Weigh your words—memory-based lemmatization for Middle Dutch

Mike Kestemont
Institute for the Study of Literature in the Netherlands (ISLN) and University of Antwerp, Belgium

Walter Daelemans and Guy De Pauw
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Abstract

This article deals with the lemmatization of Middle Dutch literature. This text collection—like any other medieval corpus—is characterized by an enormous spelling variation, which makes it difficult to perform a computational analysis of this kind of data. Lemmatization is therefore an essential preprocessing step in many applications, since it allows the abstraction from superficial textual variation, for instance in spelling. The data we will work with is the *Corpus-Gysseling*, containing all surviving Middle Dutch literary manuscripts dated before 1300 AD. In this article we shall present a language-independent system that can 'learn' intra-lemma spelling variation. We describe a series of experiments with this system, using Memory-Based Machine Learning and propose two solutions for the lemmatization of our data: the first procedure attempts to *generate* new spelling variants, the second one seeks to implement a novel string distance metric to better *detect* spelling variants. The latter system attempts to rerank candidates suggested by a classic Levenshtein distance, leading to a substantial gain in lemmatization accuracy. This research result is encouraging and means a substantial step forward in the computational study of Middle Dutch literature. Our techniques might be of interest to other research domains as well because of their language-independent nature.

Correspondence:
Mike Kestemont,
Universiteit Antwerpen,
Stadscampus, Prinsstraat 13,
Room D.118, 2000
Antwerpen, Belgium.
E-mail:
mike.kestemont@ua.ac.be

1 Spelling Variation in Middle Dutch

Middle Dutch is a typical example of a historical language displaying a considerable amount of spelling variation (Van der Voort van der Kleij, 2005; Ernst-Gerlach and Fuhr, 2006; Kestemont and Van Dalen-Oskam, 2009; Souvay and Pierrel, 2009). Especially before the advent of the printing press, there existed no standard language variety of Dutch, let alone a standard spelling. As such, medieval Dutch spelling was generally highly phonological and 'personal' in nature, since it would represent each writer's own dialectal pronunciation and local spelling habits. That is why even highly frequent words could be spelled in very different ways, reflecting the abundant variety of dialects and local substandards then found in the Low Countries (Fig. 1).

This spelling variation makes it difficult to process medieval texts in any computational application. For instance for authorship attribution, it

Literary and Linguistic Computing, Vol. 25, No. 3, 2010.

doi:10.1093/llc/fqq011 Advance Access published on 4 August 2010