How to use TRANSKRIBUS – a very first manual

A simple standard workflow for humanities scholars and volunteers (screenshots below)

0.1.7, 2015-06-15

1. **Introduction**
    a. Transkribus is an **expert tool**. As with other feature-rich software it is designed to meet the needs of users who "know what to do and how".
    b. Our main objective is to support you in being able to generate high-level scientific transcriptions of handwritten (and printed) documents based on state-of-the-art technology in the most efficient and user friendly way. We believe we are on a good way, but still a lot of things need to be done. So **this is still work in progress.**
    c. Be aware that it will take you some time until you explore all options and get familiar with the behaviour of Transkribus. Of course we are happy to support you in the best way we can (don't be shy in contacting us, or use the bug report and feature request button within Transkribus).

2. **Basics**
    a. You can only transcribe text **if there are text regions and lines segmented on the image**. In this way a direct link between your transcribed text and the image is generated. (The nice effect: In this way also an automated Handwritten Text Recognition engine can be trained which may support you after a while in transcribing a document). Therefore every document in Transkribus first needs to go through a "segmentation" or "layout analysis" step. Note: the segmentation is either part of an automated process, or can be done manually (more below).
    b. Though for many users a good **TEI (Text Encoding Initiative) document** will be the most important "end-product" and though Transkribus supports you in generating such a TEI, our philosophy ist that TEI is **just one format among others**, and that there are other formats for other user groups and purposes which need to be supported as well, such as **METS/ALTO, PAGE, RTF, PDF**, etc. Transkribus is therefore more than a TEI editor, but also less (we will not support all peculiarities of TEI but just those which are necessary to create a good, standardized transcription).

3. **Register at the website**
    a. http://transkribus.eu/
    b. The University of Innsbruck is offering this service as a research infrastructure.
    c. Read our **User agreement** (will soon come in English!) – we will respect your privacy and use the data only to **improve our**

services and **support research in humanities and computer science**!

4. **Download TRANSKRIBUS from the website**
   a. Our **expert tool** allows you to work with your documents in a professional way.
   b. Download the tool – it is Java based and was tested on Windows, MacOS and Linux. The minimum Java version required is 1.7.
      Note: Unzip the file – you cannot start the tool from the zip File.
   c. Start the tool with
      i. Transkribus.exe (as Windows user)
      ii. Transkribus.command (as Mac user) or
      iii. Transkribus.sh (as Linux user).

      Notes for Mac and Linux users:
      You may have to add the executable permission flag to the start script. To this end, open up a terminal, change into the program folder and type the following command:

      chmod u+x Transkribus.command (or Transkribus.sh on Linux)

      In the future, we are planning to provide simpler start programs for MacOS and Linux too.

5. **Some sample documents**
   a. In order to give you a feeling of the behaviour of TRANSKRIBUS right from the start, we have prepared some test documents in the **Transkribus Cloud Collection**. Play around, these documents are just for fun.
   b. Use the “Login Button” (second button at the top left) to login and access documents from the TRANSKRIBUS server.
   c. Open the Transkribus Cloud Collection and double click on one of the documents
   d. HTR Reichsgericht
      i. A document from the early 20th century from Germany. Recognized automatically with the HTR engine. Probably the first document with Kurrent script.
   e. Bentham Test 1 – Test 3
      i. Several documents from the Bentham collection, all recognized automatically.
   f. OCR Sample Document
      i. Several pages with Gothic Font recognized with ABBYY FineReader 11.

6. **Work with own documents locally**
   a. Transkribus supports you to work with your own documents as well.
      i. You may open your own document in local mode using the “Open local folder button” (folder icon in the top toolbar). The images of a document must be contained in a folder.

      ii. The program automatically creates two sub-directories:
         - 'page' (XML files) containing the transcription and segmentation information

- 'thumbs' (thumbnails) containing thumbnail versions of your images

Also, a “metadata” XML file containing some basic metadata of the document is created.

iii. Note: Several operations of Transkribus cannot be carried out in local mode, e.g. most “Tools” will run only in the server mode, due to the several different architectures and preprocessing steps the tools require.

iv. You can start working offline and upload the document to the server later

7. Digitisation of documents

a. You can either scan a document yourself, or use images which you have downloaded from the Internet.
b. If you scan a document yourself, use a **flatbed scanner or an office copying machine**. Their quality is usually sufficient for all operations!
c. But take care that the scans are complete, that no pages are missing, that the borders are as straight as possible, no warping at the binding, no shadows of fingers…
d. The image format is not problematic, you can work with **300 dpi, JPEG compressed**, no need for large TIFF files.
e. You may also **download documents from the Internet**. Many libraries and archives follow Open Access policies and are therefore encouraging further usage of their collections – just ask archives and libraries directly!

8. Upload your documents to your private collection

a. Use the **“Upload” button** ('Documents' tab; the button with a folder and a small plus) in TRANSKRIBUS to transfer the images from your computer to the platform.
b. Note, that the images have to reside in a separate folder!
c. Be aware that the upload of several hundred megabytes **may be difficult due to our (currently) rather slow server connection.**
d. You may use a file sharing system, such as **Dropbox or WeTransfer** and afterwards contact us directly, respectively send us the link. We will include your documents into your private collection.
e. When uploading a document create your own collection. Only users who are authorised by you will be able to access your documents, it is you who has **full control** on all your documents.
f. Use the “**Collection Manager**” (pen button in the 'Documents' tab beside the collection selector) to add users to your collection. You can only invite users who are already registered for TRANSKRIBUS.

9. **Segment your document into text blocks and baselines**

a. Segmentation / Layout Analysis

The transcription process requires a so-called '**segmentation**' (i.e. layout analysis) of the page. In this way, a direct link between your transcribed text and the image is generated and an HTR (Handwritten Text Recognition) engine can be trained.

The segmentation is structured **hierarchically**. Each page consists of the following elements:

Text Regions (0 or more)
|---------- Text Lines (0 or more)
|---------- Baseline (0 or 1)
|---------- Words (0 or more)

Text regions are the top elements of each page. They are typically created for each coherent part of a text (e.g. paragraph or heading).

Inside each text region, **text lines** must be created for each line of the text.
Additionally, a **baseline** can be specified for each line.

Important note*:* to simplify the process of line generation it is possible to draw the baseline (which is enough information for the HTR engine) directly into the text regions and a parent text line will be created automatically. Note, that the shape of the text line is not important to the HTR engine so you don't have to correct those automatically created lines if they do not perfectly match the line in the image.

Inside each text line, **words** can then be segmented to link each individual word of the text with the image. This process is not necessary for a transcription however and can be carried out automatically after the transcription on line level is completed.

The segmentation can be done either **manually** using the buttons on the top centre toolbar: 'TR' for text region, 'L' for line, 'BL' for baseline, 'W' for word. You can decide whether to create a shape as rectangle (default) or polygon. Click once into the image to draw a point and press enter or double-click to finish the shape.

It is also possible to use the **automatic** 'Layout Analysis' tools in the 'Tools' tab on the right (however, all automatic tools produce errors that the user will have to correct)

Layout Analysis Tools:

1. **Detect regions**
    a. Detects blocks of handwritten or printed text on a page. Runs well with simple layout, has problems with more sophisticated layout.
    b. Usually it is better to just draw the region by hand.
    c.
2. **Detect lines and baselines**
    a. Detects lines and baselines of a text region in one step. For the transcription process, only the baselines are needed, the line regions do not have to be perfect!
    b. Runs well with straight lines, may cause problems with short lines and long ascenders

and descenders. Benefits a lot from good text regions (manually drawn)

3. **Detect baselines**
    a. In some rare cases there might be already line regions, with this tool the baselines are added.
    b. Usually not necessary.

Each time you start one of these tool a new **job** will be performed on the server – you can see the list of jobs using the 'Jobs' tab on the left. When the tool has finished, the corresponding page will automatically reload and display the result.

Note, that although the segmentation process might feel complicated and annoying the first time round, it can be carried out rather quickly once you got the “best practise”. A 100 percent manual segmentation of a typical page using text regions and baselines takes about 2-5 minutes which is quite fast compared with the effort of actually transcribing the text.
Besides to the input to the HTR, the user also benefits in another way from a segmentation as a full text search can display search results more accurately (e.g. searching in a PDF will highlight the exact matching words).

10. Start your transcription

a. Once there are text blocks and (base-)lines visible on your image you are able to **write text into the text-field**.
b. Display of image and text are synchronised this will make it easier for your eyes when transcribing the text: if you click on a line in the image area, the corresponding line in the text-field will be highlighted and vice versa. Double-click on either a shape in the image area or a line in the transcription widget will focus the selected line or region.
c. Use the **Structure tab** on the left hand side to navigate through the page, one click highlights the element, double-click zooms the element.
d. Special characters can be found in the „Virtual Keyboard“ on the right hand side. You can configure the virtual keyboards using the “virtualKeyboards.xml” file in your program folder. There you can either specify single special characters directly or a whole Unicode range.

11. Save and export your transcription

a. Press the “Save” button to save a page. When working online, each previous version of the file is maintained and can be recovered. Different versions can be accessed via the 'Versions' tab on the left.
b. You are able to export the whole document at any time of the process using the export button at the top toolbar (folder icon with green arrow pointing to the right).
c. In the upcoming dialog you can choose between different versions to export. In addition to the export of images including PAGE XML transcription files and a METS file, we currently also support **RTF**

(Rich Text Format), **PDF** (with the text in the background) or **TEI** (Text Encoding Initiative).
Note: RTF and TEI export are currently in alpha stadium with not much functionality available. The PDF export is more advanced but also still in development mode too.

