The application of HTR to early-modern museum collections: a case study of Sir Hans Sloane's Miscellanies catalogue

Marco Humbel and Julianne Nyhan,

University College London

Research context

Handwritten Text Recognition (HTR) is “the ability of a computer to transform handwritten input represented in its spatial form of graphical marks into an equivalent symbolic representation as ASCII text.” (Romero et al., 2012: 5) What is the state of the art of the application of HTR to early modern manuscripts? With what level of accuracy can HTR models automate their transcription? What is known about how HTR currently accommodates manuscript text that shows changing writing styles, hands and text in multiple languages? We will explore these questions with reference to the wider literature and a case study of the first HTR model to be created for the hand of Sir Hans Sloane (1660-1753).

Optical Character Recognition on documents with perfectly machine-printed characters can reach an accuracy level of more than 99% (Cao and Natarajan, 2014: 336–37). However OCR is often problematic for historical documents (Smith and Cordell, 2019: 5). It also cannot be used for handwritten documents, since the space between characters and words is inconsistent. *Holistic segmentation-free off-line* HTR technology works at a line level and can deal with cursive characters, slanted words and irregular calligraphy, but it must be trained for a specific handwriting (Alabau and Leiva, 2012: 2274; Sánchez et al., 2014: 111–12). HTR is not accurate enough to replace human expertise, however it holds the potential to bolster the transcription process (Toselli et al., 2018: 174;176).

‘Enlightenment Architectures: Sir Hans Sloane’s Catalogues of his Collections is a Leverhulme-funded collaboration between the British Museum and UCL. It studies 5 of the manuscript catalogues of Sloane,[1] and is encoding them in TEI to understand the information architectures they use. In 2017, selected catalogues were transcribed to a high level of accuracy by the company AEL Data Service in Chennai, India. We thus had high quality transcriptions of Sloane’s manuscript materials, in addition to images of his catalogues, available for use in the training of an HTR model for Sloane’s hand.

The HTR model discussed here was trained using the software Transkribus. The aim of the e-Infrastructure project READ (Recognition and Enrichment of Archival Documents) is to make archival sources more accessible through technological development. The centrepiece of READ is the service platform and application Transkribus, which enables the automatic recognition and transcription of handwritten documents and the ability to search within them (READ project, 2018a; READ project, 2018b).

Methodology

To train an HTR model with Transkribus, one has to provide it with training data (digital surrogates of the original folios and their transcriptions). This is known as *ground truth* or *reference data.* The segmentation of the document into its elements, in particular the baselines, and the actual transcription is crucial for creating an adequate HTR model. The ground truth data must consist of a representative sample of a collection’s documents and also respect the original appearance of the script, e.g. special characters, as closely as possible. With Transkribus, this serves the purpose of training the HTR model, and also the evaluation

[1] The catalogue of Miscellanies, two of his Natural History catalogues (Fossils vol. I and vol. V) and two of his library catalogues (Sloane MS 3972C vol. VI and Sloane MS 3972B).

of its accuracy. Between 75 and 100 pages (around 15,000 to 20,000 words) of training data are necessary for an effective HTR model. A randomized selection of documents is recommended (READ project, 2018c: 3–4; READ project, 2017: 10). We determined that the first sub-section of the Miscellanies catalogue (folio 2-152v) would give enough training and test data to evaluate the model because it contains important characteristics of the whole collection of catalogues, such as annotations and a complex layout.[2]

For this research, five different HTR models were created to allow a comparison between their changing accuracy. This includes one pre-test model. Training started with 75 folios and was then increased to 100 and 125 folios. For the last model, in addition to the 125 folios of training data, a base model was added.

Results

The quality of an HTR transcription can be evaluated according to a Word Error Rate (WER) and Character Error Rate (CER) (Romero et al., 2012: 93). Transkribus allows both measures (READ project, 2018c: 5). WER is […] "the minimum number of words that need to be substituted, deleted or inserted to convert a sentence recognized by the system into the corresponding reference transcription, divided by the total number of words in the reference transcription […]" (Romero et al., 2012: 55). CER is the minimum number of single characters which need to be corrected, divided by the total number of characters in the reference text (Romero et al., 2012: 55). Transkribus also allows the evaluation of the general accuracy of a model with a learning curve visualisation and the accuracy of a model on the page level to be specified via the compute accuracy function (READ project, 2018d: 9–12). According to READ (2018d: 10), a model with an accuracy rate of 90% can be regarded as an effective automated transcription.

The evaluation showed that our current model of 20,803 words reached a CER of 12.73% without the base model. The transcription has not reached a level of accuracy that is sufficient for academic research without further human input. The model has problems transcribing names (persons and places), abbreviations, double letters (e.g. ee), punctuation, Latin text and the numbers in the margins correctly.

