

Transkribus (/transkribus/) >

How-To Guides (https://readcoop.eu/transkribus/resources/how-to-guides/) >

How To Transcribe Documents with Transkribus - Introduction

How To Transcribe Documents with Transkribus - Introduction

♀ Basic overview ☐ Transkribus Expert Client

Last update 6 months ago

« How-To's overview (https://readcoop.eu/transkribus/resources/how-to-guides/)

About Transkribus ()

Table of Contents

^

How To Transcribe Documents with Transkribus - Introduction

Introduction

- 1. Upload documents to Transkribus
- 2. Segmentation Layout Analysis

Viewing profiles

Automatically detect text regions, lines and baselines

Correcting the results of automated segmentation

A line has been missed or added by mistake

A marginal note needs to be split into a separate text region

Remove a region which is not needed

Merge two regions

Correct baselines

3. Transcription

Simple transcription - for HTR training

Advanced transcription – for a scholarly edition

Diacritics and ligatures

Punctuation marks

Working in a team - adding other users to your collection



(https://readcoop.eu)



This guide explains how to transcribe documents with Transkribus to create training data for the automated recognition of your specific documents or to create a transcription for a scholarly edition.

If you want to have a more general overview, have a look at our 10-Steps Guide (https://readcoop.eu/transkribus/howto/use-transkribus-in-10-steps/).

Transkribus is a platform for the automated recognition, transcription and searching of historical documents using Handwritten Text Recognition (HTR) technology.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

- Used to train a Handwritten Text Recognition (HTR) model, which is capable of automatically recognising printed or handwritten documents;
- · As the basis for digital scholarly editions.

If you already have transcribed documents available and would like to use them as training data for HTR, please consult our How To Use Exisiting Transcriptions (https://readcoop.eu/transkribus/howto/how-to-use-existing-transcriptions-to-train-a-handwritten-text-recognition-model/) guide.

Introduction

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

- 1. Uploading: upload your documents to the Transkribus platform;
- 2. Segmentation: run the automated segmentation tool to create baselines for your document.;
- 3. Transcription: transcribe the text in the segmented lines.

This form of simple transcription is sufficient for training Handwritten Text Recognition (HTR) technology. Note that HTR can work on both handwritten and printed documents. The efficiency of a model will depend on the quality of the training material (your manual transcription), the quality of the images and on how neat or messy the writing is.

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.

1. Republic ad documents to Transkribus



(https://deadcoepsile) 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.

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.

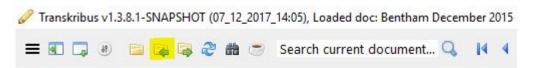


Figure 1. Upload files to your personal collection

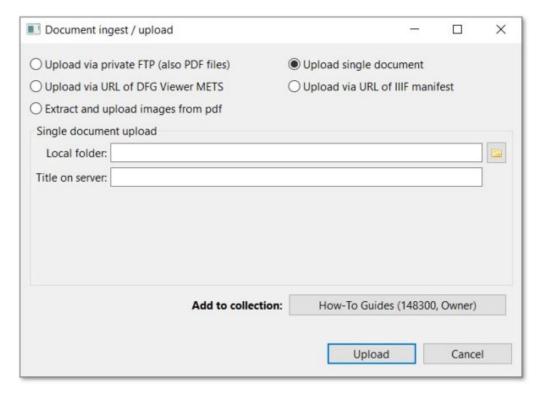


Figure 2. Select "Upload single document" for documents up to 500 MB

You have five options to upload documents:

1. Upload single document from a local folder:

This option allows you to upload documents up to 500 MB. In order to choose this option, choose "Upload Single Document". Please make that the files to be uploaded reside in an extra folder. When choosing the files for the upload, you won't be able to see the files in the folder. That is normal in this case. Just mark the folder and confirm with "OK".

REPAID via FTP:

c This sulutable if you want to upload several large documents. You can upload image (https://filesigoopyellips PDF-documents with this option. Please make sure the PDF-files are not inside a folder when using the FTP-upload.

