Handwritten text recognition applied to the manuscript production of the Carthusian Monastery of Herne in the Fourteenth Century

Haverals, Wouter

wouter.haverals@uantwerpen.be
University of Antwerp, Belgium

Kestemont, Mike

mike.kestemont@uantwerpen.be
University of Antwerp, Belgium

Cultural archives are characterised by hyper-diversity, especially when it comes to premodern, handwritten documents that abound in miscellaneous languages, writing systems and scribal hands. Handwritten text recognition (HTR) offers a promising, emergent technology to digitise and make available the text contained in these materials (Muehlberger et al. 2019). Modern HTR systems typically rely on supervised machine learning (e.g. neural networks) to perform automated transcription: models are trained on a gold standard of manually provided transcriptions and can then be applied to unseen target material. It is well established that for such a system to perform well, the distribution of training data should maximally approximate that of the target documents. Moving across different writing systems and scribal hands will therefore inevitably challenge models and practitioners should consequently consider carefully which training material they should invest valuable annotation time in. As a result, a conundrum presents itself: ground truth data should capture as much as possible the diversity of a (corpus of) document(s), yet at the outset it is often unclear what that diversity actually comprises. Consequently, the selection process of ground truth data renders the digitization process prone to the so-called Matthew effect, where existing inequalities regarding subsets of the material (e.g. availability, quality, ...) risk being reproduced and even enlarged. Whereas scholarly work in HTR has exploded in recent years (Nockels et al. 2022), there exist few studies that shed light on the practical limitations imposed by the archival hyper-diversity in this respect. This paper aims to contribute to the state of the art by assessing the feasibility of various set-ups in the environment of a single case study, where diversity can be studied in a controlled fashion.

The Carthusian monastery of Herne (near Brussels) was a proverbial hotspot in the translation, copying, and creation of Middle Dutch literature in the fourteenth century. The output of this monastery is unparalleled in the cultural history of the Low Countries. Many unique and salient texts survive from Herne (Haverals & Kestemont 2020). A milestone in the recent research has been the dissertation of Erik Kwakkel (2002) in which he was able to situate the production of a large number of codices in this charterhouse. He identified many collaborating hands in these documents, which shed an unusual light on this vibrant community. A unique feature is that the manuscripts often contained detailed conversations in the margins. Because the Carthusians were a silent order, much of the internal discussions among collaborating scribes had to happen through such marginal notes. Some of the scribal oeuvres in Herne are so sizable that they might be among the largest attested in medieval Europe. Interestingly, these scribes not only acted as authors and translators themselves, but they also worked under the patronship of wealthy outsiders. They were able to actively change their writing mode, producing clean and highly readable products for the outside world ("high style") but much less accessible, dense documents for internal usage ("low" and "middle style"). A challenging aspect, for human and computational readers alike, of the latter writing modes is the exceptionally high density of abbreviations for vernacular texts.

Over the past years, we obtained high-quality, digital facsimiles of the entire manuscript collection that is currently associated with the Herne monastery. Additionally, we have produced sizable sample transcriptions of nearly every document in the corpus using the Transkribus platform (Kahle 2017). This renders it possible to obtain an automated transcription of the entire corpus. In this talk, we shall discuss the performance of a "Grand Model", which is trained on our inherently diverse corpus (containing multiple scribal hands, various handwriting styles, spelling profiles, textual genres, etc). The overarching question here is: to what extent does variation aid HTR-models to produce accurate automatic transcriptions? And – more importantly – when does the amount of variation become too large, making the model's accuracy suffer from it? To investigate this, we will confront our Grand Model with a structured patchwork of smaller case studies that shed light on the effect of various experimental conditions. For this, we will construct different train-target combinations in the available material.

Apart from more general issues, such as the effect of size and diversity of the training material, the Herne case enables us to study detailed issues, such as the effect of switching between different hands and writing modes, but also the directionality effect when moving between direct copies of texts. The results of a preliminary experiment are offered in Figure 1, which reports on the Character Error Rates (CERs) obtained by three different models. A first model was trained on a Middle Dutch text transcribed by scribe 1 (ca. 28k words), while another model was trained on the same text copied off by a fellow monk, scribe 2 (ca. 29K words). A third model was trained on the full corpus; this is the so-called 'Grand Model' (ca. 230K words). From these results, we can observe that applying a model trained on a single hand to the same text by a different scribe is penalised with an increased CER of ca. 6%. Despite the high diversity in the full corpus, the Grand Model performs well in the controlled environments of single scribal hands (we observe only a slight increase of ca. 0.5% CER).

Figure 1. Comparison of Character Error Rates (% of incorrectly recognized characters) for different train-test setups.

		Trained on…		
		Scribe 1	Scribe 2	Full corpus
Tested on…	Scribe 1	2.56	8.34	3.22
	Scribe 2	8.74	2.86	3.09
	Full corpus	10.14	11.77	3.19

In this paper, we tackle these experiments from the point of view of quantitative performance scores, however, we will also include a qualitative discussion of the results using insightful evaluation tools. Ultimately, our paper will contribute to the assessment of various – possibly impacting – factors during the collection of ground truth data for training HTR-systems. By scrutinising different parameters, users will be able to better assess perhaps the most frequently asked question in digital humanities: more data is always better, but when do I have enough?

Bibliography

Haverals, Wouter / Kestemont, Mike (2020): “Silent voices: A Digital Study of the Herne Charterhouse Scribal Community (ca. 1350-1400)”, in: Queeste 27(2), 186–195.

Kahle, P. / Colutto, S. / Hackl, G. / Mühlberger, G. (2017): “Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents”, in: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 4, 19–24.

Kwakkel, Erik (2002): Die dietsche boeke die ons toebehoeren: De kartuizers van Herne en de productie van Middelnederlandse handschriften in de regio Brussel (1350-1400). Leuven: Peeters.

Muehlberger, G. / Seaward, L. / Terras, M. et al. (2019): “Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study”, in: Journal of Documentation 75(5), 954–976.

Nockels, J. / Gooding, P. / Ames, S. / Terras, M. (2022): “Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research”, in: Archival Science 22(3), 367–392.