

# Layout und (Para-)Text: Erprobung hybrider Ansätze und Heuristiken zur Erforschung von Werkausgaben des 18. Jahrhunderts

**Bogdanović, Arsenije**

arsenije.bogdanovic@ilw.uni-stuttgart.de  
Universität Stuttgart, Deutschland

**Lange, Liesen-Sophie**

lilange@uni-mainz.de  
Johannes Gutenberg-Universität Mainz, Deutschland

**Ajouri, Philip**

pajouri@uni-mainz.de  
Johannes Gutenberg-Universität Mainz, Deutschland

**Viehhauser, Gabriel**

gabriel.viehhauser@univie.ac.at  
Universität Wien, Österreich  
ORCID: 0000-0001-6372-0337

## Einführung

In der digitalen Literaturwissenschaft werden meist weitumspannende Konzepte wie Genres oder Epochen in den Blick genommen. Die dort entwickelten Methoden können jedoch auch durchaus gewinnbringend zur Untersuchung von buchgeschichtlichen Fragestellungen zum Einsatz gebracht werden. Das in Kooperation von Buchwissenschaft und Digital Humanities durchgeführte DFG-Projekt „Scalable Reading von ‚Gesammelten Werken‘ des 18. Jahrhunderts, exemplarisch durchgeführt an Friedrich-von-Hagedorn-Werkausgaben“ widmet sich dementsprechend der Erforschung des Buchtypus der Gesamtausgabe, der sich im 18. Jahrhundert etabliert hat und zu einem zentralen Medium des Buchmarkts geworden ist, das für das Selbstverständnis von Autor\*innen sowie für Leser\*innen, Verlage und Bibliotheken große Bedeutung erlangt hat.

Die Untersuchung der Ausbildung dieses Typus lässt sich mit quantitativen Methoden unterstützen, im Projekt werden etwa die – mengenmäßig durchaus umfangreichen – Werkausgaben Friedrichs von Hagedorn exemplarisch untersucht. Dabei gilt es, die Unterschiede in der Zusammenstellung dieser Werkausgaben mit einem Scalable-Rea-

ding-Ansatz (Mueller 2014) in den Blick zu bekommen, um so Aussagen über die Ausprägung des Buchtyps treffen zu können. Dafür wird im Projekt ein zweistufiger Workflow anvisiert: Zunächst werden unterschiedliche Methoden von Layout-Erkennungen eingesetzt, um möglichst automatisch die einzelnen Textteile der Werkausgaben identifizieren zu können. Diese werden in einem zweiten Schritt mit Methoden der Text-Reuse-Detection miteinander aligniert und auf Abweichungen bzw. Parallelen hin untersucht. Der vorliegende Beitrag widmet sich dem ersten Schritt dieses Workflows, der Layouterkennung der Werkausgaben des 18. Jahrhunderts.

## Historische Werkausgaben



Abb. 1: Eine Seite mit voller paratextueller ‚Ausstattung‘ aus der Bohn-Ausgabe der „Oden und Lieder“ Hagedorns von 1769. Digitalisat bereitgestellt durch die ULB Sachsen-Anhalt (VD18-Nr.: 11043512)

Wie der Blick auf ein exemplarisches Faksimile einer Hagedorn-Ausgabe in Abbildung 1 verrät, ist es zunächst die ausgeprägte Materialität und Paratextualität von historischen Werkausgaben, die unweigerlich ins Auge fällt.

Dies spiegelt sich u.a. in der Auswahl von Format/Papier, Druckschrift, Layout; der Ausgestaltung von Titelseiten, Kupferstichen, Vignetten; dem Umfang, der Autorschaft und Tendenz von Titeln, Vor- und Nachworten, ‚Beigaben‘ und ‚Nachlesen‘. Bei Hagedorn interessieren darüber hinaus dessen ausgiebigen Anmerkungen (Münster 1999, 10), welche mit der Zeit durch editorische Eingriffe immer weiter aus dem Blickfeld des Publikums verbannt wurden.

