Maria Kallio
the National Archives of Finland
maria.kallio@arkisto.fi

**Handwritten Text Recognition and 19th century Court Records**

This paper will demonstrate how the READ project is developing new technologies that will allow computers to automatically process and search handwritten historical documents. These technologies are brought together in the Transkribus platform, which can be downloaded free of charge at https://transkribus.eu/Transkribus/. Transkribus enables scholars with no in-depth technological knowledge to freely access and exploit algorithms which can automatically process handwritten text. Although there is already a rather sound workflow in place, the platform needs human input in order to ensure the quality of the recognition. The technology must be trained by being shown examples of images of documents and their accurate transcriptions. This helps it to understand the patterns which make up characters and words. This training data is used to create a Handwritten Text Recognition model which is specific to a particular collection of documents. The more training data there is, the more accurate the Handwritten Text Recognition can become.

Once a Handwritten Text Recognition model has been created, it can be applied to other pages from the same collection of documents. The machine analyses the image of the handwriting and then produces textual information about the words and their position on the page, providing best guesses and alternative suggestions for each word, with measures of confidence. This process allows Transkribus to provide the automatic transcription and full-text search of a document collection at high levels of accuracy.

For the quality of the text recognition, the amount of training material is paramount. Current tests suggest that models for specific style of handwriting can reach a Character Error Rate of less than 5%. Transcripts with a Character Error Rate of 10% or below can be generally understood by humans and used for adequate keyword searches. A low Character Error Rate also makes it relatively quick and easy for human transcribers to correct the output of the Handwritten Text Recognition engine. These corrections can then be fed back into the model in order to make it more accurate. These levels also compare favorably with Optical Character Recognition, where 95-98% accuracy for early prints is possible.

Of even more interest is the fact that a well-trained model is able to sustain a certain amount of differences in handwriting. Therefore, it can be expected that, with a large amount of training material, it will be possible to recognize the writing of an entire epoch (e.g. eighteenth-century English writing), in addition to that of specific writers.

The case study of this paper is the Finnish court records from the 19th century. The notification records which contain cases concerning guardianships, titles and marriage settlements, form an enormous collection of over 600 000 pages. Although the material is in digital form, the usability is still poor due to the lack of indices or finding aids. The National Archives of Finland started to produce transcripts from a small part of the collection in cooperation with the Genealogical Society of Finland in 2017. Around 30 volunteers have been producing ground truth for training of Handwritten Text Recognition. The first HTR model was trained based on 75,000 words of training data. The result was very good, with an average character error rate (CER) of only 12 per cent. The genealogists have continued to work on the material and the goal is to collect as much training data as possible in order to reach a CER of 5–10 per cent. Even with the current result of CER 12% it is possible to search and use the material with the Key Word Spotting -tool in Transkribus, which already improves the research opportunities of the court records considerably. The goal is to provide at least part of the notification records in computer readable form by the end of the project and thereby enable new versatile possibilities for research.