Polemics Visualized: Morphological Analysis for Syriac

Hannes Vlaardingerbroek
Vrije Universiteit Amsterdam
hannes@vlaardingerbroek.nl

Marieke van Erp
Vrije Universiteit Amsterdam
marieke.van.erp@vu.nl

Wido van Peursen
Vrije Universiteit Amsterdam
w.t.van.peursen@vu.nl

Syriac, a language from the Aramaic family, has been the lingua franca of the Middle East for centuries. Many important theological documents from the period of the formation of the early church have been written in Syriac. These texts form a considerably large corpus, the published works of Ephrem the Syrian for example already exceed 500,000 words. The theological study of textual corpora of such size would benefit greatly from computational analysis of these texts.

The purpose of the Polemics Visualized pilot project is to explore the possibilities of computational linguistics and natural language processing for use in theological research of Classical Syriac texts. More specifically, we would like to answer the question whether Ephrem the Syrian, who wrote extensive polemics against Bardaisan, a theologian living two centuries earlier, was indeed discussing the same issues as Bardaisan addressed in his only remaining work.

So far, we have successfully trained a tokenizer using Apache OpenNLP and the annotated Syriac resources available at the Eep Talstra Center for Bible and Computer. With the resulting model we succeeded in recognizing 96% of word boundaries in the test data (Vlaardingerbroek, Van Erp, and van Peursen 2015). We then used our tokenizer on the unannotated text of Ephrem, and with the resulting data and that of Bardaisan's annotated work we trained an LDA topic model. The resulting topic model yields some sensible topic-document relations, but not sufficiently useful to aid in finding answers to questions such as the example mentioned above. We now aim to improve the results by morphological and part of speech tagging of the data, which would allow more efficient filtering of the input data for the topic analysis algorithm.

Since the development of a new part of speech tagging algorithm is beyond the scope of the current project, we are now looking into the possibilities to adapt solutions from existing projects to our implementation. NLP software has been developed for other Semitic languages, such as Hebrew and

Arabic. However, these approaches rely on large contemporary annotated textual corpora, which are not available for Syriac (Zitouni 2014: 52-55). The only other Syriac NLP software project so far, Syromorph, aims to facilitate the annotation of a large corpus of Classical Syriac, with morphological annotation, links to dictionary entries, and morphological attributes, using a joint pipeline model (McClanahan et. al. 2010). We are now working on the adaptation of the joint pipeline model of Syromorph for morphological analysis and part of speech tagging, in order to improve the results of the tokenizer model and topic analysis.

References

Hannes Vlaardingerbroek, Marieke van Erp, Wido van Peursen (2015) Polemics Visualised: Experiments in Syriac text comparison. Computational Linguistics in the Netherlands, Antwerp, February 2015.

Zitouni, Imed. (2014) *Natural language processing of Semitic languages*, Heidelberg: Springer.

McClanahan, P., Busby, G., Haertel, R., Heal, K., Lonsdale, K., Seppi, K., and Ringger, E. (2010) 'A probabilistic morphological analyzer for Syriac'. In: *Proceedings of the 2010 conference on empirical methods in natural language processing*, Massachusetts: MIT pp. 810-820.