Guidelines zur Volltexdigitalisierung von Dramen des 17 bis 19. Jahrhunderts mit OCR4all.

Katrin Dennerlein
Martin Rupnig
Nadine Kastenhofer

Stand: 12. August 2024

1 Einleitung

Die Bearbeitung der Dramen erfolgt über den Browser. Hierzu wird erst eine Verbindung zum Netzwerk der Uni Würzburg hergestellt (VPN) und anschließend die entsprechende Instanz mit den Dramen geöffnet:

> URL: `http://ocr4all-internal.informatik.uni-wuerzburg.de:1244/ocr4all/` Benutzer: `user` Passwort: `dennerlein2023`

Nach dem Einloggen kann unter Project das Drama mit der entsprechenden Nummer ausgewählt und mit einem Klick auf den Button `LOAD PROJECT` geladen werden:

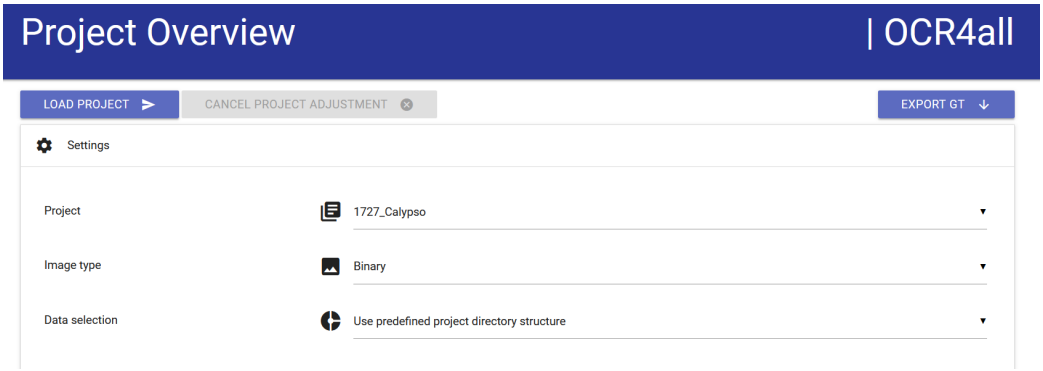


Abb. 1: Project Overview

Beim erstmaligen Öffnen des Projektes erscheint die Meldung `"Attention... some or all of these files..."`. Hier kann `"Convert files directly"` ausgewählt werden. Dadurch werden alle Scans von .jpg in .png umgewandelt und umbenannt.
Einen allgemeinen User Guide für OCR4all gibt es übrigens auch unter diesem Link.

2 Vorbereitung der Dateien und Preprocessing

In einem weiteren Schritt muss das `Preprocessing` durchgeführt werden. Dabei werden Binär- und Graustufenbilder produziert. Durch einen Klick auf das Menü-Symbol in der linken oberen Ecke öffnet sich eine Lasche, in der Preprocessing ausgewählt werden kann. Nun öffnet sich die Übersicht für weitere Optionen. Standardmäßig werden alle gescannten Seiten verarbeitet. Am rechten Rand gibt es jedoch die Möglichkeit, einzelne Seiten auszuwählen. Es wird empfohlen, die voreingestellte Auswahl beizubelassen und alle Seiten zu verarbeiten. Nachdem alle zu verarbeitenden Seiten ausgewählt wurden, kann durch einen Klick auf EXECUTE das Preprocessing gestartet werden.

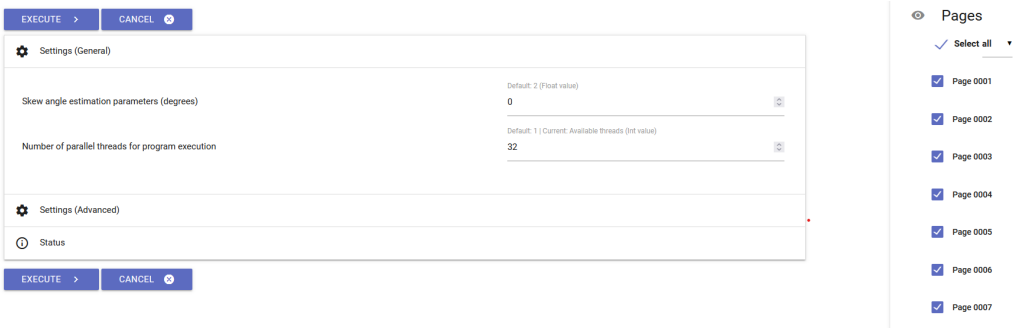


Abb. 2: Preprocessing

Je nach Seitenzahl des Dramas und der Leistung des Rechners kann dies mehrere Minuten dauern. Während der Prozess läuft wird `Ongoing` angezeigt, danach `Completed`. Nach Fertigstellung werden die Seiten in der `Project Overview` in der Spalte `Preprocessing` durch grüne Häkchen markiert.
Der im Menü angezeigt Arbeitsschritt `Noise Removal` wird für dieses Projekt nicht verwendet und kann ignoriert werden.