12. Use of automated Handwritten Text Recognition (HTR) for your documents

a. Currently the recognition of documents is an offline process. But you can rather simply trigger this process: **Transcribe at least 100 pages of your document and segment the remaining pages into blocks and baselines – the machine will do the rest!**
b. Once you have transcribed 100 pages – just drop us an email, we will care about the training of the HTR engine and apply it to your document.

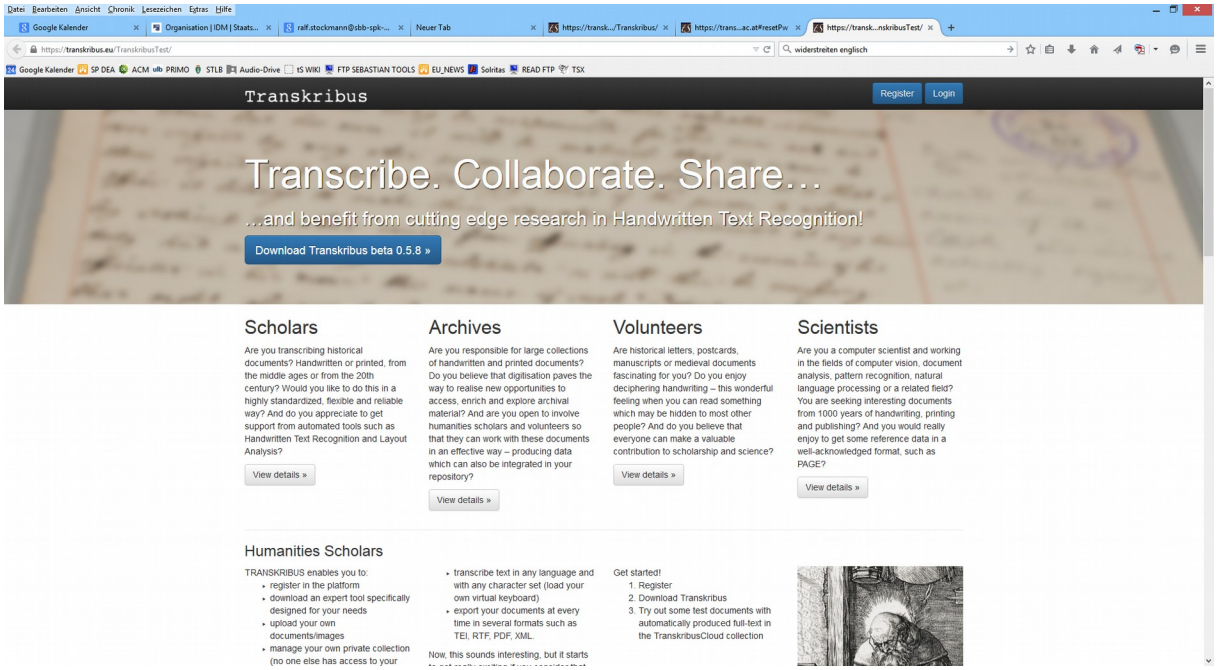
13. Advanced uses

a. Collection Manager
    i. Here you can manage not only the users of your documents, but also create new collections, or make documents available to other collections, etc.
b. Viewing settings
    i. You can change the thickness and density of the borders of segments in the image canvas. Click on the “Home” button and select “Change Viewing Settings”.
c. Structural mark-up
    i. Documents consist of page numbers, headings, footnotes, etc. You may be interested to tag text blocks using the 'Metadata' tab on the right. Just select a segmentation element in the image area and select or type the structure type of the element.
d. Tagging
    i. You may also be interested to tag parts of the text with specific tags, such as „person“, „date“, „textStyle” or „landscape“. This can be accomplished using the 'Tagging' tab on the right hand side.
    ii. To add a tag, first select the corresponding text in the transcription widget. Then select the tag from the list of tags in the tagging widget, edit its properties on the bottom and press the add tag button (a green plus beside the tag) to add the tag to the transcription. It will then be highlighted inside the transcription widget with the colour shown in the Tags table.
    iii. To delete a specific tag, place the cursor inside its definition in the transcription widget. It will then be listed inside the “Tags under cursor” table in the tagging widget where you can delete it using the delete button in the table.
    iv. You can also add new tags to the list inside the program.
    v. The list of persistent tags, including their properties is defined in the 'config.properties' file in the program directory. The attribute 'tagNames' there is a list of predefined tags including a list of their attributes in curly braces

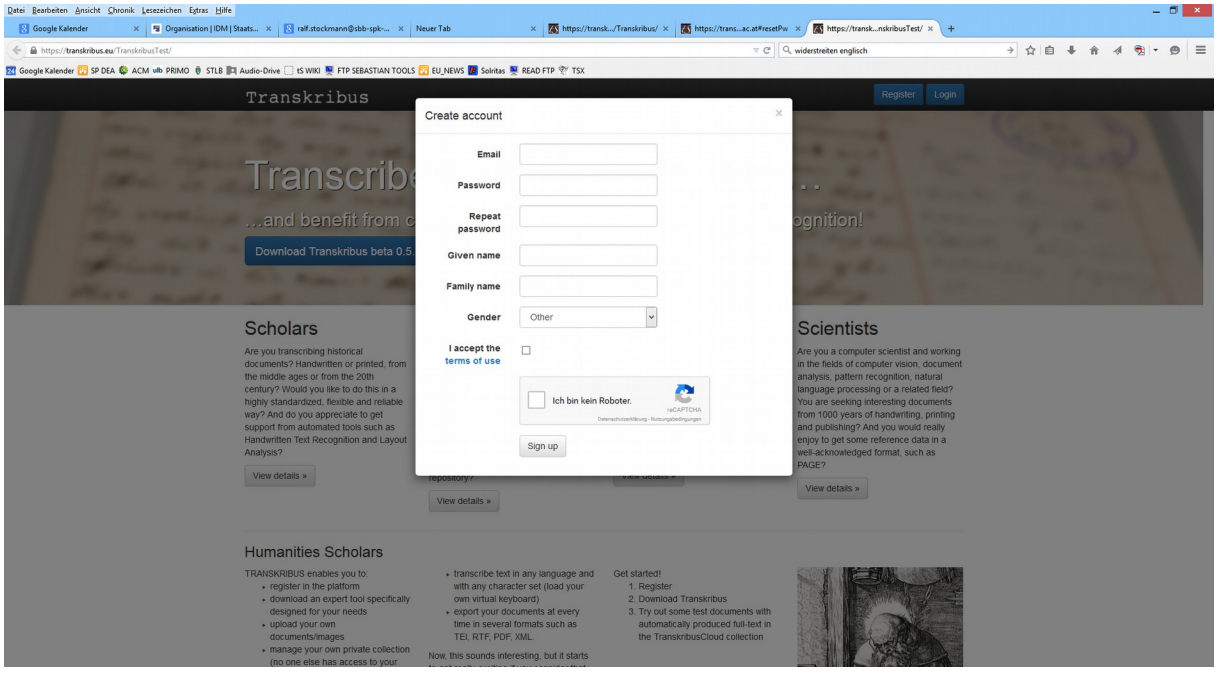
e. Configuration files
    i. In the Transkribus folder you will find the “config.properties” and the “virtualKeyboards.xml” files which you can configure for your purposes.

