LinkSyr: Linking Syriac Data

Wido van Peursen (w.t.van.peursen@vu.nl)
Hannes Vlaardingerbroek (hannes@vlaardingerbroek.nl)
Mathias Coeckelbergs (mathias.coeckelbergs@student.kuleuven.be)

Eep Talstra Centre for Bible and Computer

How do the Biblical heritage and Hellenistic culture interact in the oldest documents of Syriac Christianity? The Eep Talstra Centre for Bible and Computer (ETCBC) investigates this question in the CLARIAH pilot project LinkSyr (2017–2018), using linguistic data processing, especially topic modelling. The Syriac Book of the Laws of the Countries (BLC), written by the 2nd/3rd-cent. Syriac philosopher Bardaisan is compared with the ancient Syriac translation of the Bible ("Peshitta"), other sources from the ancient Mespotamian and Hellenistic world, and later authors (especially the 4th-cent. author Ephrem the Syrian) who react to Bardaisan's teachings. The analysed texts are exposed as Linked Open Data and related to the lexicographical and encyclopedic resources of Syriaca and SEDRA. The former presents the URIs for a large number of place names and person names for Syriac heritage, whereas the latter contains dictionary information for a list of more than 50,000 lexemes.

Data Preparation

The Data Preparation part of this project involves the preparation of language model training sets from our tagged corpora on the one hand, and the preparation of untagged corpora for automatic analysis on the other hand. For the first step, we have conducted experiments with OpenNLP and NLTK, acquiring accuracy with combined segmentation and PoS-tagging of 82%. Our current aim is to improve that result with a Brill-tagger, which uses transformation rules to optimize the PoS-analysis. We hope that a transformation rule-based approach will give better results than the more common approach using Hidden Markov Models, given the small size of our tagged corpora (ca. 150,000 tagged words). The second step, the preparation of untagged corpora, consists mostly of converting digital texts to a uniform data format, proofreading the texts for obvious errors and structuring the texts into sentences. Converting the texts to a uniform data format can mostly be done automatically, but structuring and especially proofreading also require more labour-intensive manual inspection. The final step will be word-sense disambiguation and entity recognition, in order to map the analyzed forms to the SEDRA and Syriaca databases.

Linked Data

The first completed task on the linked data aspect of the project involves collecting the data stored in the Syriaca and SEDRA databases. Once these data were brought together and converted into standard formats developed by the W3C (RDF and JSON-LD), we performed term matching experiments on the BLC text, hence taking first steps at developing a platform where we can input a text, whereafter its words can be reconciled with the information in Syriaca and SEDRA. Currently, we are exploring to what extent events from a new text can be encoded using the URIs we already have at our disposal, so that a database of facts can be created semi-automatically. The main difficulty lies in identifying salient relationships among the URIs, which afterwards can be queried. Our ongoing research focusses on this identification of relationships, as well as on the question of incorporating additional knowledge, such as linguistic information, which then in turn can form the basis of an ever-widening set of questions exploring the interconnections of texts.