University of Stuttgart • Institut für Literaturwissenschaft • Neuere Deutsche Literatur I
Peggy Bockwinkel, M.A. • Keplerstr. 17 • Stockwerk 2b, Zi. 2.055 • 70174 Stuttgart Tel.: +49 (0)711-685-82279 • peggy.bockwinkel@ilw.uni-stuttgart.de

Documentation: Clemens Richter Financing: www.uni-stuttgart.de/dda + QSM

Content

Inhalt

1) Prepare document

Advantage of "PDFscissors" 1: all pages are superimposed transparently so that you can see exactly where the center is without cutting each page individually. In addition, the different sides can overlap, which is not possible with "Finereader".

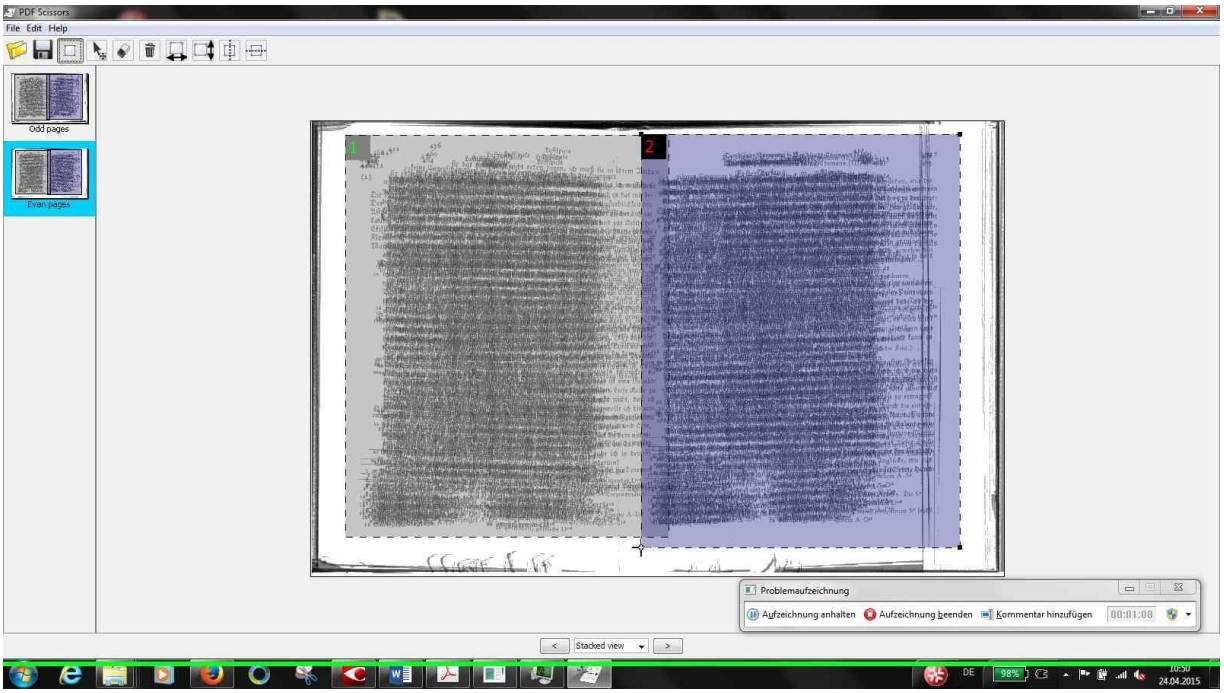
a) With "PDFscissors".
   i) Open "PDFscissors
   ii) Confirm security warning
   iii) File?? open
   iv) Open? select? select file? open?? ok?
   v) Dragging a rectangle for the first page
   vi) Click the rectangle icon to the right of "save-icon" and click "Draw an area for cropping".

[1]Direct download link: http://sites.google.com/site/pdfscissors/pdfscissors.jar
pdfscissors-offline. jnlp"file can be used to start the program. This will download the. jar file. The. jar file, which contains the actual program, can also be executed directly. In any case,"java" must be installed on the computer.

