Recent developments in the *Diplomata Belgica*-project. Case study: NER applied to a corpus of Middle Dutch charters

Jeroen Deploige / Marijke Beersmans

The database *Diplomata Belgica* was launched in 2015. Then since, the Belgian Royal Historical Commission continued to invest in the further development of its software end the enrichment of its contents. At the same time, the database also became the object of both traditional research and computational experiments. Finally, in recent years, the question arose how national projects like *Diplomata Belgica* might be integrated in bigger international datasets. In this paper, we will focus on two topics. First, we will discuss the challenges that emerged within the *Diplomata Belgica*-project since 2015, with particular attention to the possibilities and modalities for integration of the project in bigger international initiatives. Second, we will present a recent example of computational experimentation within the project, and more precisely of Named Entity Recognition (NER). NER, or the automatic extraction of personal names, place names and other rigid designators, is often considered a first, but important step in information extraction. For historic texts, automating the extraction of Named Entities can greatly improve their accessibility and as such facilitate research. We will report on the annotation of a small corpus of fourteenth-century Middle Dutch charters (155 documents/50 000 tokens) and the training and testing of two neural NER-models, one using the state of the art NLP-library Spacy and one Bi-LSTM model. We found that while the models were nowhere near state of the art, both were able to detect named entities relatively successfully (surpassing a macro-f1 score of 70). Our presentation will indicate that NER on Middle Dutch charters is feasible and definitely worth pursuing with a larger dataset.