Automatic morphosyntactic and dependency annotation of the Anglo-Norman text database

Achim Stein and Yela Schauwecker (University of Stuttgart)

This paper discusses the grammatical annotation of an Anglo-Norman corpus. Anglo-Norman (AN) is the variety of Old French (OF) spoken and written in medieval England for about 400 years, until well after 1400. AN is also called Anglo-French, since its origins are not limited to the continental Norman variety of OF. We present the first attempt to provide an automatic grammatical analysis of a 3 mio word AN corpus, the *ANDhub text-database*[1], compiled to support the *Anglo-Norman Dictionary* project (Rothwell and Trotter, 2005).

The automatic analysis of AN is a challenge. On top of presenting the irregularites in e.g. spelling, inflection and word order that are also characteristic of OF, AN developed particular spelling variants, shows even less consistent case marking (e.g. in the accusative-dative distinction), and shows considerable variation between the earliest (c.1112) and the latest (c.1440) texts of the corpus.

In the approach taken here we applied machine-learning methods combined with lexicon-driven tools trained on existing OF resources. In order to solve the issue of specific AN spelling variants, we applied graphical "normalisation" rules that mapped as many forms as possible to a list of OF forms.[2] Although this method is tantamount to mapping one non-standardised language (AN) to another (OF), we expect to obtain better results with tools that were previously trained on OF corpora.

Apart from normalisation, the quality of the results heavily depends on the choice of the tools for lemmatisation, part-of-speech tagging, and syntactical parsing, as well as on the way they are combined, since e.g. parsing accuracy is dependent on tagging accuracy, and lemmatisation may provide the parser with more data for particular forms in the training process. A sample annotation process consists of the following steps:

1. We trained parsers on a manually annotated OF dependency treebank, the *SRCMF*[3]. Depending on the type of parser, this may require previous lemmatisation, and part-of-speech tagging (e.g. with the *mate tools* graph-based parser, Bohnet 2010) or else a joint morphological and syntactic analysis (e.g. with the joint transition-based parser, Bohnet et al. 2013).

2. We lemmatised the inflected forms using
   - either the *TreeTagger* trained on the OF *Nouveau Corpus d'Amsterdam* using the OF lexicon mentioned above
   - or with the lemmatiser contained in the mate tools distribution.

   Optionally we used further taggers with higher precision scores to improve the analysis at word level.

Technical aspects, in particular the pros and cons of using pure machine-learning vs. lexiconbased tools vs. combinations of both types for the analysis at word level will be one focus of this talk. The second focus will be on the results. Due to the absence of an annotated gold standard for Anglo-Norman, we are unable to provide a global, quantitative evaluation of the applied methods. We will therefore present the results from a linguistic point of view by highlighting selected examples and selected grammatical structures like argument structure and left dislocations.

So even if the more practical aspects of this talk reflect work that is still in progress, we intend making a contribution to the methodological question of how syntactic research, and

[1] http://www.anglo-norman.net

[2] This lexicon is part of the *Medieval-French Language Toolkit*, freely available at https://github.com/sheiden/Medieval-French-Language-Toolkit.

[3] *Syntactic Reference Corpus of Medieval French*, http://srcmf.org

more specifically historical syntax, can profit from unsupervised (and necessarily error-prone) corpus annotation methods.

References

Bohnet, B. 2010. “Top Accuracy and Fast Dependency Parsing is not a Contradiction”. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 89–97. Beijing, China: Coling 2010 Organizing Committee.

Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F. and Hajic, J. 2013. “Joint Morphological and Syntactic Analysis for Richly Inflected Languages”. In *TACL* 1 , 415–428.

Rothwell, W. and Trotter, D., (eds). 2005. *Anglo-Norman Dictionary 2*. Online Version . London: MHR.