

Zum Aufbau digitaler Dramenkorpora. OCR4alltoDraCorTEI als Baustein für die Edition von maschinenlesbaren Versionen historischer Dramendrucke

Dennerlein, Katrin

katrin.dennerlein@uni-wuerzburg.de

Julius-Maximilians-Universität Würzburg, Deutschland

ORCID: 0000-0003-0059-9597

Rupnig, Martin

martin.rupnig@stud-mail.uni-wuerzburg.de

Julius-Maximilians-Universität Würzburg, Deutschland

Reul, Christian

christian.reul@uni-wuerzburg.de

Julius-Maximilians-Universität Würzburg, Deutschland

Motivation und Ziel

Die Computational Literary Studies (CLS) können nur so gut sein, wie die Korpora, die ihnen zur Verfügung stehen. Insbesondere für die Geschichte des deutschsprachigen Dramas vom 17. bis 19. Jahrhundert repräsentieren diese bislang jedoch fast nur die hochkanonischen Texte. Die Präferenz liegt, wie bereits in den kodifizierten Literaturgeschichten der Germanistik auf original deutschsprachigen Sprechtheaterwerken und dabei vorwiegend auf Tragödien (Alt, 1994, Meid, 2009: 327-501, Schulz, 2007). Hingegen bleiben Libretti, populäre Komödien und generell Übersetzungen und Dramen von Frauen zumeist gänzlich unberücksichtigt, obwohl sie die Mehrheit der gedruckten und gespielten Werke ausmachen (Jahn, 1996, Krämer, 1998, Dennerlein, 2021, Kord, 1992). Dadurch ist die Geschichte des deutschsprachigen Dramas nicht nur äußerst lückenhaft, sondern entbehrt auch zahlreicher populärer und wegweisender Werke. Die Textauswahl für einzelne Autor:innen, Genres, Textgruppen wie Repertoires oder Sammlungen ist jeweils so klein, dass quantitative, genre- und periodenvergleichende Studien nur sehr eingeschränkt durchgeführt werden können. Für eine Erforschung der Gesetzmäßigkeiten der literarischen Evolution ist die gezielte Korpuserweiterung deshalb unabdingbar.

Auf edierte und normalisierte Neuausgaben, wie sie der Digitalen Bibliothek und den auf sie aufsetzenden Projekten Textgrid und GerDracor zu Grunde lagen, kann für diese Erweiterungen allerdings nicht zurückgegriffen werden, weil die fehlenden Dramentexte nicht neu ediert wurden.¹ Digitalisierungsvorhaben zu Dramen des 17.–19. Jahrhunderts müssen deshalb von Bilddigitalisaten von Fraktur-exemplaren ausgehen, wie sie inzwischen in großem Umfang zur Verfügung stehen.

Dabei stellen sich sowohl editorische Fragen der Transkription und Normalisierung als auch die Fragen der Automatisierung des TEI-Taggings. Im Folgenden sollen einige Vorgehensweisen zur Edition, Volltextdigitalisierung und Textauszeichnung historischer Dramentexte mit und im Anschluss an die Open Access-Software OCR4all² beschrieben werden, die als vollständig kostenfreie Digitalisierungssoftware für jedermann zugänglich ist. Vor allem soll ein Skript vorgestellt werden, mit dem man die Ergebnisse der OCR-Erkennung, die PAGE-XMLs, automatisch mit Elementen des DraCorTEI-Tagsets auszeichnen kann (Reul et al., 2019; Fischer et al., 2019). Dieses Skript kann im Anschluss an jede OCR-Software verwendet werden, solange die Ergebnisse als valides PAGE ausgegeben bzw. in dieses konvertiert werden können. Die ausführlichen und vollständigen Guidelines und das Konvertierungsskript sind auf Zenodo zu finden (Dennerlein/Rupnig/Kasenhofner, Dennerlein/Rupnig).

Segmentierung und Transkription historischer Dramendrucke

Das Layout von Dramendruckten um 1800 ist nicht normiert und variiert von Drucker zu Drucker, die prototypische Struktur ist jedoch wie folgt aufgebaut: Titelseiten enthalten Angaben zu Titel, Untertitel, Verfasser:in, Druckerei und zumeist auch zum Erscheinungsjahr. Es folgt das Personenverzeichnis inklusive der Figurenaufzählung, gefolgt von dem Beginn des Drameninhaltes mit folgender Struktur: Akt/Aufzug > Szene/Auftritt > Ortsangabe > Figurenaufzählung > Figurenname > Dialogtext > Regieanweisung. Je früher ein Dramentext erschienen ist, desto wahrscheinlicher ist es auch, dass er eine Vorrede vor dem Personenverzeichnis enthält. Üblicherweise entspricht die Reihung der Dramenelemente im Druck der tatsächlichen Lesereihenfolge. Jedoch gibt es Fälle, die dieser Logik nicht folgen. Ein Beispiel sind am Ende der Seite abgedruckte Fußnoten im Drama „Dido“ (1794) von Charlotte von Stein, die bestimmte Textstellen kommentieren (vgl. Abb. 1).

