Universität 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

Dokumentation: Clemens Richter Finanzierung: www.uni-stuttgart.de/dda + QSM

Inhalt

1) Dokument vorbereiten

Vorteil von „PDFscissors“[1]: alle Seiten werden transparent übereinandergelegt, so dass man sehen kann, wo genau die Mitte ist, ohne jede Seite einzeln zu schneiden. Außerdem können sich die verschiedenen Seiten überlappen, was mit „Finereader“ nicht möglich ist.

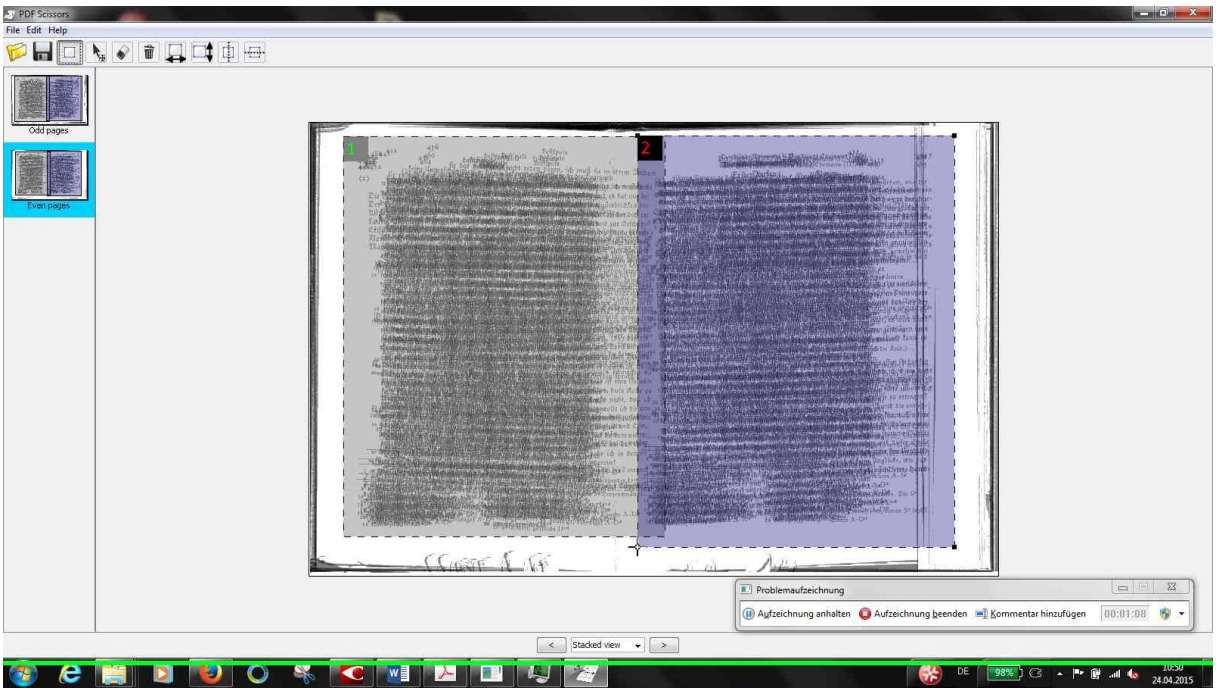
a) Mit „PDFscissors“
   i) „PDFscissors“ öffnen
   ii) Sicherheitswarnung bestätigen
   iii) File → open
   iv) Open → Datei auswählen → öffnen → ok
   v) Rechteck für erste Seite ziehen
   vi) Rechteck-icon rechts neben “save-icon” klicken „Draw an area for cropping“

[1] Direkter Downloadlink: http://sites.google.com/site/pdfscissors/pdfscissors.jar
„pdfscissors-offline.jnlp"-Datei kann genutzt werden, um das Programm zu starten. Diese lädt dazu einmalig die .jar Datei herunter. Die .jar Datei, welche das eigentliche Programm enthält, kann aber auch direkt ausgeführt werden. In jedem Fall muss auf dem Rechner „java“ installiert sein.