Page Identifier	Preprocessing	Noise Removal	Segmentation	Line Segmentation	Recognition	Ground Truth
0001	✓	✗	✓	✗	✗	✗
0002	✓	✗	✓	✗	✗	✗
0003	✓	✗	✓	✗	✗	✗
0004	✓	✗	✓	✗	✗	✗
0005	✓	✗	✓	✓	✓	✓
0006	✓	✗	✓	✓	✓	✓
0007	✓	✗	✓	✓	✓	✓
0008	✓	✗	✓	✓	✓	✓
0009	✓	✗	✓	✓	✓	✓
0010	✓	✗	✓	✓	✓	✓

Abb. 3: Project Overview

3 Segmentation

Die Segmentierung dient dazu, Textabschnitte in bestimmte Regionen zu unterteilen, um sie später identifizieren, zuordnen und verarbeiten zu können. Hierbei geht es darum, verschiedene Teile des Textes zu unterscheiden und ihren Funktionen bzw. Regions zuzuweisen. Stehen auf einer Seite beispielsweise eine Überschrift, Sprechernamen und Sprechertexte, so haben diese unterschiedliche Funktionen im Drama und müssen dementsprechend markiert werden.
Die Bearbeitung erfolgt in LAREX. Um es zu öffnen, muss links im Menü Segmentation > LAREX ausgewählt werden.

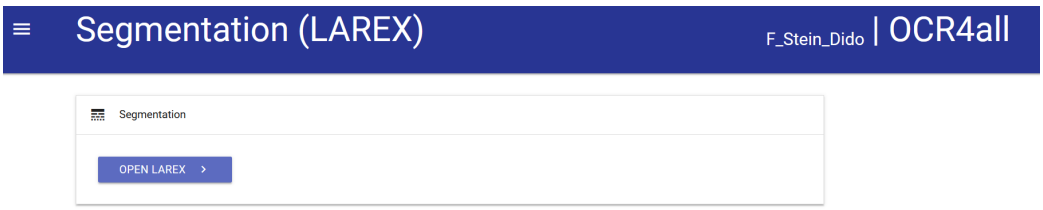


Abb. 4: Segmentation

Durch einen Klick auf `OPEN LAREX` öffnet sich das Tool, in welchem die manuelle Bearbeitung stattfindet.

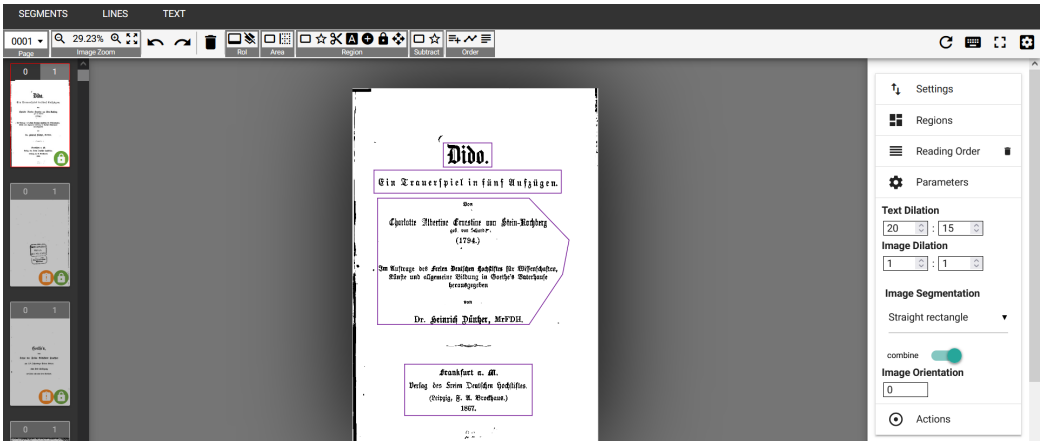
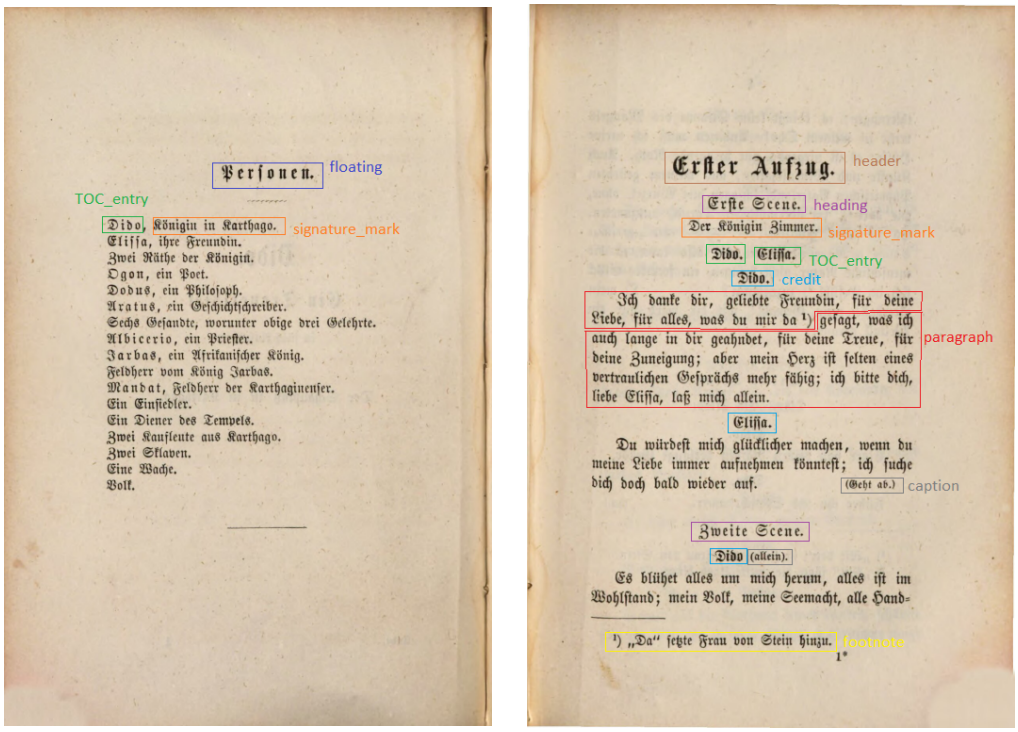