University of Stuttgart • Institut für Literaturwissenschaft • Neuere Deutsche Literatur I
Peggy Bockwinkel, M.A. • Keplerstr. 17 • Stockwerk 2b, Zi. 2.055 • 70174 Stuttgart Tel.: +49 (0)711-685-82279 • peggy.bockwinkel@ilw.uni-stuttgart.de

Documentation: Clemens Richter Financing: www.uni-stuttgart.de/dda + QSM

i) Draw rectangle for second side



ii) Select File→ srop & save→ location (default is folder of the source file with "_scissored" appended to the name)→save

b) With „Finereader“

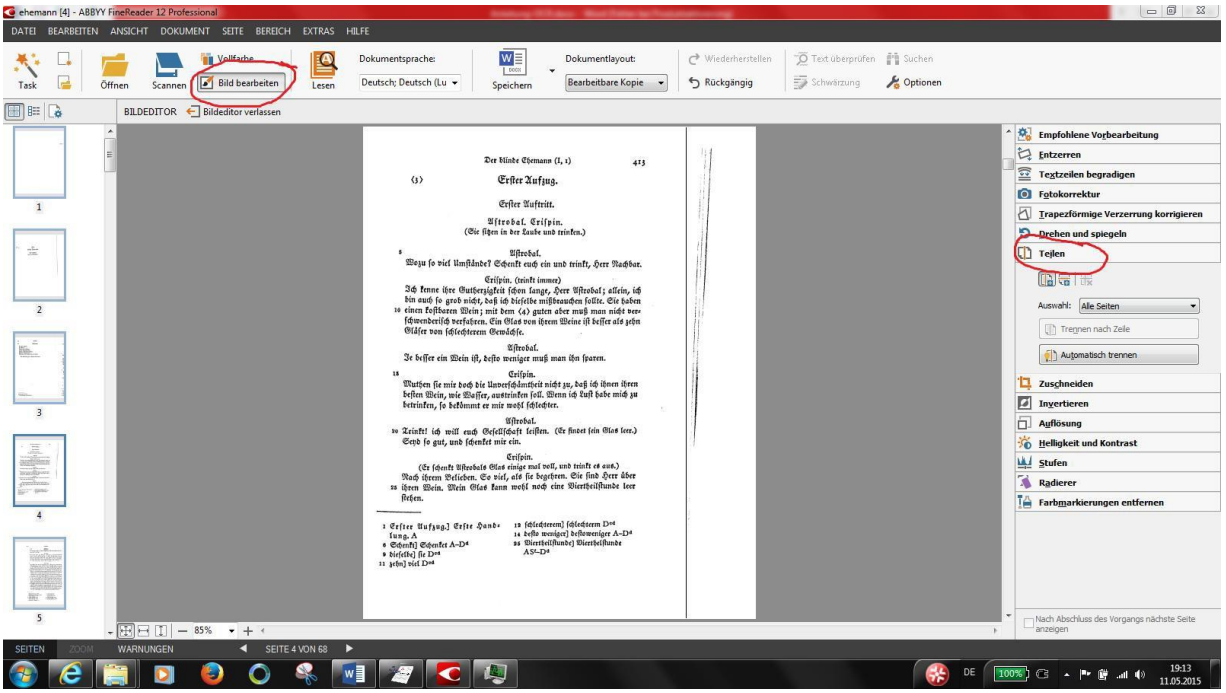
i) „Finereader“- Open document (if none is open yet)

ii) „Edit image“

University of Stuttgart • Institut für Literaturwissenschaft • Neuere Deutsche Literatur I
Peggy Bockwinkel, M.A. • Keplerstr. 17 • Stockwerk 2b, Zi. 2.055 • 70174 Stuttgart Tel.: +49 (0)711-685-82279 • peggy.bockwinkel@ilw.uni-stuttgart.de

Documentation: Clemens Richter Financing: www.uni-stuttgart.de/dda + QSM

iii) „Dividing“



iv) Select whether to split all/straight/odd or only the current page (if multiple pages are split at the same time, make sure that al-le pages that are split have the center at the same place. Unfortunately, this cannot be checked visually like with "PDFscissors".

v) Click on the position on the page where you want to cut.

vi) „separate by line“

2) Create new project / open project

a) Create new "Finereader" document

file → new "Finereader" document

File→ open PDF file or image→ select file

b) open „Finereader“-document

file →„Finereader“-open document (strg+umschalt+N)

c) Create new pattern file/import pattern

Each "finereader" document has its own pattern, these can be exported and imported together with the other documents specific settings. A new pattern should be created for each character set.