Screenshots

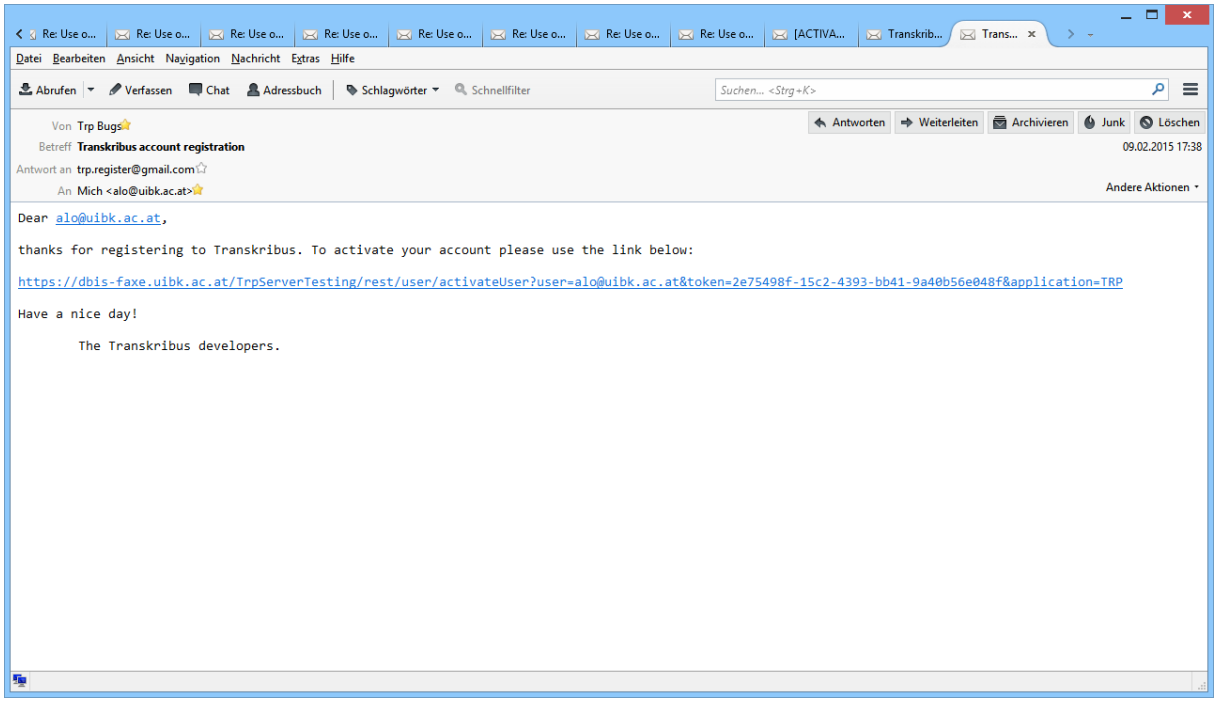
Website: http://transkribus.eu/



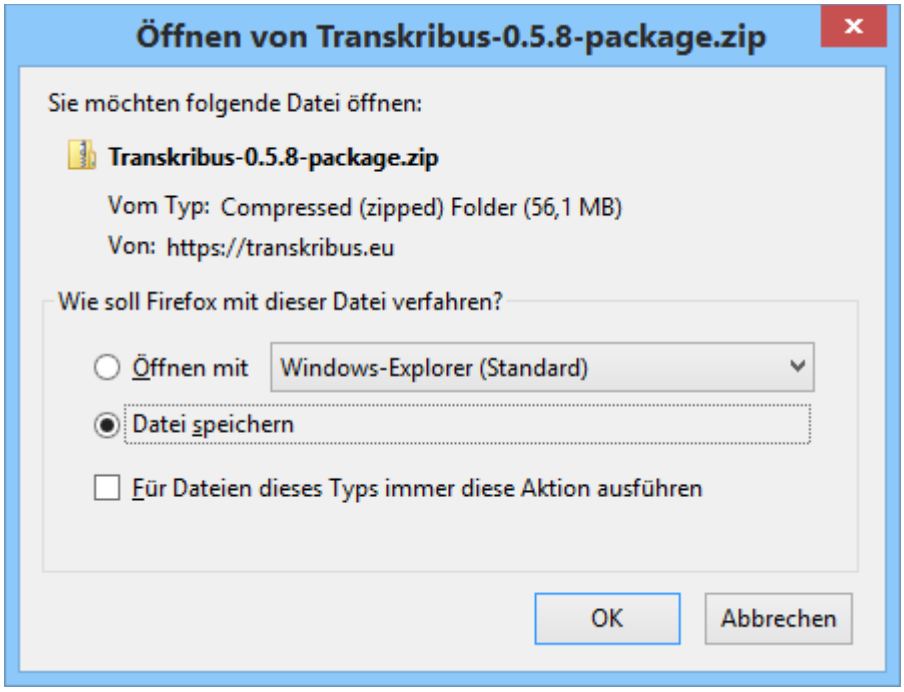
Register



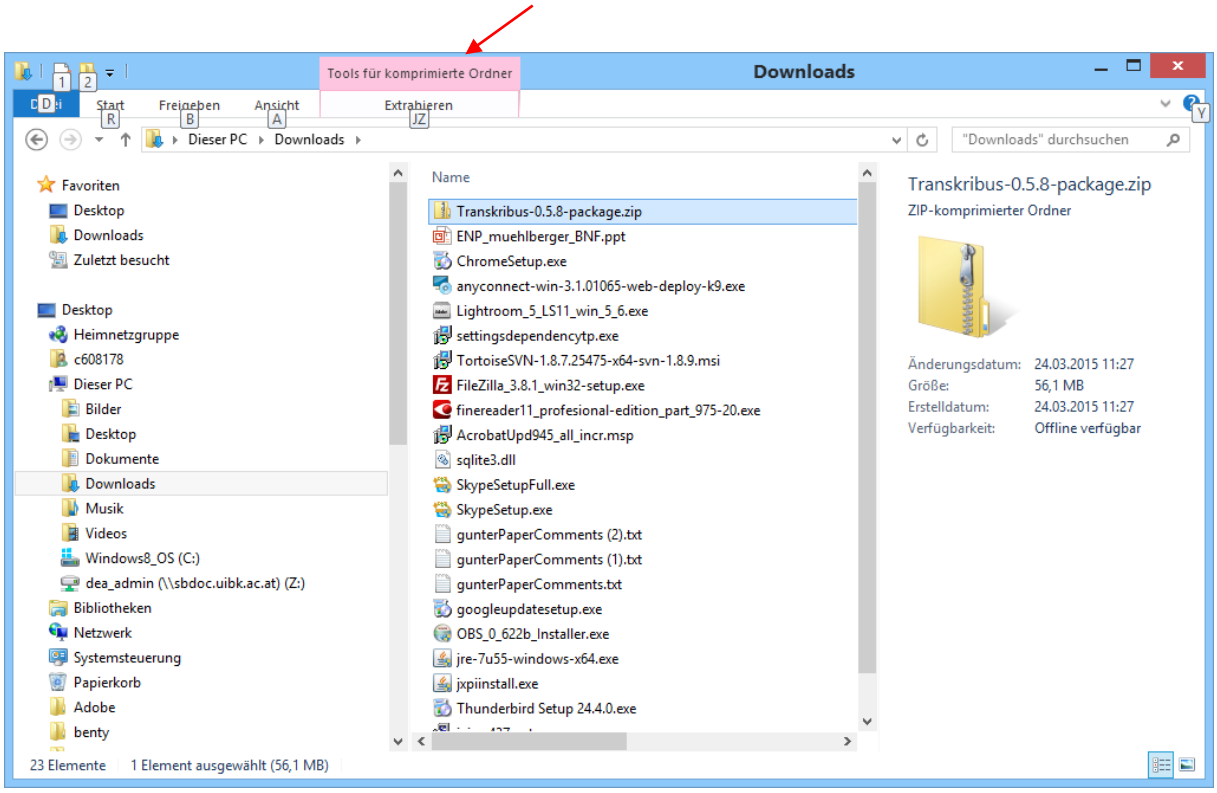
Activate your account



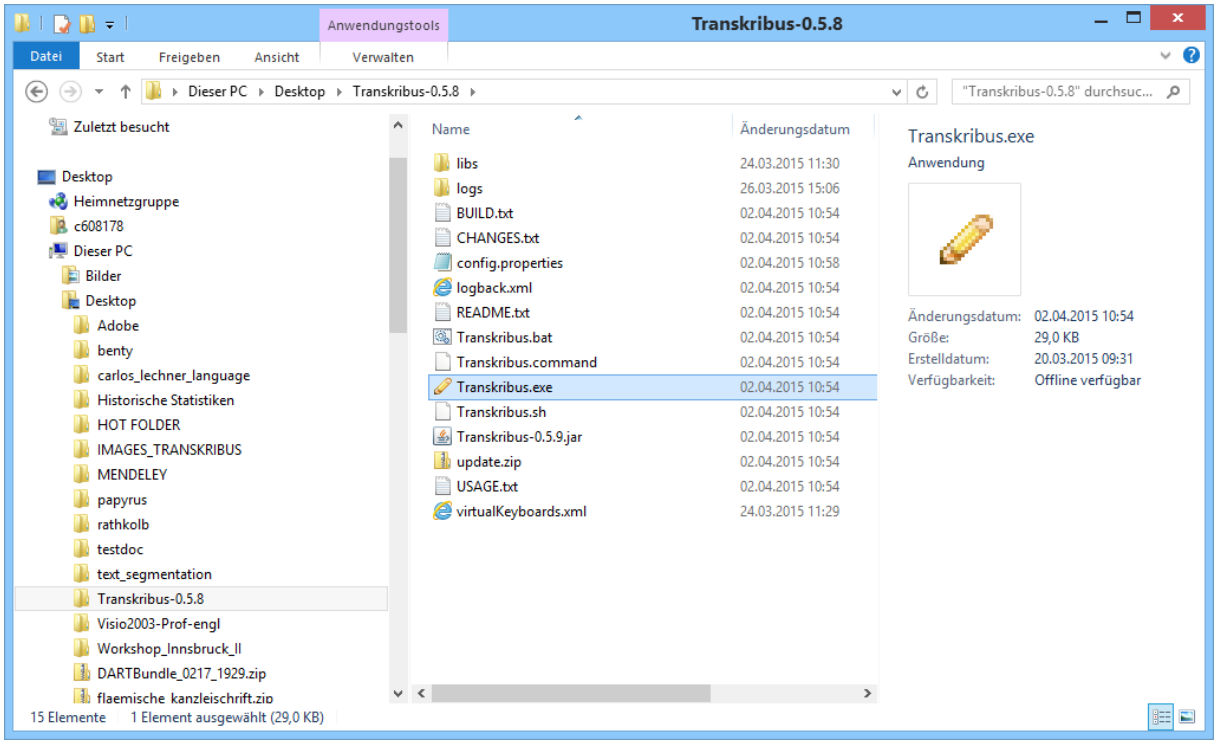
Download Transkribus



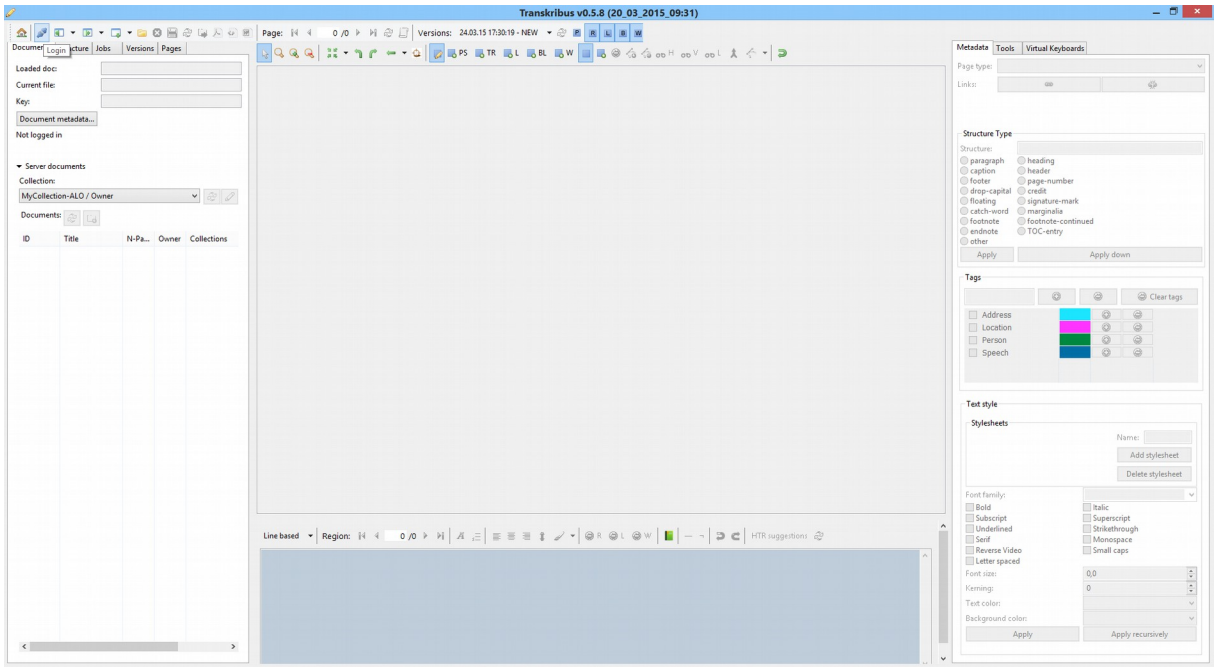
Unzip

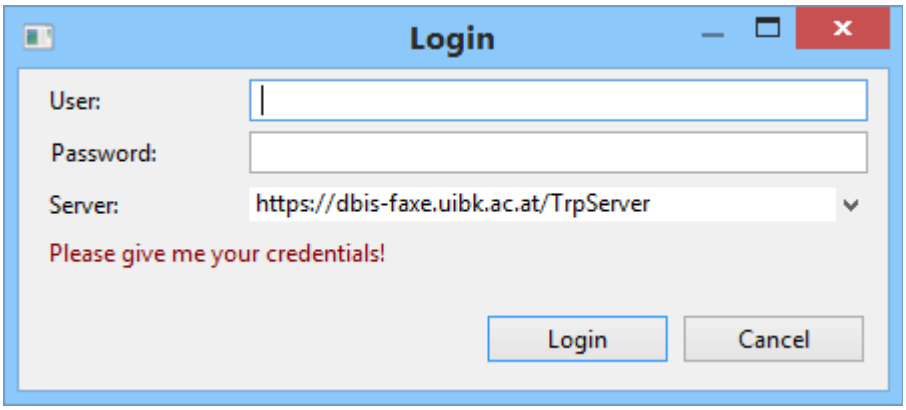


Start with .exe (Windows), .command (Mac) or .sh (Linux)

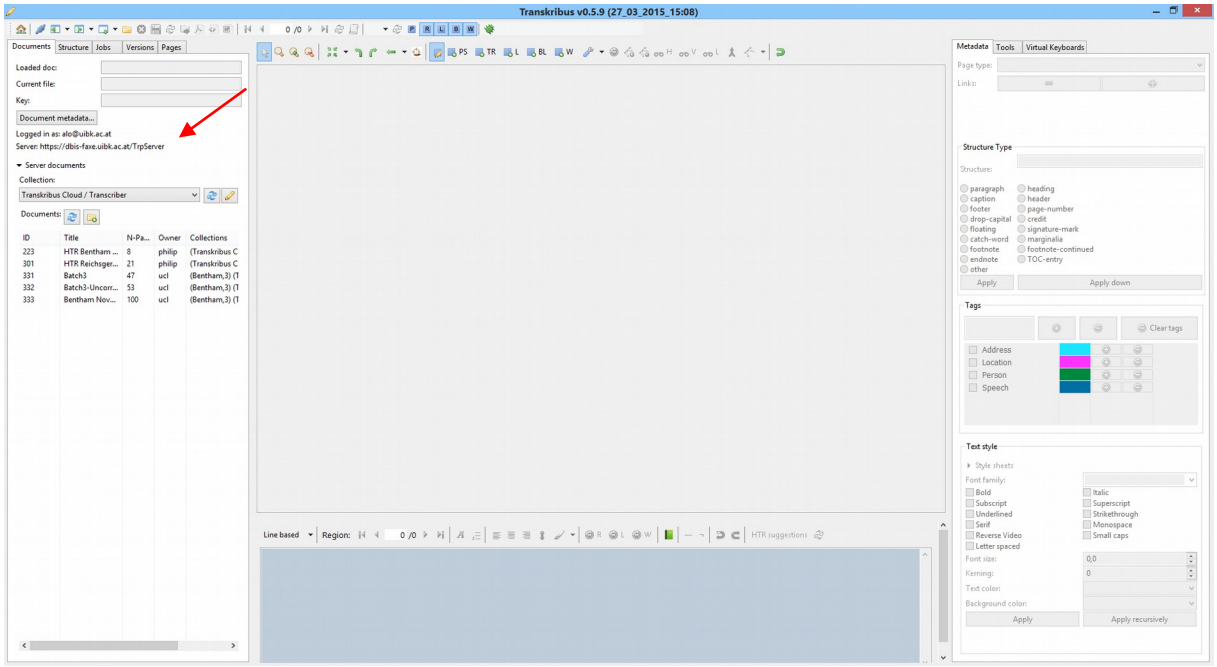


Login

