Towards a General Purpose Tagger-Lemmatizer for Pre-Modern Dutch

Mike Kestemont, Guy de Pauw, Renske van Nie and Walter Daelemans

Historic Dutch texts are characterized by large amounts of variation in spelling and spacing, due to the lack of a written standard language and orthography in the pre-modern period. This variation renders it difficult to automatically process historic text material in computational applications, ranging from plain searching to advanced text categorization tasks. Therefore the automatic normalization of historic Dutch, for instance via lemmatization and part of speech tagging, is an important preprocessing step for many applications in Digital Humanities. In this paper we present the general-purpose tagger-lemmatizer we have developed for pre-Modern Dutch. Our architecture efficiently deals with historic spelling variation for pre-modern, and in particular medieval Dutch. Our source code will be freely available and an online interface will be provided where users can upload their own texts and have them annotated in the cloud. An innovative feature of our system is that it draws on all available training data sets for pre-modern Dutch, amongst which the Corpus-Gysseling (literary and legal texts), the CRM Charter corpus, the Repertory for Proper Nouns in Middle Dutch Literary texts, ... It is the first time that a state of the art normalization system is trained on all these resources simultaneously.