3. 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.

4. Upload via URL of IIIF manifest:

Insert the URL of the IIIF manifest in the provided field and click "Upload".

5. Extract and upload images from PDF:

This option is suitable for images you want to upload which reside in a PDF-document. If those PDF-documents are of big size, please use the FTP-upload. Also, if you have JP2000 images in your PDFs it makes sense to go for the FTP-option.

To add pages to an already existing document in Transribus: load the document you would like to add pages to in Transkribus. Open the Document Manager and select the document again by clicking on its name in the "Document Manager"-window. Click on the green circle icon next to "Add new page(s)" and search and add the new pages via the directory.

To delete documents from your collection: select the document in the collection overview within the "Server"-tab. Click on the folder icon with the small red circle "Delete the selected documents from Transkribus". The deleted document(s) will reside in the recycle bin (icon "contains deleted documents") for two weeks. If you have deleted a document by mistake, you can contact us (info@readcoop.eu), and we will be able to reactivate the document within these two weeks. After this, the document will be permanently deleted.

2. Segmentation - Layout Analysis

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 and baselines, and for the HTR to work, the **text and image need to be connected**.

All the segmented elements, such as print space, text region, line region or baseline, are stored in the PAGE file with their coordinates.





(https://repdofdpseane 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.

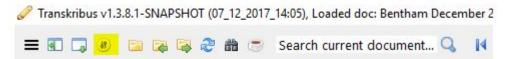
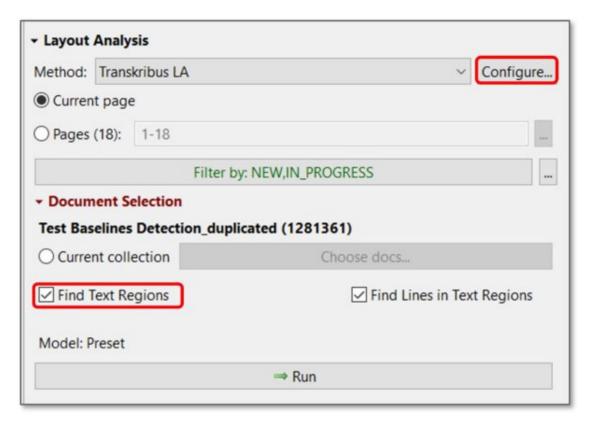


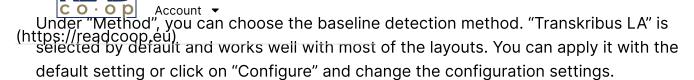
Figure 3. Viewing profiles for segmentation and transcription tasks

Automatically detect text regions, lines and baselines

To automatically run the Layout Analysis, go to the "Tools" tab in the Managing & Tools Bar (on the left side of the screen). The section we are interested in is named "Layout Analysis".



Figration in the "Tools" tab



=

In the Layout Analysis Configuration window, the setting you can configure are:

- <u>Model</u>: leave the "Preset" model if you haven't trained a specific baselines model on the layout of your documents.
 - The Preset Transkribus LA model works well for most document typologies. Only if your documents have a complex layout and the preset model is unsatisfactory, you can train a Baselines model specific to your document typology, as explained here (https://readcoop.eu/transkribus/howto/how-to-train-baseline-models-in-transkribus/).
- <u>Minimal baseline length</u>: it indicates the minimum length of baselines in pixels. Baselines shorter than this length will not be detected.
- Baseline accuracy threshold: in the first stage of the layout analysis, each pixel is labelled as baseline, separator or other. The baseline accuracy threshold applies to the baseline labelling at this stage. It ranges between 0 and 255, and higher values enforce higher accuracy in the detected baselines.
 If you have low-resolution images and no or only a few baselines are detected, try to reduce the value. Bear in mind, however, that the results can get noisy for lower thresholds.
- <u>Separator threshold</u>: separators are small vertical lines drawn beside each baseline; they mark the beginning and end of each baseline (do not confuse them with actual separators in printed document images). As for the baseline accuracy threshold, the separator threshold refers to the first stage, when pixels are labelled.
 The separator threshold ranges between 0 and 255: 0 means that separators are not used at all; with a higher value, separators are used, thus, nearby baselines tend not to be merged.
 - Usually, low values are sufficient to prevent a connection between nearby baselines. Use, for instance, 1 to use separator information "sometimes" and larger values to use them pretty much all the time, for instance, when text lines are close together but have to be separated because belonging to different columns.
- <u>Max-dist for merging</u>: in the second stage, the algorithm tries to merge nearby baselines but only when their distance is smaller than the set value. The value is not measured in pixels but is a fraction of the image's width. By default, it is set to 0.01:

Property of baselines are closer than the 0.01 fraction of the image width, they will be perged if Abeyrare more distant than this value, they will not be merged. According to the image's width, you can increase the fraction value to merge more distant lines or reduce it to prevent nearby baselines from being merged together.

<u>Max-dist for clustering</u>: this value refers to the text region creation: after the baselines are detected, they are clustered in text regions based on their distance. The Max-distance for clustering is a fraction of the image's width: baselines that are closer than this fraction are clustered together in a text region.
 If too many text regions are created with the default settings, you can try to increase the value so that more baselines are clustered together. If it is set to -1, no region clustering will be performed, and only one text region will be produced as the bounding box of all lines.

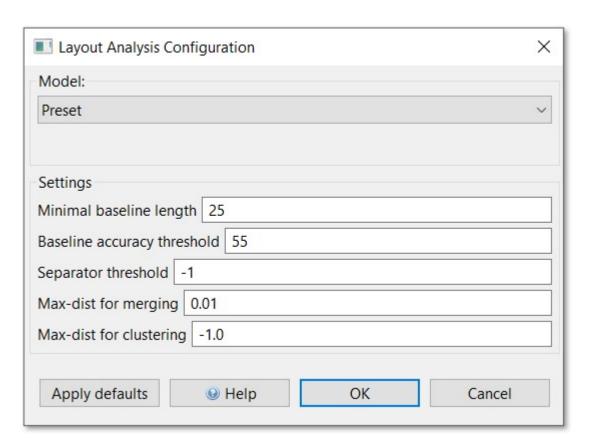


Figure 5. Layout Analysis Configuration

For more information about the Transkribus LA algorithm and setting, consult this page (https://readcoop.eu/transkribus/docu/layout-analysis-help/).

To start the automatic Layout Analysis, select if you want to process only the current page, distinct pages, or the whole document. Make sure "Find Text Regions" is selected and click the "Run" button.

If y draw the text regions by hand and then search for the baselines in these regions; until k the find Text Regions"-option before starting the layout analysis.



(https://readcoop.eu)

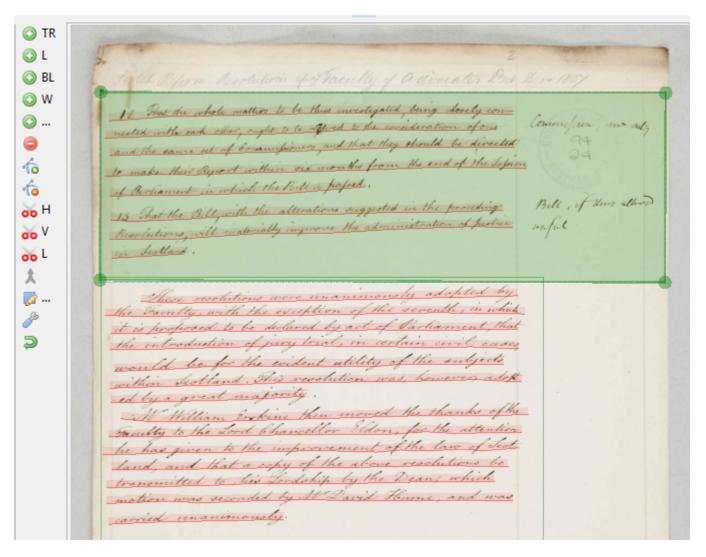
Correcting the results of automated segmentation

It could happen that the automatic Layout Analysis needs some manual correction because some baselines are missing or you want to merge/slip the text regions.

