Universität Stuttgart
Institut für Literaturwissenschaft
Abteilung *Digital Humanities*

Segmentierung & Semantic Labeling von Paratexten für (Hagedorn-)Werkausgaben des 18. Jh.s

Arsenije Bogdanović

DFG-Projekt: „Scalable Reading von ‚Gesammelten Werken' des 18. Jahrhunderts, exemplarisch durchgeführt an Friedrich-von-Hagedorn-Werkausgaben"

- **Kooperation:** Uni Stuttgart (Digital Humanities) und Uni Mainz (Buchwissenschaft)
- **Projektteil 1:** Ausgabenzusammenstellung der Werkausgaben Friedrich von Hagedorns aus dem 18. Jh., erforscht mit den Mitteln der Document (Layout) Analysis
- **Projektteil 2:** Textänderungen in Werkausgaben Friedrich von Hagedorns aus dem 18. Jh., erforscht mit den Mitteln von Text-Reuse und Sequence Alignment

Friedrich von Hagedorn
(von Dominicus van der Smissen (1704–1760), Hamburger Kunsthalle)
https://de.wikipedia.org/wiki/Friedrich_von_Hagedorn#/media/Datei:Hagedorn.jpg

Eckdaten Korpus

Bestand: ca. 90 Bde. von Hagedorns WA (davon werden nur die posthumen aktuell bearbeitet, ca. **70**); erschienen bei insg. **15** Verlegern (exkl. Koverleger-Paare);

Publikationsmodus: Meistens in **drei**, seltener in **fünf** Bde. aufgeteilt; ca. 150-200 S. pro Bd.

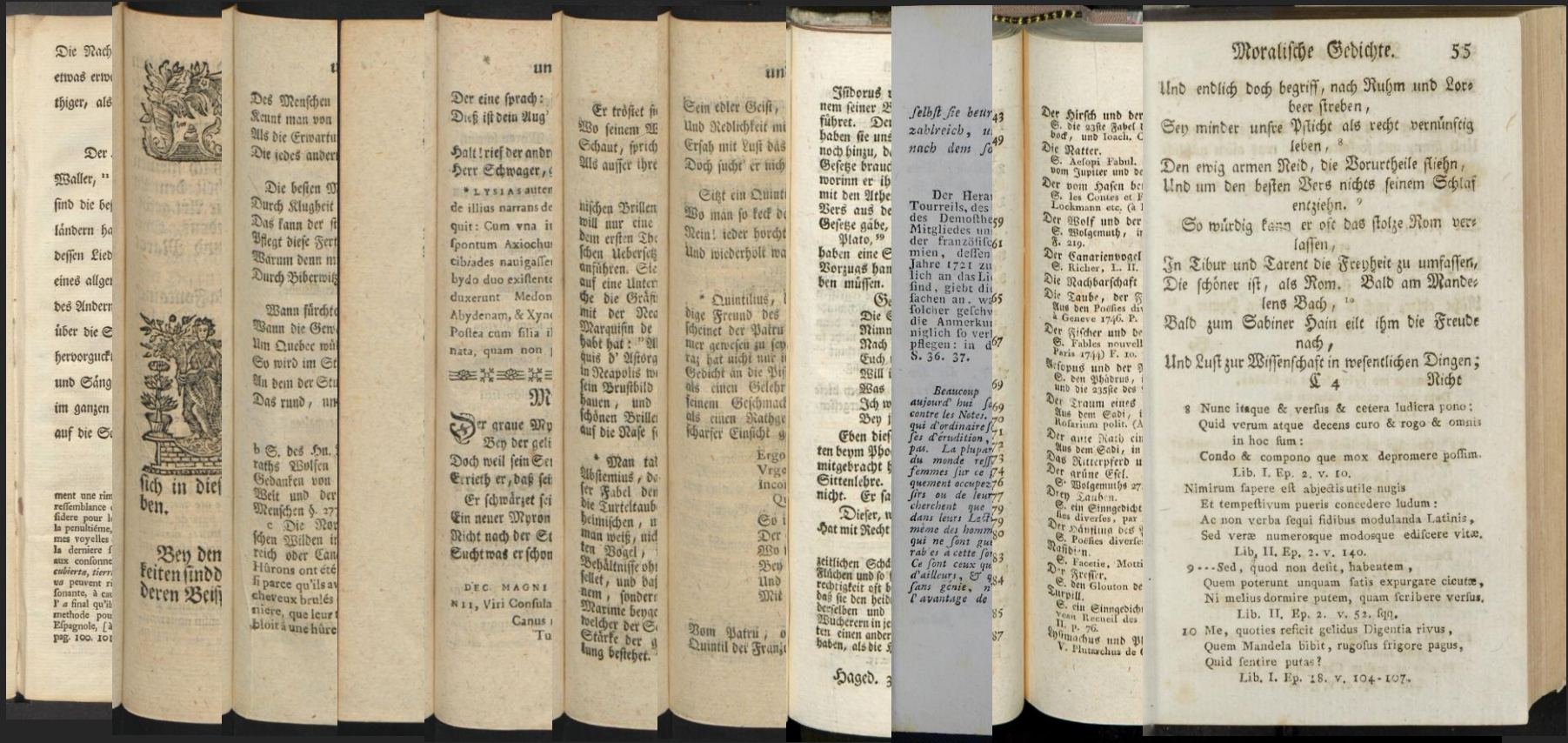
Layout allgemein: Gedichtsatz dominant, vereinzelt auch Briefsatz (etwa bei Widmungen; manche Ausgaben enthalten zusätzliche Zeugnisse und „Beigaben“…); Registersatz nur bei Inhaltsverzeichnissen; keine Tabellen und keine Marginalien.