Universität 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

Dokumentation: Clemens Richter Finanzierung: www.uni-stuttgart.de/dda + QSM

vii) Rechteck für zweite Seite zeichnen



viii) File → srop & save →Speicherort wählen (Standard ist Ordner der Quelldatei mit „_scissored“ an den Namen angehängt) → speichern

b) Mit „Finereader“

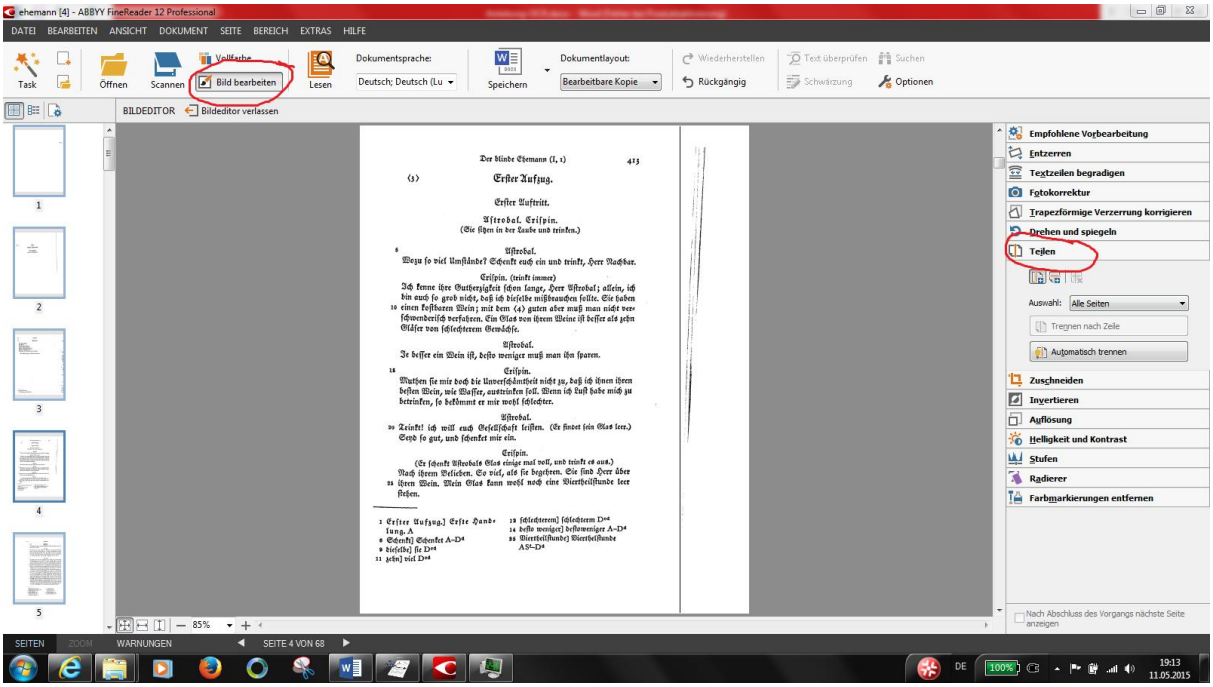
i) „Finereader“-Dokument öffnen (wenn noch keins offen ist)

ii) „Bild bearbeiten“

Universität 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

Dokumentation: Clemens Richter Finanzierung: www.uni-stuttgart.de/dda + QSM

iii) „Teilen“



iv) Auswählen ob alle/gerade/ungerade oder nur die aktuelle Seite geteilt werden soll (wenn mehrere Seiten gleichzeitig geteilt werden, muss darauf geachtet werden, dass alle Seiten, die geteilt werden, die Mitte an der gleichen Stelle haben. Leider kann dies hier nicht wie bei „PDFscissors“ visuell überprüft werden