If you are training a HTR model, the text regions do not need to be corrected, and the reading order of the text is not relevant. What is important is that the characters of the line rest upon the baseline and the descenders extend below and that there is a correspondence between the line in the image and the transcribed line.

All tools for corrections on the layout analysis can be found in the "Canvas"-menu left of the image. You can check their functionality by hovering over the icon.

A line has been missed or added by mistake



ig refined 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 (https://readcoop.eu) it to the existing text region, click inside the region so that it is highlighted and drag the border of the text region as needed. To draw the baseline, click the "+BL" button in the Canvas menu: click once to start drawing your baseline and double-click to finish your line.

A marginal note needs to be split into a separate text region

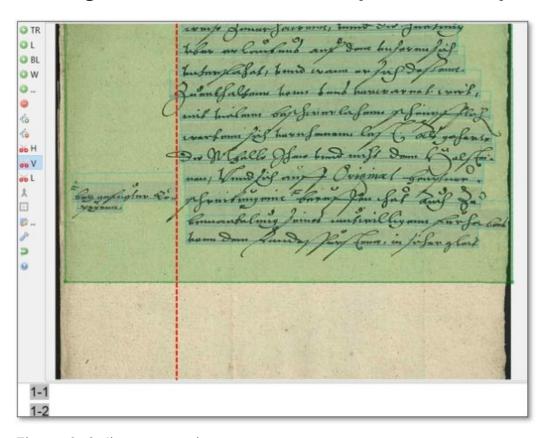


Figure 6. Split a text region

If you need to split one region into two, you can do this with buttons in the Canvas menu. 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 a customisable line. Remember always first to select the text region you want to split.

Remove a region which is not needed

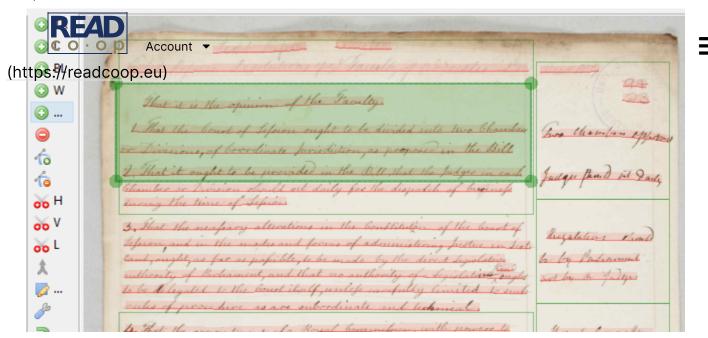


Figure 7. Remove region

In the example above, two regions overlap, 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.

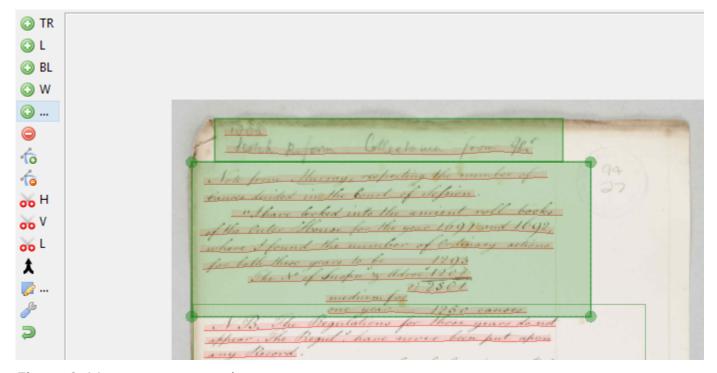


Figure 8. Merge two text regions





(https://leadcitop.ego 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.