Besonderheit: Sehr grob betrachtet, unterscheiden sich WA v.a. in der Auswahl, Abfolge und im Layout der dargebotenen Texte. Ein und dasselbe Gedicht kann durch 20 Ausgaben hindurch unterschiedlich ‚iteriert‘ werden, bleibt im Kern/Inhalt größtenteils identisch.

1)

Problem- und Zielstellung

Vielfalt und Komplexität von Werkausgaben



Moralische Gedichte. 55

Und endlich doch begriff, nach Ruhm und Lorbeer streben,
Sey minder unsre Pflicht als recht vernünftig leben, [8]
Den ewig armen Neid, die Vorurtheile fliehn,
Und um den besten Vers nichts seinem Schlaf entziehn. [9]
So würdig kann er oft das stolze Rom verlassen,
In Tibur und Tarent die Freyheit zu umfassen,
Die schöner ist, als Rom. Bald am Mandelens Bach, [10]
Bald zum Sabiner Hain eilt ihm die Freude nach,
Und Lust zur Wissenschaft in wesentlichen Dingen;

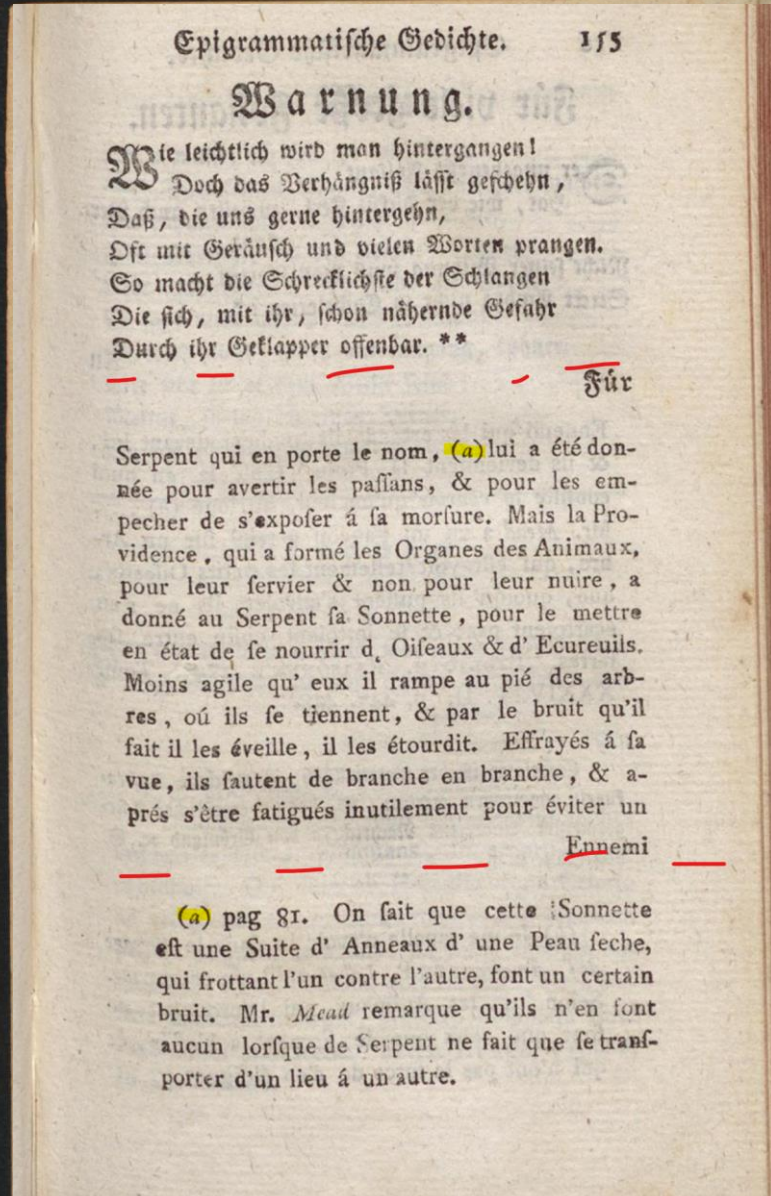
L 4 Nicht

8 Nunc itaque & versus & cetera ludiera pono:
Quid verum atque decens curo & rogo & omnis in hoc sum:
Condo & compono que mox depromere possim.
Lib. I. Ep. 2. v. 10.
Nimirum sapere est abjectis utile nugis
Et tempestivum pueris concedere ludum:
Ac non verba sequi fidibus modulanda Latinis,
Sed veræ numerosque modosque ediscere vitæ.
Lib. II. Ep. 2. v. 140.
9 --- Sed, quod non desit, habeatem,
Quem poterunt unquam satis expurgare cicutæ,
Ni melius dormire putem, quam scribere versus.
Lib. II. Ep. 2. v. 52. sqq.
10 Me, quoties reficit gelidus Digentia rivus,
Quem Mandela bibit, rugosus frigore pagus,
Quid sentire putas?
Lib. I. Ep. 18. v. 104-107.

Vielfalt und Komplexität von Werkausgaben

Hohe Variabilität im Erscheinungsbild von Ausgabe zu Ausgabe:

- **Format und Layout**:
  - ein- und zweispaltig inkl. Kombination;
  - uneinheitliche Konventionen auch innerhalb derselben Ausgabe;
  - (para-)textlastig inkl. Ambiguitäten (Zitat vs. Fußnote);
  - uneinheitlicher Durchschuss
- **Typographie**: mehrere Schriftarten (inkl. Auszeichnungsschriften) – Fraktur vs. Antiqua; auch multilingual
- **Materialität/Bildqualität**: Bleedthrough, Schmutz/Verfärbungen, diverse Schäden, handschriftliche Notizen; misslungene/doppelte Aufnahmen, Artefakte, uneinheitliche Nachbearbeitungen



Epigrammatiſche Gedichte. 155

Warnung.

Wie leichtlich wird man hintergangen!
Doch das Verhängniß läſſt geſchehn,
Daß, die uns gerne hintergehn,
Oft mit Geräuſch und vielen Worten prangen.
So macht die Schrecklichſte der Schlangen
Die ſich, mit ihr, ſchon nähernde Gefahr
Durch ihr Geklapper offenbar. **

