READ
Recognition and Enrichment of Archival Documents

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

Version v1.9.1
Last update of this guide: 07/01/2020

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. The Text to Image Tool is now implemented into the Transkribs Expert interface and this guideline explains how you can match your images and existing transcriptions yourself.

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. 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. Please be aware that this technology won't be able to provide an error-free transcription. It is based on a HTR-model, which has got a certain kind of error rate itself. Some manual corrections will be needed. If you will need a perfect transcription without any mistakes, it might be quicker to copy the existings transcriptions directly into Transkribus.

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.

File-preparation

- 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.

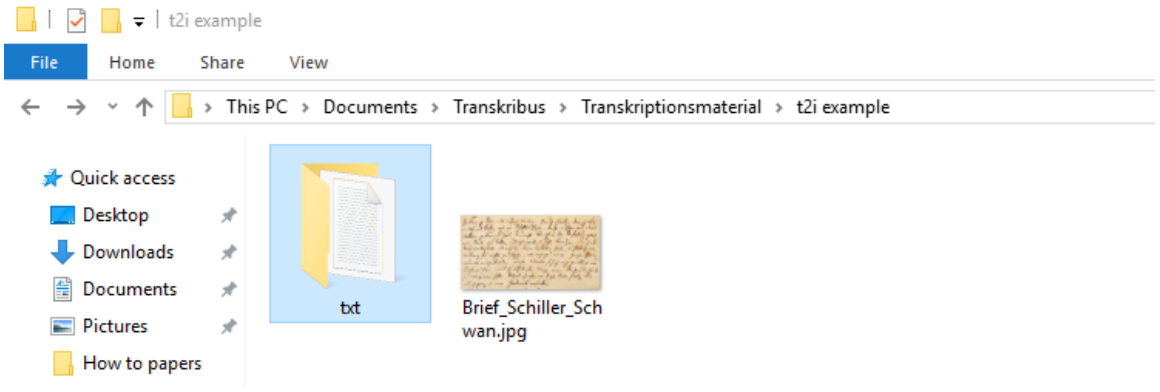

Figure 1 How files need to be split

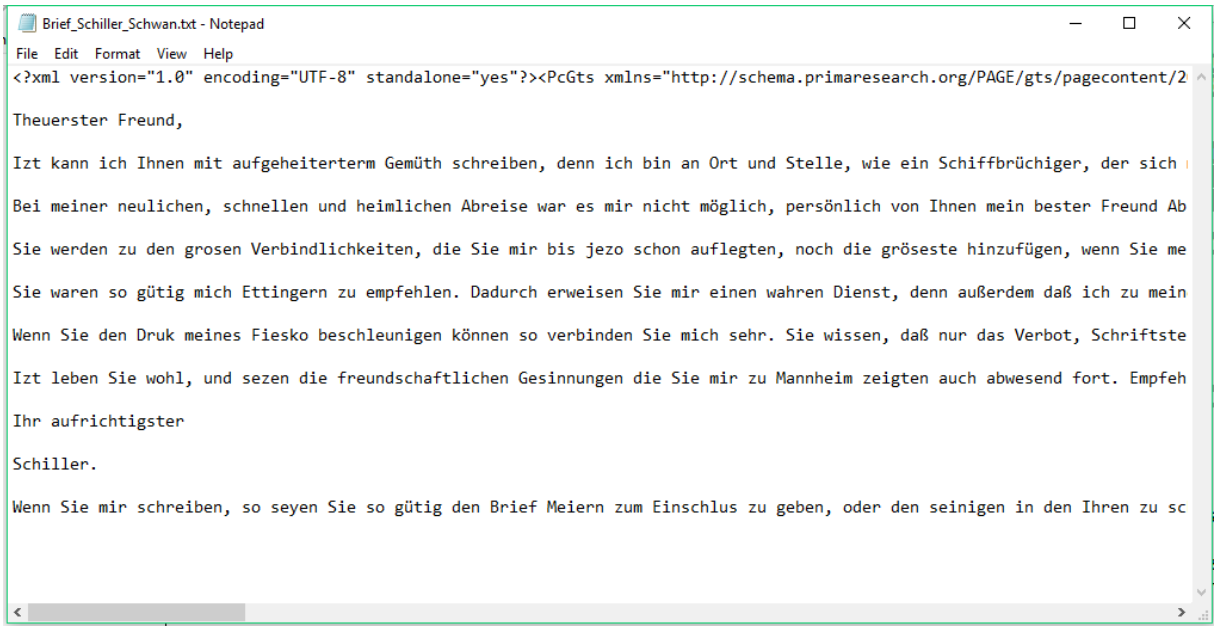

Figure 2 TXT file

- **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.