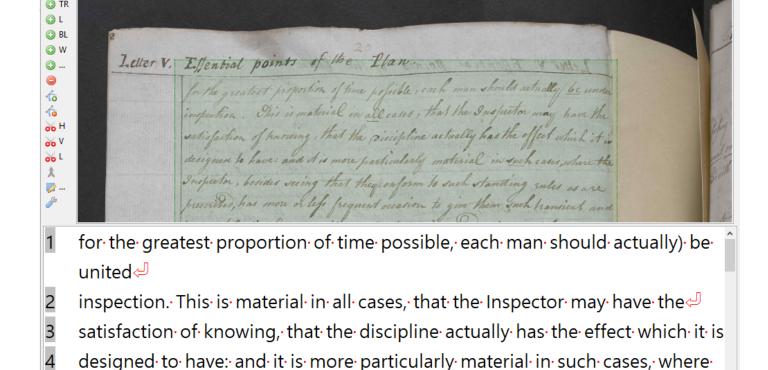
You can also delete a baseline and draw a new one from scratch. Click the "+BL" button in the Canvas menu. Click once to start drawing your baseline, and double-click to finish your line.

3. Transcription

Simple transcription – for HTR training

To transcribe your document, 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.



Inspector, besides seeing that they conform to such standing rules as are

the⊄

Fig scribe your document

You can have more than one person working on a document, but they should not (https://readcoop.eu) simultaneously work on the same page. You can let other Transkribus users see your documents by clicking the "User Manager" button in the "Server" tab.

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. Read here (https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/) how to train your HTR model.

Advanced transcription – for a scholarly edition

Once a document has been segmented into text regions, lines and baselines, you may need to think about the **reading order** of the text (this is not relevant if the transcription should only serve as training material). 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.

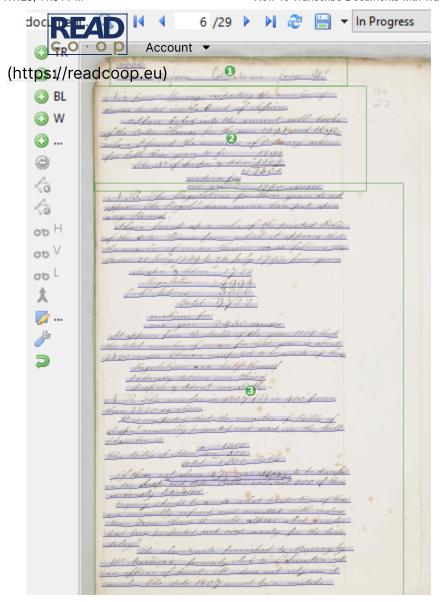
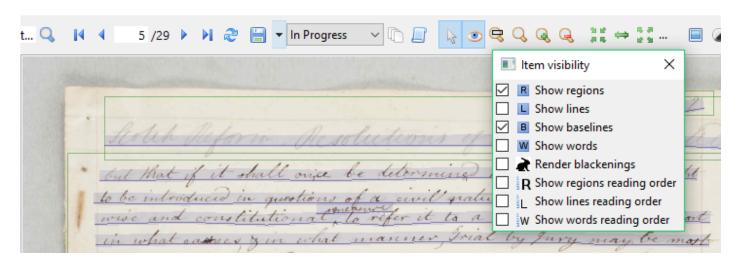


Figure 10. Reading order of text regions – numbers can be reordered

This mechanical reading order can be changed: 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).



Figration visibility" button displays the logical order of segmentation elements

Once you choose to show the reading order of text regions or lines, numbers will be (https://readcoop.eu) 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.

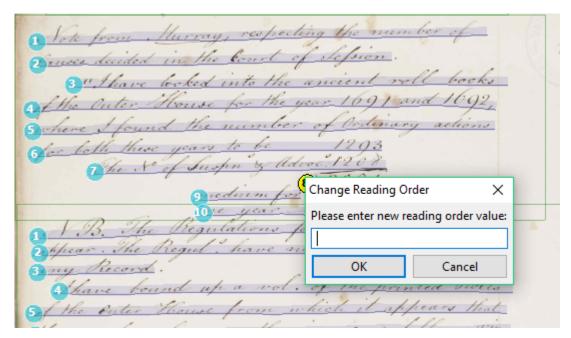


Figure 12. Edit the reading order by clicking on the digit and entering a new number

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 the 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.