Für

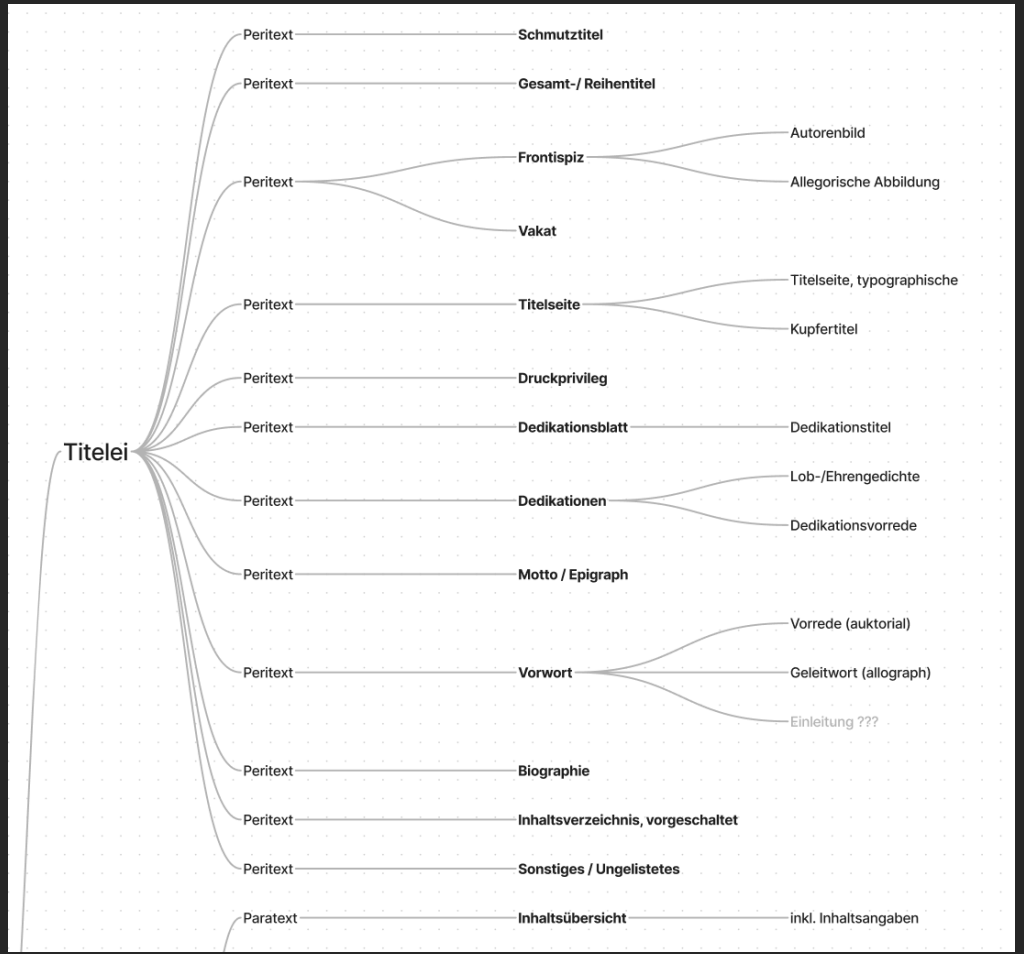
Serpent qui en porte le nom, (a) lui a été donnée pour avertir les paſſans, & pour les empecher de s'expoſer á ſa morſure. Mais la Providence, qui a formé les Organes des Animaux, pour leur ſervier & non pour leur nuire, a donné au Serpent ſa Sonnette, pour le mettre en état de ſe nourrir d. Oiſeaux & d' Ecureuils. Moins agile qu' eux il rampe au pié des arbres, oú ils ſe tiennent, & par le bruit qu'il fait il les éveille, il les étourdit. Effrayés á ſa vue, ils ſautent de branche en branche, & aprés s'être fatigués inutilement pour éviter un

Ennemi

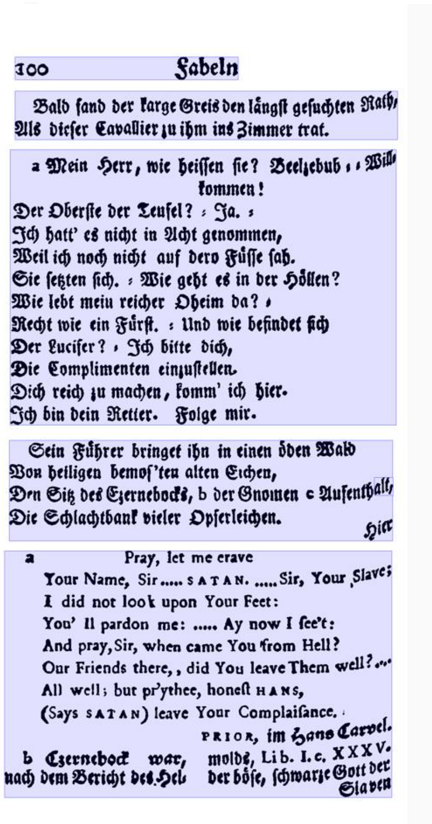
(a) pag 81. On ſait que cette Sonnette eſt une Suite d' Anneaux d' une Peau ſeche, qui frottant l'un contre l'autre, font un certain bruit. Mr. *Mead* remarque qu'ils n'en font aucun lorſque de Serpent ne fait que ſe tranſporter d'un lieu á un autre.

Forschungsinteresse

- **Ausgangspunkt**: Taxonomie von Textsorten und Buchteilen, die für WA relevant sind (vgl. McConnaughey et al 2017; Underwood 2013);
- **Ziel**: Rekonstruktion zusammenhängender Textsequenzen (Textsorten) anhand von Layoutinformationen auf Seitebene (semantisch ausgezeichnete Regionen);
- **Haupt-Task**: Trennung von Kern- und Paratext-Regionen;
- **Nice-to-have**: Unterscheidung zwischen Gedicht- und Prosatexten, sowie Erkennung von Grafiken/Schmuck…
- **Ideal-Output**: TEI-Datei
- **Größtes Hindernis**: Fehlerhafte Segmentierung und folglich falsche Lesereihenfolge; fehlendes/ungenaues semantisches Labeling;



Testlauf und Vergleich



100 Fabeln

Bald fand der karge Greis den längst gesuchten Rath,
Als dieser Cavallier zu ihm ins Zimmer trat.