HTR-model

- In order to run the t2i you will need an HTR-model, consistent with your document.
- We already have a couple of existing models, you can check if one of those is suitable.
- Otherwise you can prepare your own model for the t2i. For this you will have to copy in the transcription into Transkribus for some pages and then use this as training material. More information on the training of model you can find in this guideline: Modell Training in Transkribus

T2i in Transkribus

Uploading scans and transcripts together

- If you upload scans and transcripts together, follow the instructions above and subsequently use the “normal” Transkribus import, which you can find in the main menu.

Uploading scans and transcripts separately

In case you have already uploaded the images at an earlier point of time without the text files, please proceed as follows:

- Open the images in Transkribus
- Save the text files in a separate folder on your computer
- Click on “Main Menu” in Transkribus (top left)
- Click on “Document”
- Choose “Sync local text files with doc”
- Choose the text files in the directory
- The following window will open:

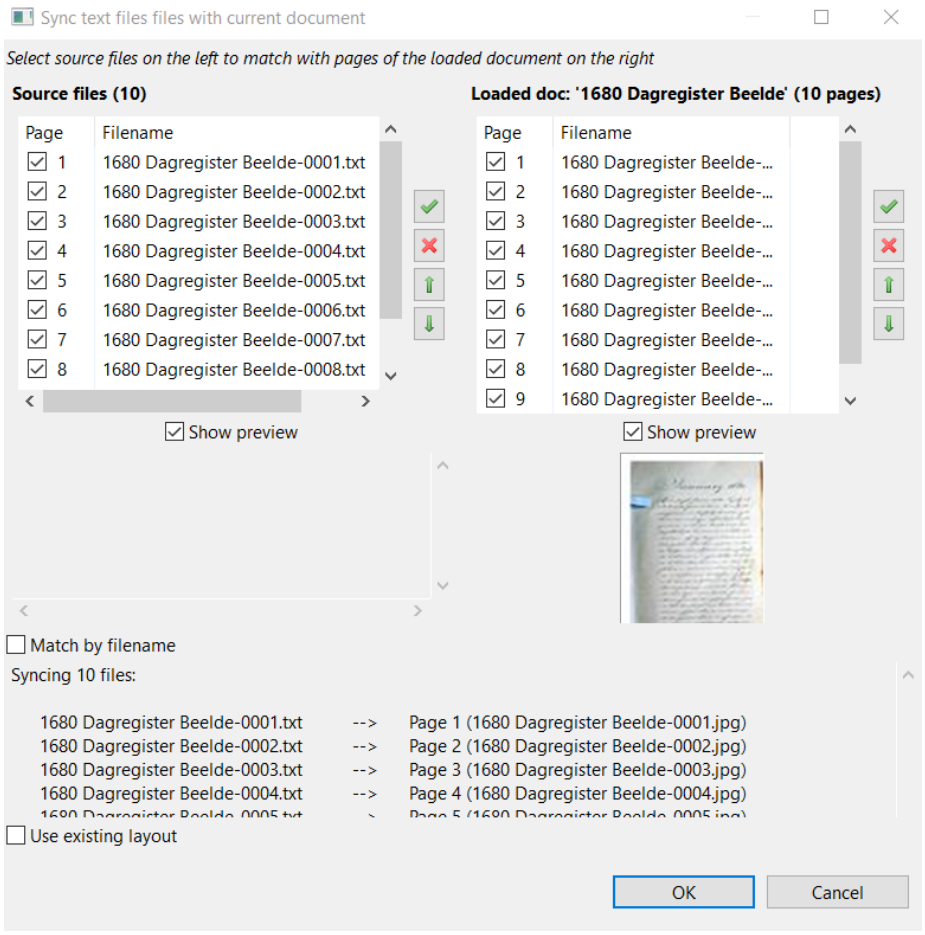

Figure 3 Sync text files with doc

- **„Use existing layout“:** normally the t2i will start a new layout analysis for the document, if you do not want this, you can deselect this option.
    - **Advantages of using the already existing layout:** you can correct it afterwards by bringing the lines to the right position with “Control” and “Enter”.
    - **Risk when creating the layout in the course of the t2i:** it can happen, that lines are missed out.
- **„Match by filename“:** select to synchronize the files by name
- Confirm with “OK”

Matching in Transkribus

- Import the documents into Transkribus with one of the options described above.

- Open the “Tools”-tab in Transkribus. Within the “Other Tools” section you can find the t2i-tool. If you click on it the following window will open:

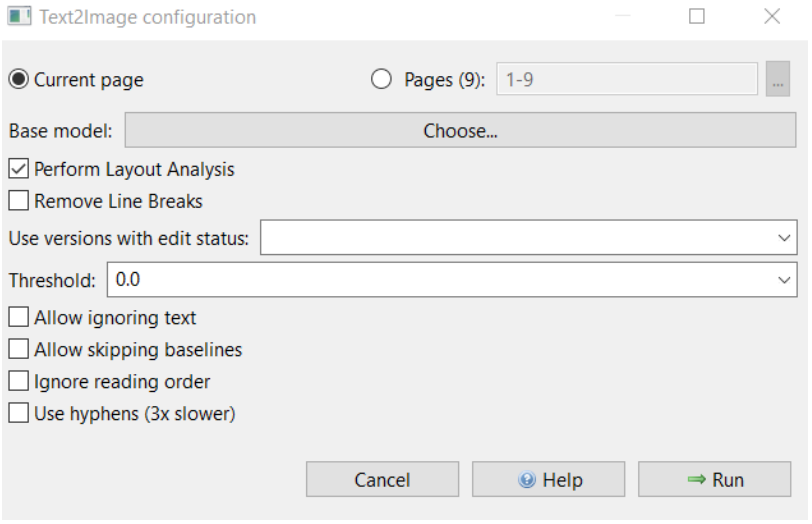

Figure 4 t2i configuration

- Choose the pages, which should be matched.
