How To Transcribe Documents with Transkribus-
-For training Handwritten Text Recognition technology
-For Scholarly Editions

Transkribus is a platform for the automated recognition, transcription and searching of historical documents, using Handwritten Text Recognition (HTR) technology.

Transcripts generated with Transkribus can be:
- Used to train a neural network (“model”) which is capable of automatically recognising printed or handwritten documents
- Enriched and marked-up to serve as the basis for digital editions of documents.

This introduction enables you to either quickly create training data for the automated recognition of your specific documents or to create a transcription for a scholarly edition.

If you already have transcribed documents available and would like to use them as training data for HTR, please consult our HowToUseExistingTranscriptions guide.

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

Introduction .................................................................................................................. 4
Upload documents to Transkribus ............................................................................. 4
Segmentation .................................................................................................................. 5
  Viewing profiles ......................................................................................................... 5
  Automatically detect text regions, lines and baselines ............................................ 6
  Correcting the results of automated segmentation .................................................. 6
Simple transcription – for HTR training ..................................................................... 9
Train a HTR model ....................................................................................................... 10
Advanced transcription - for a scholarly edition ....................................................... 10
  Reading order ........................................................................................................... 10
  Reading order: Interline additions ............................................................................ 13
  Reading order: Additions as extra notes ................................................................... 14
Transcription and Virtual Keyboards .......................................................................... 16
Diacritics and ligatures ............................................................................................... 17
Punctuation marks ........................................................................................................ 18
References .................................................................................................................. 18
Credits ......................................................................................................................... 19

The READ project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 674943.
Introduction
This guide explains the process of transcribing documents in Transkribus.

These transcripts can be used:

- As training data for a Handwritten Text Recognition (HTR) model which is capable of automatically transcribing your documents.
- As the basis for a digital scholarly edition.

There is a simple three-step process for transcribing a document in Transkribus:

Step 1: Uploading
- Upload your documents to the Transkribus platform

Step 2: Segmentation
- Run the automated segmentation tool to create baselines for your document.

Step 3: Transcription
- Transcribe the text in the segmented lines.

This form of simple transcription is sufficient for training HTR technology. Note: HTR can work on both handwritten and printed documents.

There are also advanced transcription options for those working on scholarly editions. You can adjust the reading order of the text, use historical characters, add tags and metadata, expand abbreviations and more.

Upload documents to Transkribus

- In order to be able to run the necessary tools on your documents they need to reside on the Transkribus server. This means that you need to upload them to Transkribus.
  - Note: All collections and documents in Transkribus are private. Only users authorised by you are able to see your documents. They are not made available to the public.
- To upload click on the “Import Documents” button in the Main menu.

- You have four options:
  - Upload single document from a local folder:
    - This option allows you to upload documents up to 500 MB
  - Upload via FTP
    - This is suitable if you want to upload several large documents
  - Upload via URL of DFG Viewer METS
This allows you to upload documents directly from repositories which support the DFG (Deutsche Forschungsgemeinschaft – German Science Funds) Viewer

- **Extract and upload images from PDF**

  - Figure 2 Select "Upload single document" for documents up to 500 MB

**Segmentation**

- Once you have uploaded your documents to Transkribus, you are ready to start segmentation.
- In order to transcribe your documents in Transkribus, they must be segmented into text regions, lines and baselines.
- For the HTR to work, the text and image need to be connected.

**Viewing profiles**

- Viewing profiles are available to help you with the tasks of segmentation and transcription.
- You can select between viewing profiles for “**Segmentation**” and “**Transcription**” by clicking the “Profiles” button in the Main menu.
- The “Segmentation” profile means that baselines are displayed in red, making it easier to spot any errors resulting from the automated segmentation process.
- The “Transcription” profile means that the Text Editor field will be displayed, allowing you to transcribe your document.
- Of course you can simply use the “default” profile to perform either task.
Automatically detect text regions, lines and baselines
- Select the “Segmentation” viewing profile from the Main menu.
- Select the “Tools” tab on the left side of the screen and go to the “Layout Analysis” section.
- Under “Method:” select “CITlab Advanced”.
- Select if you would like to run the layout analysis only for the current page, for distinct pages, or for the whole document.
- Make sure you select both, “Find Text Regions” and “Find Lines In Text Regions”.
- Click the “Run” button.

![Figure 4 Perform automated segmentation in the “Tools” tab](image)

Correcting the results of automated segmentation
- Note: if you are training a HTR model, the position of text regions does not need to be completely exact and the reading order of the text is not relevant.
- If you are working on a scholarly edition where a higher degree of accuracy is required, it is possible to manually correct the text as in the examples below.
A line has been missed or added by mistake

Figure 5 Add a line to an existing text region

- In the example above the first line had been missed by the program. If you would like to add it to the existing text region:
  o Click inside the region so that it is highlighted.
  o Drag the border of the text region as needed.

A marginal note needs to be split into a separate text region
If you need to split one region into two, you can do this with buttons in the Canvas menu.

- As shown in Figure 6, the “H-button” splits a text region horizontally.
- The “V” button splits a text region vertically.
- The “L-button” allows you to split a text region with customisable line.