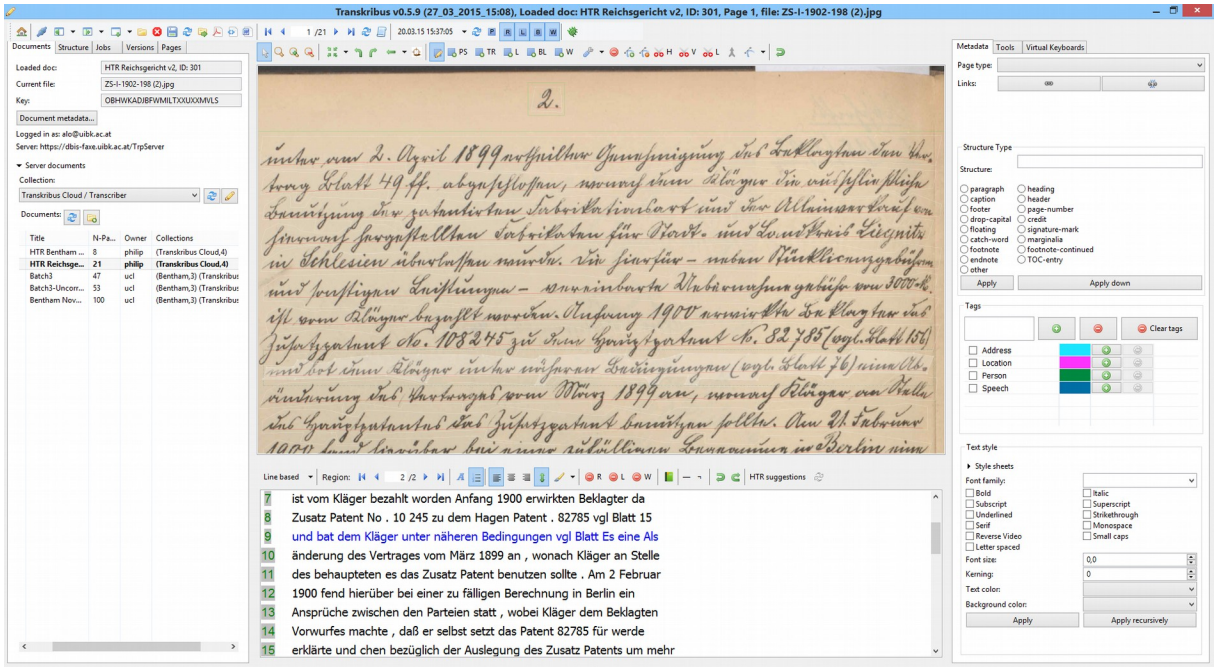

Transkribus v0.5.8 (20_03_2015_09:31)
Login
Loaded doc:
Current file:
Key:
Document metadata...
Not logged in
Server documents
Collection:
MyCollection-ALO / Owner
Documents:
ID
Title
N-Pa...
Owner
Collections
Metadata
Tools
Virtual Keyboards
Structure Type
Tags
Address
Location
Person
Speech
Text style
Stylesheets


Login
User:
Password:
Server:
https://dbis-faxe.uibk.ac.at/TrpServer
Please give me your credentials!
Login
Cancel