Create new pattern file/Import pattern:

New: Extras→ pattern editor→ new

Import: Extras→ Options→ Tab: read→ load from file

3) Start detection -> Settings (with/without training)

a) „ Analyze pages “

In the step "Analyzing pages" you can see which areas of the page contain text and what kind of text it is. Text elements that belong to the overlying page (if double pages were separated) are usually ignored. If not, it has to be repaired.

Select the desired pages in thumbnail view on the left→ right click on one of them→"parse selected page" (or ctrl+e)

When adding new pages to a "Finereader" document, they are automatically analyzed (this can be changed in the settings)

b) Check and correct detected areas

By right-clicking on an area, you can change the area type or delete the area.

Footnotes as table→ line breaks are kept (right click on area→ change area type to→ table)

Or remove them completely if they are not needed (right click on area→ delete)

In order to improve the recognition of antiqua and elevations in footnotes, the integrated patterns can be used in addition to the user patterns.

In "Extras→ Options→ tab: read" you can select under Training which patterns should be used.

c) „ Read pages “

With "Read pages" the text recognition (OCR) is started. When pattern training has been tivotated, the pattern must now be trained (→ chapter "Pattern training").

Select the desired pages in thumbnail view on the left→ right click on one of them→"read selected pages" (or ctrl+r)

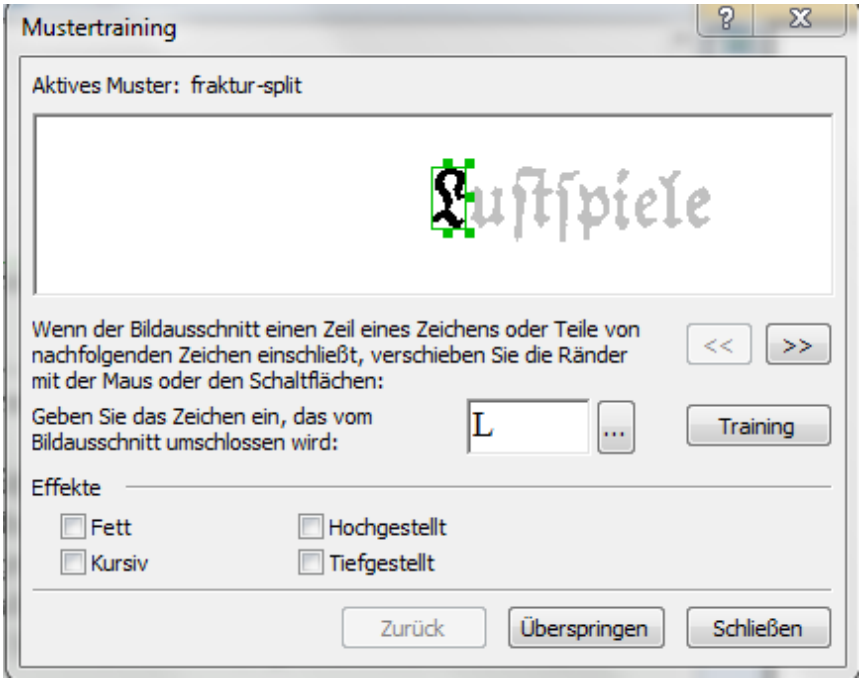
4) Training patterns

i) Switch training on and off

Extras→ Options→ Tab: read→Reading with training

This setting only applies to the current "Finereader" document, but is included in the configuration's ex port/import.Trainieren