- **„Base Model“:** choose a suitable base model for the document.
- **„Perform Layout Analysis“:** normally the t2i will start a new layout analysis for the document. If you do not want this, deselect this option.
- **„Remove Line Breaks“:** choose this option, if the line breaks in the text files are not put. With this option you choose if line breaks should be considered or not.
- **„Use versions with edit status“:** in case you do not want to use the latest version of the document for the t2i, you can choose another version here. This option refers to the status, which has been assigned to the document in Transkribus.
- **„Threshold“:** indicates, how strict the accordance should be in order to have a match. The default value is 0.0 because of the fact, that wrong matches can be corrected afterwards quite easily. The lower the threshold value, the more tolerant the matching.
- **„Allow ignoring text“:** if there is text in the text files, which is not represented in the image.
- **„Allow skipping baselines“:** choose this option, if there are missing lines in the text files.
- **„Ignore reading order“**: with this option the t2i will ignore the line-order, which was defined in the course of the layout analysis. This option can be helpful for complicated layouts (e.g. if there is vertical as well as horizontal writing in one document) and for writings, which are read from right to left.
- **„Use hyphens“:** with this options you define, that the following punctuation marks will cause a line break: - = : ¬

Correcting results

- After t2i-processing is finished, wrongly matched lines can be corrected.
- A good way might be to jump from text region to text region and check the first and last line.
- To correct the position of lines you can move them downwards with “Control” and “Enter”, upwards with “Return”, and then you can of course simply delete or add text in the text editor.
- If you would like to delete lines or regions it can be helpful to do this within the “Layout”-tab, where you can find an overview of the layout of the document.

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.