READ
Recognition and Enrichment of Archival Documents

How to use existing transcriptions to train a Handwritten Text Recognition (HTR) model

Version v1.3.8.1-Snapshot /(07_12_2017)
Last update of this guide: 22/07/2018

This is a short introduction for those users who have existing transcriptions and would like to use them to train a Handwritten Text Recognition (HTR) model. It is especially useful for users who already have at least 500-1000 pages of transcribed material.

Download the Transkribus Expert Client, or make sure you are using the latest version:

- https://transkribus.eu/

Consult the Transkribus Wiki for further information and other How to Guides:

- https://transkribus.eu/wiki/

Transkribus and the technology behind it are made available via the following projects and sites:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Contact

- The Transkribus Team: email@transkribus.eu

Contents

The READ project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 674943.

Introduction

Getting started

The Transkribus platform allows users to train a Handwritten Text Recognition (HTR) model to automatically process a collection of documents. The model needs to be trained to recognise a certain style of writing by being shown images of documents and their accurate transcriptions.

Over the past 20 years, thousands of scholarly transcription projects have been carried out and many are still running. A vast amount of documents have been transcribed and are now available in electronic form All these transcriptions can be used in a simple and straightforward way as training material for HTR.

The t2i tool, developed by the CITlab team at the University of Rostock, creates training data from existing transcriptions. It uses an algorithm to automatically match transcriptions with images of handwritten material and processes them to create a HTR model.

The tool is especially useful for users who already have at least 500-1000 pages of transcribed material.

Instead of producing training data for HTR manually in Transkribus (cf. How to Transcribe Documents with Transkribus – Introduction), users can simply use their existing transcriptions to try out the technology. In this way, reliable transcriptions can be generated without any need to change the workflow or editing interface used by a project.

Preparation

Introduction

- If you would like to work with the t2i tool, you need to have access to digitised images and transcripts of your documents.
- These files also need to be prepared according to the below instructions before they can be processed with t2i.

Number of pages

- We recommend that you start the training process with at least 20,000 (or around 100 pages) of transcribed material.
- T2i works especially well if a larger number of transcripts are already available, e.g. 500 or more pages.
- This technology is able to process a high quantity of transcripts (100 000 pages and more).
- The neural networks in HTR learn quickly and the more training data, the better the results can be.

Image files

- All kinds of images can be processed.
- The images should have a resolution of at least 200 ppi, or – if the images come from a camera – as a rule of thumb the x-height of a single character should be represented by at least 15-20 pixels.
- Of course, the accuracy of the HTR is somewhat related to the quality of the images. Nevertheless, with enough training data more difficult material from microfilms or bitonal scans can be processed.

Transcript files

- All transcripts should be saved in the form of simple text (TXT) files.
- If your transcriptions are available as TEI (Text Encoding Initiative), Word, XML, or HTML files, you should convert them into TXT files, i.e. by copying and pasting the transcripts into Notepad.
- Transcriptions should be saved on page level, i.e. one TXT file for each page image.
    - If you are familiar with TEI, your can create the TXT files with a "Text Export".
    - If you are creating TXT files manually, you may find it quicker to copy and paste your transcripts directly into Transkribus, line by line. See How to Transcribe Documents with Transkribus – Introduction for information about how to do this.

Transcriptions

- Transcriptions should be free of all mark-up.
- If your transcription contains line breaks, these can be retained. However, it is not necessary to include line breaks at the end of each line of text.
- The t2i tool can also handle cases where a word split over two lines has been transcribed in full without a hyphen.
- If there is an illegible word in your transcript, it is best to simple delete the entire line in which that word appears. This line will then not be used for training the HTR.
- Transcriptions do not need to be complete. If words are missing from the transcript, they will not be used for training the HTR.
- It is possible to work with all kinds of Unicode characters, including Arabic and Hebrew writing.
- In some cases, transcriptions where abbreviations have been extended can be used for t2i and HTR training as well (abbreviations will be extended automatically).

Naming files

- The files containing your images and transcriptions should be clearly linked.
- To achieve this, each image file should saved with the exact same name as its corresponding TXT file.

Help with preparing files

- If you do not have the resources to prepare your files in this way, the Transkribus team can help. Send us an email (email@transkribus.eu) to find out more.

Delivery of files

- Once you have prepared your images and transcripts, you need to put them into the right structure:
    - Name of document
        - TXT
- You can upload your files directly to Transkribus. For the upload, the TXT files should be included in an extra folder called 'txt', within a folder of images.