Paratexte gliedern und framen das ‚Werk‘ somit immer aufs Neue, steuern die Rezeption und ermöglichen – einmal systematisch in ihrem Wandel erfasst – anderweitige kulturgeschichtliche Rückschlüsse und Hypothesenbildungen (Ajourri 2017).

## Taxonomie von Buchteilen

Um eine so gegebene Komplexität weitestgehend aufzufangen, wurde im Projekt zunächst eine Taxonomie gängiger Kern- und Paratextformen erstellt, wobei Umfang und Granularität anpassbar bleiben. Ankerpunkt hierfür waren die Arbeiten von McConnaughey et al. (2017) und Underwood (2014), nicht zuletzt die Erläuterungen zur formalen/inhaltlichen Erschließung von Volltexten des DTA-Basisformats (<https://www.deutschestextarchiv.de/doku/basisformat/>) Daran schloss eine buchwissenschaftliche Recherche an, begleitet durch die Sichtung repräsentativer Werkausgaben des Autors und seiner Zeitgenossen. Somit sind die mesoanalytischen Einheiten auf Seiten-Ebene (z.B. Textsorten wie Vorwort, Widmung, Lobgedicht u.v.m.) und ihre Spielarten begrifflich erschlossen und können im Weiteren nachmodelliert werden. Eine Differenzierung von Kerntexten nach zumeist irreführenden Genrebezeichnungen wurde zugunsten einer möglichst formalen, typographischen Aufteilung unterlassen.

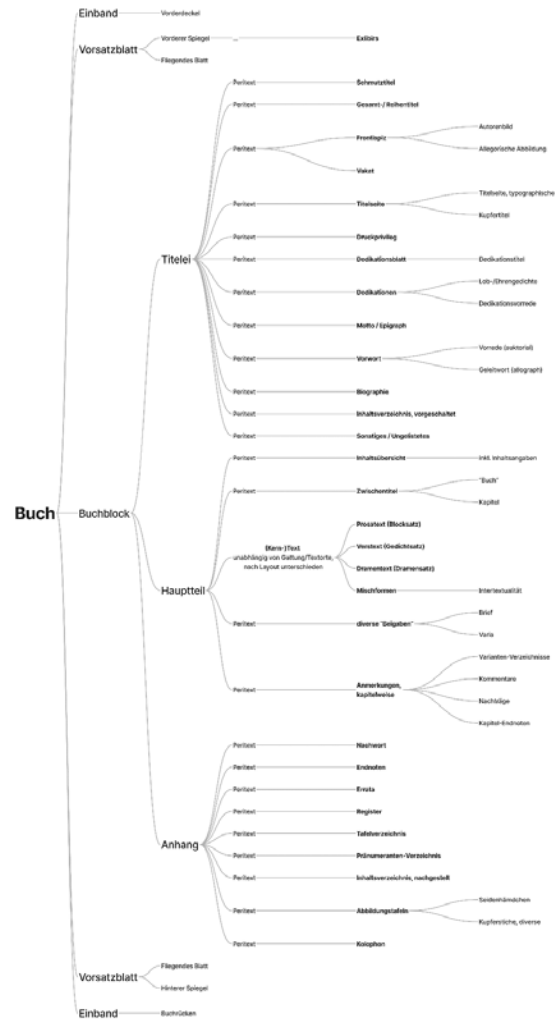
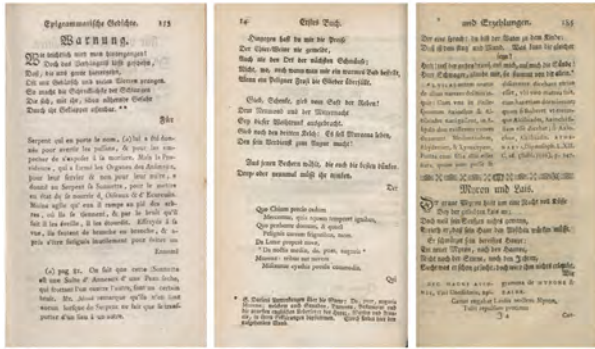


Abb. 2: Taxonomie von Buchteilen nach aktuellem Stand.