v) An die Stelle auf der Seite klicken, an welcher getrennt werden soll

vi) „trennen nach Zeile“

2) Neues Projekt erstellen / Projekt öffnen

a) Neues „Finereader“-Dokument erstellen

Datei → neues „Finereader“-Dokument

Datei →PDF Datei oder Bild öffnen →Datei auswählen

b) „Finereader“-Dokument öffnen

Datei →„Finereader“-Dokument öffnen (strg+umschalt+N)

c) Neue Muster-Datei erstellen/Muster importieren

Jedes „Finereader“-Dokument hat eigene Muster, diese können gemeinsam mit den anderen Dokumenten spezifischen Einstellungen ex-und importiert werden. Für jeden Zeichensatz sollte ein neues Muster erstellt werden.

Neue Muster-Datei erstellen/Muster importieren:

Neu: Extras → Mustereditor → neu

Import: Extras → Optionen → Tab: lesen → laden aus Datei

3) Erkennung starten -> Einstellungen (mit/ohne Training)

a) „Seiten analysieren“

Im Schritt „Seiten analysieren“ wird erkannt, welche Bereiche der Seite Text enthalten und welche Art von Text das jeweils ist. Textelemente, welche zum Beispiel zu der gegenüberliegenden Seite gehören (wenn Doppelseiten getrennt wurden), werden meist ignoriert. Falls nicht, muss nachgebessert werden.

In der Miniaturansicht links die gewünschten Seiten auswählen → Rechtsklick auf eine davon → „gewählte Seite analysieren“ (oder strg+e)

Beim Hinzufügen von neuen Seiten in ein „Finereader“-Dokument werden diese automatisch analysiert (dies kann in den Einstellungen geändert werden)

b) Erkannte Bereiche prüfen und korrigieren

Durch einen Rechtsklick auf einen Bereich kann der Bereichstyp geändert werden oder der Bereich gelöscht werden.

Fußnoten als Tabelle → Zeilenumbrüche werden behalten
(Rechtsklick auf Bereich → Bereichstyp ändern zu → Tabelle)

Oder ganz entfernen, wenn sie nicht benötigt werden
(Rechtsklick auf Bereich → löschen)

Um die Erkennung von Antiqua und Hochstellungen in Fußnoten zu verbessern, können zusätzlich zu den Benutzermustern die integrierten Muster genutzt werden.

In „Extras → Optionen → Tab: lesen“ kann unter Training ausgewählt werden, welche Muster genutzt werden sollen.

c) „Seiten lesen“

Mit „Seiten lesen“ wird die Texterkennung (OCR) gestartet. Wenn das Mustertraining aktiviert wurde, muss jetzt das Muster trainiert werden (→ Kapitel „Muster trainieren“).

In der Miniaturansicht links die gewünschten Seiten auswählen → Rechtsklick auf eine davon → „gewählte Seiten lesen“ (oder strg+r)