Cloud Collection

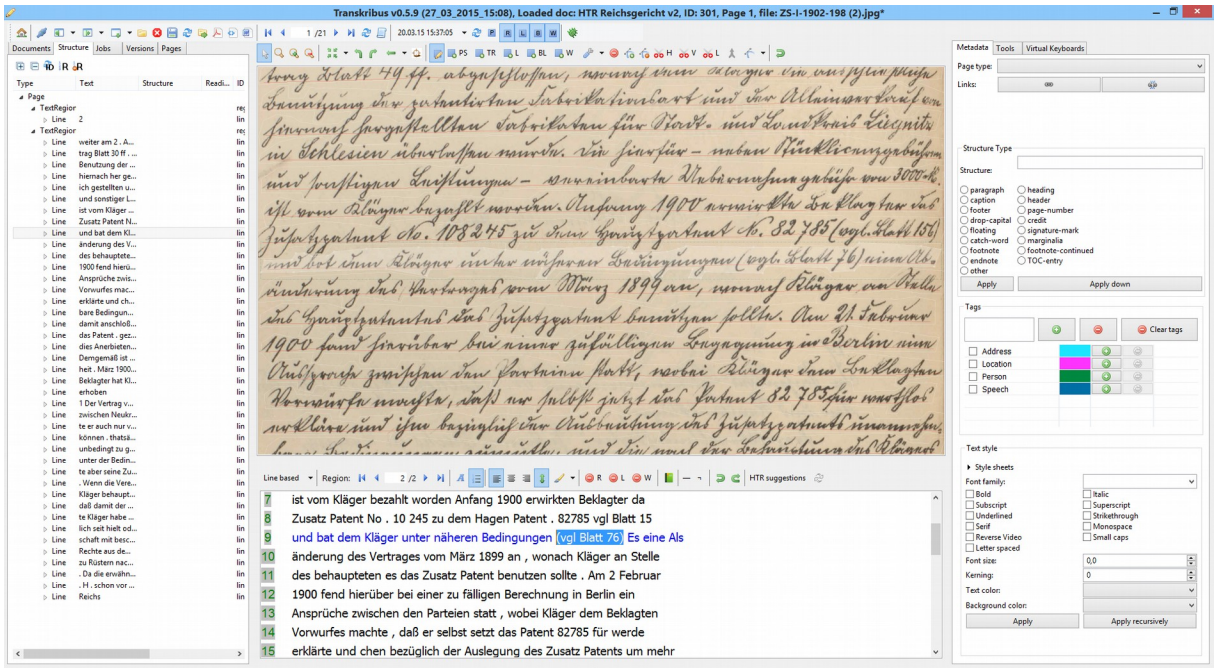


Cloud Collection – View Sample Documents – Text was generated automatically!

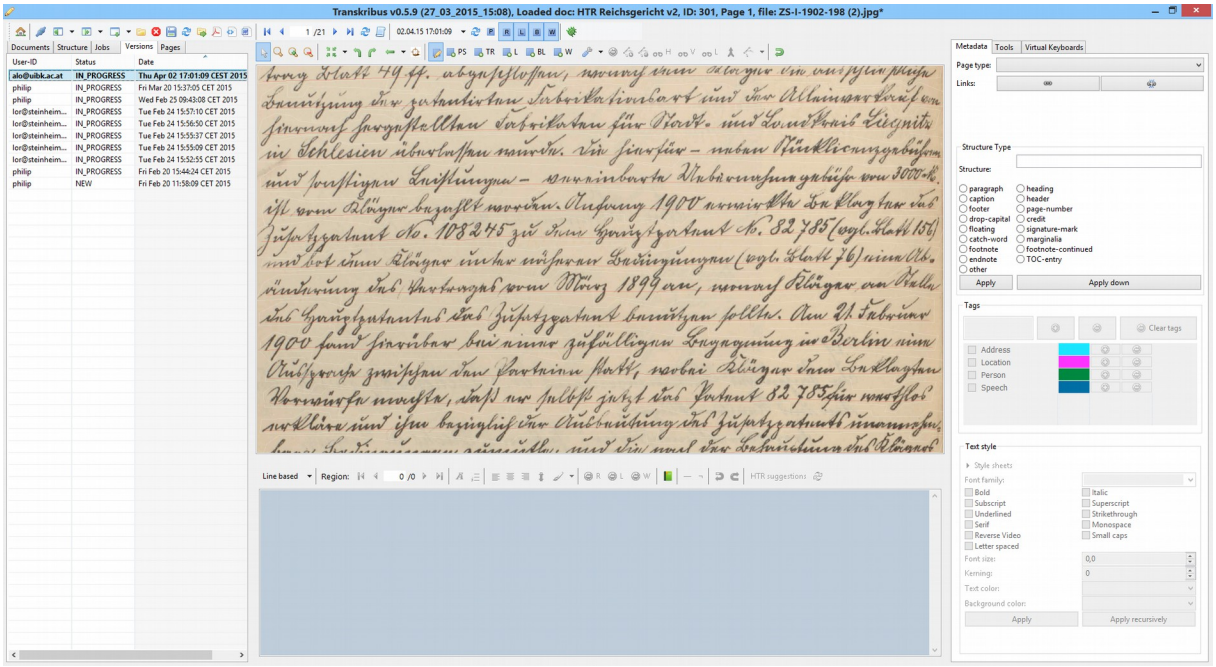


Use “Structure Tab” to navigate, double click to zoom the line and to highlight it in the text window.

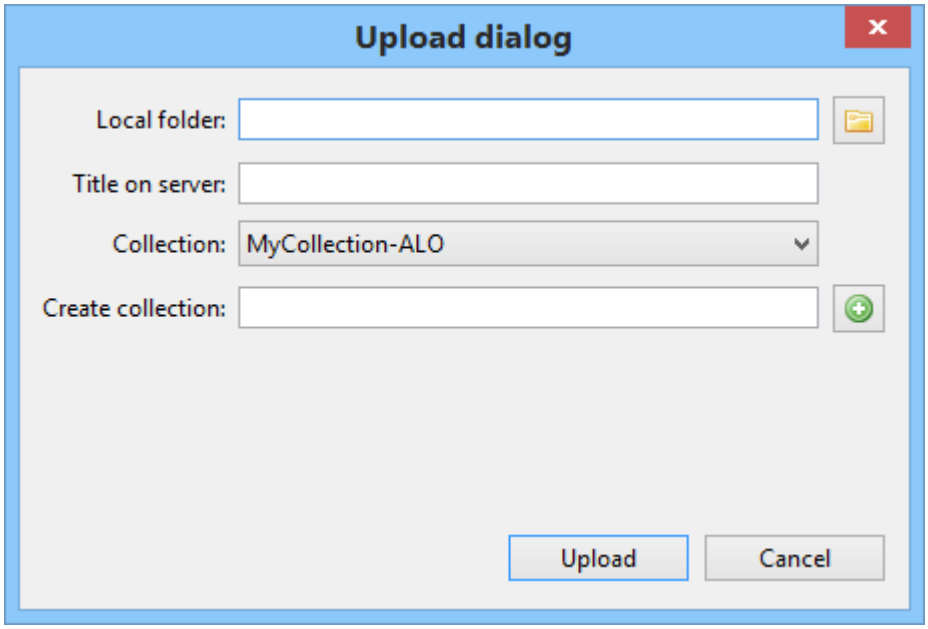
Correct text in the text window, “save” button will generate a new version



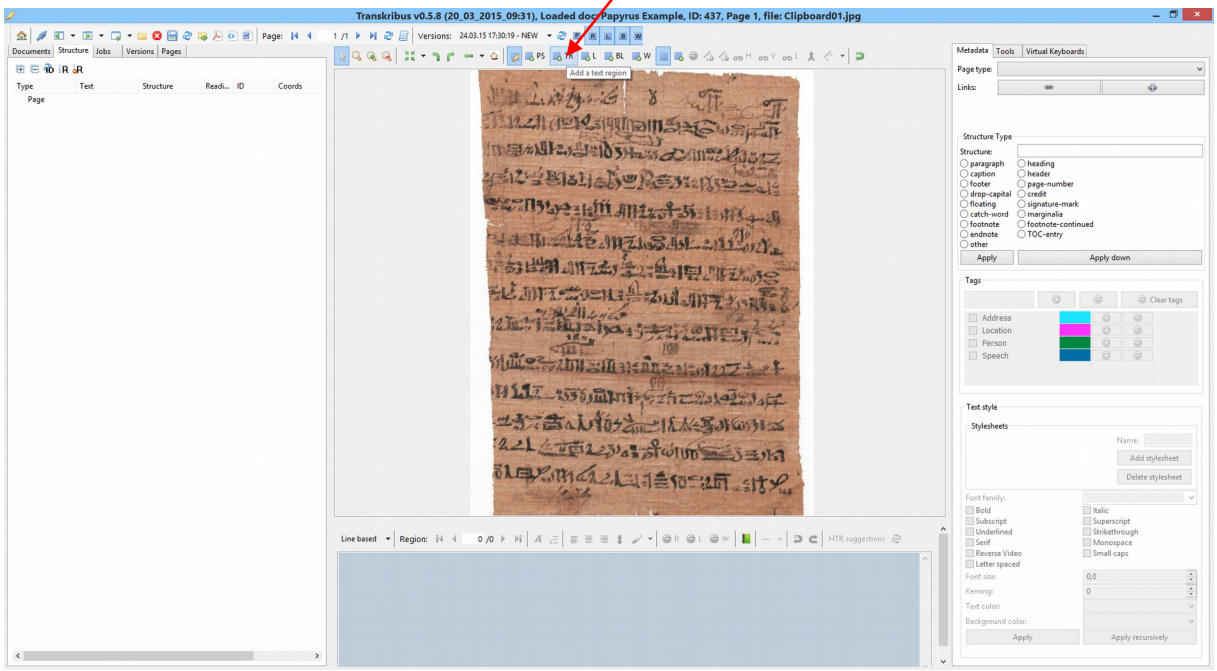
Use “Versions” tab to view previous versions of the text