Abb. 5: LAREX

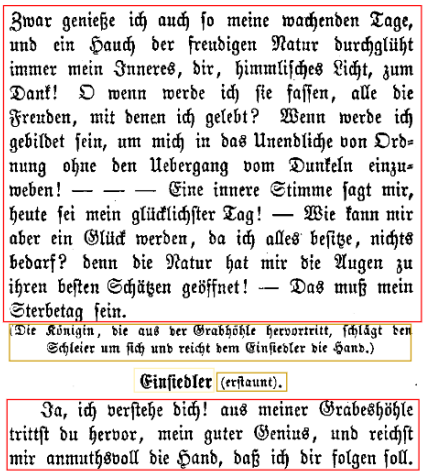
Layoutelement	**Layoutregion**
Text (gesprochener Text)	paragraph
sonstiger Text (z.B. Text in Vorreden)	other
Sprecherangabe (Name der Figur)	credit
Sprecherangabe bei mehreren SprecherInnen (z.B. Chor)	drop_capital
Regieanweisung innerhalb des Dialogtextes (Regieanweisungen, die sich auf die einzelne Figur beziehen)	caption
Fußnote	footnote
Akt oder Aufzug	header
Szene oder Auftritt	heading
Überschrift (Gesang, Gedicht, z.B. „Aria“, „Ballade“)	floating
Strophennummerierung	endnote
sonstige Überschriften, alles auf dem Titelblatt	catch_word
Figurennamen im Personenverzeichnis, Aufzählung der Personen am Szenenbeginn in der Regieanweisung	TOC_entry
Figurenbeschreibung im Personenverzeichnis, Regieanweisung zum Setting oder zu den Figuren, kann am Szenenbeginn oder -ende stehen oder zwischen zwei Repliken (Bsp: Replik Figur a, Regieanweisung z.B. „Figur b geht ab, Replik Figur b)	signature_mark
Bild, Holzschnitt, Diagramm, Tabelle, Initiale, Zierinitiale, Formel, ...	Image



3.1 TextRegions

LAREX führt im Vorfeld für jede Seite automatische Segmentierungen durch. In einem ersten Schritt müssen diese gelöscht werden (Strg + A + Entf). Danach kann die manuelle Bearbeitung beginnen. Über den Shortcut 3 können Rechtecke um die Textelemente gezogen werden. Für komplexere Textstellen empfiehlt sich der Shortcut 4 um ein Polygon zu erstellen. Die Elemente müssen dann per Rechtsklick ihrer entsprechenden Bezeichnung zugeordnet werden. Standardmäßig wird ein Element als `paragraph` definiert. Für andere Textteile müssen die entsprechenden Elemente gewählt werden. Sollten die Elemente nicht aufgelisten sein, müssen diese erst erstellt werden.

Ein paar wichtige Prinzipien sind dabei zu beachten: Es ist notwendig, den Dramentext zumindest oberflächlich mitzulesen, da gesprochener Text von Regieanweisungen unterbrochen sein kann und diese Textteile unterschiedlich markiert werden müssen. Steht ein Sprechertext alleine, kann der ganze Abschnitt als ein `paragraph` markiert werden. Ist dieser jedoch von einer Regieanweisung unterbrochen, muss auch der Sprechertext geteilt werden.

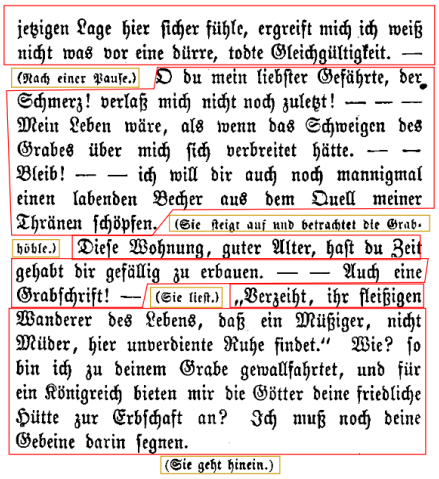

Zwar genieße ich auch so meine wachenden Tage, und ein Hauch der freudigen Natur durchglüht immer mein Inneres, dir, himmlisches Licht, zum Dank! O wenn werde ich sie fassen, alle die Freuden, mit denen ich gelebt? Wenn werde ich gebildet sein, um mich in das Unendliche von Ordnung ohne den Uebergang vom Dunkeln einzuweben! — — — Eine innere Stimme sagt mir, heute sei mein glücklichster Tag! — Wie kann mir aber ein Glück werden, da ich alles besitze, nichts bedarf? denn die Natur hat mir die Augen zu ihren besten Schätzen geöffnet! — Das muß mein Sterbetag sein.

(Die Königin, die aus der Grabhöhle hervortritt, schlägt den Schleier um sich und reicht dem Einsiedler die Hand.)

Einsiedler (erstaunt).

