Language adaptability and performance evaluation of historical text normalization tools VARD2 and TICCL

Iris Hendrickx (1) and Martin Reynaert (1,2)

(1) Centre for Language Studies, Radboud University Nijmegen
(2) TiCC research group, Tilburg University

i.hendrickx@let.ru.nl, reynaert@uvt.nl

Historical texts are digitized either by scanning old books and applying OCR on the scanned images or by manual transcription. If the former were a faultless process, both would result in what amounts to a diplomatic text transcription, preserving all original spelling variation. Such diplomatic editions are important for research in historical language use and change. However, when searching for a particular word form in digitized collections of historical texts, many potential matches will not be retrieved as spelling variants will not match with the modern word form. Furthermore, applying natural language processing tools will be much easier on a normalized version of the historical text as these tools were developed for modern text.

Here we investigate the performance of two statistical tools for spelling normalization of historical texts. We compare the tools VARD2 and TICCL. The VARD2 tool has been originally developed for normalizing Early Modern English while TICCL was developed for English and Dutch. The VARD2 tool was explicitly developed for historical data, while TICCL originally aimed to handle spelling and Optical Character Recognition variation in very large corpora of digitized 19th and 20th century text.

In this study we compare the performance of the two tools on historical Spanish and Portuguese texts. In previous work we already looked at Portuguese, but here we extend this comparison to Spanish and Dutch.

To evaluate the two tools we use data collected in the Post Scriptum Project. In this European project 7000 personal letters (3500 in Spanish and 3500 in Portuguese) are being collected from different historical archives. These letters are manually transcribed to an online digital format in TEI XML encoding. For Spanish we use an evaluation set of about 200 letters from the time period 1550 to 1830. For Portuguese we have 200 letters from 1550 until 1911.

VARD2 requires manual adaptation and training to work on another language. TICCL learns from the language data. We demonstrate this by setting it to work on historical Dutch texts in the framework of NWO project Nederlab. To evaluate TICCL on 17th century text, we use a selection of verses from the 1637 edition of the State Bible for which a gold standard modern Dutch transcription from 2010 is available. For evaluation on 18th century text, we evaluate on a book that was manually OCR-corrected and transcribed into both historical and modern gold standards: `Kort begrip der waereld-historie voor de jeugd' by Martinet, 1789. We specifically measure the contribution to performance made by the historical lexicon and name list compiled by the Institute for Dutch Lexicology in this task.

We discuss the performance of the tools, and their strengths and weaknesses for these different data sets spanning different time periods and different languages.