MAHEC – S2-M8ii
08/03/2022

Dr. N. Janz
Reading and studying war letters of WWII – From collecting, digitizing, transcribing to reading and understanding of ego-documents of Luxembourgish conscripts

TRANSCRIPTIONS

- Rather to understand the text, we have to read it and to understand and to follow the content

- Various forms of transcriptions can be found in cultural heritage institutions, such as handwritten copies, transcriptions of spoken language in parliamentary minutes or transcriptions made with the aim of an edition.

TRANSCRIPTIONS

What is a transcription?

- Texts, writings only unique books and transcriptions
- Preservation of knowledge and stories

To transcribe means to transfer a source document into a target document by means of rules and conventions.

Transcription is intended to capture the text

Only through the continuous transcription work of copyists and translators over the centuries can we today receive texts and the pre-modern cultures on which they are based.

With the advent of printing in Europe during the Renaissance:

- typesetting was added as a new form of transcription in which the effort required to make copies, in contrast to manual copying, continued to decrease as the number of copies to be produced increased.
- The consequences were a significant expansion of cultural exchange, the rediscovery of texts previously manifest only in isolated manuscripts and the shaping of many scientific disciplines as we know them today. Since the advent of microchip technology and the internet, we find ourselves in the next cultural-technological evolution, in which IT-supported procedures are used at all levels of transcription work.

Transcription is the process in which historical artifacts are turned into editable text, and in this case, into digitally editable text.

Not only copy the text – but to produce an automatic transcription or copy of the document

Digital transcriptions?

AUTOMATIC TEXT RECOGNITION - GENERAL

OCR – optical looks mainly on the individual characters

- Standard front, clear backgrounds – printed text

HTR – context of text recognition

- It goes up and down – learns the order of symbols from a language
- Not just character based but line based (context)
- Mathematical data – human readable textual information (best-guesses, alternatives and confidence ratings)

Ideal Procedure:

- The process of text recognition is usually divided into several phases: a pre-processing of the basic raster graphics, the actual character recognition and a post-processing of the recognized text.
- In the pre-processing, an attempt is first made to optimize the raster graphics as a template for the OCR. This includes the compensation of possible distortions or rotations of the original, whereby a good image digitization can and ideally should anticipate this step.
- each pixel of the raster graphic is assigned either text or background, which the OCR software decides on the basis of an algorithm. Smooth transitions, impurities or disintegration of the original (in the case of historical documents) can cause problems, as can insufficient resolution of the digital copy.
- This is followed by an analysis of the layout and segmentation of the original. The aim is to identify the areas that contain any text at all. Other elements such as pictures or graphics are marked as such, but are not included in the further text recognition. Lines and often words are then separated from each other in the text areas.
- Modern OCR software stores the layout information (coordinates) of text blocks, individual lines, words or even lines together with the captured text, so that the position of the hit in the digital image can be visually highlighted in a later full text search.
- Modern OCR software also tries to capture the document in its logical structure, for example, to separate headers, footers, marginalia and other pretexts from the actual text, to recognize headings and paragraphs, to mark embedded images and graphics as such or to reconstruct a table structure.
- This is initially done purely on the basis of typographical criteria, such as spacing measurements between text blocks or the change of fonts, and can therefore ultimately only serve as an indicator of a logical-content structure of the text, which usually needs to be corrected. [1]

- The goal of **digital transcription** is not always a critical edition.
- Modern OCR and Handwritten Text Recognition (HTR) procedures are based on Deep Learning (DL) and NN method.
- The aim of these methods is to automate the transcription of image-digitised works by teaching the computer to recognise the text in the image line by line. These procedures must first be trained with manually transcribed line-image line-text pairs

[1] Jannidis, Fotis, Hubertus Kohle, and Malte Rehbein. *Digital Humanities. Eine Einführung*, 2017. P. 194