Ja, ich verstehe dich! aus meiner Grabeshöhle trittst du hervor, mein guter Genius, und reichst mir anmuthsvoll die Hand, daß ich dir folgen soll.

Abb. 6: ohne Unterbrechung


jetzigen Lage hier sicher fühle, ergreift mich ich weiß nicht was vor eine dürre, todte Gleichgültigkeit. —

(Nach einer Pause.) O du mein liebster Gefährte, der Schmerz! verlaß mich nicht noch zuletzt! — — — Mein Leben wäre, als wenn das Schweigen des Grabes über mich sich verbreitet hätte. — — Bleib! — — ich will dir auch noch mannigmal einen labenden Becher aus dem Quell meiner Thränen schöpfen. (Sie steigt auf und betrachtet die Grabhöhle.) Diese Wohnung, guter Alter, hast du Zeit gehabt dir gefällig zu erbauen. — — Auch eine Grabschrift! — (Sie liest.) „Verzeiht, ihr fleißigen Wanderer des Lebens, daß ein Müßiger, nicht Müder, hier unverdiente Ruhe findet.“ Wie? so bin ich zu deinem Grabe gewallfahrtet, und für ein Königreich bieten mir die Götter deine friedliche Hütte zur Erbschaft an? Ich muß noch deine Gebeine darin segnen.

(Sie geht hinein.)

Abb. 7: mit Unterbrechung

3.2 Reading Order

Nachdem alle `Text Regions` ausgezeichnet wurden, muss die `Reading Order` definiert werden. Das ist besonders wichtig für die Weiterverarbeitung der Dateien nach dem gesamten OCR-Verfahren. Grundsätzlich wird die Reihenfolge entsprechend der Lesereihenfolge ausgezeichnet. Ausgenommen ist jener Text, der nicht in der Tabelle aufgeführt ist (z.B. Seitenzahlen). Auf manchen Seiten können Fußnoten auftreten. Diese werden in der Reading Order nach jener Stelle eingefügt, an der sie im Druck referenziert werden (siehe Abbildung 9). Nachdem die Seite bearbeitet wurde, sollten durch den Shortcut Strg + S alle Bearbeitungsschritte gespeichert werden, damit keine Änderungen verloren gehen können.