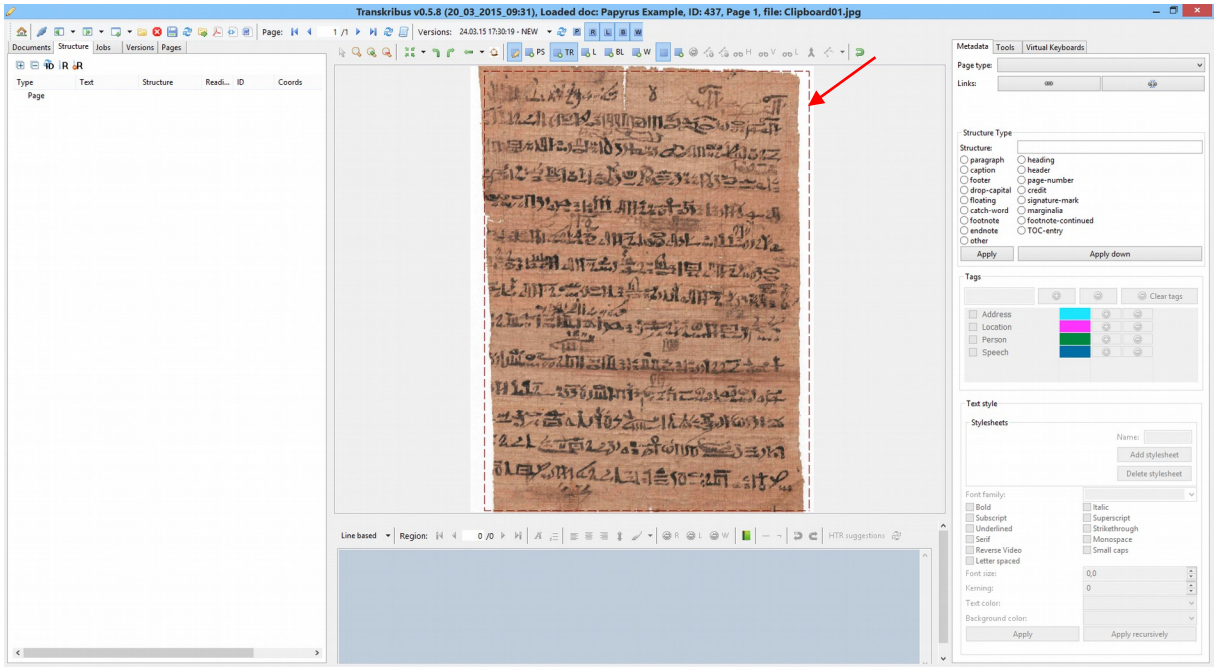
Upload document



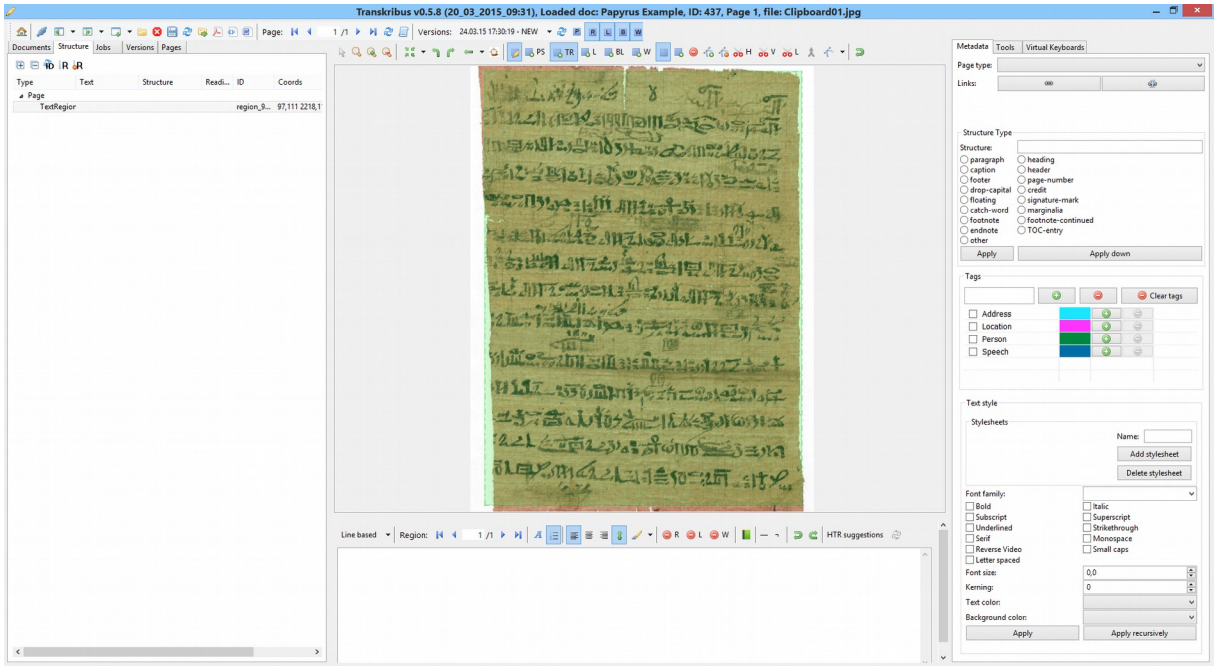
Segment document manually – Add a text region +TR