Figure 13. Set reading order according to coordinates

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 and select "Show lines reading order" (as explained above)
- 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

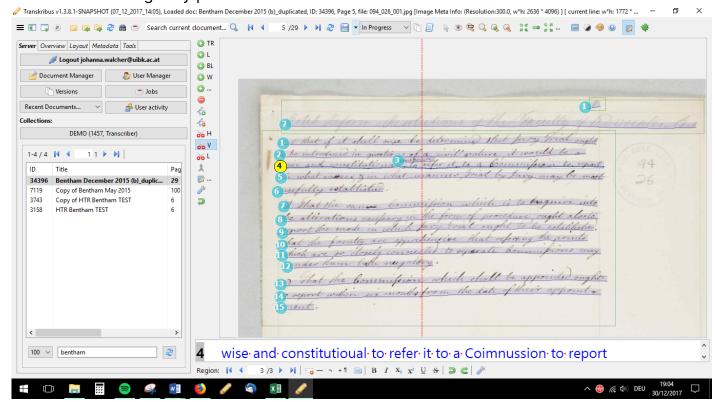


Figure 15. Apply the "V" button to split the line region

 Edit the reading order so that it is correct. Click on the number associated with each baseline and then type the correct one.

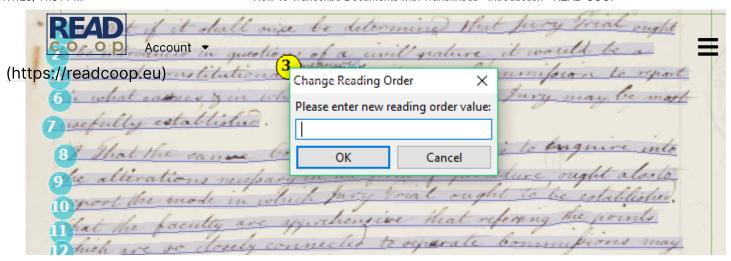


Figure 16. Add correct reading order: 4 (=first part of the line) becomes 3; 3 (=interline addition) becomes 4 and 5 (second part of the line) stays as 5.



Figure 17. Correct reading order after manual editing

Additions which appear as **extra notes** (e.g. at the margins of a page) should be handled in a similar way to interline additions. There are three options to deal with marginal notes:

- Option 1: The text region can be expanded so that all baselines of the addition are also part of the respective text region. You can use either rather large rectangular 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. 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 that the linking will be part of the XML file (PAGE) but is currently not supported in the other export formats.



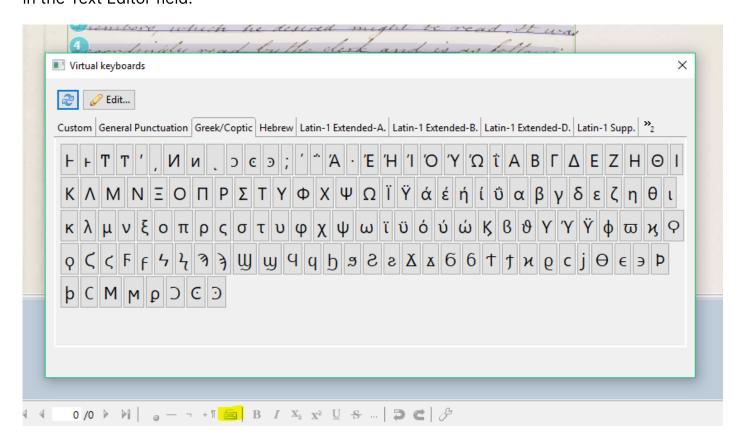
Figure 19. Link two shapes

up/) guide.