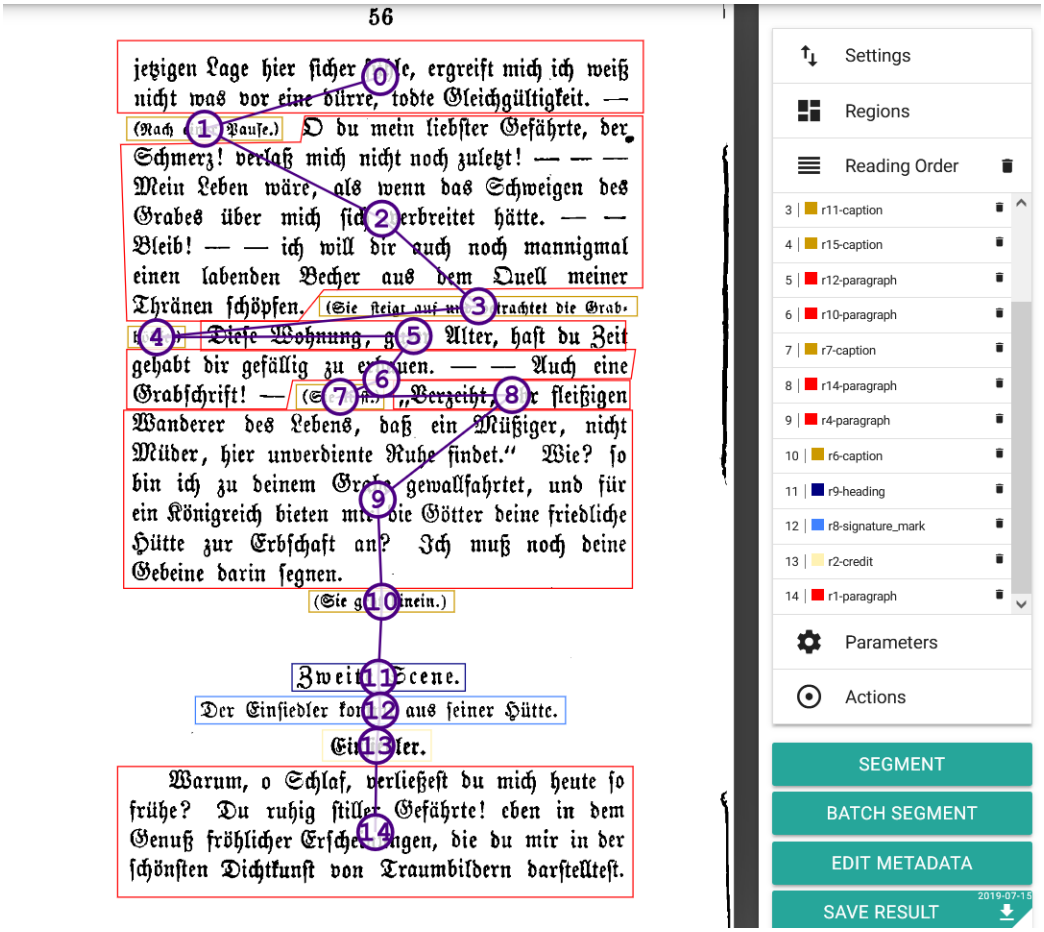


Abb. 8: ReadingOrder

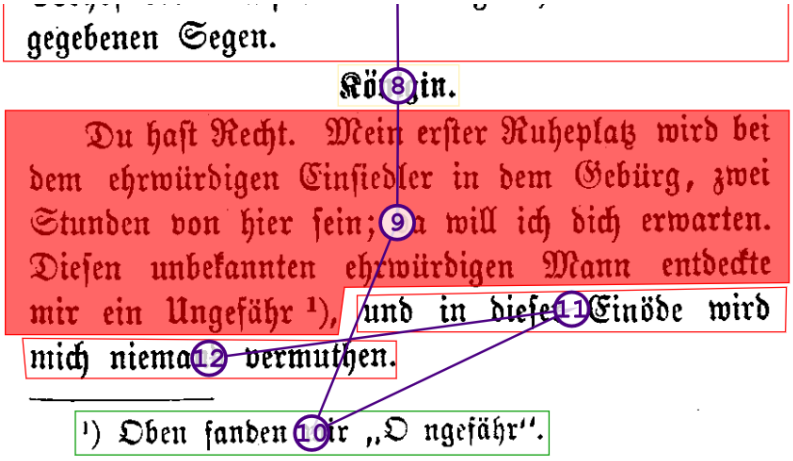


Abb. 9: Sonderfall Fußnote

4 Line Segmentation

Wenn die Segmentierung abgeschlossen ist und die Reading Order richtig definiert wurde, erfolgt in einem weiteren Schritt die `Zeilensegmentierung`. In diesem Vorgang werden innerhalb der zuvor definierten Text Regions die einzelnen Zeilen erkannt, um sie für die Texterkennung vorzubereiten. Hierzu wählt man im Menü Line Segmentation.

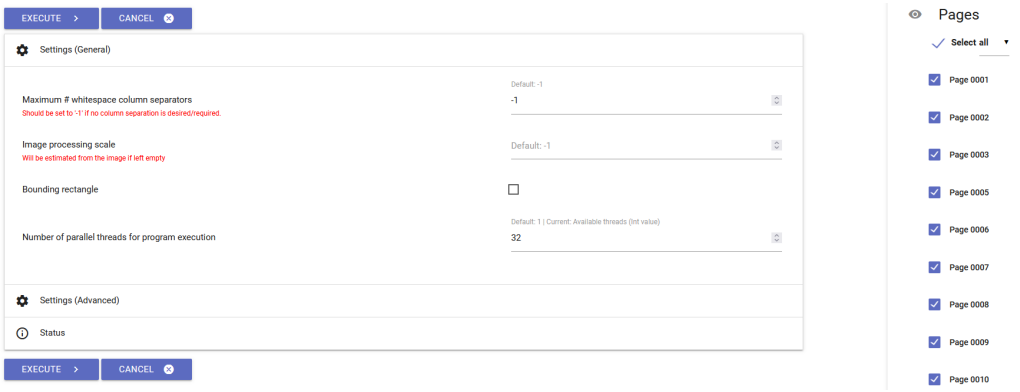


Abb. 10: Line Segmentation

Wie schon beim Preprocessing wird auch die Line Segmentation automatisch durchgeführt. Unter Settings (General) können verschiedene Parameter ausgewählt werden. In der Regel sind die Voreinstellungen jedoch ausreichend. In der Übersicht rechts werden nur die Seiten angezeigt, für die auch

eine Segmentierung stattgefunden hat. Hier gibt es wieder die Option, alle oder nur bestimmte Seiten auszuwählen. Mit einem Klick auf EXECUTE wird der Prozess gestarten. Sollte es zu Problemen kommen, kann es helfen, den Regler `Minimum scale permittet` unter Limit auf -2/-3 zu stellen und den Prozess erneut zu startet. Je mehr Seiten verarbeitet werden, desto länger dauert der Prozess. Wichtig ist dabei, dass der Rechner währenddessen eingeschaltet und mit dem Netz der Universität Würzburg verbunden bleibt. Unter Project Overview kann der Status der Zeilensegmentierung überprüft werden. Etwaige Fehler bei der Zeilensegmentierung können in LAREX behoben werden. In der Regel werden die Zeilen korrekt erkannt und nummeriert. Es kann jedoch vorkommen, dass Zeilen oder einzelne Zeichen (meist Initialen) getrennt und als zwei Zeilen erkannt werden. In diesem Fall gibt es zwei Optionen: Wenn z.B. nur die obere Hälfte der Initiale abgeschnitten ist und als Zeile 0 erkannt wird, die zweite Zeile jedoch in der Gänze erkannt wird, kann der zusätzlich als Zeile erkannte Teil einfach gelöscht werden.

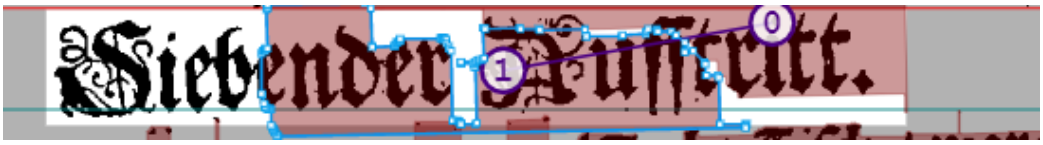


Abb. 11: getrennte Zeilen

In diesem Fall wurde eine Zeile als zwei Zeilen erkannt, zudem fehlt die Hälfte des ersten Wortes. Hier müssen beide Zeilen gelöscht und manuell neu gezogen werden (Shortcut 3).

