<tr>
<td>daz left mal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Manuel Alignment

1. Cognates (1:1) (chind : kint ‘child’)
2. Synonyms (1:1) (mal : brot ‘meal’)
3. Coreferent expressions (n:m) 
   (mein son : er ‘my son’ : ‘he’)
4. [ Complex equivalent expressions ] (n:m) 
   (mein lieber son : mein kint ‘my dear son’: ‘my child’)

H1: do [ mein lieber son ihesusz ] [ dasz nachtmal ] mit sienen iüngern
M4: do [ mein kint ] mit seinen iüngern . . . [ das iüngst essen ]
Similar and less similar texts

\[
\text{(IAA: } \kappa = .98) \quad \text{and} \quad \text{(IAA: } \kappa = .76)\]

Stefanie Dipper
Dialect comparison in the Anselm Corpus
Geographical classification (Wegera 2014)

Manually:

- Based on **extra-linguistic hints**, e.g. assumed place of origin
- Based on selected **characteristic forms**, e.g. spellings with *kch, ch, k*
Geographical classification: automatic

Pilot study

- Can the manually determined locations be replicated by automatic means?
- Features: uni-, bi- and trigrams of:

<table>
<thead>
<tr>
<th>Original</th>
<th>Sant aunʃhalm waz von herczen fro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters</td>
<td>sant aunshalm waz von herczen fro</td>
</tr>
<tr>
<td>(simplified) Ex:</td>
<td>#sa, san, ant, ...</td>
</tr>
<tr>
<td>Phones</td>
<td>zant aUnshalm vats fOn he:6tSEn fRo:</td>
</tr>
<tr>
<td>(SAMPA; txt2pho,</td>
<td></td>
</tr>
<tr>
<td>MBROLA)</td>
<td>(normalized + RFTagger, Schmidt &amp; Laws 2008)</td>
</tr>
</tbody>
</table>
Similarity of documents

- Each document: represented by a vector of the occurring ngrams and their frequencies
- **Cosine similarity**: computes angles between the vectors (pairwise)
- Build **cluster** based on the cosine values
- Finally: compare these clusters with the manual classification
The cluster of Low German

Neighbor Joining (with phylip)
Results

- The features characters/phones very well cover the large regions Upper / Central / Low German
- Certain subclusters correspond almost perfectly to the manually-determined classes
- Best feature: bigram phones
Visualizing user-determined features
(Tool by Florian Petran)
Visualizing user-determined features
(Tool by Florian Petran)
Summary

- In comparing different dialects or time periods, we take into account all data.
- Parallel data (Anselm corpus) facilitates comparison:
  - similar vocabulary, register, …
- Automatic tools: still rather low performance:
  - older stages (MHG, OHG): less variance, easier to process.


Search interfaces: links

- REM’s Simple Search: http://smokehead.linguistics.rub.de/simplesearch/
- ANNIS: http://smokehead.linguistics.rub.de/annis3/rem
- REMVIS: http://smokehead.linguistics.rub.de/remvis/