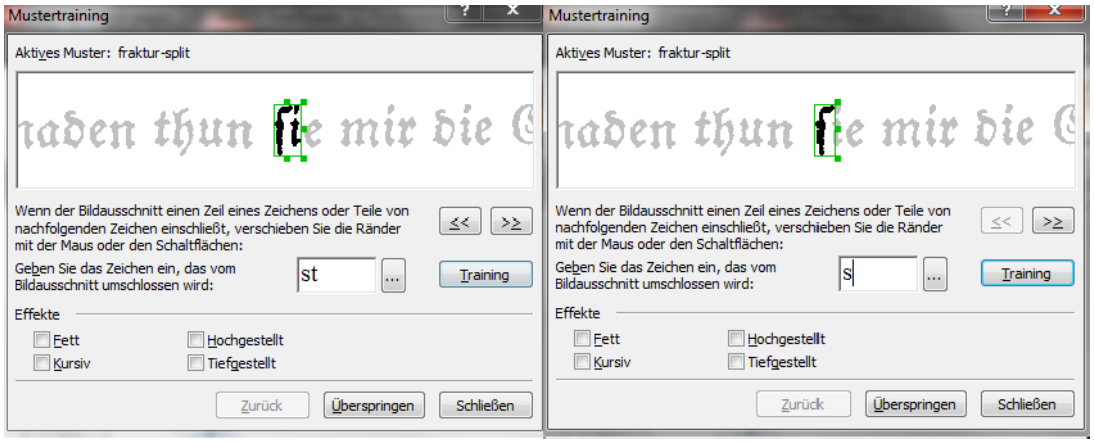
Trainiert wird immer, wenn Training aktiv ist und „gelesen“ wird (3 c).



(1) Check the green rectangle around the character and adjust it if necessary

Use << and >> buttons to zoom in/out or manually drag the rectangle at the corners up/down.

Example:



In this example, the "<<" key was used

(2) Type in letter(s)

If this letter has already been recognized, the input field can already contain a letter. However, this letter may be incorrect, in which case you must correct the letter. Although incorrectly trained letters can be corrected, you should be careful not to make any mistakes here, as this will greatly impair recognition.

(3) Click "train" if signs have been properly detected, otherwise skip.
Badly recognizable signs should not be trained and skipped, as this can worsen the recognition.

Examples:

Here the "h" was not separated cleanly from the "c". This sign should be skipped. If this occurs more frequently in this text, it should be noted that

can be entered as a ligature. (To do this,"ch" must be marked with the green rectangle,"ch" must be typed into the input field and "train" clicked, then the system automatically asks whether a ligature is to be created.

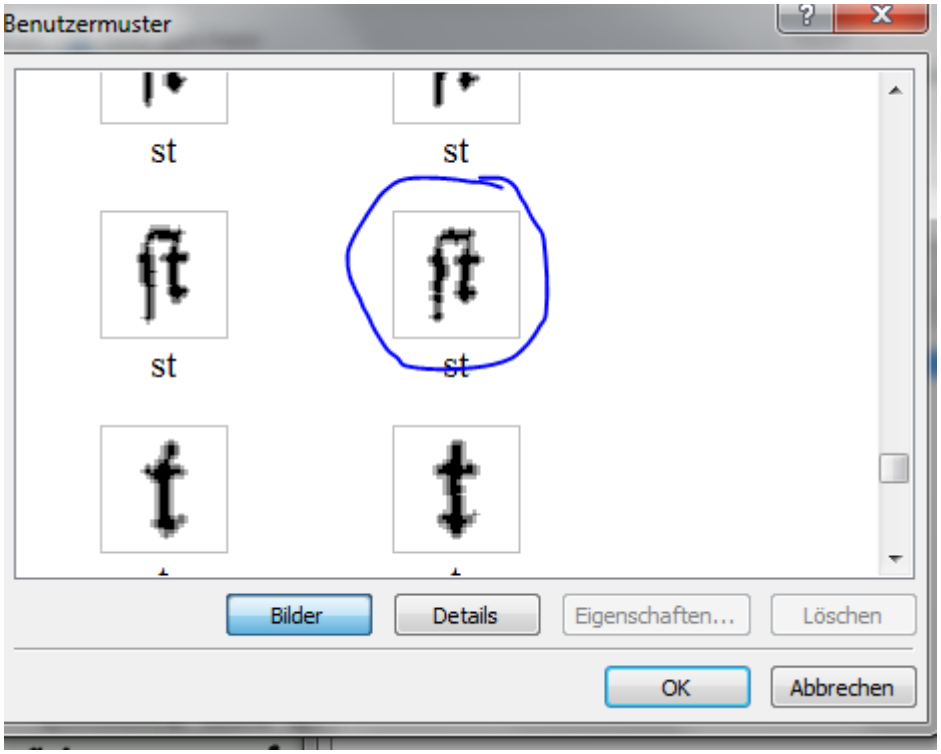
This badly resolved and unclean "st" should not be trained and skipped.