Bei Einheiten unterhalb der Seitenebene wächst erwartungsgemäß der Ambiguitätsgrad, weshalb sie iterativ neu verhandelt werden. So etwa die genaue Einordnung von (a) rekursiven Anmerkungen; (b) Zitaten, die mal indiziert in den Apparat eingebunden, mal selbstständig stehen; oder (c) ‚schwebenden Fußnoten‘, die an Gedichttexte anschließen können und einem weiteren Kerntext voranstehen, aber keine Endnoten sind (s. Abb. 2).



a) b) c) Abb. 3: Unterschiedliche Ausprägungen von Paratextualität bei Hagedorn, hier v.a. bezogen auf den Anmerkungsapparat. Digitalisate bereitgestellt durch die SUB Göttingen (VD18-Nr.: 10857397), ULB Sachsen-Anhalt (VD18-Nr.: 90332156) und SB-Berlin (VD18-Nr.: 11676728).

Um auch hier die Komplexität einzudämmen, wurde eine basale Typologie beobachteter Layout-Konstellationen erstellt.

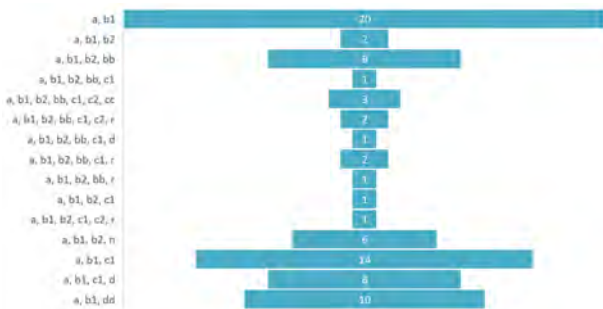


Abb. 4: Übersicht auftretender Kombinationen von Seitenlayouts bezogen v.a. auf den Anmerkungsapparat innerhalb eines Werkausgabenexemplars. Legende: a = einspaltiger Haupttext; b1 = einspaltige Fußnote, b2 = zweispaltige Fußnote, bb = zweispaltige Fußnote neben einspaltigem Zitat; c1 = 'schwebende' einspaltige Anmerkungen/Zitate, c2 = 'schwebende' zweispaltige Anmerkungen, cc = 'schwebende' zweispaltige Anmerkungen neben einspaltigem Zitat; d = Text-Endnoten (können kontextfrei mit b1 und c1 zusammenfallen), dd = Buch-Endnoten; n = Notensatz; r = Registersatz (bei Inhaltsverzeichnissen).

Daraus lassen sich tendenziell auch typographische Vorlieben der jeweiligen Verleger ablesen, was über das Hagedorn-Korpus hinaus von Interesse sein dürfte, da es sich hierbei um im gesamten deutschsprachigen Raum tätige Akteure handelte (insgesamt 18 Verleger bzw. Ko-Verleger, s.u. Abb. 5). Das Hagedorn-Korpus hat damit ein hohes exemplarisches Potential, dessen Übertragung auf andere Gesamtausgaben lohnend erscheint.

Die Layouttypen lassen sich weiter auf einzelne Seitentypen zerlegen und mit anderen Eigenschaften kombinieren (Format, Druckschrift), um neue Cluster sichtbar zu machen. Seitentypen lassen sich wiederum weiter auf Komponenten runterbrechen und zur Generierung theoretisch möglicher Layouts verwenden, was in Zeiten kostspieliger Ground Truth seine Vorteile haben kann (vgl. Fleischhaker 2024).