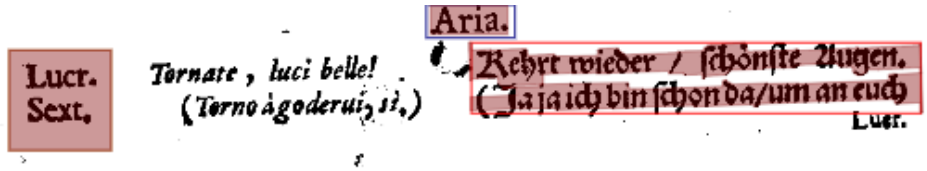


Abb. 12: fehlende Zeilen

Dasselbe gilt für Zeilen, die nicht als solche erkannt wurden. Hier wird wiederum mit Shortcut 3 manuell die Zeile markiert. In dem Beispiel sollten eigentlich die Sprecher in zwei Zeilen aufgeteilt sein, wurden aber nur als eine Zeile erkannt.

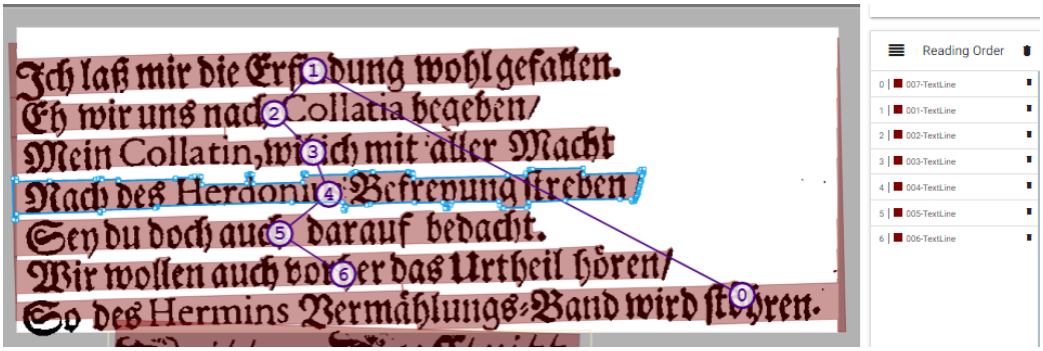


Abb. 13: fehlerhafte Zeilennummerierung

Diese Art von Fehler kann leicht übersehen werden. Die letzte Zeile wurde fälschlicherweise als erste Zeile erkannt. Hier muss die Reading Order korrigiert werden. Hierfür können rechts in der Box die Zeilen per Drag and Drop verschoben werden. Alternativ kann die Reading Order auch gelöscht und mit Shortcut R neu definiert werden.

5 Recognition

Bei der Recognition wird der Text der bereits erstellten Zeilen erkannt. Auch dieser Schritt wird über das Menü gestartet, indem zunächst in der Übersicht Recognition ausgewählt wird. Danach können wieder alle bereits zeilensegmentierten Seiten ausgewählt werden und der Prozess mit `EXECUTE` gestartet werden. Damit der Prozess gestartet werden kann, müssen die entsprechenden Schriftmodelle ausgewählt werden, mit denen das Programm den vorliegenden Text abgleichen soll. Für die zu bearbeitenden Dramen eignet sich am besten das Modell *default/deep3_fraktur-hist/0*. Damit die Recognition starten kann, müssen mindestens fünf Modelle ausgewählt werden, in diesem Fall also *default/deep3_fraktur-hist/0* - *default/deep3_fraktur-hist/4*. Nur wenn fünf Modelle aus der Liste ausgewählt wurden, lässt sich der Arbeitsschritt starten.