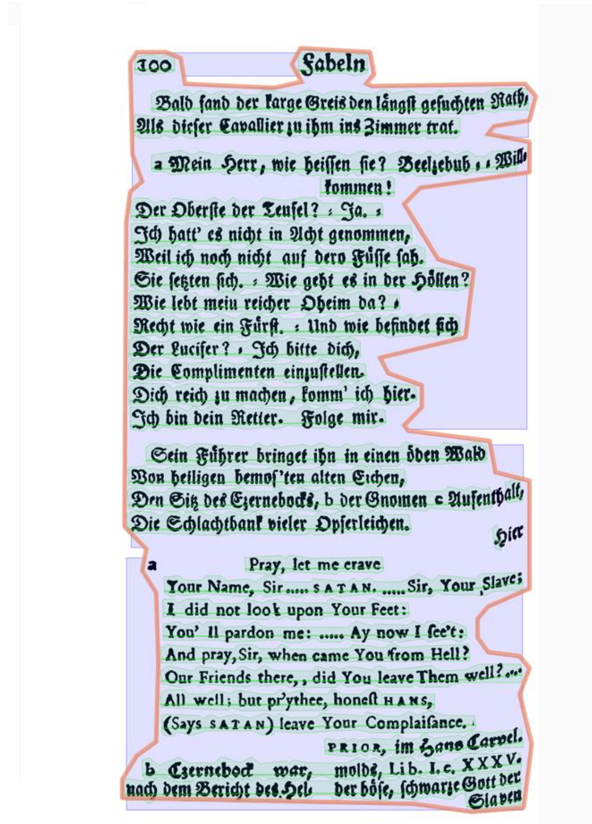
a Mein Herr, wie heissen sie? Beelzebub, , Will-
kommen!
Der Oberste der Teufel? , Ja. ,
Ich hatt' es nicht in Acht genommen,
Weil ich noch nicht auf dero Füsse sah.
Sie setzten sich. , Wie geht es in der Höllen?
Wie lebt mein reicher Oheim da? ,
Recht wie ein Fürst. , Und wie befindet sich
Der Lucifer? , Ich bitte dich,
Die Complimenten einzustellen.
Dich reich zu machen, komm' ich hier.
Ich bin dein Retter. Folge mir.

Sein Führer bringet ihn in einen öden Wald
Von heiligen bemos'ten alten Eichen,
Den Sitz des Czernebocks, b der Gnomen c Aufenthalt,
Die Schlachtbank vieler Opferleichen.
Hier

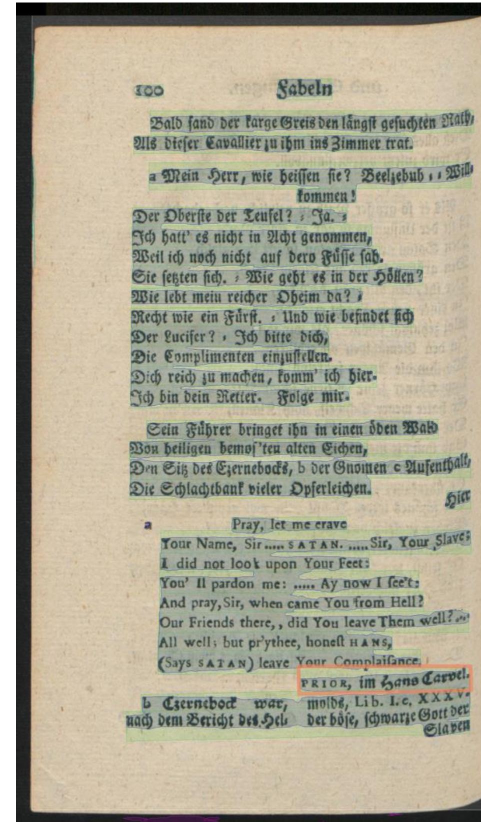
a Pray, let me crave
Your Name, Sir..... SATAN. .....Sir, Your Slave;
I did not look upon Your Feet:
You' ll pardon me: ..... Ay now I see't:
And pray, Sir, when came You from Hell?
Our Friends there,, did You leave Them well?...
All well; but pr'ythee, honest HANS,
(Says SATAN) leave Your Complaisance.
PRIOR, im Hans Carvel.

b Czernebock war, nach dem Bericht des Hel- molds, Lib. I. c. XXXV. der böse, schwarze Gott der Slaven

Tesseract



Kraken



eynollah

2)