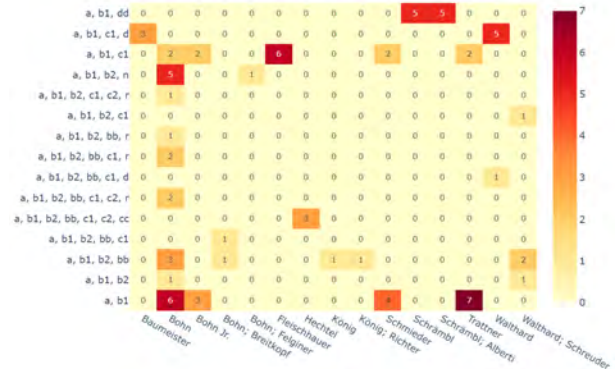


Abb. 5: Heatmap zur Verteilung von Layouttypen nach (Ko-)Verlegern.

## Gewinnung von Strukturdaten

In einem weiteren Schritt gilt es nun, TEI- oder vergleichbaren XML/JSON-Dateien zu erzeugen, in denen die Texte nach den Maßgaben dieser Taxonomie ausgezeichnet sind. Dadurch wären sie leicht auf ihre Position, Inhalt, Umfang und Abfolge bzw. die daraus folgenden (para-)textuellen Konstellationen hin beforschbar.

Eine manuelle Erstellung derart detaillierter Ausgaben erscheint jedoch freilich zu aufwändig; in Ausnahmefällen, wo solche vorliegen (dreibändige Bohn-Ausgabe im DTA-Archiv, [https://www.deutschestextarchiv.de/book/show/hagedorn\\_sammlung01\\_1742](https://www.deutschestextarchiv.de/book/show/hagedorn_sammlung01_1742)), sind sie jedenfalls als Ground Truth-Ergänzung willkommen. Eine potenzielle Hilfestellung bieten die für VD18-Digitalisate vorgesehenen Inhaltserschließungen in METS XML, die sich allerdings auf die Seitenebene beschränken und mit unterschiedlicher Granularität und Konsequenz eingepflegt wurden. Allenfalls liegen für die Hälfte der rund 80 Hagedorn-Digitalisate Kapiteleinteilungen, für ein Viertel zusätzlich Gedichtüberschriften vor, was u.a. eine erste Exploration von vorkommenden Buchteilen ermöglicht. Für eine genauere Untersuchung müssen jedoch Methoden eruiert werden, die zur automatischen Auszeichnung der Teile beitragen.

## Herausforderungen des Layouts für die OCR und Dokumentenanalyse

Aus den genannten Gründen müssen OCR-Verfahren – genauer die der *Document Layout Analysis* (DLA) herangezogen werden, um die avisierten Einheiten aus vorhandenen Buchscans – wenigstens ansatzweise – rekonstruierbar zu machen.

Die DLA ist eine unentbehrliche Vorstufe der Zeichenerkennung (OCR im engeren Sinne), steht aber im Gegensatz zu dieser weiterhin vor nicht wenigen Schwierigkeiten. Defizite und Desiderata im Bereich historischer Drucke sind in der Community allgemein bekannt und be-

treffen Segmentierung wie *Semantic Labeling* (z.B. Girdhar et al. 2024) Offenkundig wird das etwa, wenn im Rahmen von Massendigitalisierungsbemühungen entstandene Volltexte eingesehen werden. Je nach angewandter Software kann es sich um mehr oder minder unstrukturierter Output handeln. Daraus extrahierte Texte geben etwa durch fehlerhafte Spaltenentrennung – um nur eine berückichtigte Fehlerquelle zu nennen – auch bei verhältnismäßig hoher OCR-Genauigkeit unentwirrbare Wortgeflechte wieder. Hinzu kommen fehlende semantische Unterscheidungen, wodurch etwa Absatz- und Anmerkungsgebiete ineinanderfließen; grafische Elemente bleiben zumeist komplett ausgeblendet oder werden fälschlicherweise als Textteile erkannt. All dies kann nachfolgende Analyseschritte stark beeinträchtigen (Bartsch et al. 2023; Laramée 2019). Berücksichtigt man die oben dargestellte Variabilität gewinnt die Layoutproblematik für das Hagedorn-Korpus zusätzlich an Gewicht.

