CorA-XML Utils: Processing Diplomatic Transcriptions in Historical Corpora

Adam Roussel[1], Fabian Barteld[1,2], Katrin Ortmann[1]

[1] Ruhr-Universität Bochum, [2] Universität Hamburg

adam.roussel@ruhr-uni-bochum.de, fabian.barteld@ruhr-uni-bochum.de, katrin.ortmann@ruhr-uni-bochum.de

Donnerstag,
05.03.2020
10:30–11:15
ESA1 W Foyer

CLP

When annotating non-standard texts, annotators often need to change the tokenization (while retaining information about whitespace and line breaks in the original text) and/or the text itself (e.g. when correcting errors in the digitization), however this feature is not found in many annotation tools. One exception is CorA (Bollmann et al. 2014), which was developed specifically for the annotation of historical German for the Reference Corpus of Early New High German, though it has also been used for CMC (Beißwenger et al. 2016). To retain flexibility, CorA does not interpret the text itself: Tokenization and the conversion of texts into its document model is handled by external scripts. To simplify the creation of these scripts, we present CorA-XML Utils (https://github.com/comphist/coraxml-utils).

CorA-XML Utils supplements the CorA document model with an ontology of character types. These enable CorA-XML Utils to convert data in the CorA-XML format, in which the token strings are not interpreted at all, into other useful formats, e.g. TEI-XML (https://tei-c.org/) which can make use of this information, e.g., in order to mark deleted passages in the transcription. Apart from this basic differentiation between actually transcribed characters and meta-characters, the character ontology adds further information, for instance marking some characters as having been "difficult to recognize" or "completed from a published edition". Such characters can be selectively masked in an exported diplomatic representation of the document. Other meta-characters in the transcription may mark the token boundaries used for annotation. In this case, Cora-XML Utils can then tokenize the document according to the whitespace in the original text or according to the additional boundaries, as required.

CorA-XML Utils already contains parsers for the transcription format of the ReM (Klein et al. 2016), ReN (Ren-Team 2019), ReF, ReDI, and Anselm (Dipper et al. 2018) corpora, all of which are or will be available in CorA-XML format. Therefore it can be readily used to convert these corpora into other formats. To use CorA-XML Utils with corpora that employ other transcription conventions, the addition of new tokenizers and parsers is possible. And with the addition of new importers and exporters, CorA-XML Utils is extendable to new import and export formats as well.

References: M. Beißwenger et al. (2016). EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. Proceedings of the 10th Web as Corpus Workshop. Berlin: Association for Computational Linguistics, 44–56. DOI: 10.18653/v1/W16-2606. URL: https://www.aclweb.org/anthology/W16-2606. M. Bollmann et al. (2014). CorA: A Web-Based Annotation Tool for Historical and Other Non-Standard Language Data. Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH). Gothenburg, Sweden, 86–90. URL: https://aclweb.org/anthology/W14-0612. S. Dipper et al. (2018). The Anselm Corpus. Version 1.0. Ruhr University Bochum. URL: http://islrn.org/resources/568-178-806-856-4/. Th. Klein et al. (2016). Referenzkorpus Mittelhochdeutsch (1050–1350). Version 1.0. URL: http://islrn.org/resources/332-536-136-099-5/. ReN-Team (2019). Referenzkorpus Mittelniederdeutsch/Niederrheinisch (1200–1650). Version 1.0, August 14, 2019. Hamburger Zentrum für Sprachkorpora. URL: http://hdl.handle.net/11022/0000-0007-D829-8.