Republic HTR paper

*Ronald Sluijter (ronald.sluijter@huygens.knaw.nl), Rutger van Koert (rutger.van.koert@di.huc.knaw.nl), Michael Baars (michael.baars@*huygens.knaw.nl*), Marja Swüste, Michel van Gent (michel.van.gent@huygens.knaw.nl,), Esther van Gelder (esther.vangelder@kb.nl, Jesse Hollestelle, Ger Ruigrok (ger.ruigrok@huygens.knaw.nl), Ida Nijenhuis, Joris Oddens (joris.oddens@huygens.knaw.nl),*

Abstract

Using annotation software provided through the Transkribus[1] Platform we annotated scans, concerning mostly 17th century handwritten documents from the National Archive of the Netherlands, with their textual transcriptions. The resulting ground truth was used to train a machine learning model yielding very accurate results. The ground truth was made available as an open access dataset.

Introduction

After receiving a large grant from the Dutch Research Council (NWO), the REPUBLIC project has started in 2019. The project is carried out by the Huygens Institute in collaboration with the National Archives of the Netherlands, which is providing the scans of the resolutions.[2] The goal is to publish the resolutions of the States General between 1576 and 1796 online and make them searchable via a user application and API's.

The resolutions of the assembly in which the seven provinces of the Republic negotiated their common interests are considered one of the most important historical sources of Dutch political history. Some of the resolutions have been previously published in traditional book form (1576-1625) and as an online XML edition (1626-1630), but for the most part as summaries in modern Dutch.[3] The rapid development of modern technologies such as OCR and HTR now makes it possible to publish this important archival resource at a much faster pace, and also completely, i.e. not limited to abstracts. The first phase of the REPUBLIC project concerns the production of the machine-readable text of the handwritten and printed resolutions. This paper is the companion text to the publication of the ground truth of the handwritten text.

The dataset

REPUBLIC'S handwritten text corpus consists of 225,000 scans of the resolutions of the States General between 1576 and 1796. Because these are historical documents, the handwriting of the seventeenth-century resolutions in particular differs greatly from that of modern manuscripts. Consequently, it is largely illegible to untrained readers. The fair copies of the resolutions used

[1] Transkribus, https://readcoop.eu/transkribus/
[2] Nationaal Archief, Den Haag, Resoluties van de Staten-Generaal, nummer toegang [1.01.02]
[3] https://resources.huygens.knaw.nl/besluitenstatengeneraal1576-1630 (3 March 2023).

for the project were written by some eighty different clerks, each in their own handwriting. Moreover, handwriting style in general underwent a development in the early modern period that is reflected in the corpus. Despite the age of the documents, on the whole they are in very good condition, although some sections have suffered water damage and there are signs of wear in all of them. The dataset we provide is a subset of 515 scans with their ground truth in the PageXML[4] format. The transcriptions of the entire corpus is made available as open data[5].

Transcribing Republic

The first step in the transcription process was to add segmentation to each scan, which is necessary to connect the text image to the transcription. Transkribus has a module for automatic layout analysis, but it was initially not very suitable for resolutions. Resolution pages have a specific layout with several elements that need to be distinguished from each other because each has its own meaning. These elements are the actual resolution text, page number, meeting date, attendance list and marginalia. In addition, many volumes also contain an index, which has its own layout, usually in two columns per page. Because Transkribus' automatic layout analysis could not properly distinguish these elements, many had to be corrected manually. This also applied to baselines, which in many cases were not long enough, were split into two sections on one line or were missing altogether.

The data for the training set were selected semi-randomly using random numbers, taking into account the different individual handwriting styles and the general development of handwriting style during the study period. We selected approximately one thousand pages (515 scans) that we believe represent a representative cross-section of handwritten documents during this period. These scans were carefully transcribed in an iterative process. At first, we transcribed from scratch. After about fifty scans, we trained an HTR model and repeated this process. The performance of the model improved rapidly; the second model appeared to perform well enough to change the transcription process. After this, we used the HTR model to transcribe the text automatically and then correct it manually. Except for one iteration, each model performed better than the previous one (see Table 1).

Table 1 Overview of trained HTR-models with CER-percentage on validation set

	CER %
Model 1 (Republic_1):	8,79
Model 2 (Republic_2):	3,95
Model 3 (Republic_3):	4,30
Model 4 (Republic_4):	4,01
Model 5 (Republic_5):	3,64

[4] https://github.com/PRImA-Research-Lab/PAGE-XML
[5] https://zenodo.org/record/7695131

Model 6 (Republic_6)	3,32
Model 7 (Republic_7)	2,99

Unfortunately, Transkribus currently no longer offers the HTR+ software used to create the latest models. For this reason, we publish only the ground truth, and not the HTR model.

The basic transcription rule we followed was simple: as literal as possible. This was done to get the best training results from the HTR. Nevertheless, the historical handwriting regularly led to some challenges. For example, the use of capital characters was very different from our current habits. In many cases it is not clear whether the clerk intended to use a capital character or not. Sometimes a character shape looks like a capital character but is written in lowercase, sometimes characters appear in more than two forms: capital, lower case, and an intermediate form. In these cases, a choice had to be made by the transcribers, and it is likely that this was not done consistently across the whole corpus of transcriptions. Also, the documents contain characters for which there is no equivalent on modern keyboards. In these cases, we have used characters that were most similar, or used one character for several characters in the text that seemed to have one specific meaning, such as "et cetera". Another example are characters used to indicate abbreviations, such as a horizontal bar above the (last part of a) word, or a curl after the last character. These additions have been ignored. However, we have tagged and spelled the abbreviations as such using the tool built into Transkribus for this purpose.

Experiments

The set of scans we used for training and validation of our latest model, Republic_7, consists of over 236,000 words. Using the Transkribus platform, we split the data into a training and validation set at a ratio of 90-10%. We trained for 500 epochs and achieved a Character Error Rate (CER) of 2.99% on the validation set. We are currently using the crowdsource platform 'Vele Handen' to correct HTR transcripts on a larger set of scans. Part of this will be used to try to improve the Republic_7 model even further.

Conclusion

HTR software can provide good quality transcriptions when training with accurate ground truth transcriptions. Adding more ground truth helps the machine learning, but for each additional percent reduction in CER significantly more data is needed to achieve the same reduction in errors.

Acknowledgements:
We kindly thank NWO and the KNAW for funding to make this project possible.