## Computer Vision- und hybride Ansätze zur Layouterkennung

Tabelle 1 gibt eine Übersicht über den allgemeinen IoU (Intersection over Union) für Textregionen, einem nicht ganz unproblematischen aber indikativen Standardmesswert (Subramanyam 2021), für drei ausgewiesene Open-Source-Engines. Je nach Bearbeitungsebene (Regionen vs. Zeilen) weisen diese zudem unterschiedliche Schwächen und Stärken (Performance vs. Computationszeit etc.) sowie Ausbaumöglichkeiten (Modell-Training) auf. So ist etwa die hier nicht abgebildete Zeilenerkennung das eigentliche ‚Kerngeschäft‘ von OCR-Engines, auf die eher zu bauen ist.

**Tabelle 1:** Evaluation für Textregionen-Erkennung an einem 100-seitigem Probestock ‚schwieriger‘ Layout-Typen, u.z. mit: Tesseract 5.3.4 ( <https://github.com/tesseract-ocr/tesseract> und [https://github.com/OCR-D/ocrd\\_tesseractocr](https://github.com/OCR-D/ocrd_tesseractocr) ), Eynollah 0.2.2, und Kraken 4.3.13 ( <https://github.com/mittagessen/kraken> und [https://github.com/OCR-D/ocrd\\_kraken](https://github.com/OCR-D/ocrd_kraken) ). Trotz höherer Werte schneidet Kraken am schlechtesten bei der Spaltenentrennung ab; das eher rudimentäre Standardmodell lässt sich allerdings am userfreundlichsten nachtrainieren, was derzeit auch erprobt wird. Die weitaus besten Ergebnisse erzielte Eynollah (Rezanezhad et al. 2023). Durchgeführt wurden die Tests mit OCR-D ( <https://ocr-d.de/de> ), aus Gründen der Flexibilität bei der Parametersetzung und Massenverarbeitung per Kommandozeile, einer Vielzahl an integrierten OCR-Engines und Tools sowie gewährleisteter Interoperabilität auf allen Workflow-Ebenen.

Threshold 0.5–0.95	Tesseract	Eynollah	Kraken
Precision	0.016	0.130	0.042
Recall	0.073	0.320	0.137

Obwohl nicht für historische Dokumente vorgesehen, beabsichtigen wir zusätzlich Probeläufe mit sog. Instance-Based-Engines wie YOLO und Detectron2 durchzuführen, die für grobe Zoneneinteilungen, ergänzende Bilderkennung (z.B. Vignetten), oder synergisch mit den o.g. Pixel-Classifiers einsetzbar sind (vgl. Najem-Mayer/Matteo 2022).

Sich auf Computer Vision allein zu stützen resp. vom ‚flachen‘ Text ausgehend nur auf NLP-Methoden zu setzen – dies gilt bei perfekter Transkription (vgl. Pagel et al. 2021), geschweige denn qualitativ stark schwankender OCR –,

scheint angesichts der skizzierten Herausforderungen wenig erfolgversprechend. Daher empfehlen sich hybride bzw. holistische Ansätze, die entweder 1) multimodal vorgehen, also zeitgleich Positionierung und Inhalt von Textteilen auf der Seite berücksichtigen, oder 2) eine Nachbearbeitung von anderweitigen DLA-Outputs durchführen. In die erste Kategorie fällt die LayoutLM-Modellfamilie ( <https://github.com/microsoft/unilm/blob/master/layoutlm/>), für die u.a. adäquate deutschsprachige Modelle fehlen. Andernfalls wäre ein Klassifikator zu trainieren, der bereits vorliegende geometrische Daten auf Zeilenebene (Höhe, Breite, Stellung etc.) auswertet und gleichzeitig das (Nicht-)Vorhandensein vordefinierter Keywords beachtet und entsprechend gewichtet (vgl. Gutehrle/Atanassova 2022; Fischer et al. 2023).