Remove a region which is not needed

- In the example above two regions are overlapping, so one can be deleted.
- Click on the text region you wish to delete, and click the red “Remove a shape” button.

Merge two regions

- Sometimes the program creates two text regions where only one is needed. In this case you can easily merge the two together.
- Hold down the “CTRL” button on your keyboard and click on both text regions.
- Click the “Merges the selected shapes” button in the Canvas menu.

**Correct baselines**
- Of course it is also possible to correct the baselines in your document.
- As with the text regions, click on a baseline and you can drag the parts of the line, split a line into two or merge two lines together.
- You can also delete a baseline and draw a new one from scratch. Click the “+BL” button in the Canvas menu. Click once to start drawing your baseline and double-click to finish your line.
- Note: Baselines are most important for HTR; line regions do not need to be corrected.

**Simple transcription – for HTR training**
- Select the “Transcription” viewing profile from the Main menu.
- You will see the Text Editor field below the image: For each line/baseline in the image you will find a corresponding line in the Text Editor. The image and the text are connected in this way.
- Transcribe the text according to the language of your source document. Use the characters of your keyboard.

- **You can have more than one person working on a document but they should not work on the same page simultaneously.** You can let other Transkribus users see your documents by clicking the “User Manager” button in the “Server” tab.

**Train a HTR model**

- If you wish to train a HTR model to recognise your documents, this simple transcription is sufficient.

- We recommend that you start the training process with between 5,000 and 15,000 words (around 25-75 pages) of transcribed material. If you are working with printed rather than handwritten text, a smaller amount of training data is usually required.

- Once you have transcribed enough pages, just drop us a short email (email@transkribus.eu) and we will get back to you about the training of your model. You can also find out how to train a model yourself in the How To Train a Handwritten Text Recognition model guide.

**Advanced transcription - for a scholarly edition**

**Reading order**

- Once a document has been segmented into text regions, lines and baselines, you may need to think about the reading order of the text.

- Many handwritten documents include corrections and additions added by the author, or someone else.

- In a scholarly edition you want to keep the reading order and maybe also express the fact that this text was an addition.

- For this purpose all segmentation elements can be ordered according to a user-defined order.

- The default reading order follows the topology of the text or line regions. All shapes are ordered according to the coordinates of the top left corner of a text or line region.
- This mechanical reading order can be changed:
  o Click on the “Item visibility” button in the Main menu, and you can then choose to show the reading order of text regions, lines, baselines (or words).
Once you choose to show the reading order of text regions or lines, numbers will be displayed on the image of your document.

By clicking on one of the numbers marking the reading order, it is possible to type in a new number and change the reading order accordingly. The same can be done by moving the segmentation elements in the “Layout” tab.

- In cases where the reading order of a page is completely incorrect, it is possible to reorder the text
  - Make the line reading order visible as described above
  - Click on the “Layout” tab on the left side of the screen
  - Select the page or text region that you wish to reorder
  - Click the “R” button
  - The reading order will be rearranged according to the coordinates of the top left corner of a text or line region. After that, the lines should be in right order.
  - There can be issues with the reading order of newspaper columns and similar documents. E.g. the programme assigns a reading order based on the horizontal layout of lines on a page, rather than putting the lines in order by column. To fix this issue, use the “V” button in the Canvas menu to split the text region on the page into separate regions for each column. Once there is a separate text region for each column, the reading order should automatically update and be correct.
Reading order: Interline additions

- Interline additions are a frequent way in which text is added to a document.
- In order to generate the correct reading order, the following steps need to be performed manually:
  - Click the “Item visibility” button in the Main menu
  - Select “Show lines reading order”

  ![Figure 14](image)
  *Click the “Shape Visibility” button, then choose to show baselines and the reading order of lines.*

  - **Select the baseline below the addition** (if the addition is above the line).
  - Split the line region with the “V” button in the Canvas menu exactly where the addition should be logically placed

  ![Figure 15](image)
  *Apply “V” button to split the line region*

- Edit the reading order so that it is correct. Click on the number associated with each line region and then type the correct one.
Reading order: Additions as extra notes
- Additions which appear as extra notes (e.g. at the margins of a page) should be handled in a similar way to interline additions.
  - Note: Often such extra notes (or marginalia) are not part of the reading order but are “comments” and as such are on a different level to the primary reading order.
  - It will therefore be sufficient to mark them as “marginalia” in the Metadata tab. Instructions on marking-up text can be found in the How to enrich transcribed documents with mark-up guide.
- But if the extra note is really an addition to the running text and needs to be added in the reading order then it can be done in the following ways:
  - Option 1: The text region can be expanded so that all baselines of the addition are also part of the respective text region.
    - Note: You can use either rather large text regions, or you may use polygonal text regions. For this purpose select the “Add point to selected shape” button from the Canvas menu.
Following the movement of your mouse pointer you can add points to the original text region and expand the shape so that it also includes the addition.

Afterwards the additional lines/baselines can be renumbered according to their correct reading order.

