Towards gold standards for personal names

Gerrit Bloothooft, UiL-OTS Utrecht University, The Netherlands
(contact, g.bloothooft@uu.nl)
David Onland, UiL-OTS Utrecht University, The Netherlands
Martin Reynaert,TiCC Tilburg University, The Netherlands
Katrien Depuydt, INT Leiden, The Netherlands
Tanneke Schoonheim, INT Leiden, The Netherlands

The enormous variation in the spelling of personal names in historical documents hampers search tasks and nominal record linkage. The Clariah pilot project NAMES [1] aims to come to grips with this variation and attempts to map 598.000 surnames and 190.000 first names from the 19th century Dutch vital registration onto gold standards. Since ambiguities and uncertainties complicate such an enterprise, the focus is on a level of standardization which reduces the search space sufficiently and effectively with optimized recall rather than precision. This implies that there is no need to arrive at standards that are fully etymologically or genealogically justified.

Three approaches are followed. 1: In the earlier LINKS project [2], name variant pairs were derived on the basis of highly confident but inexact record matches [3]. This allowed for an initial clustering of 127.000 surnames into 15.114 base names and 48.000 first names into 2.772 base names (gender-dependent, with a further reduction to 926 gender-independent base names). 2: Expert review of the previous cluster results. 3: Application of TICCL [4] (Text-Induced Corpus Clean-up, a tool developed for post-processing of OCR results) to learn statistics of spelling variation from the name variant pairs, to suggest and test clustering of variants against the experts' views, and finally to cluster the remaining names.

Major observations are 1) a low edit distance may indicate a variant pair, but a genuinely distinct pair as well, especially for short names, 2) while in some cases a spelling difference can be judged on a single character (reading errors like *s* and *f* or known spelling variation like *c* and *k*), often the context of the spelling difference is important to decide for or against a variant, 3) experts have difficulties to make systematic decisions on this context (and to develop a rule system for it), 4) even though we have an extensive corpus, for rare variant types there can be insufficient training material for machine learning.

Typical variation in first names is found in suffixes, resulting in large edit distances which cannot easily be learned automatically (*Jan – Johannes*). These are usually present, however, in proven variant pairs, and are quite easily recognized by experts. Parsing of compounded family names (*Zwijnen-burg*) would facilitate systematic analysis, but is hard under conditions of spelling variation. In family names, attached prefixes (*De Meulemeester – Demeulemeester*) pose additional problems, again especially in short names.

Our approach is to iteratively apply the three methods and to review intermediate results. At the cluster level (of standards) this implies optimization of intra-cluster variant pairs against inter-cluster variant pairs. This analysis takes into account a) frequency of occurrence of a variant pair, b) frequency of the individual variants in the 19th century corpus, and c) frequency of occurrence of the individual variants in the modern vital registration (which indicates a genuine name).

Final results will become available as linked open data, and by means of the INT lexicon service.

References:

[1] https://www.clariah.nl/projecten/research-pilots/names/names

[2] https://socialhistory.org/en/hsn/linking-system-historical-family-reconstruction-links

[3] Bloothooft, G. and Schraagen, M. (2015),' Learning name variants from inexact high-confidence matches', in Bloothooft et al. (Eds.), *Population Reconstruction*, Springer, Switzerland

[4] Reynaert, M. (2010). Character confusion versus focus word-based correction of spelling and OCR variants in corpora. International Journal on Document Analysis and Recognition , 14:173–187.