If 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 (https://readcoop.eu/transkribus/howto/how-to-enrich-transcribed-documents-with-mark-

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.



tual Keyboard

With the "Edit..." button, it is possible to add shortcuts for frequently used characters and to (https://readcoop.eu) 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.

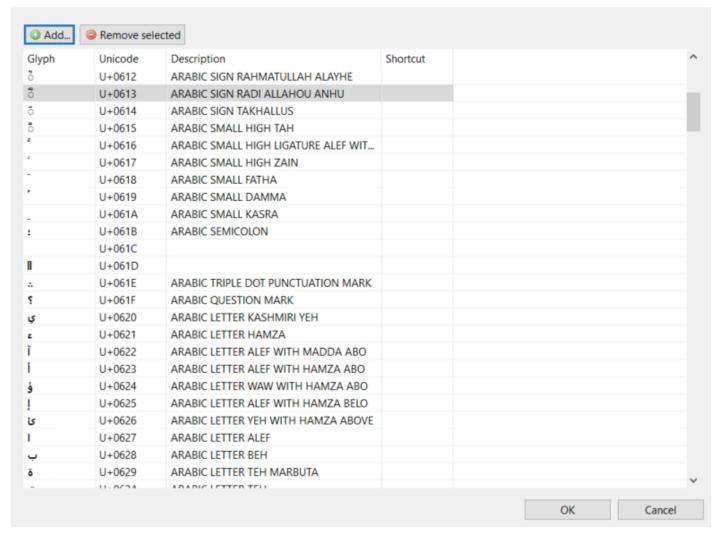


Figure 21. Adding Unicode characters and shortcuts

In the text editor, you can use "Backspace" to move text one line up and "Ctrl" + "Return" to move text one line down.

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:

1. Slight normalisation according to the 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.

(https://deacretical.eigh, indicating the history of this character coming from ii or ij. Therefore you find two dots or a something similar looking above the "y".

In simple transcripts, you will transcribe this as LATIN SMALL LETTER Y since the base character is clearly visible.

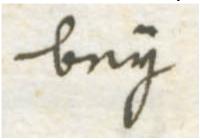
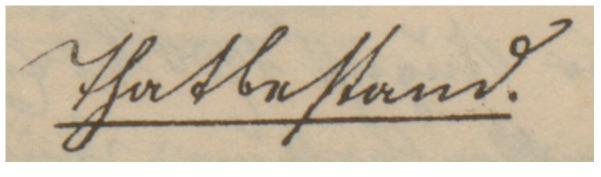


Figure 22. German Kurrent Script: "bey". Note: y is written as LATIN SMALL LETTER Y since the base character is still clearly visible

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.



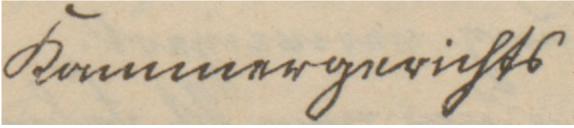
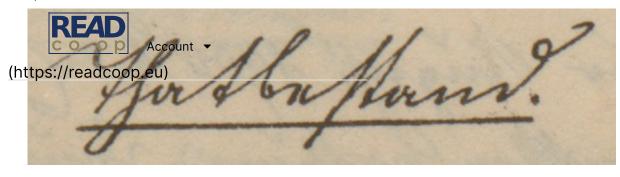


Figure 23. "Thatbestand." vs. "Kammergerichts": LATIN SMALL LETTER LONG S vs. LATIN SMALL LETTER S



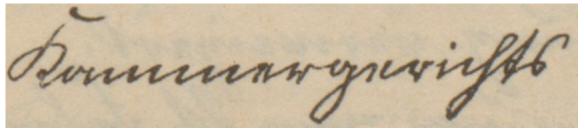


Figure 24. Palaeographic transcription: Thatbeltand vs. Kammergerichts

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 much more work than slightly normalized transcription.

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 that 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 keeping 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.