- **Option 2:** You can generate just one large text region for the whole page and do the line/baseline segmentation manually in the correct order. In this way you will get the correct reading order right from the beginning.
  - Note: this may be the best option if you are dealing with a document which has a sophisticated layout with many additions, notes and deletions.

- **Option 3:** You can connect the extra text region which contains the addition to the line where the addition belongs. To do this, select both text regions and then click the “Links two shapes” button in the “Structural” tab, within the “Metadata” tab.
  - Note: The linking will be part of the XML file but is currently not supported in the export formats.
Transcription and Virtual Keyboards

- A transcription which will serve as a **basis for a scholarly edition** should make more data **explicit to the user and offer more contextual data** than a simple transcription. In this case not only **machine readability** (i.e. training data for the HTR engine) but also **human readability** of the text will play an important role.
- You can add special characters and Unicode symbols using the “**Virtual keyboards**” button in the Text Editor field.
- With the “**Edit...**” button it is possible to add shortcuts for frequently used characters and to add new Unicode characters.
- To create a shortcut, you just need to type it in the “**Shortcut**” column.
- To add new Unicode characters, you use the green plus button.

![Figure 20 Virtual Keyboard](image)
Diacritics and ligatures

- The correct transcription of diacritics and ligatures requires some expert knowledge. There are two main options for handling the correct transcription of these characters:
  - Option 1: **Slight normalisation** according to dictionary
    - The main rule to be applied here is the following: As long as you can clearly see the base character of a glyph and as long as the base character is also the one which is used in the dictionary to express this glyph, keep to the base character.
    - **Example 1**: LATIN SMALL LETTER Y will appear in many documents with an extra diacritical sign, indicating the history of this character coming from ii or ij. Therefore you find two dots or a something similar looking above the “y”.
    - ![Figure 22 German Kurrent Script: “bey”. Note: y is written as LATIN SMALL LETTER Y since the base character is still clearly visible](image)
    - In simple transcripts you will transcribe this as LATIN SMALL LETTER Y since the base character is clearly visible.
    - **Example 2**: LATIN SMALL LETTER S is expressed with two graphemes in most European historical scripts. We find therefore a clear distinction between LATIN SMALL LETTER S and LATIN SMALL LETTER LONG S.
But although there is a clear distinction, a simple transcription would use LATIN SMALL LETTER S in both cases.

**Option 2: Palaeographic Transcription**

- Philologists or palaeographers are not only interested in the correct transcription, but also in the historical appearance and development of graphemes. Therefore it might also be interesting to transcribe the above examples with full support of the Unicode character set or even by utilizing the private area of Unicode.

Note: Please take into account that this is an important decision and will affect the usability of the text in many ways. If you decide to go for a palaeographic transcription it will cause a lot more work than with a slightly normalized transcription.

**Note**: In printed texts (which can also be transcribed in Transkribus) the transcription of ligatures may play a role. Again the same rule can be applied: Though specific combinations of letters, such as “ft” are expressed with a specific grapheme where two graphemes are matched together, and though such ligatures can also be expressed with specific Unicode letters, we recommend transcribing them according to the dictionary.

**Punctuation marks**

- Punctuation marks are transcribed in the same way as characters. Use the appropriate character on your keyboard and do not normalize or add punctuation marks. Typical punctuation marks are:
  - modern characters such as dot, comma, semicolon, colon: ".", ",", ";":
  - historical characters such as virgule (slash), or line fillers, etc.
  - Note: Colons in historical texts are often used to mark abbreviated words. These should be transcribed as a colon.

- In contrast to many transcription rules where punctuation marks are added and omitted according to a modern understanding we recommend to keep to the original punctuation marks.

- If you want to add punctuation marks which do not appear in the original document you may use the “supplied” tag in the “Tagging” tab, within the “Metadata” tab to indicate that the punctuation mark was added by yourself.

**References**

To get an overview on scripts from Unicode: [http://www.unicode.org/charts/](http://www.unicode.org/charts/)

For historical transcriptions the following extensions are of interest:

- Contains e.g.:
  - Non-European and historic Latin
  - Phonetic and historic letters
  - Additions for Slovenian and Croatian
  - etc.

Latin Extended-C: [http://www.unicode.org/charts/PDF/U2C60.pdf](http://www.unicode.org/charts/PDF/U2C60.pdf)

- Contains e.g.:
  - Orthographic Latin additions
  - etc.


- Contains e.g.:
  - Medievalist additions
  - Insular and Celtic letters
  - Ancient Roman epigraphic letters
  - etc.

MUFI (Medieval Unicode Font Initiative)

- This initiative has collected and systematized about 1512 characters which are especially recommended for the transcription of medieval documents. Note: Some of them are still in the “private” section of Unicode, therefore not officially available.
- [http://folk.uib.no/hnooh/mufi/](http://folk.uib.no/hnooh/mufi/)
- [http://folk.uib.no/hnooh/mufi/specs/MUFI-Alphabetic-4-0.pdf](http://folk.uib.no/hnooh/mufi/specs/MUFI-Alphabetic-4-0.pdf)

**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.