Eine weitere Besonderheit stellt die uneinheitliche Gestaltung von Figurenaufzählungen dar. Üblicherweise beginnen Szenen mit einer Aufzählung aller in der Szene auftretenden Figuren gefolgt von den Dialogen. Einige Dramen verzichten jedoch in einzelnen Szenen auf die Aufzählung ganz oder integrieren die Nennung der Figuren in die anfängliche Regieanweisung. Um alle diese Elemente berücksichtigen zu können, sollte eine Digitalisierungsumgebung

gewählt werden, die eine differenzierte semantische Auszeichnung von Segmenten erlaubt. Da es bei den knappen Ressourcen im wissenschaftlichen Bereich unabdingbar ist, eine kostenfrei nutzbare Digitalisierungsumgebung zu verwenden, die dennoch bestmögliche Ergebnisse liefert und stetig gewartet und aktualisiert wird, bietet sich OCR4all an.³ Integriert in OCR4all ist die Bearbeitungsumgebung LAREX, in der, basierend auf dem PAGE-Schema, die folgenden Regionsbezeichnungen für die jeweiligen Layoutelemente verwendet werden können:

Tabelle 1: Zuordnung von Layoutelementen in Drucken zu Layoutregionen in LAREX

Layoutelement	Layoutregion
Haupttext (gesprochener Text)	paragraph
sonstiger Text (z. B. Text in Vorreden)	other
Sprecher:angabe zum Dialogtext (Name der Figur)	credit
Sprecher:angabe zum Dialogtext bei mehreren Sprecher:innen (z. B. <i>Chor</i>)	drop_capital
Regieanweisungen innerhalb des Dialogtextes (Regieanweisungen, die sich auf die sprechende Figur beziehen)	caption
Fußnote	footnote
Akt oder Aufzug	header
Szene oder Auftritt	heading
Überschrift (Gesang, Gedicht, z. B. "Aria", "Ballade")	floating
Strophenummerierung	endnote
sonstige Überschriften, alle Angaben 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 (z. B.: Replik Figur a, Regieanweisung z. B. „Figur b geht ab“, Replik Figur b)	signature_mark
Bild, Holzschnitt, Diagramm, Tabelle, Initiale, Zierinitiale, Formel...	Image

Einige Elemente werden nicht als Layoutregionen ausgezeichnet und werden deshalb bei der späteren Texterkennung nicht berücksichtigt. Dazu zählen insbesondere Seitenzahlen, Seitentitel und Kustoden. Auch die Seitenumbrüche gehen verloren, nicht jedoch die Zeilenumbrüche, die automatisch bei der Zeilensegmentierung in LAREX erkannt werden. Zentrale Eigenschaften des Drucks gehen auf diese Weise verloren, dafür wird der Segmentierungsprozess bzw. die händische Nachkorrektur der Segmente etwas beschleunigt und das Hauptziel – maschinenlesbare, für die Zwecke der CLS verwendbare Dramentexte zur Verfügung zu stellen – erreicht. Repliken werden nicht durch Seitenumbrüche, Seitenzahlen oder Kustoden unterbrochen und dadurch unbrauchbar für Stilometrie, Topic Modelling oder Sentiment bzw. Emotion Analysis (vgl. Dennerlein et al., 2023). Ziel ist es nicht, eine diplomatische Transkription zu erstellen, sondern die Datenpublikation zu gewährleisten (vgl. Sahle 2013, Teil 2: 256-266). Daher ist auch ein bestimmter Umgang in OCR4all mit druckspezifischen Zeichen wie Superskripte oder die Verwendung von Schaft-S zu gewährleisten. In OCR4all können unterschiedliche Ergebnisse erzielt werden, je nachdem welches Modell für die Texterkennung angewendet wird. Mit dieser Methodik ist es unerheblich, ob das Modell auf die exakte Erkennung der konkreten Zeichen trainiert ist oder bereits eine Normalisierung ausgewählter Zeichen berücksichtigt. Die Normalisierung ausgewählter Zeichen für die DraCor-

TEI-Datei kann in einem weiteren Schritt über das Konvertierungsskript vorgenommen werden.

