Translation of Historical Dutch in the Nederlab Project

Erik Tjong Kim Sang
Meertens Institute, Amsterdam

The goal of the project Nederlab (2013-2017) is to provide access to a large diachronic corpus of texts written in Dutch with syntactic annotation [1]. A challenge in achieving this is that the language Dutch has changed significantly in the past centuries which makes older text harder to process for our annotation tools. To overcome this problem, we first translate the older text to modern Dutch before we apply the tools for syntactic annotation [3]. After processing the text we link the annotations of the modern text back to the original tokens.

We focus on the variants of Dutch spoken in the Dutch Golden Age: the seventeenth century. For translating texts from this period to modern Dutch, we have built an ad hoc system which contains the following components:

1. A parallel word lexicon extracted from two versions of the Dutch Statenvertaling bibles (1637 and 1888) with MGIZA [2]
2. Another parallel word lexicon extracted from two versions of the Dutch Statenvertaling bibles (1888 and 2010) with MGIZA
3. A list of modern Dutch words extracted from the 2010 Statenvertaling bible
4. A parallel word-to-token lexicon of other words extracted from the INL lexicon service (not used in Run 1)
5. A list of character replacement rules extracted from the two parallel lexicons mentioned under 1 and 2

The translation system has been optimized on the Blankaart text provided by the CLIN2017 shared task for good BLEU scores. It employs a context-insensitive backoff strategy which starts with looking up a token in the two parallel lexicons (1, 2), then looks for remaining words missing in the modern word list (3) in the INL lexicon service (4) and, for words missing there as well, applies the character sequence replacement rules (5).

References

[1] Hennie Brugman, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang, and Antal van den Bosch. Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora. In *Proceedings of LREC 2016*. ELRA, Portoroz, Slovenia, 2016.

[2] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[3] Erik Tjong Kim Sang. Improving part-of-speech tagging of historical text by first translating to modern text. In Bozic, Mendel-Gleason, Debruyne, and O'Sullivan, editors, *2nd IFIP International Workshop on Computational History and Data-Driven Humanities*. Springer Verlag, 2016.