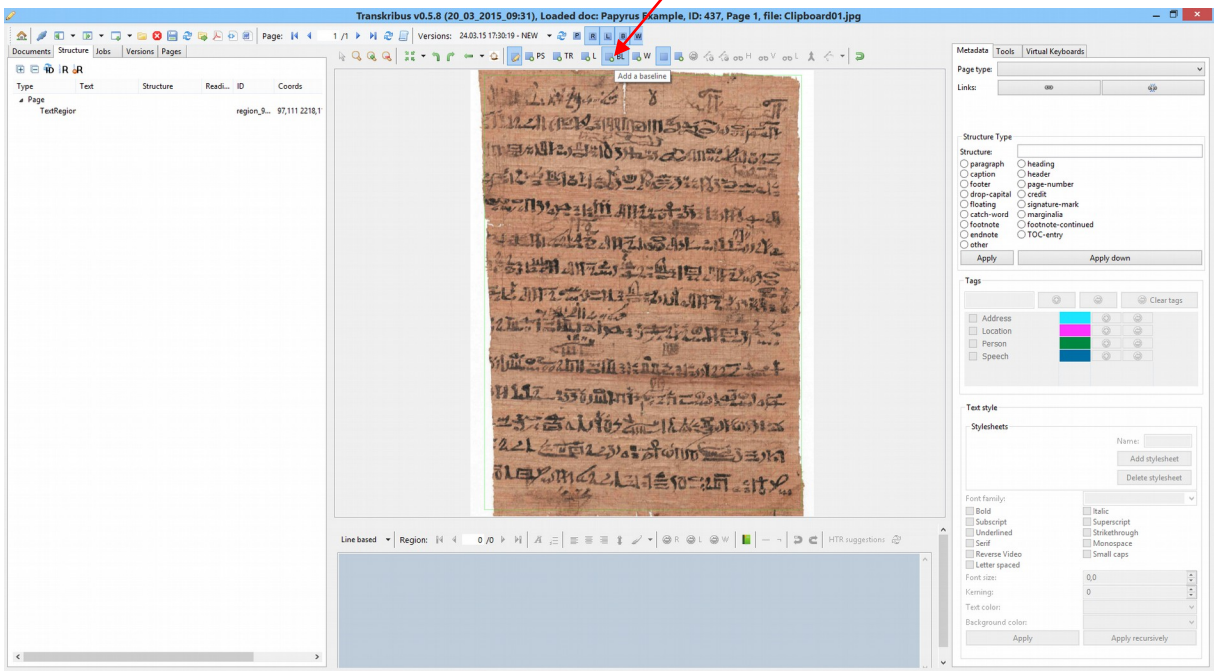
Draw rectangle



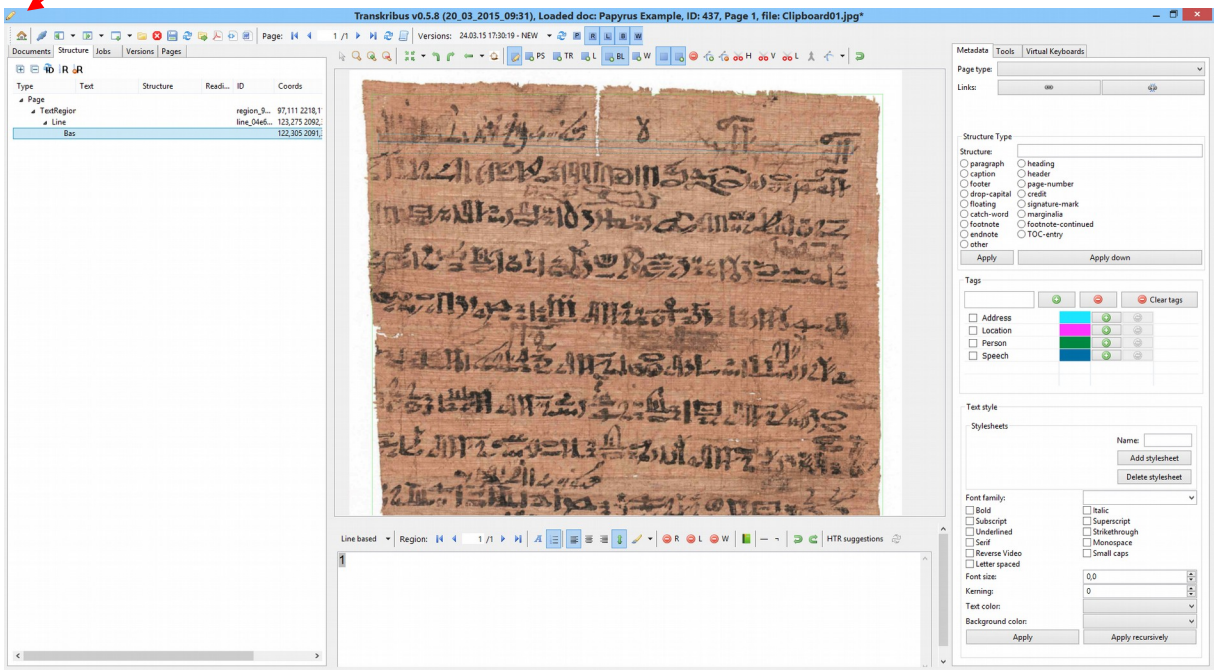
Display of text region – here in green



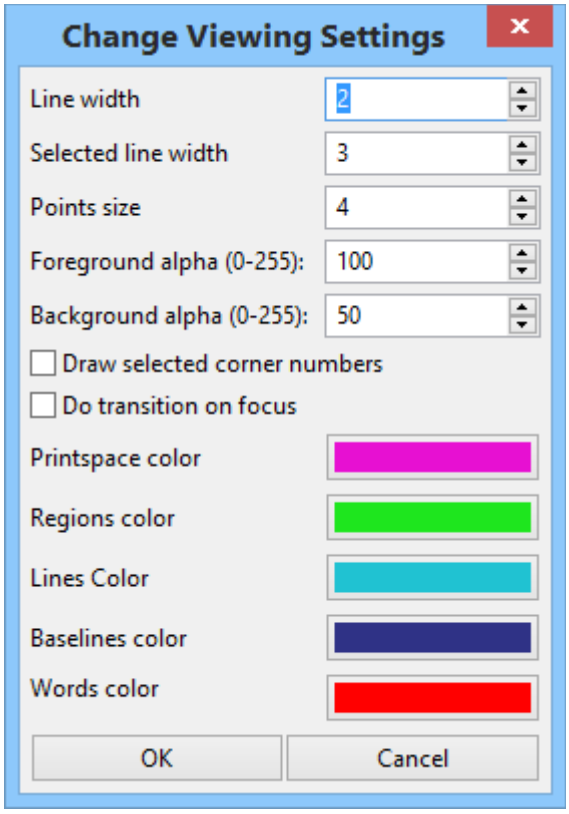
Add a baseline

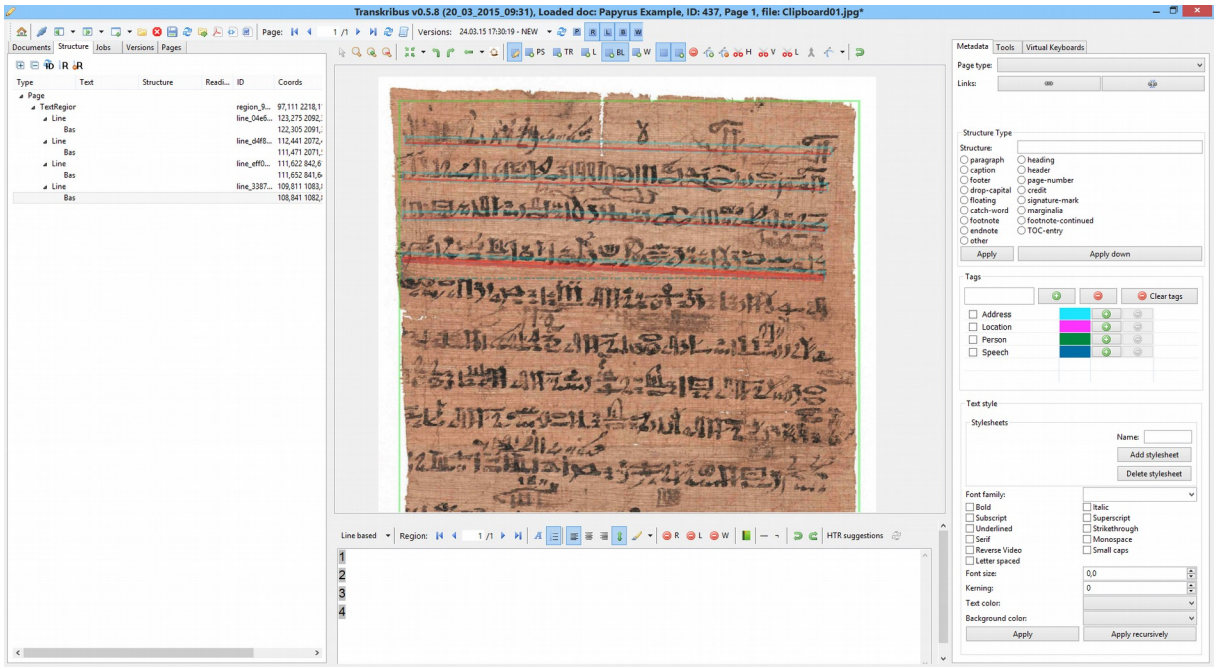


Display of regions – too light? Change them according to your needs

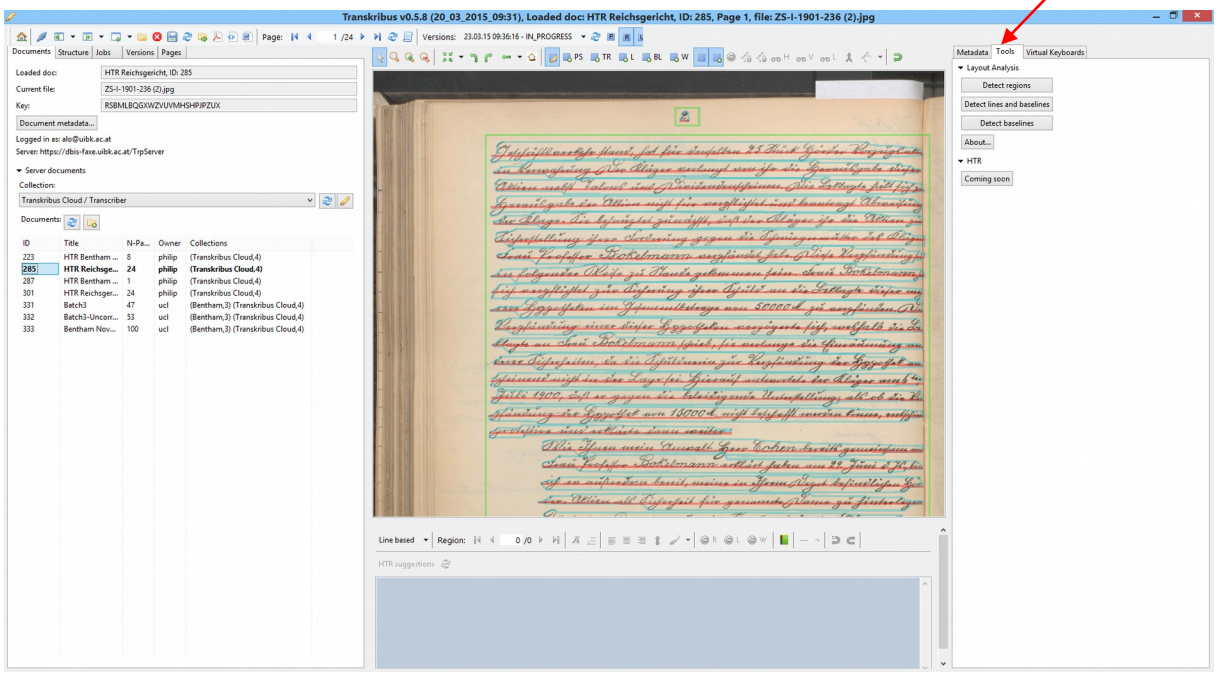


Home button – Change Viewing Settings

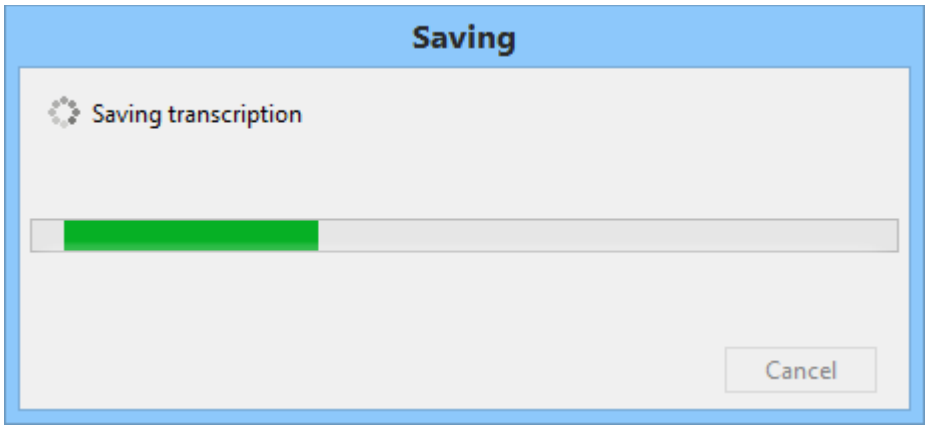




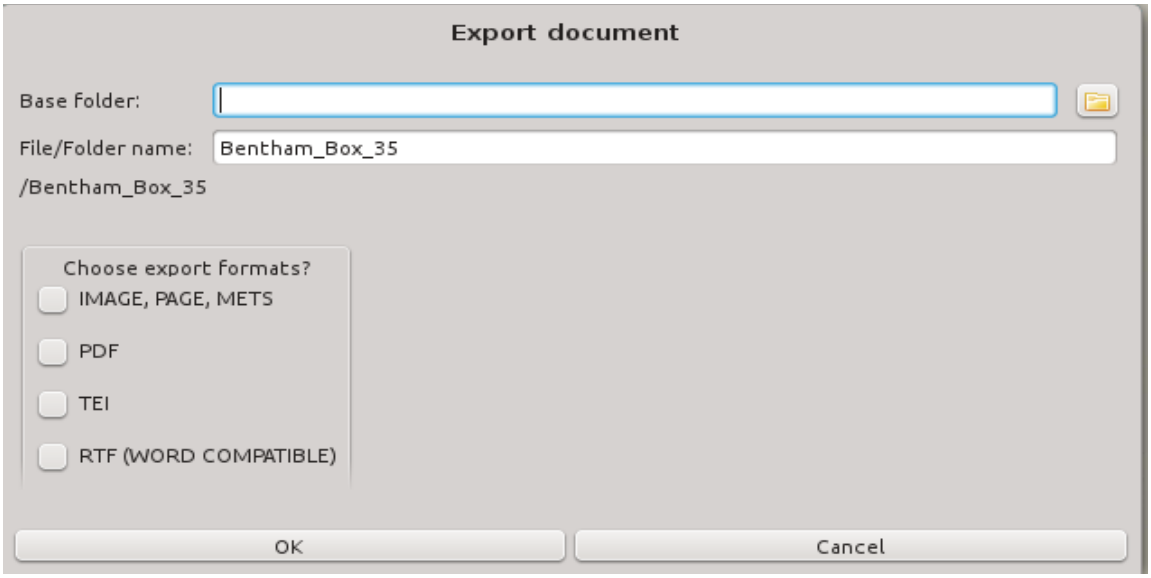
Automated line and baseline detection (works only when logged in). Attention: Select “Detect lines and baselines”.