## Annotationsfortschritt

Da sich das Projekt aktuell in der Annotationsphase befindet, wird eine Ground Truth iterativ in PAGE XML hergestellt (Pletschacher/Antonacopoulos 2010), einem Format, das aufgrund hoher Flexibilität und Interoperabilität alle angerissenen Lösungsszenarien bedienen kann. Je nach gewähltem Ansatz muss allerdings abgewogen werden, ob und wie man mit fehlerhaften Vorergebnissen umgeht, bzw. ob man diese unkorrigiert belässt; gleiches gilt für die nachgeschaltete OCR-Erkennung, wobei hier eine ‚schmutzige‘ OCR bewusst in Kauf genommen wird. Das Annotationskorpus zählt aktuell ca. 200 S. und wurde in Anlehnung an die OCR-D-Richtlinien ( <https://ocr-d.de/de/gt-guidelines/trans/>) erstellt. Das Regionen-Repertoire wurde mit Blick auf den hagedornschen Anmerkungsapparat etwas erweitert und beinhaltet teilweise Auszeichnungen für strukturell relevante Zeilentypen (etwa Gedichtanfang). Sind genügend repräsentative Seitenexemplare nach o.g. Typologie ausgezeichnet, kann sodann die nächste Phase der Feature-Extraktion anfangen.

## Ausblick

Obwohl die Layouterkennung für historische Drucke weiterhin vor vielen Hürden steht und v.a. jene Forschende, die aus materialitätsbezogenen Fragestellungen neue Erkenntnisse schöpfen wollen, oft enttäuscht zurücklässt, gilt es die angerissenen und sich stets weiterentwickelnden Ansätze zu erproben und auszubauen, um eine möglichst genaue Isolierung fokussierter Buchteile für nachgeschaltete Analyseschritte zu ermöglichen. Bereits eine Übersicht über bevorzugte Layouts gibt Anlass genug, um solche vermeintlich unscheinbaren Konstellationen weiter qualitativ wie quantitativ – also skalierbar – zu befragen.

Das Wissen um Verteilungen von Seitentypen und deren Komponenten nach Verlegern u.a. Metadaten kann weiter praktisch bei der Zusammenstellung des Annotationskorpus für OCR-relevante Tasks oder auch für die automatisierte Erstellung von Trainingsdaten herangezogen werden.

Darüber hinaus ist der Buchtypus ‚Werkausgabe‘ – und insofern die beachtliche Menge an Hagedorn-Exemplaren – ein interessanter Anwendungsfall für die Erprobung von layoutgerechten Methoden: Hier wird ein relativ überschaubarer (Para-)Textkorpus in relativ vielen ‚Fassungen‘ visuell realisiert, was zu einer weitergehenden Ergänzung von Computer-Vision- und textbasierten Ansätzen einlädt. Dies betrifft zuvorderst die zahlreichen Anmerkungen – ein Phänomen, dessen genauere Erfassung bislang eine eher untergeordnete Rolle spielte oder auf weniger variable Erscheinungsformen fokussiert war.

Alle relevanten im Projekt erarbeiteten Forschungsdaten, Tools und Modelle werden abschließend im Einklang mit den FAIR-Prinzipien u.a. auf *GitHub* und *Zenodo* bereitgestellt. Dies umfasst auch die Struktur-Ground-Truth, die je nach Forschungsinteresse anderweitig modifizierbar bleibt.