4) Muster trainieren

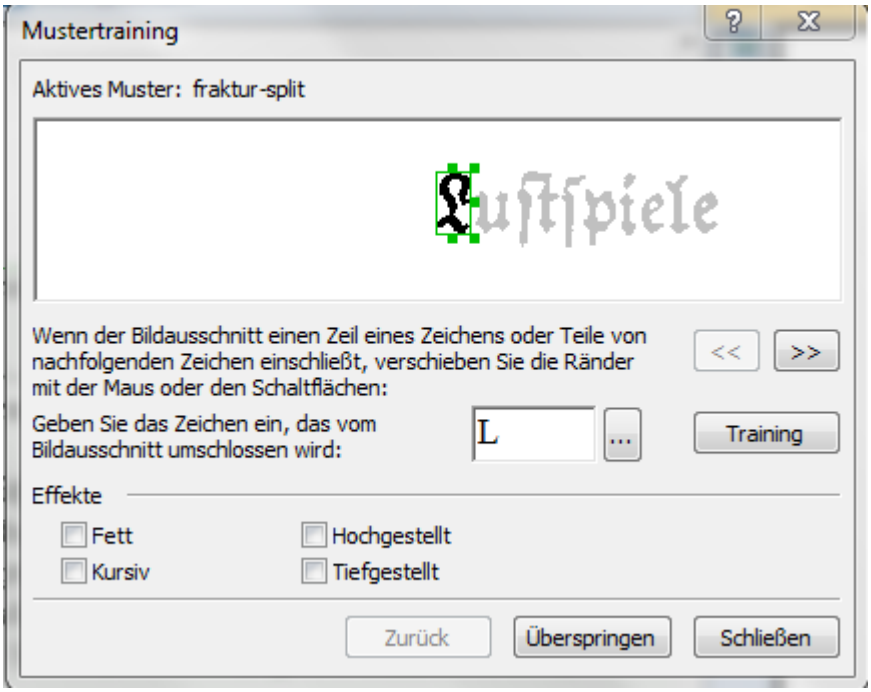
i) Training an- und abschalten

Extras → Optionen → Tab: lesen → Lesen mit Training

Diese Einstellung gilt nur für das aktuelle „Finereader“-Dokument, wird aber beim Export/Import der Konfiguration mit einbezogen.

ii) Trainieren

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

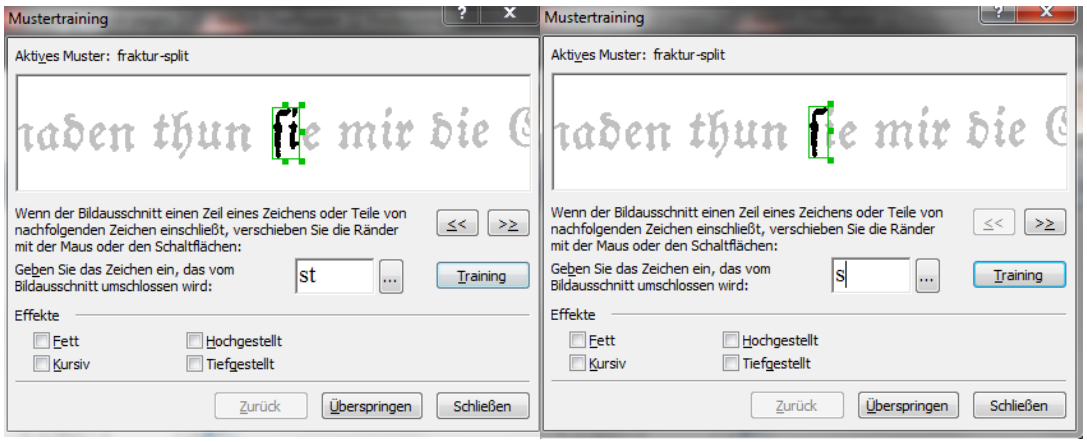


Universität Stuttgart • Institut für Literaturwissenschaft • Neuere Deutsche Literatur I
Peggy Bockwinkel, M.A. • Keplerstr. 17 • Stockwerk 2b, Zi. 2.055 • 70174 Stuttgart
Tel.: +49 (0)711-685-82279 • peggy.bockwinkel@ilw.uni-stuttgart.de

(1) Grünes Rechteck um das Zeichen kontrollieren und eventuell anpassen

Mit << und >> Tasten Einzug vergrößern/verkleinern oder manuell das Rechteck an den Ecken größer/kleiner ziehen.

Beispiel:



In diesem Beispiel wurde die „<<“ Taste genutzt

(2) Buchstabe(n) eintippen

Wenn dieser Buchstabe schon einmal erkannt wurde, kann in dem Eingabefeld schon ein Buchstabe stehen. Dieser Buchstabe kann jedoch falsch sein, in welchem Fall Sie den Buchstabe korrigieren müssen. Obwohl falsch trainierte Buchstaben korrigiert werden können, sollten Sie hier aufpassen, keine Fehler zu machen, da dies die Erkennung stark verschlechtert.