Korpuserstellung & Auswahlkriterien

Arbeitshypothesen und Clustering

Verleger:in = Layout: Ausgaben und folglich Layouts lassen sich am einfachsten nach Verleger:innen gruppieren

Type-Token-Verhältnis: Ausgaben mit (fast) identischem oder sehr ähnlichem Satz/Typographie wurden zusammengeführt und auf 11 repräsentative Ausgaben reduziert

Seitentypologie: Diese werden weiter geclustert nach vorkommenden Layouteigenschaften

Enthält ein Bd. …

- Zweispaltige Fußnoten, dann „b2“
- Ein- sowie zweispaltige Fußnoten auf separaten Seiten, dann „b1-b2“
- Ein- sowie zweispaltige Fußnoten gleichzeitig (im „Apparat“ auf ders. S.), dann „bb“
- Zweispaltige „hängende“ Fußnoten/Endnoten – „c2“
- …

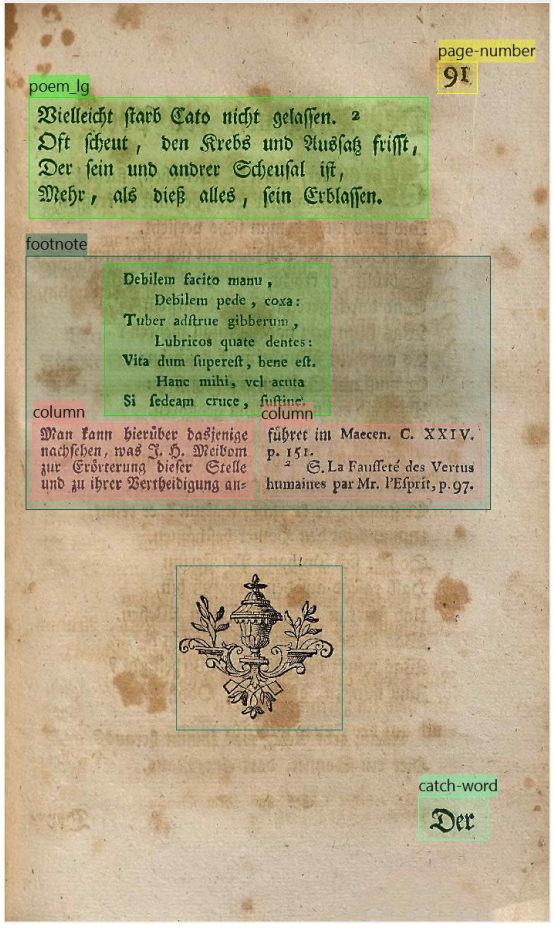
Einfacher Satz: 8 Ausg, davon 4 mit „Quasi“-Fußnoten;

Zweispaltiger Satz: 3 Ausg.; davon 1 sehr schwierig.

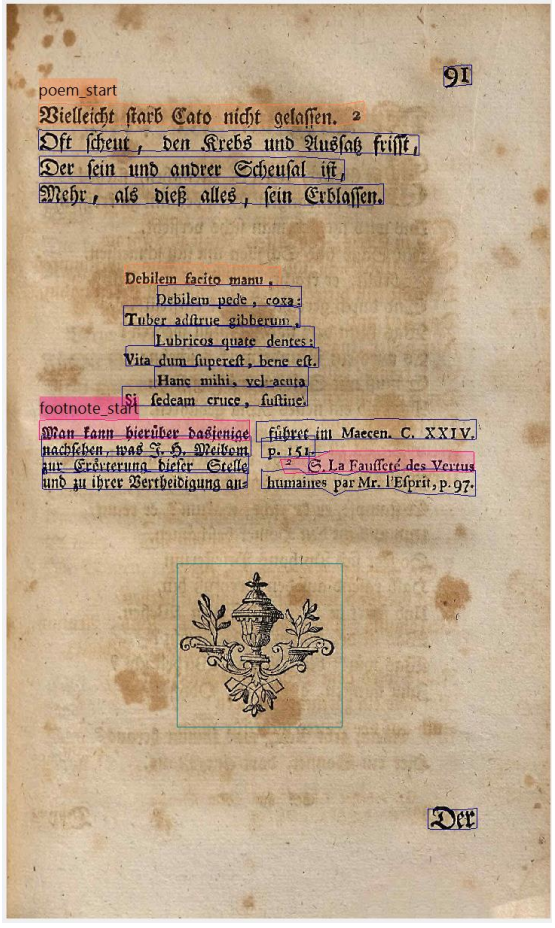
3)