WREAD gin a team – adding other users to



In Transkribus, it is also possible to work on collections and documents together with other Transkribus users. You can add somebody else to your collection via the "User Manager" to be found in the "Server Tab". First, you will need to search for the other user via email or name down right, then select the right line above, then choose "Add user" down left and then finally add the authorisations, which come with the user role. In the screenshot below, you can check the rights of each user role:

	Reader	Crowd Transcriber	Transcriber	Editor	Collection Owner	Document Owner/Uploader	Comment
View Documents	✓	√	√	✓	✓	-	
Transcribe	X	✓	✓	1	√	B	Edit & Save Layout/Text
Delete Page Transcript	X	✓	√	✓	✓	-	
Add/Remove Credits	X	✓	✓	✓	✓	-	Credit packages must be owned by user
Export Docs.	X	✓	✓	✓	✓	5	
Run HTR/LA	X	✓	✓	1	√	<u>+</u>	
Start Training	X	√	✓	✓	✓	-	
View Coll. Users	X	✓	✓	√	✓	-	
Manage Coll. Users	X	X	X	✓	✓	H	Add/Remove users in collection, change roles
Duplicate Docs.	X	X	X	1	√	<u>+</u>	
Edit Doc. Metadata	X	X	X	√	✓	-	
Edit Collection Metadata	X	X	X	✓	✓	-	
Edit Doc. Pagination	X	X	X	✓	✓	5	Add/Move/Delete pages in documents, replace image of a page
Ingest Docs.	X	X	X	1	√	<u>+</u>	
Add/Remove HTR models	X	X	X	✓	✓	-	
Set transcription status Final/GT	X	X	X	✓	√		
Link/Unlink Docs.	X	X	X	X	✓	✓	
Manage Coll. Owners	X	X	X	X	√	81	Assign/unassign "Owner" role
Delete collection	X	X	Χ .	X	✓	-	
Delete document	X	X	X	X	X	✓	
Subscribe to Crowd Collections	✓	✓	✓	✓	✓	7	User becomes "Crowd Transcriber" after subscribing
Unsubscribe from Crowd Collections	X	✓	X	X	X	_	

Figure 25. User roles

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:

Latin Extended-B: http://www.unicode.org/charts/PDF/U0180.pdf (http://www.unicode.org/charts/PDF/U0180.pdf)

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



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

Latin Extended-D: http://www.unicode.org/charts/PDF/UA720.pdf (http://www.unicode.org/charts/PDF/UA720.pdf)

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

The COOP

About us (https://readcoop.eu/about/)

Join us! (https://readcoop.eu/join/)

Our Members (https://readcoop.eu/members/)

Success Stories (https://readcoop.eu/success-stories/)

Work with us (https://readcoop.eu/work-with-us/)



(https://readcoop.eu)

Products & Services

Transkribus (https://readcoop.eu/transkribus/)

Transkribus lite (https://transkribus.eu/lite)

Read&search (https://readcoop.eu/readsearch/)

ScanTent (https://readcoop.eu/scantent/)

Useful information

News (https://readcoop.eu/news/)

Download Transkribus (https://readcoop.eu/transkribus/download/)

Public Models (https://readcoop.eu/transkribus/public-models/)

Payment and shipping (https://readcoop.eu/payment-and-shipping/)

Helpful resources

Resource center (https://readcoop.eu/transkribus/resources/)

How-to Guides (https://readcoop.eu/transkribus/resources/how-to-guides/)

Getting started with Transkribus (https://readcooff:tt/transkribus/start/)

FAQs (https://readcoop.eu/transkribus/questlons/)/W

Videos (https://readcoop.eu/transkribus/resdwces/Wdellowit

Copyright © 2023 READ-COOP SCE



Terms & conditions (https://readcoop.eu/terms-and-conditions/).

Account Privacy Policy (https://readcoop.eu/privacy-policy/)

(https://readcoop.eu)

Contact (https://readcoop.eu/contact/)

Imprint (https://readcoop.eu/imprint/)

ΕN

