CorA: A web-based annotation tool for historical and other non-standard language data

Marcel Bollmann, Florian Petran, Stefanie Dipper, Julia Krasselt

Motivation & Purpose

CorA (short for “Corpus Annotator”) is a web-based annotation tool. It is designed for the annotation of non-standard language data, e.g., historical texts, computer-mediated communication (CMC), or texts from language learners. Tokenization is often problematic with these types of texts. For historical text, there can be OCR or transcription errors that are only detected during annotation. Also, automatic (pre-)annotation is often difficult due to the sparseness of training data.

CorA tries to alleviate these issues by providing options to edit the source document during annotation (including tokenization changes). Annotation software can be integrated, with options for retraining on newly annotated passages, and reannotating passages in already imported texts.

We plan to make CorA available as open source, including an English user interface. CorA is based on JavaScript, PHP 5, and MySQL.

Features

- Web-based
- XML-based import & export
- CorA XML can be imported into the ANNIS corpus search tool (Zeldes et al., 2009)

Annotation Types

- Part-of-speech and morphological tags
- Lemmatization
- Lemma part-of-speech
- Normalization
- Modernization and modernization category

Tagger Integration

- Integration of external annotation tools (e.g., POS taggers)
- Call taggers directly from the web interface
- Retraining taggers on existing annotations within the same project

Editing the Document

- Edit, add, or delete tokens in the source document
- Custom embedded scripts can perform validity checks, UTF-8 conversion, tokenization, etc.

Projects using CorA

- A reference corpus of Early New High German (Projekt Referenzkorpus Frühneuhochdeutsch, n.d.)
  - Annotated with POS, morphology, and lemma
  - Integrated lemma suggestions and linking to an external dictionary
- A parallel corpus of Early New High German (Anselm project; see below)
  - Annotated with POS, normalization, and modernization (including category)
  - German chat corpus
  - Annotated with POS
  - German learner corpus
  - Focuses on categorization of spelling variation

Future Work

- Internationalization of the user interface (soon)
- Export to TEI format
- Closer integration with the ANNIS corpus tool for late error correction
- Performing inter-annotator agreement calculations

Use Case: The Anselm Project (Dipper & Schultz-Baluff, 2013)

1. From the historical manuscript, a diplomatic transcription is created manually by specialists for German medieval studies.

2. The transcribed document is converted to CorA XML format, preserving different layers of tokenization.

3. The XML file is imported into CorA. Only the modernized tokenization is shown for annotation purposes.

From the historical manuscript, a diplomatic transcription is created manually by specialists for German medieval studies.