(3) „trainieren“ klicken, wenn Zeichen sauber erkannt wurden, ansonsten überspringen.

Schlecht erkennbare Zeichen sollten nicht trainiert und übersprungen werden, da dies die Erkennung verschlechtern kann.

Beispiele:

Hier wurde das „h“ nicht sauber von dem „c“ getrennt. Dieses Zeichen sollte übersprungen werden. Wenn dies häufiger in dem vorliegenden Text vorkommt, sollte „ch“ als Ligatur eingetragen werden. (Dazu muss „ch“ mit dem grünen Rechteck markiert, „ch“ ins Eingabefeld getippt und „trainieren“ geklickt werden, dann wird automatisch gefragt, ob eine Ligatur erstellt werden soll.)

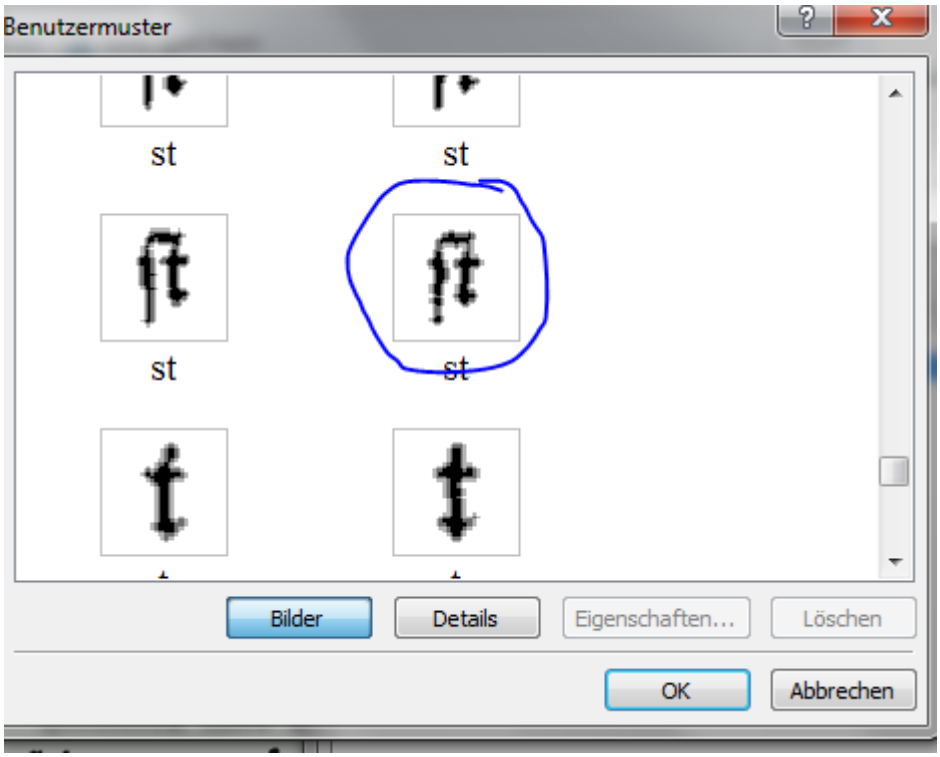
Dieses schlecht aufgelöste und unsaubere „st“ sollte nicht trainiert und übersprungen werden.

iii) Muster prüfen und schlecht erkannte Zeichen löschen

Extras → Mustereditor → Muster auswählen (zzt:fraktur-split) → „bearbeiten“

Nach unsauberen Zeichen suchen und diese löschen.