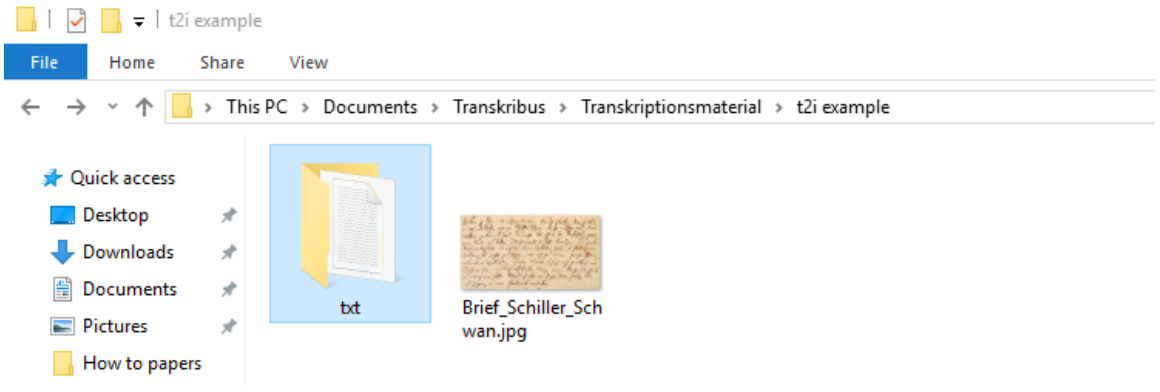

Figure 1 How files need to be split

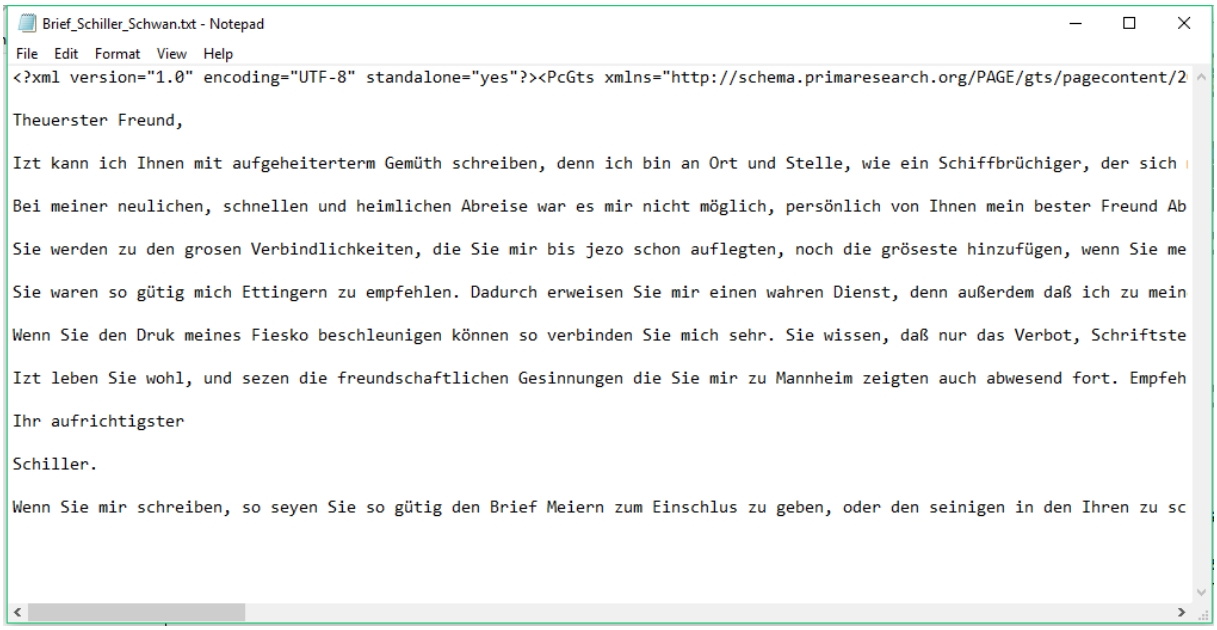

Figure 2 TXT file

- Alternatively, you can send your files to the Transkribus team (email@transkribus.eu) using a file-sharing system like WeTransfer.
- In both cases, drop us an email (email@transkribus.eu) to let us know that the files are ready.
- We will get back to you about the training of a model.
- **Note:** when the image and text files are uploaded to Transkribus, the text lines are not matched with the image lines right away. At first, one text line is generated for each image. The Transkribus team will then include existing HTR models in the T2I process to match the lines of image and text together.
- **Note:** the t2i tool is not perfect yet! Normally 50-75% of the lines are matched correctly straight away. Where lines of image and text fail to match up, some manual corrections will be needed.
- Of course we are working on the efficiency of the tool ☺

Credits

We would like to thank the many users who have contributed their feedback to help improve the Transkribus software.

Transkribus is made available to the public as part of H2020 e-Infrastructure Project READ (Recognition and Enrichment of Archival Documents) which received funding from the European Commission under grant agreement No. 674943.