Christel Annemieke Romein, Sara Veldhoen and Michel de Gruijter

Short Project Introduction: The enhancement of the search-ability of OCRed Texts. HTR, Segmentation and meta-dating Early Modern Ordinances

Keywords: Legal History (Normative texts); HTR - Gothic font; Text segmentation (article segmentation); Categorisation (Labelling)

From May till November 2019, the KB National Library of the Netherlands (The Hague) hosts the Researcher-in-Residence project 'Entangled Histories of Early Modern Ordinances'. This 'short paper'-presentation reports on the ongoing project. The project's hypotheses read that: when problems arose, small 'states' had to act swiftly; thus, governments may have adopted – parts of – successful legislation from neighbouring areas. These 'entangled histories' may become obvious while studying the political-institutional activities through legislation. Hence, the project uses printed books of ordinances from the Low Countries (1500-1800s) to answer this hypothesis. These sources used for this project are early modern norms (ordinances) printed in a gothic font (25%) and roman font (75%). The techniques that will be developed through 'entangled histories' will have a widespread positive effect. These techniques coincide with the project's three steps:

1. Segmentation of the texts, going from sentence-recognition (Lonij, Harbers, 2016) to article-segmentation of larger texts; this requires that the computer is trained to recognise the beginning and end of texts, either as a chapter or as an individual text within a compilation of texts. We apply the P2PaLA-tool from the University of Valencia, which is integrated into Transkribus to recognise the different sections in our corpus.
2. Improving the OCR-results of digitised sources. By using machine learning techniques, such as the Handwritten Text Recognition suite Transkribus we will reprocess the files which currently have poor quality OCR. By treating the printed with hand-carved letters as very consistent handwriting we hope to obtain a higher quality of recognition (CER <5%).
3. We want to create automatic metadata by training a computer to recognise the conditions that categories were based upon. This will be based on an already developed tool called Annif (created by the Finnish National Library). After supervised training, the computer can then suggest, apply and supplement categories to other texts based on the idea of topic modelling (Leydesdorff, 2017). This is a pilot that will prove the applicability of the tool to other languages as well.'

We will use NLP such as NER technologies to identify dates, titles, and persons. Due to its significance for such a broadly studied range of sources, we hope to make the output of the tagged entities available as RDF with the intention of making the output available in the Dutch national infrastructure With the data generated in this project, visualisations of the development of

laws across the Netherlands – and possibly Europe - could become possible. These normative texts (laws) this data is based on, contain indications of how governments of burgeoning states dealt with unexpected threats to safety, security, and order through home-invented measures, borrowed rules, or adjustments of what was established elsewhere. While the thousands of texts become more easily accessible, it will become possible to look for entangled histories with neighbouring states, due to synchronic and diachronic comparisons. Firstly within the Low Countries (from one province to another), from which we will be using 108 books of ordinances; secondly, it will be possible to look for connections with German-speaking lands as the same metadata is applied as the Max-Planck-Institute für europäische Rechtsgeschichte has done in several projects. The connection with the German dataset will be established through (numbered) URI's and Linked Data, as a step after the initial project.

The ordinances within the books of ordinances are frequently consulted by researchers of various disciplines (e.g. history, law, political sciences, linguistics) to unravel rules for controlling complex societies. Having the possibility of a longitudinal search, based upon contents rather than the index or title, as well as having an overview based upon several states has so far been impossible due to the impenetrable amount of scanned texts.

• Leydesdorff, L., and A. Nerghes. (2017) "Co-word maps and topic modelling: A comparison using small and medium-sized corpora (N< 1,000)." Journal of the Association for Information Science and Technology 68(4): 1024-35.
• Lonij, J., Harbers, F. (2016), Genre classifier. KB Lab: The Hague http://lab.kb.nl/tool/genre-classifier
• Finnish National Library, Annif 2.0, https://github.com/NatLibFi/Annif/