Conclusion

In the paper we will reflect on how our methodology and model might be refined in order to improve the CER, in line with the experiences of other projects (for example Hodel, 2017 or Prell, 2018). We will give particular attention to questions like 'Where in particular does recognition fail?'. 'How much training data is necessary to create a model with an accuracy of at least 90%?' and 'how might external resources like gazetteers and name authority lists be integrated into Transkribus and used in conjunction with the HTR model in order to increase the accuracy of the transcription of named entities? Our responses to questions like this are likely to be transferable to other projects who seek to build HTR models for the transcription of early-modern manuscript materials.

Although our model reached a relative high level of accuracy, is it not good enough to be used for scholarly work. We will therefore also reflect on scenarios where the model could still be used, such as Authorship Attribution (Franzini et al., 2018) or Named Entity Recognition (Carbonell et al., 2018; Toledo et al., 2019).

[2] We wish to thank the members of the Enlightenment Architectures team for their assistance in making this selection and for their wider advice about this case study.

References

Alabau, V. and Leiva, L. (2012). Transcribing handwritten text images with a word soup game. *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. Austin, Texas: ACM Press, pp. 2273–78 doi:10.1145/2212776.2223788.

Cao, H. and Natarajan, P. (2014). Machine-Printed Character Recognition. In Doermann, D. and Tombre, K. (eds), *Handbook of Document Image Processing and Recognition*. London: Springer London, pp. 331–58 doi:https://doi.org/10.1007/978-0-85729-859-1_44.

Carbonell, M., Villegas, M., Fornes, A. and Llados, J. (2018). Joint Recognition of Handwritten Text and Named Entities with a Neural End-to-End Model. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. Vienna: IEEE, pp. 399–404 doi:10.1109/DAS.2018.52.

Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J. K., Franzini, E., Byszuk, J. and Rybicki, J. (2018). Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm. *Frontiers in Digital Humanities*, **5**(4) doi:10.3389/fdigh.2018.00004.

Hodel, T. (2017). Sending 15th-Century Missives through Algorithms: Testing and Evaluating HTR with 2,200 Documents *Schrift Im Kloster* https://solascriptum.wordpress.com/2017/07/11/imc-leeds-paper-sending-15th-century-missives-through-algorithms-testing-and-evaluating-htr-with-2200-documents/ (accessed 27 March 2019).

Prell, M. (2018). Frühneuzeitliche Briefe als Herausforderung automatisierter Handschriftenerkennung: Ein Transkribus-Projektbericht. doi:10.22032/dbt.34849. https://www.db-thueringen.de/servlets/MCRFileNodeServlet/dbt_derivate_00041045/Transkribusbericht_2018_06_02.pdf (accessed 27 March 2019).

READ project (2017). How To Transcribe Documents with Transkribus: Simple Mode https://transkribus.eu/wiki/images/a/ad/HowToTranscribe_SimpleMode.pdf (accessed 27 March 2019).

READ project (2018a). Services https://read.transkribus.eu/services/ (accessed 27 March 2019).

READ project (2018b). About https://read.transkribus.eu/about/ (accessed 27 March 2019).

READ project (2018c). How To Prepare Test Projects with Transkribus - for Archives and Libraries https://transkribus.eu/wiki/images/8/81/HowToPrepareTestProjects.pdf (accessed 27 March 2019).

READ project (2018d). How To Train A Handwritten Text Recognition Model In Transkribus https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf (accessed 26 March 2019).

Romero, V., Toselli, A. H. and Vidal, E. (2012). *Multimodal Interactive Handwritten Text Transcription*. Vol. 80. (Series in Machine Perception and Artificial Intelligence). Singapore: World Scientific Pub.

Sánchez, J. A., Bosch, V., Romero, V., Depuydt, K. and Does, J. de (2014). Handwritten text recognition for historical documents in the transcriptorium project. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*. Madrid, Spain: ACM Press, pp. 111–17 doi:10.1145/2595188.2595193.

Smith, D. A. and Cordell, R. (2019). *A Research Agenda for Historical and Multilingual Optical Character Recognition*. Northeastern University https://ocr.northeastern.edu/report/ (accessed 10 March 2019).

Toledo, J. I., Carbonell, M., Fornés, A. and Lladós, J. (2019). Information extraction from historical handwritten document images with a context-aware neural model. *Pattern Recognition,* **86**: 27–36 doi:https://doi.org/10.1016/j.patcog.2018.08.020.

Toselli, A. H., Leiva, L. A., Bordes-Cabrera, I., Hernández-Tornero, C., Bosch, V. and Vidal, E. (2018). Transcribing a 17th-century botanical manuscript: Longitudinal evaluation of document layout detection and interactive transcription. *Digital Scholarship in the Humanities*, **33**(1): 173–202 doi:https://doi-org.libproxy.ucl.ac.uk/10.1093/llc/fqw064.