ii) check samples and delete badly recognized characters

Extras→ Pattern Editor→ Select pattern (currently: fracture-split) →"Edit"

Search for unclean characters and delete them.

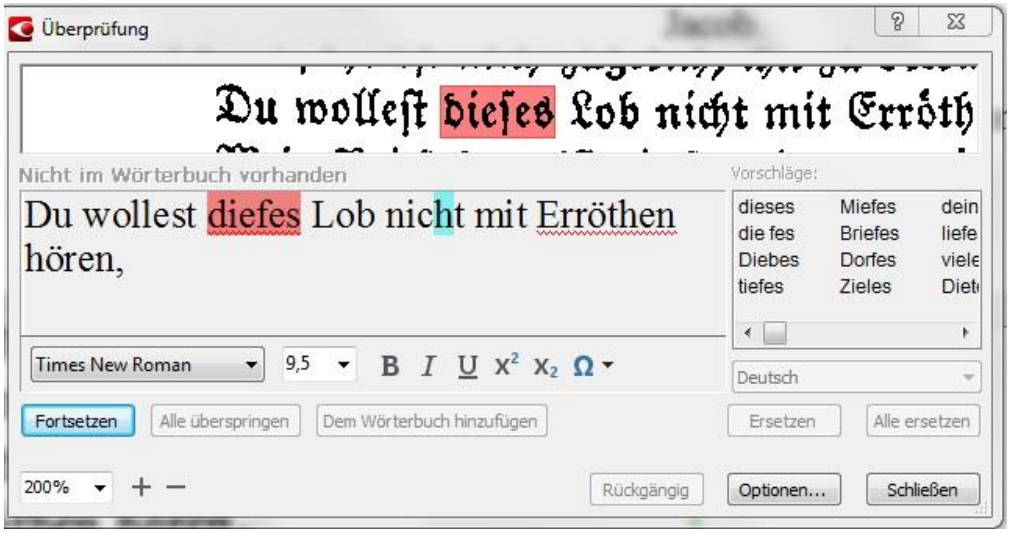
Here you can also correct errors from the previous step.



5) Check and correct recognized text

After all pages have been read, the recognized text must be corrected. Finereader "helps to do this by going through badly recognized words and words that are not in the dictionary. Font and font size can also be adjusted here.

Tools -> Review



In this case, a long "s" was wrongly detected. This can be corrected by clicking on "this" on the left side of the suggestions and then on "replace all". Finereader "then automatically jumps to the next problem area.

Correctly recognized words can be added to the dictionary or skipped.

6) Save/export project

a) Save *"Finereader"* document

These documents can only be opened with "Finereader" and save all pages and settings.

Datei →„Finereader"-Dokument speichern

b) Export document options and patterns

Extras→ Options→ Tab: read→"save to file"

All settings and patterns of the current document are saved in these. fbt files and can be imported into another or new "Finereader" document (? chapter 2)

c) Export result document
File -> Save document as→ PDF/doc/txt/....

7) Notes

- To further improve the detection rate, the scan resolution can be increased. For example,"n" and "u" will then be easier to distinguish, since the 300dpi bridge is sometimes only a few pixels wide, which can lead to confusion in recognition. For a higher scan resolution, however, a new pattern should be trained.

- The pattern "fracture-split" is the one I have trained most so far. The pattern is called "split", because I have separated the pages in this pattern before learning.
- http://www.finanzer.org/blog/2009/02/09/fraktur-ocr-mit-“Finereader“/