**THE ANSELM PROJECT**

Tools for Automatic Analysis of a Parallel Corpus in Early New High German

Marcel Bollmann, Stefanie Dipper, Julia Krasselt, Florian Petran
(bollmann,dipper,krasselt,petran@linguistics.rub.de)

---

**The Project**

In the Anselm project, we develop tools for the processing and automatic analysis of the medieval German treatise "St. Anselmi Fragen an Maria" (St. Anselm's questions to Mary), in collaboration with historical linguists from Ruhr-Universität Bochum.

**Goals**

- Diplomatic transcription of all residual manuscripts and printings in German
- Annotation with bibliographic and linguistic information, e.g. POS tagging
- Alignment on paragraph, sentence, and word level
- Linguistic analysis of dialectal variation
- Publication of a digital edition

**Transcription**

- Diplomatic transcription with twofold collation
- Medieval texts use spelling conventions no longer in use today, e.g. long 's' (ſ) or macrons (˛)
- Graphematic peculiarities are potentially of interest for users of the corpus (e.g. for dialectal analysis) and should therefore be retained in transcriptions

**Normalization**

We map ENHG wordforms to New High German (NHG) wordforms, with the idea of...

- Applying already existing annotation tools for German (e.g. TreeTagger for POS tagging) to our data
- Facilitating the alignment process by standardizing the spelling and
- Measuring the distance between old and modern wordforms for purposes of linguistic analysis.

**Manual**

- Manual normalization done by student assistants for evaluation purposes.
- Two levels of normalization are distinguished:
  1. Modern wordforms that are closest to the historical form (Vener Frauen → unser Frauen)
  2. Wordforms adjusted with regard to infection and semantics to form grammatical NHG phrases (Vener Frauen → unvereinig)

**Automatic**

- Idea: Learn characteristic spelling variations
- Character rewrite rules, sensitive to context

\[
v \rightarrow \text{u} / \text{u} \_ \_ \_ \text{n} \quad \quad \text{(historical 'v' becomes modern 'u' between a word boundary and 'n', e.g. in the mapping 'vnd' → Unmd)}
\]

- Rules learned from an aligned Luther bible corpus
  Based on the 1545 version and a modernized equivalent from 1892, both freely available
- Alignment on paragraph, sentence, and word level
- Evaluation showed performance superior to a pure dictionary-based word substitution method
- Example mappings generated by our method:

  - etlichen → der
  - wie → unser
  - zlt → zittern (instead of the correct zeit)

---

**The Corpus**

In the texts, Anselm of Canterbury asks questions to the Virgin Mary concerning the Passion of Jesus Christ. She answers him in the form of longer monologues.

- 44 German manuscripts and prints
- Written between 14th to 16th centuries, in Early New High German (ENHG)
- Verse and prose versions of different lengths
- Comparable content, logical structure, and even (semi-)parallelism in sentence structure and wording
- No fixed spelling conventions
- Dialectal variations in graphematics, phonology, morphology, and syntax

**The Tools**

- Browser-based
- No installation required by the user
- Centrally maintained

**OTTto (Online Transcription Tool)**

- Adds the diplomatic transcription of documents
- Special characters can be represented with user-definable escape sequences (e.g. "\text{'v' for 'v'

**CorA (Corpus Annotator)**

- Supports semi-automatic normalizations and annotations of various types

---

**Alignment**

As the texts are semi-parallel, alignment is often possible even at the word level.

**Purpose**

- Projection of normalizations and annotations from one text over the whole collection
- Comparative linguistic and literary studies on the texts

**Problems for traditional alignment methods:**

- Missing or inserted sections
- Spelling differences hinder recognition of corresponding wordforms
- No sentence boundary markings

**Solution**

- Extract a cognate dictionary
- Construct alignments from sequences closely occurring in translation pairs
- Gradually expand the sequences to find the longest possible sequence of alignment pairs
- Assign a confidence score and keep only the highest scoring sequence in case of conflict

Although it is a work in progress, this approach already performs better than traditional alignment methods based on heuristic text segmentation.

---

Sample from an Anselm manuscript (Upper German dialect), Würm, Hrab: Anna Amalia Bibliothek, Inv. No. 4, fol. 1r

---

An excited teacher by the name of Anselm asked our lady, the mother of God, for a long while...

Eyn hohher lerher hel'd Anheime der pat Vener Frauen die muter gottes lange timeless...

St. Anselm asked our dear lady of the heavens for a long time...

---

This project is supported by Deutsche Forschungsgemeinschaft, Grant 111344-1.