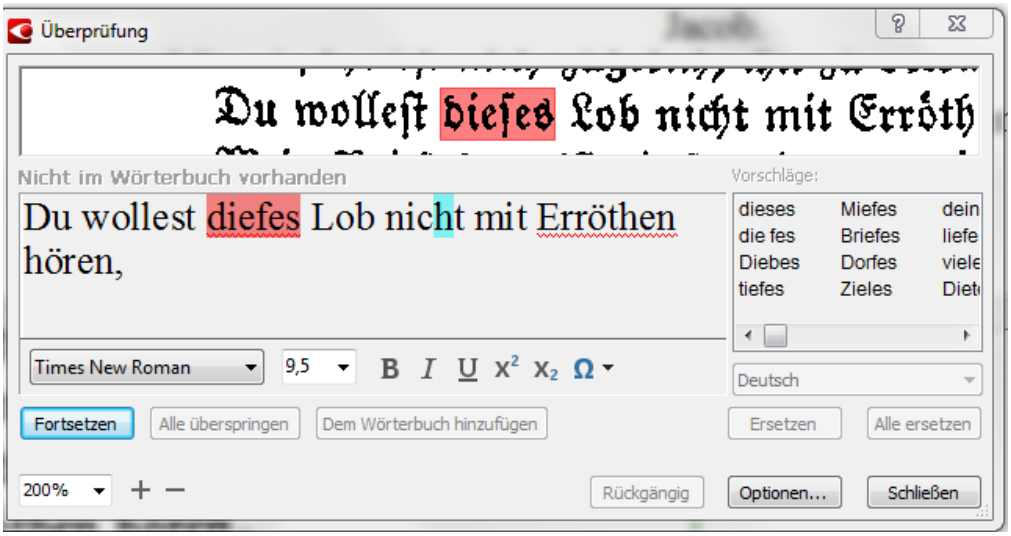
Hier können auch Fehler aus dem vorherigen Schritt behoben werden.



5) Erkannten Text prüfen und korrigieren

Nachdem alle Seiten gelesen wurden, muss der erkannte Text korrigiert werden. „Finereader“ hilft dabei, indem schlecht erkannte Wörter und Wörter, die sich nicht im Wörterbuch befinden, nacheinander durchgegangen werden. Hier können auch Schriftart und Schriftgröße angepasst werden.

Extras -> Überprüfung



In diesem Fall wurde ein langes „s“ falsch erkannt. Dies kann mit einem Klick auf „dieses“ links in den Vorschlägen und danach auf „alle ersetzen“ korrigiert werden. „Finereader“ springt dann automatisch zur nächsten Problemstelle.

Richtig erkannte Wörter können hier zum Wörterbuch hinzugefügt oder übersprungen werden.

6) Projekt speichern/ exportieren

a) *„Finereader“-Dokument* speichern

Diese Dokumente lassen sich nur mit „Finereader“ öffnen und speichern alle Seiten und Einstellungen.

Datei → „Finereader“-Dokument speichern

b) Dokumentenoptionen und Muster exportieren

Extras → Optionen → Tab: lesen → „speichern in Datei“

In diesen .fbt Dateien sind alle Einstellungen und Muster des aktuellen Dokuments gespeichert und können in ein anderes bzw. neues „Finereader“-Dokument importiert werden (→ Kapitel 2)

c) Ergebnisdokument exportieren

Datei -> Dokument speichern als → PDF/doc/txt/…

7) Hinweise

- Um die Erkennungsrate weiter zu verbessern, kann die Scan-Auflösung erhöht werden. Zum Beispiel werden dann „n“ und „u“ besser unterscheidbar, da der Steg bei 300dpi manchmal nur wenige Pixel breit ist, was bei der Erkennung zu Verwechslungen führen kann. Für eine höhere Scan-Auflösung sollte jedoch ein neues Muster trainiert werden.

- Das Muster „fraktur-split“ habe ich bis jetzt am meisten trainiert. Das Muster heißt „split“, da ich bei diesem Muster, vor dem Einlernen die Seiten getrennt habe.
- http://www.finanzer.org/blog/2009/02/09/fraktur-ocr-mit-“Finereader“/