GT-Erstellung

Vorgehensweise Annotation



Regionenauszeichnung



Zeilenauszeichnung

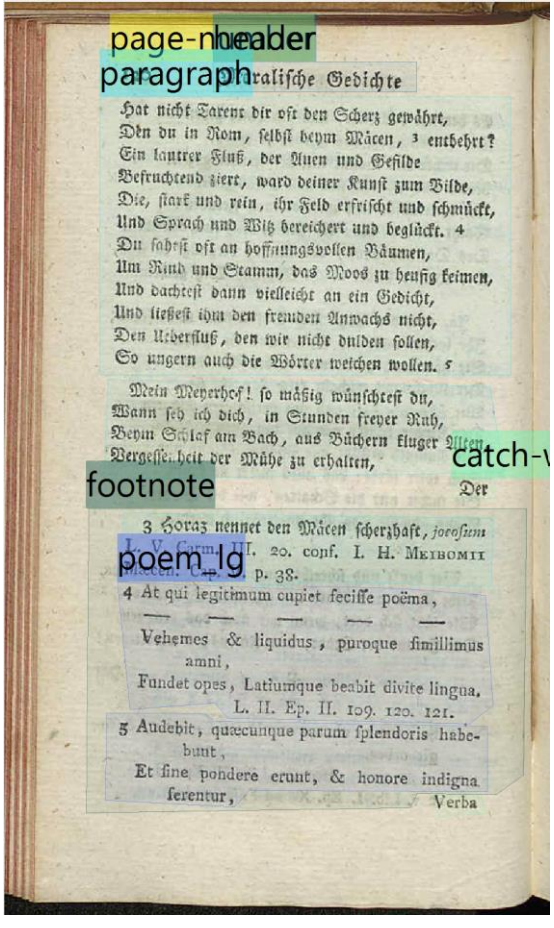
In Anlehnung an Gutehrlé/Atanassova (2023) und Baránek (2024) „zweigleisige“ Annotation auf:

- Regionenebene (Pase 1): manuelle Seg. + Ausz.
- Zeilenebene (Phase 2): auto. Seg. (Transkribus) + manuell Ausz.

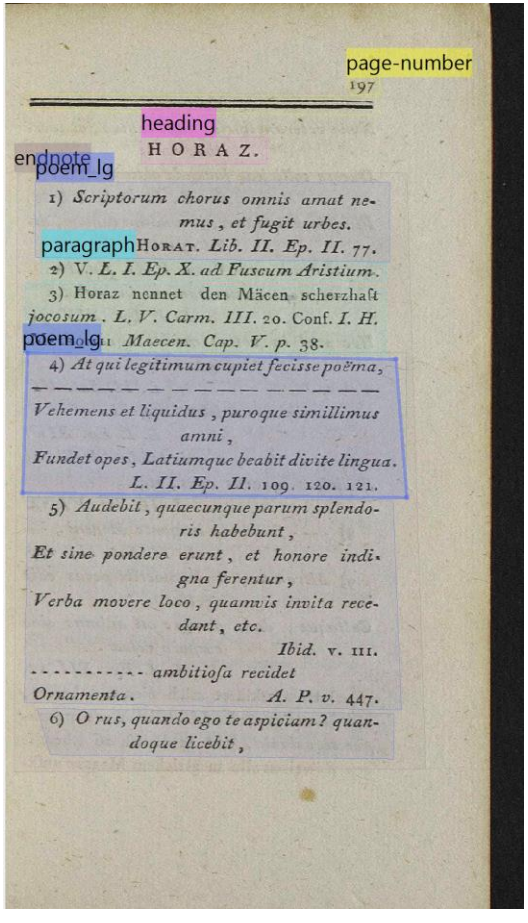
Custom Tags:

- Modifizierte OCR-D-Richtlinien (Level 2 mit gewissen Abstrichen)
- Regionen
  - Spalte = column
  - Strophe = poem_lg
- Zeilen
  - Erste Zeile Strophe = strophe_start
  - Erste Zeile Absatz = paragraph_start
  - Erste Zeile Fußnote = footnote_start

Kollidierende Zuordnungen



Regionenauszeichnung



Zeilenauszeichnung

Semantische Ambiguität:

- Fuß- vs. Endnote, oder einfach ‚Apparat‘?
- Regionen- und Zeilen-Zuordnung deckt sich nicht immer (Gedichtzeile = Fußnotenzeile)
- Weitere Zeilen-Tags: allgemein für Fußnoten, Verse, Autorangaben???

Weitere Fragen…

Allgemein

- Werk- vs. layoutspezifisches Training?
- Welches Splitting?
- Wie weit OCR-D-konform bleiben?
- Interoperabilität: Transkribus mit Kraken/eScriptorium
- Vorverarbeitung (Binarisierung etc.?), Wasserzeichen?