Handwritings

HTR

- Applying the same process to handwritten text is far more challenging because of the far greater range of variability in every stage of the writing process.
- Different periods, cultures, and languages, writing with different implements, fashion the letters of the alphabet in different ways; and so do different writers, and indeed even the same writer in different circumstances. As a consequence, progress in the challenging task of automatic transcription of handwritten text has had to await the arrival of OCR fortified by sophisticated forms of artificial intelligence, namely by the application of a neural network approach to the field of HTR.[2]

TRANSKRIBUS[3]

- Transkribus uses machine learning (neural networks) to train 'models' irrespective of different languages and styles of handwriting. In an iterative process of interaction between the scholar (who creates the transcription) and the software (that 'learns' from the subsequent corrections), a so-called 'model' is created that feeds the specific documentation back into the overall program.
- The training pro- cedure itself consists of the upload of images, the recognition of baselines and regions of the text on the image, and the manual or automated line-by-line transcription of its content
- Once a sufficient quantity of manually transcribed text has been generated by the user (typically about fifty pages), the training process can be started: by pressing a button, the scholar enables the program to 'learn' – that is, to recognize and extract patterns automatically – from what has been fed into the system regarding the scholarly validated relationship between the scanned image and the text[4]

HOW

- using machine learning algorithms in order to teach new writing styles to the system. The system can transcribe the text in any language and handwriting type. After a user transcribes part of the text manually, the software engine learns to identify the characters and then finishes the task automatically with impressive accuracy. Thus, the idea behind the platform seems exceptionally simple and pioneering. All the user needs to do give an image to the software and a part of the corresponding text and based upon this text; the software can learn the handwritten script and similar fonts.

---

2 Hotson, Howard, and Thomas Wallnig. Reassembling the Republic of Letters in the Digital Age. Reassembling the Republic of Letters in the Digital Age, 2019. https://doi.org/10.17875/gup2019-1146. P 244

[3] https://readcoop.eu/transkribus/?sc=Transkribus Kahle, Philip, Sebastian Colutto, Gunter Hackl, and Gunter Muhlberger. "Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents." Proceedings of the International Conference on Document Analysis and Recognition, ICDAR 4 (2018): 19–24. https://doi.org/10.1109/ICDAR.2017.307.

[4] Hotson, Howard, and Thomas Wallnig. Reassembling the Republic of Letters in the Digital Age. Reassembling the Republic of Letters in the Digital Age, 2019. https://doi.org/10.17875/gup2019-1146. P 245

Upload image
Analyze text segment – where the text is

- After uploading the images to Transkribus, the next step is “segmentation” or “layout analysis”, i.e. Transkribus identifies the different elements on the page (heading, footer, marginals, paragraphs, etc.) and the lines of text within these elements. Having processed more than 500 pages, my conclusion is that this step cannot be automated completely. In most cases, you will have to use the “segmentation mode” in Transkribus to manually correct the results of the automated segmentation

OUTLOOK

- Automatic character recognition (OCR) is still not a 100% reliable technique, but requires control of the results and rework.
- The results achieved are largely of higher quality with typographically uniform and clean texts than with font mixtures, column divisions and wavy paper (the OCR-D project aims to address these weaknesses: http://ocr-d.de/). Despite the increased use of machine learning methods (Hidden Markov Models, Deep Learning), automatic handwriting recognition does not yet achieve results that would significantly save manual work when integrated into digital processing. Worth mentioning in this context is the recently developed tool Transkribus (https://transkribus.eu/), which has a focus on Sütterlin and earlier German scripts. Transkribus can produce increasingly better results, but is still far from being able to replace manual handwriting transcription[5]

5 Baillot, Anne. “2.16 Digitalisierung und ihre Einflüsse auf den Umgang mit alten wie neuen ‚Briefen‘ in deutscher wie internationaler Perspektive.” Handbuch Brief, 2020, 387–98. https://doi.org/10.1515/9783110376531-025.