TRANSKRIBUS. Research Infrastructure for the Transcription and Recognition of Historical Documents – Workshop

Günter Mühlberger,
Sebastian Colutto, Philip Kahle
Digitisation and Digital Preservation group
(University of Innsbruck)

Objectives

- Enable humanities scholars
  - to create highly standardized digital scholarly editions of handwritten and printed documents in a Virtual Research Environment
  - to carry out the work in a standardized and transparent way, with or without automation support
  - to involve colleagues and volunteers
  - to export documents in various formats so that they fit to further workflows, e.g. for archives, libraries, digital humanities tools, etc.

Main steps in the workflow

- Capture documents
- Upload documents to Transkribus
- Build your team
- Set up your editorial rules
- Segment images into blocks and lines
- Annotate segmentation
- Run OCR or HTR (if available)
- Transcribe (or correct) text
- Annotate text with Named Entities
- Export text in various formats
- Display documents in a digital library

Capture documents

- The better the scan, the less work afterwards!
  - A good scan is one of the most important success factors for HTR and OCR
- A good scan
  - Has 300 ppi or comes from a good camera
  - Needs not to produce uncompressed TIF files, but JPG or PNG are fully sufficient
  - Has straight lines (horizontally and vertically)
  - Shows the text in the binding and has no warping
  - Is sharp
- Transkribus
  - Has no image correction functionality. Should be done in beforehand.
  - Stores the original file and several working copies.

Upload documents

- Single documents
  - Can be done via Transkribus
  - It is currently not recommended to upload documents larger than 100 images
- Batch upload
  - Will be implemented
  - Currently the simplest solution is to upload them to a file sharing platform (WeTransfer, Dropbox) and we will upload them
  - Collections of documents should be organized by directories

Build your team

- User management
  - All registered users are able to create their own collections and to upload documents into their own collections
  - As collection owner every user is able to add other Transkribus users and to give them roles (owner, editor, transcriber)
- Crowd-users
  - Every collection is also accessible via a web-interface (TSX) which is meant to support especially transcription and encourage involvement of volunteers
  - Volunteers also need to be registered in Transkribus

Set up your editorial rules

- Detailed decisions are needed for every kind of scholarly edition
  - E.g. usage of s and long s, or how to handle person names if they are sometimes written with capital letters in the source document, sometimes not
  - Or how to handle abbreviations and their extensions, etc.
  - Very important for the consistency of the edition
  - Many of these detailed decisions will only appear after the actual work has started
- Editorial declaration (coming soon)
  - A database where these decisions can be documented
  - Values (features) will exported so that the Editorial Declaration is always part of the document

Segment images

- In Transkribus transcribed text always needs to be linked with the image!
  - This is done either automatically (in the case of OCR) or needs an extra workflow step (in the case of transcribing handwritten material)
  - The logic behind is that a “Text Region” has “Line Regions” and that “Line Regions” may have “Baselines”
- For feeding the HTR correct Text Regions and correct Baselines are sufficient! No correction of line regions is necessary!
  - But Baselines should be rather correct = should be near to the actual (virtual) baseline
- Current recommendation for HTR
  - Do Text Region segmentation manually
  - Run “Line and Baseline Segmentation” automatically
  - Correct erroneous baselines manually

Annotate segmentation

- Text blocks can be manually annotated e.g.
  - Page number
  - Header (running title)
  - Footnote
  - Caption
  - …
- For printed text only
  - If OCR text is available page number, running title and footnotes can be annotated with automation support

Run OCR or HTR

- OCR = FineReader 11 SDK
  - Also Fraktur (Gothic) and mixed typefaces (Fraktur/Roman Type Face) are supported!
  - License costs are covered by UIBK
- HTR
  - Training needs to be done on the basis of 100 correctly transcribed and segmented images
  - HTR needs to learn specific scribes – nevertheless it is able to deal with multi-writer documents
  - HTR engine is implemented in Transkribus but currently not made available to the user
  - HTR models will be available to everyone and may be used in the future as starting point for the training of further models
- In the current status we run HTR in the background and provide users with updated documents

Transcribe or correct text

- Once the segmentation step is done text can be either transcribed from scratch or can be loaded (if HTR or OCR was applied)
  - Text is entered on line level
  - Virtual keyboards are available and can be easily extended with own character sets or specific characters
  - What you see is what you get – WYSIWYG, e.g. bold, italic, underlined, strike through, etc. are defined
  - No need for TEI tags!
- Text can also be displayed in TSX so that the transcription can be done in a vey simple way

Annotate text

- Powerful Tagging system
  - A good transcription will not only provide a correct (diplomatic) text but also deal with
    - Abbreviations and extensions
    - Person names
    - Geo names
    - Dates
    - Etc.
  - Text can be annotated, but also normalized (=the same person may have several names within one document or collection) and standardized (=referenced with external sources)

Export text in various formats

- There is more than one format!
  - METS/PAGE XML: contains all the information and can be used for archives and libraries (their digital library systems will usually support the import of such files)
  - TEI: for further working with the transcripts in a digital humanities environment
  - RTF: for simple reading, copying text into scientific papers, exchange with collegues, etc.
  - PDF: for searching in the image and the text of the transcription
    - For reading

Display text in a digital library

- Coming soon
  - Display text in a typical digital library environment
  - Accessible only for those users who are part of the team in Transkribus
  - But documents can also be “published”
  - Main features are searches across documents and collections

Thank you for your attention!