## Bibliographie

- Ajouri, Philip.** 2017. „Wie erforscht man eine Werkausgabe? Heuristische Skizze mit Beispielen aus der Geschichte der Werkausgaben“. In *Rahmungen: Präsentationsformen und Kanoneffekte. Beihefte zur Zeitschrift für deutsche Philologie 16* hrsg. v. Ajouri, Philip, Ursula Kundert und Carsten Rohde, 201–221. Berlin: Erich Schmidt Verlag.
- Bartsch, Sabine, Evelyn Gius, Marcus Müller, Andrea Rapp und Thomas Weitin.** 2023. „Sinn und Segment. Wie die digitale Analysepraxis unsere Begriffe schärft“. *Zeitschrift für digitale Geisteswissenschaften 8*. [https://doi.org/10.17175/2023\\_003](https://doi.org/10.17175/2023_003).
- Fischer, Norbert, Alexander Hartelt und Frank Puppe.** 2023. „Line-Level Layout Recognition of Historical Documents with Background Knowledge“. *Algorithms 16* (3). <https://doi.org/10.3390/a16030136>.
- Girdhar, Nancy, Mickaël Coustaty, und Antoine Doucet.** 2024. „Digitizing History: Transitioning Historical Paper Documents to Digital Content for Information Retrieval and Mining—A Comprehensive Survey“. *IEEE Transactions on Computational Social Systems*. <https://doi.org/10.1109/TCSS.2024.3378419>.
- Gutehrle, Nicolas, und Iana Atanassova.** 2022. „Processing the structure of documents: Logical Layout Analysis of historical newspapers in French“. *Journal of Data Mining & Digital Humanities NLP4DH*. <https://doi.org/10.46298/jdmdh.9093>.
- Laramée, François Dominic.** 2019. „How to Extract Good Knowledge from Bad Data: An Experiment with Eighteenth Century French Texts“. *Digital Studies/Le Champ Numérique 9* (1) 2. <https://doi.org/10.16995/dscn.299>.
- McConaughy, Lara, Jennifer Dai, und David Bamman.** 2017. „The Labeled Segmentation of Printed Books“. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 737–747. <https://doi.org/10.18653/v1/d17-1077>.
- Martin Mueller.** 2014. „Shakespeare His Contemporaries: collaborative curation and exploration of Early Modern drama in a digital environment“. *Digital Humanities Quarterly 8* (2014), H. 3.
- Münster, Reinhold.** 1999. *Friedrich von Hagedorn, Dichter und Philosoph der fröhlichen Aufklärung*. München: Iudicium.
- Pagel, Janis, Nidhi Sihag, und Nils Reiter.** 2021. „Predicting Structural Elements in German Drama“. *Proceedings of the Second Conference on Computational Humanities Research*.
- Pletschacher, Stefan, und Apostolos Antonacopoulos.** 2010. „The PAGE (Page Analysis and Ground-Truth Elements) Format Framework“. *2010 20th International Conference on Pattern Recognition*, 257–60. Istanbul, Turkey: IEEE. <https://doi.org/10.1109/ICPR.2010.72>.
- Rezanezhad, Vahid, Konstantin Baierer, Mike Gerber, Kai Labusch, und Clemens Neudecker.** 2023. „Document Layout Analysis with Deep Learning and Heuristics“. *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, 73–78. <https://doi.org/10.1145/3604951.3605513>.
- Najem-Meyer, Sven und Romanello Matteo.** 2022. „Page Layout Analysis of Text-heavy Historical Documents: a Comparison of Textual and Visual Approaches“. <https://doi.org/10.48550/ARXIV.2212.13924>.
- Underwood, Ted.** 2014. „Understanding Genre in a Collection of a Million Volumes, Interim Report“. <https://doi.org/10.6084/M9.FIGSHARE.1281251>.
- Subramanyam, Vineeth.** 2021. IOU (Intersection over Union). *Medium*. <https://medium.com/analytics-vidhya/iou-intersection-over-union-705a39e7acef> [letzter Zugriff: 20.07.2024]
- Fleischhacker, David, Wolfgang Goederle und Roman Kern.** 2024. „Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach“. <https://arxiv.org/abs/2401.07787> [letzter Zugriff: 20.07.2024]