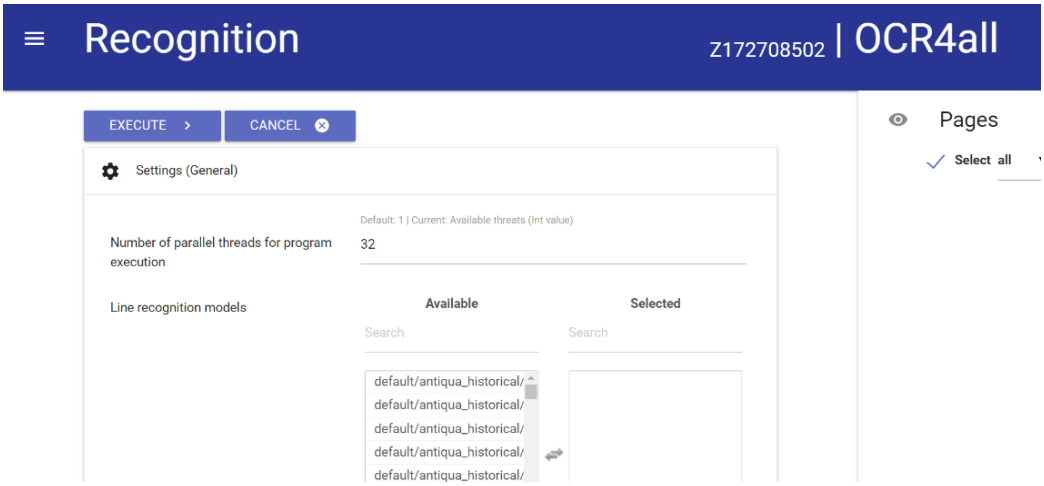


Abb. 14: Recognition

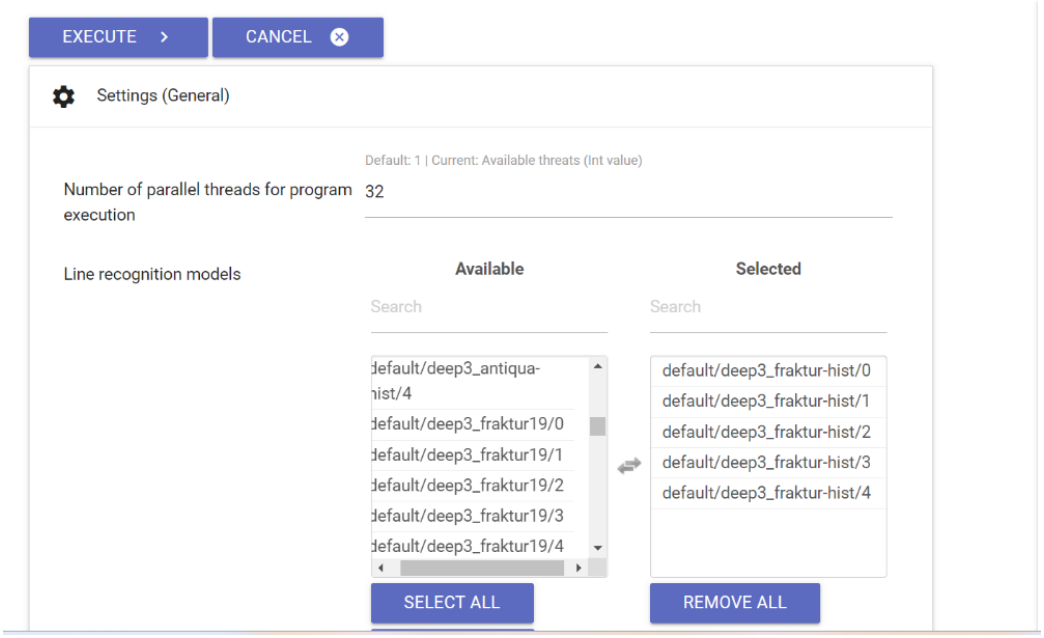


Abb. 15: Recognition

6 Ground Truth Production

Die Ground Truth Production wird zunächst über das Menü und die Schaltfläche `OPEN LAREX` gestartet. Anschließend wird für jede Zeile der vom Programm erkannte Text korrigiert. Wurde beispielsweise ein Schaft-s (ſ) fälsch-

licherweise als „f“ erkannt, wird dieser Fehler an dieser Stelle verbessert. Elementar ist, dass Schreibfehler im originalen Text, durchgestrichene und somit nachträglich veränderte Stellen oder jegliche Verunreinigungen durch suboptimale Scans so, wie sie ‘erkannt’ sind, übernommen werden. Die korrigierende Instanz soll also nicht das eigene Wissen oder Verständnis auf den Text projizieren. Häufige Fehler sind Verwechslungen zwischen g und a oder Leerzeichen. Gearbeitet wird mit zwei verschiedenen Ansichtsmöglichkeiten, `Text View` und `Page View`. Der Text View ist für die Arbeit mit Libretti vorteilhafter, da die Zeilen sehr unterschiedlich lang sein können und diese unterschiedlichen Längen im Text View übersichtlicher sind. Der Page View ist dagegen dann von Vorteil, wenn Fehler bei der Korrektur der Zeilensegmentierung übersehen wurden. Fehlerhaft erkannte Zeilen können so etwa besser zugeordnet werden.
Hinweise zur Ground Truth Production:

- Schlecht lesbare Stellen werden nach Möglichkeit ergänzt und bleiben ansonsten unkommentiert
- Anführungszeichen müssen nicht in ihrer Position korrigiert werden.