Markierung der Reading Order

LAREX bietet zudem die Möglichkeit, die Lesereihenfolge der Textregionen individuell anzupassen, sodass Sonderfälle wie Fußnoten in der Weiterverarbeitung an der entsprechenden Stelle platziert werden können. Abb. 1 zeigt die angepasste "Reading Order" in LAREX:

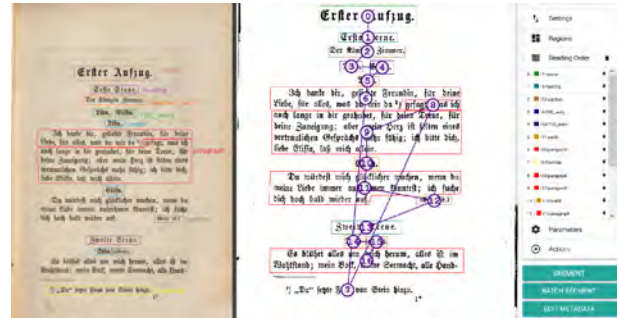


Abb.1: Geänderte Reading Order bei Fußnoten

Nach Akt, Szene, Regieanweisung und Figurenangaben folgt der erste Satz der Replik, dann ist die Fußnote eingegliedert. Sie ist damit hinterher als zugehörig zu diesem Satz identifizierbar und kann – etwa als Endnote – die spätere digitale Edition des Dramas ergänzen. Um später alle Elemente automatisch mit XML-Elementen auszeichnen zu können, lohnt sich die akribische Vorarbeit der Regionsauszeichnung, die bei entsprechender Übung nur etwa 5 Minuten pro Seite dauert. Mittelfristig sollen diese Auszeichnungen als Trainingsdaten verwendet werden, um den Automatisierungsgrad kontinuierlich zu steigern.

Anschließend werden die ausgezeichneten Regionen vollautomatisch in Zeilen zerlegt. Es folgt die eigentliche Texterkennung, bei der aus den Bildzeilen der maschinenlesbare Text mittels Modellen extrahiert wird. Hierbei können entweder existierende, „gemischte Modelle“ direkt angewendet werden oder „werksspezifische Modelle“, durch gezieltes Training auf die Erkennung einer bestimmten Drucktype hin optimiert werden. Derzeit kann bei Erkennung mit einem gemischten Modell zuverlässig eine Zeichengenauigkeit von über 98 % erreicht werden, meist sogar über 99 %. Durch das Training gemischter Modelle kann die Genauigkeit noch deutlich weiter gesteigert werden. Die dazu benötigten Trainingsdaten können ebenfalls in LAREX erstellt werden und bestehen aus Bildzeilen sowie der korrekten Transkription des darauf zu sehenden Texts.

Das Ergebnis des Digitalisierungsprozesses in OCR4all, wie auch von fast allen anderen OCR-Programmen, sind PAGE-XMLs, die neben dem gesamten Textinhalt der einzelnen Seiten weitere Informationen wie Erstellungsdatum, Metadaten, Layoutregionen und Koordinaten ent-

halten. Die Angabe der Reading-Order ist für jene Fälle wichtig, in denen die Koordinaten von Textregionen als Information nicht ausreichen, um die gewünschte Struktur in DraCorTEI abzubilden.

Konvertierungsskript

Mit EzDrama existiert bereits ein Konvertierungsskript, das Dramen, die im Plaintext in lateinischer Schrift vorliegen, teilautomatisch mit den Tags von DraCorTEI auszeichnet (Skorinkin et al., 2022). Da auch diese Dramen in der Regel nicht vollständig automatisch in TEI konvertiert werden können, wurde eine Markup-Sprache mit sehr wenigen Markups und Regeln entworfen, bei der Markierungen für bspw. Szenen, Sprecher:in oder Regieanweisungen direkt in den Text hineingeschrieben werden. In einem Colab oder mithilfe eines Jupyter-Notebooks können die so ausgezeichneten Dramen dann mit TEI-Tags ausgezeichnet werden, wie sie in DraCor verwendet werden. Dazu müssen jedoch zwei Voraussetzungen erfüllt sein: Zum einen müssen die Texte bereits maschinenlesbar in Form von lateinischer Druckschrift vorliegen, zum anderen muss Zeit für eine händische Auszeichnung der Dramen in einem proprietären Format aufgewendet werden. Das im Folgenden vorgestellte Skript versteht sich ausdrücklich als Ergänzung dieser verdienstvollen Arbeit und ist für diejenigen Fälle vorgesehen, in denen beide Voraussetzungen nicht gegeben sind.

