Adapting NLP–tools for Creating an Orthographic Layer for Early Modern Dutch Texts.

Tessa Wijckmans MA
Department of Media & Culture
University of Amsterdam
t.m.wijckmans@uva.nl

Wouter van Elburg BA
Department of Media & Culture
University of Amsterdam
w.m.vanelburg@uva.nl

The seventeenth century Dutch language did not know a standardized spelling. Because of this, many different spelling variants of the same word (i.e. *ik* and *ick*, both meaning *I*) coexist. This may largely be the result of an author's own preferences, background or the book printer he used.[1] Software to standardize, analyse or process texts is mostly developed for modern Dutch, so processing historical texts with natural language processing (NLP) tools or analysing such texts on stylometric aspects is problematic. Many Digital Humanities researches experience similar difficulties when researching historic texts of various languages.

For project COLEM (Creating an Orthographic Layer for Early Modern Dutch Texts) we want to normalize spelling differences by having digital re-speller tools form a standardized spelling variant, that could help software better understand the texts. We investigate the possibilities to provide the original text with an orthographic layer containing the normalized words. In this way, the original texts and its morfo-syntactic information are still accessible and it will be possible to search both the original text and the layer. This, for instance, will ease research to language evolution.

The Java software VARD2 (Baron 2011; Baron and Rayson, 2008) seems to be a useful tool for this purpose.[2] This tool was originally created to normalize old English texts, but can be adjusted to other languages. VARD2 will compare the words in the input text with an incorporated, but easily adaptable wordlist (a .txt-file). We replaced the default wordlist with a Dutch lexicon and we trained VARD2 on texts of two different seventeenth century Dutch prose authors: Simon de Vries and Gotfried van Broekhuizen. Texts of these authors are characterized by a significant different orthography and this could therefore help us to train normalizers and test them on different spelling forms existing within the Early Modern Dutch language.

We trained VARD2 by replacing the variants in the historical texts with a normalization. VARD2 will present suggestions for normalization by using four methods of which just one is language dependent (a modified version of the Soundex phonetic matching algorithm that is based on English phonemes). Two other methods, that of letter rules and that of known variants, we adapted by

replacing the default .txt-files 'letter rules' and 'known variants'. The last method, a normalized Levenshtein Distance, does not need to be adapted, since it is language independent.

In this presentation we will show the performances of our trained VARD2 tool. We will focus on a specific amount of problems the tools run into, like uncommon words, clitics and combined words. We will also investigate the possibilities of the Norma[3] tool (Bollman, 2012) and TICCLops[4] (Reynaert, 2014). By comparing the results that various tools offer, we will decide what tool(s) is/are most successful in dealing with these problems. This could give us ideas for follow-up research or the development of tools for normalizing Early Modern Dutch, but probably also for normalizing other languages with unstable orthographies.

References

Baron, A. (2011) *Dealing with spelling variation in Early Modern English text, Lancaster*: University of Lancaster.

Baron, A. and Rayson, P. (2008) 'VARD2: A tool for dealing with spelling variation in historical corpora', Proceedings of Postgraduate Conference in Corpus Linguistics. Birmingham: Aston University, May 2008.

Bollmann, M. (2012). '(Semi-) automatic normalization of historical texts using distance measures and the Norma tool', Mambrini, F., Passarotti M. and Sporleder C. (eds) Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ARCH-2), Lisbon, Portugal, pp. 3-14.

Reynaert, M.W.C. (2014) 'Synergy of Nederlab and PhilosTEI: diachronic and multilingual Text-Induced Corpus Clean-up', Proceedings of the Sixth International Language Resources and Evaluation (LREC'14). Reykjavik, Iceland.

---

[1] It is not known who exactly was responsible for the spelling as it was printed. But due to there being examples of texts printed by the same printer that use radically different spelling forms it is plausible that the author of the text was responsible for the spelling.

[2] VARD is an acronym for Variant Detector.

[3] Norma is written in C++11, though bindings for Python are provided as well.

[4] TICCLops v.0.2(Text Induced Corpus Clean-up online processing system) is a web application (offered in a JavaScript interface) that is intended to detect and correct typographical errors and OCR (optical character recognition) errors in text. It is usable for every language, since it bases its replacements on the input corpus by making Most Frequent Words-list. However, TICCLops is probably less usable for providing a text with an annotation layer, because the replacements take place in the texts itself, without preserving an original version of the text.