Tagging

- Ist es sinnvoller Bildregionen zusammenzuführen (Separatoren, Graphiken, Illustrationen…)?
- Wie viele @custom bzw. eigene Strukturtags sind sinnvoll?
- Sind Überschneidungen bei automatisch generierten TextLines erlaubt? Was ist mit Baslines?
- Sind eingebettete Regionen ratsam?
- Verlinkungen bei Initialen…

Literatur

Baránek, Daniel. „Kraken segmentation model for two-column prints“. Zenodo, 2024. https://zenodo.org/records/10783346

Dengel, Andreas, und Faisal Shafait. „Analysis of the Logical Layout of Documents“. In *Handbook of Document Image Processing and Recognition*, 177–222. Springer London, 2014. https://doi.org/10.1007/978-0-85729-859-1_6.

Engl, Elisabeth. „OCR-D Kompakt: Ergebnisse Und Stand Der Forschung in Der Förderinitiative“. *Bibliothek Und Praxis* 44, Nr. 2 (29. Juli 2020): 218–30. https://doi.org/10.15*Forschung*15/bfp-2020-0024.

Girdhar, Nancy, Mickaël Coustaty, und Antoine Doucet. „Digitizing History: Transitioning Historical Paper Documents to Digital Content for Information Retrieval and Mining—A Comprehensive Survey“. IEEE Transactions on Computational Social Systems, 2024, 1–30. https://doi.org/10.1109/TCSS.2024.3378419.

Gutehrlé, Nicolas, und Iana Atanassova. „Processing the structure of documents: Logical Layout Analysis of historical newspapers in French“. *Journal of Data Mining & Digital Humanities* NLP4DH, Nr. Digital humanities in... (30. Mai 2022): 9093. https://doi.org/10.46298/jdmdh.9093.

McConnaughey, Lara, Jennifer Dai, und David Bamman. „The Labeled Segmentation of Printed Books“. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/d17-1077.

Pletschacher, Stefan, und Apostolos Antonacopoulos. „The PAGE (Page Analysis and Ground-Truth Elements) Format Framework“. In *2010 20th International Conference on Pattern Recognition*, 257–60. Istanbul, Turkey: IEEE, 2010. https://doi.org/10.1109/ICPR.2010.72.

Rezanezhad, Vahid, Konstantin Baierer, Mike Gerber, Kai Labusch, und Clemens Neudecker. „Document Layout Analysis with Deep Learning and Heuristics“. In Proceedings of the 7th International Workshop on Historical Document Imaging and Processing, 73–78. San Jose CA USA: ACM, 2023. https://doi.org/10.1145/3604951.3605513.

Reul, Christian, Uwe Springmann, und Frank Puppe. „LAREX: A semi-automatic open-source Tool for Layout Analysis and Region Extraction on Early Printed Books“. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*. ACM, 2017. https://doi.org/10.1145/3078081.3078097.

Riedl, Martin, Daniela Betz, und Sebastian Padó. „Clustering-Based Article Identification in Historical Newspapers“. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 12–17. Minneapolis, USA: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/W19-2502.

Seuret, Mathias, Janne van der Loop, Nikolaus Weichselbaumer, Martin Mayr, Janina Molnar, Tatjana Hass, Florian Kordon, Anguelos Nicolau, und Vincent Christlein. „Combining OCR Models for Reading Early Modern Printed Books“, 2023. https://doi.org/10.48550/ARXIV.2305.07131.

Sven, Najem-Meyer, und Romanello Matteo. „Page Layout Analysis of Text-heavy Historical Documents: a Comparison of Textual and Visual Approaches“, 2022. https://doi.org/10.48550/ARXIV.2212.13924.

Underwood, Ted, Michael L. Black, Loretta Auvil, und Boris Capitanu. „Mapping mutable genres in structurally complex volumes“. In *2013 IEEE International Conference on Big Data*. IEEE, 2013. https://doi.org/10.1109/bigdata.2013.6691676.

Universität Stuttgart
Institut für Literaturwissenschaft
Abteilung *Digital Humanities*

Vielen Dank!

Arsenije Bogdanović

E-Mail arsenije.bogdanovic@ilw.uni-stuttgart.de
Telefon +49 (0) 711 685-81285
https://www.ilw.uni-stuttgart.de/abteilungen/digital-humanities

Universität Stuttgart
Institut für Literaturwissenschaft, Digital Humanities
Herdweg 51, 70174 Stuttgart