Das im Folgenden kurz charakterisierte Skript wandelt die PAGE-XMLs in eine gültige DraCorTEI-Datei um. Das Skript spielt eine zentrale Rolle in der automatisierten Erfassung und Verarbeitung der Dramendrucke, indem es spezifische Merkmale und Besonderheiten der Dramendrucke adressiert. Es beinhaltet vier Klassen, die folgenden Zwecken dienen:

1. Die PAGE-Klasse (page class) repräsentiert eine Seite des OCR-Dokuments und enthält Funktionen zur Verarbeitung und Sortierung der Textregionen. Erst wird eine Seite initialisiert und die Lesereihenfolge analysiert. Anschließend werden die Textregionen nach der definierten Reading Order sortiert.
2. Die TextRegion-Klasse repräsentiert eine Textregion innerhalb einer Seite und enthält Funktionen zur Verarbeitung der Textlinien. Die innerhalb der Textregionen vorhandenen Zeilen werden nach ihren Koordinaten sortiert.
3. Die TextLine-Klasse repräsentiert eine einzelne Textzeile innerhalb einer Textregion. Hier wird der Textinhalt der Textlinie extrahiert. Das Skript wählt hierzu jene Zeilen aus, die in OCR4all händisch korrigiert wurden, sollte dies der Fall sein. Liegen keine Korrekturen vor, wird der von dem Modell erkannte Text extrahiert.
4. Die Conversion-Klasse führt schließlich die Konvertierung der OCR-Daten zu DraCorTEI durch. Erst wird der Pfad zu den OCR-Daten initialisiert, danach die

Hauptkonvertierung durchgeführt und die TEI-Datei erstellt, sowie der Front- und Body-Bereich der Datei aufgebaut.

Das Skript verarbeitet der Reihe nach alle in einem Ordner liegenden PAGE-XML-Dateien und arbeitet sich von der äußersten zur innersten Ebene vor: Seite > TextRegion > TextLine > Wort > Glyph. Bei Initialisierungsfehlern bei der Verarbeitung werden Fehlermeldungen mit Angaben zur entsprechenden Stelle und dem Dateinamen ausgegeben, um Fehler in der Vorbearbeitung oder Beschädigungen in den Dateien zu finden.

Für den Fall, dass das Skript Textregionen verarbeitet, die falsch platziert sein sollten oder deren Inhalt einer falschen Textregion entsprechen, wird in der DraCorTEI-Datei an entsprechender Stelle ein Tag mit dem Inhalt "WARNING" ausgegeben, die eine notfalls mögliche händische Nachbearbeitung ermöglicht. Auch in diesem Fall soll gewährleistet werden, dass Fehler in der Vorbearbeitung in OCR4all auffindig gemacht und korrigiert werden können.

Eine der Besonderheiten, die in der Bearbeitung durch das Skript berücksichtigt werden, sind spezielle Zeichen, die für die endgültige DraCorTEI-Datei normalisiert werden müssen. Folgende regelmäßig vorkommende Zeichen werden dementsprechend normalisiert: $\text{f} \rightarrow \text{s}$, $\text{z} \rightarrow \text{z}$, $\text{æ} \text{ or } \text{ʒ} \rightarrow \text{ä}$ ö ü , etc. Sollten in etwaigen Projekten, die dieser Methodik folgen, weitere spezielle Sonderheiten in den Drucken auftreten, können diese durch kleine Anpassungen im Skript mit aufgenommen werden.

Ausblick

Mit dem hier beschriebenen Verfahren der Auszeichnung, der Transkription mit OCR4all und der XML-Kodierung, benötigt die Digitalisierung eines Dramas derzeit noch immer acht bis zehn Stunden. Verzichtet man auf die Korrektur des automatisch erfassten OCR-Textes, weil man bspw. sehr große Textmengen erfassen möchte, bei denen Fehler nicht mehr ins Gewicht fallen, verkürzt sich die Bearbeitungszeit um die Hälfte. Für andere Zeitabschnitte müsste das Verfahren zudem angepasst werden, wenn das Layout signifikant abweicht. Ein besonderer Fall sind bspw. Drucke wie die Libretti der *Hamburger Gänsemarktoper* aus der Zeit um 1700, bei denen sich die Sprecher:inbezeichnung in einer eigenen Spalte links neben dem Haupttext befinden und die Regieanweisungen oftmals rechtsbündig stehen. Für diese Fälle wurde bereits ein Verfahren getestet, bei dem die Sprecher:in in der Reading-Order zunächst an die letzte Stelle gesetzt werden und erst später wieder über die Koordinaten der Regionen vor die zugehörigen Textteile gesetzt werden.