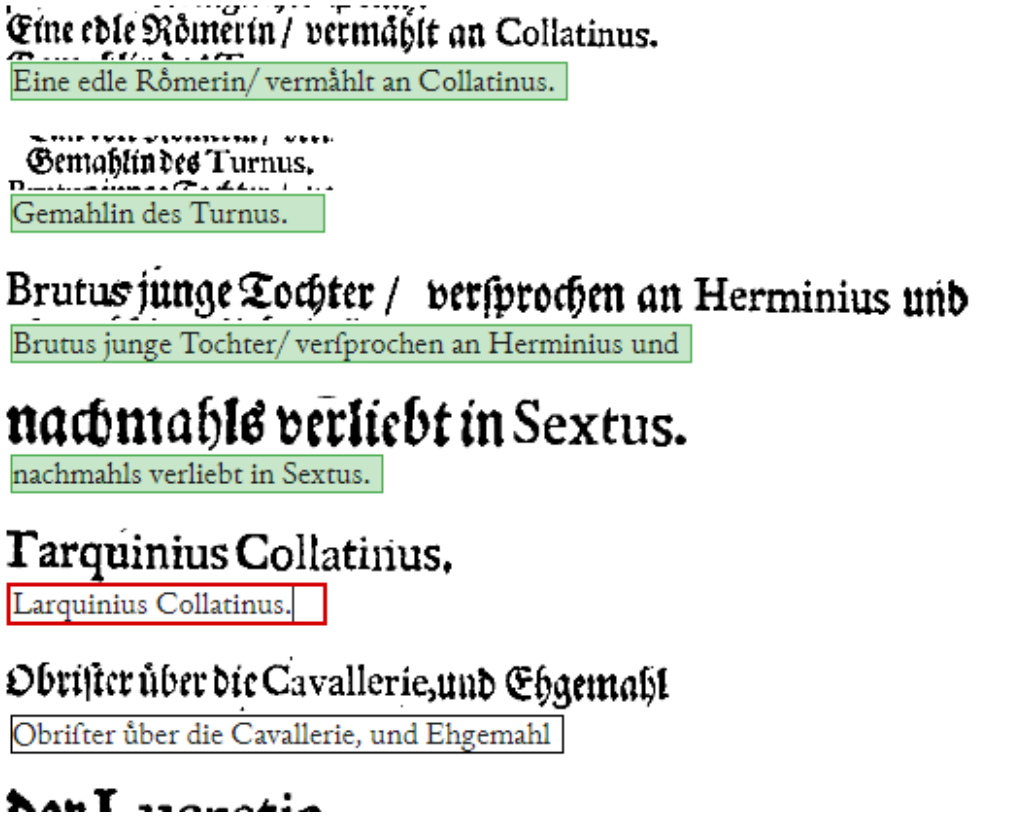


Abb. 16: Ground Truth Production

Begonnene Korrektur im Text View. Grün hinterlegt sind korrigierte und bestätigte Textzeilen, Weiß hinterlegt sind noch zu korrigierende Zeilen.

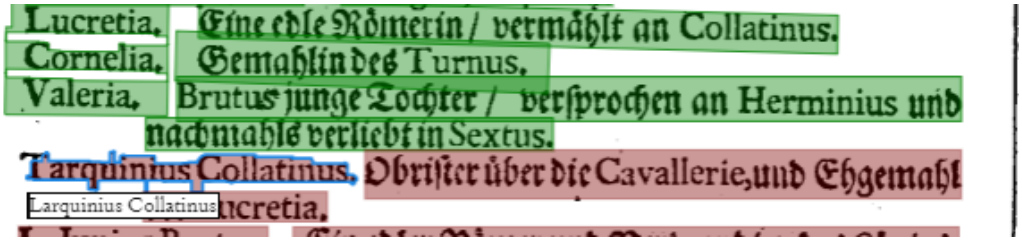


Abb. 17: Ground Truth Production

Der gleiche Text, hier im Page View. Die noch unkorrigierten Zeilen sind hier rot hinterlegt.

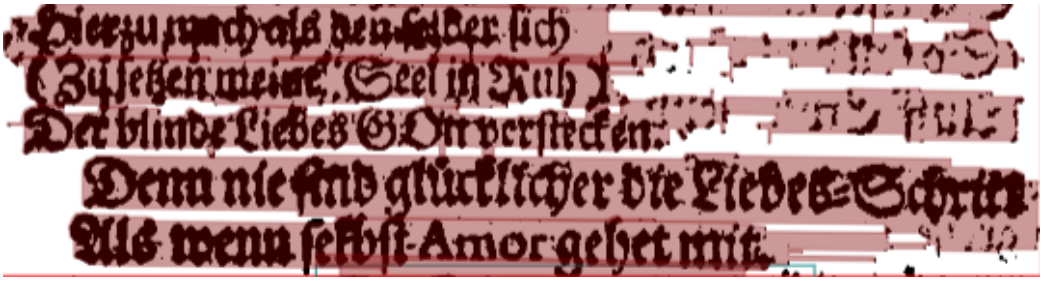


Abb. 18: Ground Truth Production

Beispiel für falsche automatische Linienziehung durch schlechte Scans – der Text der Vorseite ist durchgedrückt und wird fälschlicherweise als Teil der Zeile erkannt.

7 Post Correction

An dieser Stelle können die verschiedenen Schritte Segmente, Linien und Text im LAREX nochmals manuell angepasst werden. Zu beachten ist, dass eine Bearbeitung der höheren Ebene eine Nachbearbeitung auf niedrigerer Ebene nach sich zieht. Wird also zum Beispiel ein Segment gelöscht und neu gezogen, müssen auch die Linien und der Text neu gezogen bzw. erstellt werden.

8 Evaluation

Hierbei handelt es sich um einen sehr schnellen Schritt – er dient der Erwägung, wie erfolgreich ein Modell für den erkannten Text ist. Nachdem die fertigen Seiten ausgewählt wurden, die evaluiert werden sollen, kann man den Evaluationsprozess beginnen. Dabei wird berechnet, wie treffsicher die Transkription verlaufen ist, außerdem werden prozentuale Fehler angezeigt. Durch die Evaluation ist uns bekannt, dass die Fehlerquote bei über 3% begann und mittlerweile im Schnitt bei etwa 1,5% liegt, wobei wir auch schon mehrmals Fehlerquoten von 1,2% haben. Ziel ist ein Prozentsatz von $>1\%$.

9 Training

Mit dem Training wird das benutzte Modell werkspezifisch trainiert, mit dem Ziel, ein projektspezifisches und erfolgreiches Modell zu erstellen. Dieser

Prozess wurde bisher von Maximilian Wehner übernommen, ist aber automatisiert. Abhängig vom Werk und von Textmenge beträgt die Arbeitszeit für den Rechner etwa 1–2 Stunden. Die Auslagerung des Schrittes ist einerseits gut, da die Universitätsserver schneller sind. Allerdings kann die Absprache und das gegenseitige Warten den Prozess auch aufhalten.