Using Parallel Data to Improve Part-of-speech Tagging of 17th Century Dutch

Dieuwke Hupkes
Institute for Logic, Language and Computation, University of Amsterdam
dieuwkehupkes@gmail.com

Rens Bod
Institute for Logic, Language and Computation, University of Amsterdam
rens.bod@uva.nl

If one wants to extract information from a (historical) text, it is often useful to know the grammatical categories (or part-of-speech tags) of all the words in this text. Tools to automatically assign high accuracy POS-tags are freely available, but require large amounts of annotated training data. Automatically POS-tagging languages for which little annotated data is available has proved to be a challenging task, and when developing a POS-tagger for historical Dutch one is confronted with an additional difficulty: a large variation in spelling. Currently, digital humanities researchers working with historical Dutch texts often resort to taggers trained on contemporary Dutch (e.g., Van den Bosch et al., 2007), but the accuracy of the resulting tags is generally low.

In this paper, we explore methods for generating higher accuracy tags for historical corpora. In particular, we investigate how information extracted from a diachronic parallel corpus consisting of Dutch Bible texts from 1637 and 1977 can be used to improve POS-tagging of 17th century Dutch texts from several domains. We explore the possibility of using this corpus to rewrite/translate words prior to tagging, as well as the option of creating an annotated 17th century corpus by projecting tags from the contemporary version of the corpus to its historical counterpart and using this corpus to train a new tagger.

We show that even without applying methods to account for context dependencies, both methods result in great improvements over a baseline of tagging the texts with a tagger trained on contemporary Dutch. Furthermore, they significantly outperform using simple rewrite rules to normalise/modernise spelling before tagging. The improvement subsists across domains, but the within domain results are significantly better than the results for other domains, suggesting that incorporating knowledge about the domain of the text can lead to further improvement.

The results of this study can be of direct use to digital humanities researchers working with historical Dutch

texts of the last 3 centuries, as the tags assigned by our retrained tagger are of much higher quality than the current standard. Furthermore, it shows that using parallel data to exploit the similarities between contemporary and historical texts is a very promising path to developing diachronic taggers. Similar techniques could also be applied to other languages that have diachronic parallel corpera available, as well as to improve results on lemmatisation of historical Dutch texts. In future work, we will focus on domain adaptation techniques for a better consistency across domains.

References

Antal Van den Bosch, Bertjan Busser, Sander Canisius, and Walter Daelemans. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Computational linguistics in the Netherlands: Selected papers from the Seventeenth CLIN Meeting*, pages 99--114, 2007.