Insgesamt ist zu bedenken, dass der Prozess der Volltextdigitalisierung historischer Dramentexte verhältnismäßig komplex ist und dass man für ideale Ergebnisse deutlich mehr Zeit investieren muss als etwa für Prosatexte. Die händisch segmentierten und korrigierten Daten können jedoch als Trainingsdaten genutzt werden, so dass der Auto-

matisierungsgrad in Zukunft sukzessive gesteigert werden kann.

Fußnoten

1. [https://de.wikipedia.org/wiki/Digitale_Bibliothek_\(Produkt\)](https://de.wikipedia.org/wiki/Digitale_Bibliothek_(Produkt)) , <https://textgrid.de/> , <https://dracor.org/ger>.
2. <https://www.ocr4all.org>.
3. Im konkreten Anwendungsfall haben wir uns für OCR4all entschieden, da dieses, abgesehen von der freien Verfügbarkeit und der Tatsache, dass es am *Zentrum für Philologie und Digitalität* entwickelt wird, mit LAREX über ein sehr mächtiges Annotations- und Korrekturwerkzeug verfügt, das uns subjektiv als am besten geeignet erschien.

Bibliographie

- Digitale Bibliothek** [https://de.wikipedia.org/wiki/Digitale_Bibliothek_\(Produkt\)](https://de.wikipedia.org/wiki/Digitale_Bibliothek_(Produkt)),
GerDraCor, <https://dracor.org/ger>
TextGrid <https://textgrid.de/>
OCR4all <https://www.ocr4all.org>
PAGetoDraCorTEI: <https://github.com/dennerlein/Dramendigitalisierung-PAGetoDraCorTEI>
Alt, Peter-Andre. 1994. *Tragödie der Aufklärung: Eine Einführung*. Tübingen: Francke.
Dennerlein, Katrin. 2021. *Materialien und Medien der Komödiengeschichte. Zur Praxeologie der Werkzirkulation zwischen Hamburg und Wien von 1678–1806* (Studien und Texten zur Sozialgeschichte der deutschen Literatur 152). Berlin/New York: de Gruyter. <https://doi.org/10.1515/9783110691191>
Dennerlein, Katrin, Thomas Schmidt und Christian Wolff. 2023. "Computational emotion classification for genre corpora of German tragedies and comedies from 17th to early 19th century" *Digital Scholarship in the Humanities*, Vol. 38, Issue 4, 1466–1481. <https://doi.org/10.1093/lc/fqad046>
Dennerlein, Katrin, Martin Rupnig und Nadine Kastenhofer. Guidelines zur Volltextdigitalisierung von Dramen des 17 bis 19. Jahrhunderts mit OCR4all. "Zenodo" 2024 (DOI: <https://doi.org/10.5281/zenodo.12805233>)
Dennerlein, Katrin und Martin Rupnig. PAGetoDraCorTEI. 2024 <https://github.com/dennerlein/Dramendigitalisierung-PAGetoDraCorTEI>
Fischer, Frank, Ingo Börner, Matthias Göbel, Angelika Hecht, Christopher Kittel, Carsten Milling, Peer Trilcke. 2019. "Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama." *Digital Humanities 2019: "Complexities"* (DH2019), Utrecht. <https://doi.org/10.5281/zenodo.4284002>
Krämer, Jörg. 1998. *Deutschsprachiges Musiktheater im späten 18. Jahrhundert: Typologie, Dramaturgie*

und Anthropologie einer populären Gattung. Tübingen: Niemeyer.

Meid, Volker. 2009. "Die deutsche Literatur im Zeitalter des Barock. Vom Späthumanismus zur Frühaufklärung 1570–1740" (*Geschichte der Deutschen Literatur*, De Boor, Helmut und Newald, Richard, eds.). München: Beck.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andras Büttner und Frank Puppe. 2019. "OCR4all — An open-source tool providing a (semi-) automatic OCR workflow for historical printings" *Applied Sciences* 9(22) <https://www.ocr4all.org/>.

Sahle, Patrick. 2013. *Digitale Editionsformen - Teil 2: Befunde, Theorie und Methodik. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels*. Norderstedt: Books on Demand.

Schöch, Christof. 2017. "Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama" *Digital Humanities Quarterly*, Vol. 11, 2 <http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>

Schulz, Georg Michael. 2007. *Einführung in die deutsche Komödie*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Skorinkin, Daniil, Luca Giovanni und Peer Trilcke. 2022. *EzDrama*. <https://github.com/dracor-org/ezdrama>