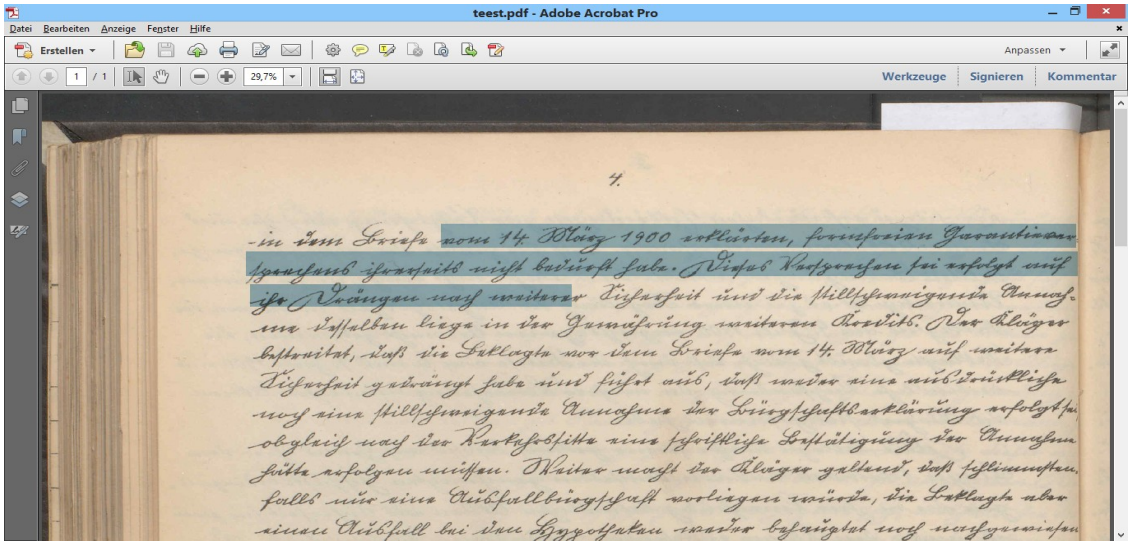


Save your transcription

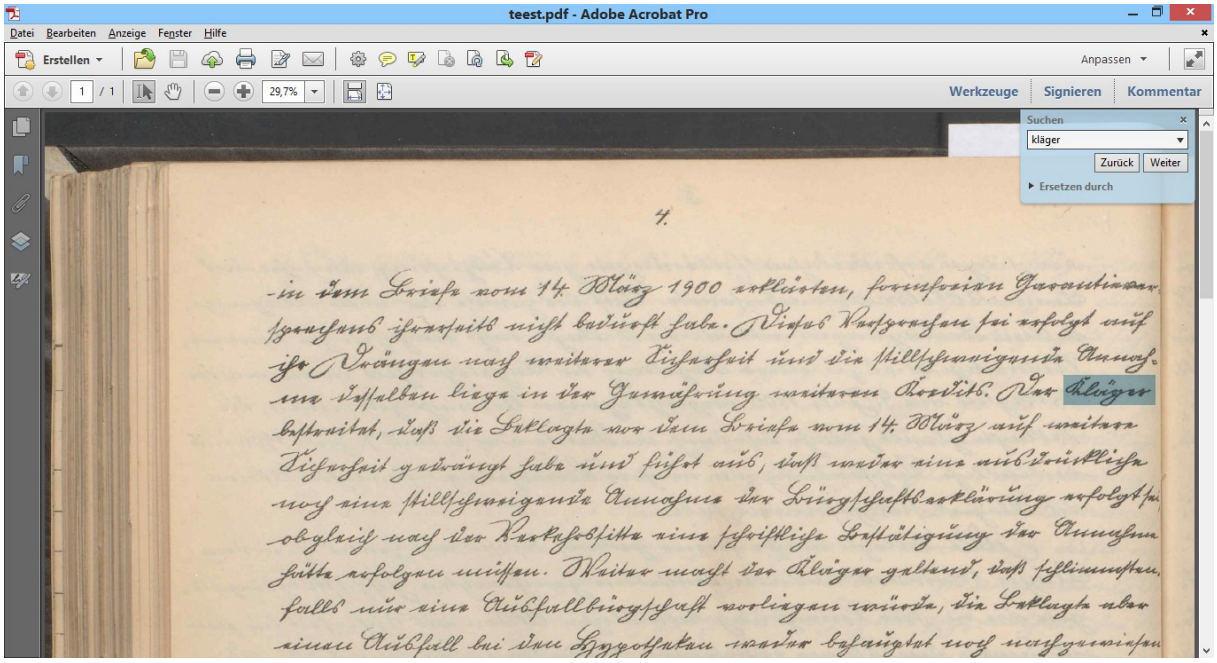


Use the Export Button to export the document

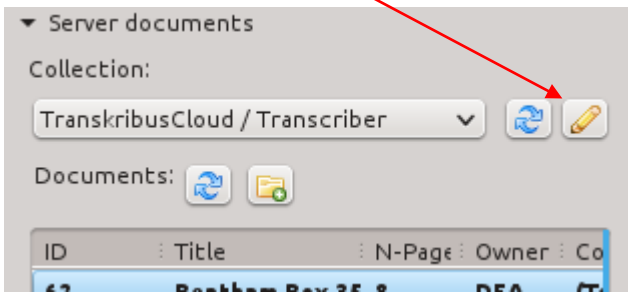


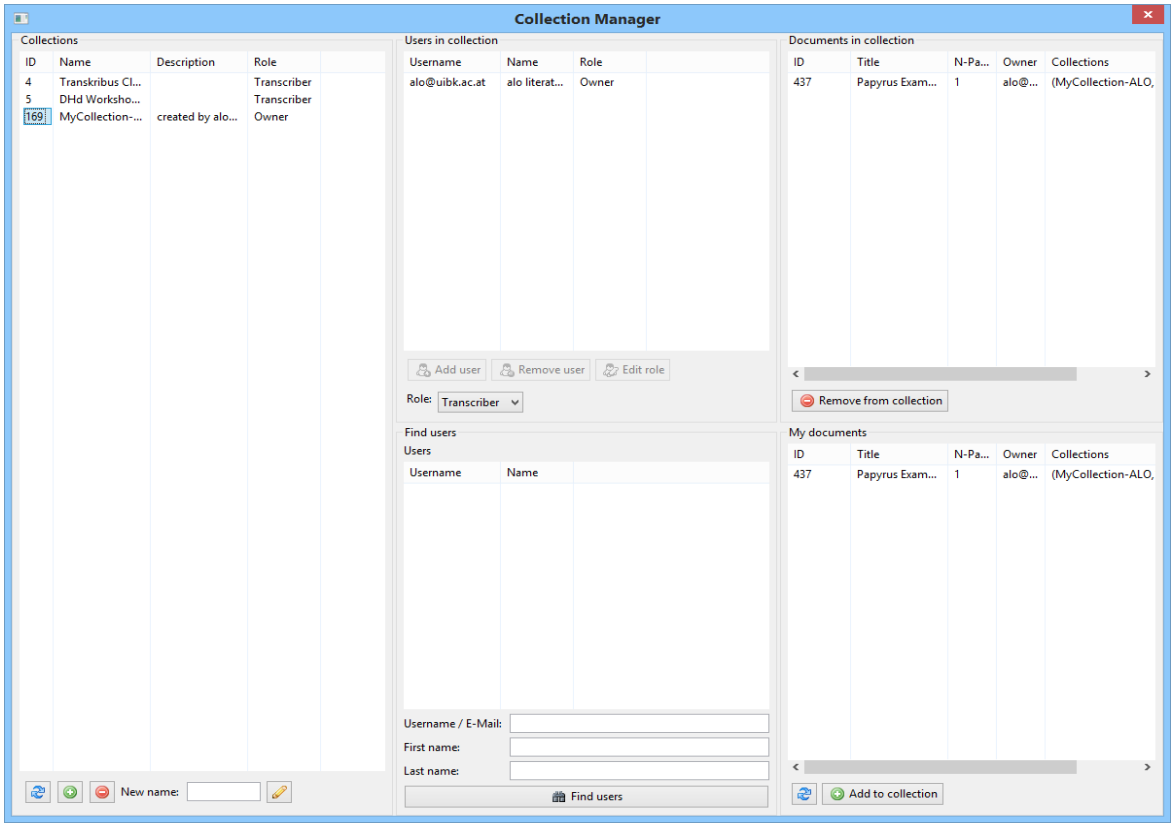


Search your transcribed text in the PDF



Collection Manager





For any further questions write an email to:
<email@transkribus.eu>