# Makroanalytische Untersuchung von Heftromanen

**Jannidis, Fotis**
fotis.jannidis@uni-wuerzburg.de
Universität Würzburg, Deutschland

**Konle, Leonard**
leonardkonle@gmail.com
Universität Würzburg, Deutschland

**Leinen, Peter**
P.Leinen@dnb.de
Deutsche Nationalbibliothek, Frankfurt a.M.

## Einleitung

Heftromane, früher als ‘Romane der Unterschicht’ (Nusser 1981) abgewertet, sind aufgrund eines weniger wertungsfreudigen Umgangs mit Populärliteratur (Hügel 2007, Kelleter 2012) in den letzten 10-15 Jahren wieder Gegenstand der Literaturforschung geworden (z.B. Nast 2017, Stockinger 2018). ‘Heftromane’ wurden immer definiert durch das eigene Publikationsformat (zumeist rd. 64 Seiten), eigene Formen der Distribution über den Zeitschriftenmarkt und nicht über den Buchhandel, und auch die Soziographie der Heftromanleser weicht deutlich von der der sonstigen Literatur ab. Im Folgenden berichten wir über erste Ergebnisse einer Auswertung von 9.000 deutschsprachigen Heftromanen aus den Jahren 2009-2017. Möglich wurde die Forschung durch eine Kooperation zwischen der Würzburger Arbeitsgruppe zur literarischen Textanalyse und der Deutschen Nationalbibliothek (DNB), die die Daten vorhält. Ziel dieser ersten, noch weitgehenden explorativen Studie, sind die Antworten auf zwei Fragen: Wie unterscheiden sich die Gattungen der Heftromane untereinander und wie unterscheiden sich die Heftromane von Hochliteratur? Im ersten Schritt gehen wir der Frage nach, wie gut sich die Gattungen klassifizieren lassen und welche Texteigenschaften dabei eine Rolle spielen. Im zweiten Schritt werden die Gattungen inhaltlich erfasst. Zuletzt geht es um die angeblich einfachere Sprache der Heftromane.

Die Analyse stützt sich auf die digitalen Texte, die an die Deutsche Nationalbibliothek abgeliefert wurden. Die Deutsche Nationalbibliothek (DNB) sammelt, archiviert, verzeichnet im gesetzlichen Auftrag die ab 1913 in Deutschland veröffentlichten Medienwerke sowie die im Ausland veröffentlichten deutschsprachigen Medienwerke, Übersetzungen deutschsprachiger Medienwerke in andere Sprachen und fremdsprachige Medienwerke über Deutschland und stellt diese der Öffentlichkeit zur Verfügung. Seit der Gesetzesnovelle von 2006 gehört auch das Sammeln von Medienwerken, die online publiziert werden, ausdrücklich zu den Aufgaben der DNB.

Der Bestand der DNB umfasst derzeit etwa 5 Millionen digitale Objekte, darunter ca. 900.000 E-Books, ca. 1,5 Millionen E-Journal Ausgaben und ca. 2 Millionen E-Paper Ausgaben. Neben dem umfangreichen physischen Bestand steht den Nutzerinnen und Nutzern der DNB damit ein wachsender Fundus von “born digital” Objekten zur Verfügung.

Die Anforderungen an die Informationsversorgung haben sich durch den digitalen Wandel insgesamt stark verändert. Die Einführung neuer Forschungsmethoden wie z.B. automatisierte Daten- und Textanalysen großer digitaler Bestände gehen mit der Notwendigkeit veränderter Formen der Bereitstellung von Beständen einher.

Die DNB sieht in der Kooperation mit den DH-Communities eine strategisch wichtige Fortsetzung[1] ihrer Dienstleistungs- und Benutzungs-Angebote. Ein Aspekt dabei ist die Unterstützung ausgewählter Kooperationspartner durch die Bereitstellung auch umfangreicher Korpora vorrangig aus born digital Objekten wie E-Books sowie einer leistungsfähigen Infrastruktur für die Durchführung automatischer Analysen. Die letzte Urheberrechtsreform eröffnet den registrierten Nutzerinnen und Nutzern der DNB diese Möglichkeit in den Räumen der DNB und auf deren Infrastruktur.

In der Vorverarbeitung wurden, soweit das aufgrund der selbst innerhalb eines Verlags sehr heterogenen Ausgangslage automatisch möglich war, Werbung, Leseproben usw. entfernt. Problematische Texte wurden nicht für die Analyse verwendet. Die unausgewogene Verteilung in Abbildung 1 kommt durch die Neupublikation älterer Hefte zustande. Die Verteilung über die Gattungen (Abbildung 2) ist sehr unausgewogen. Abb. 3 zeigt, dass manche der Gattungen von einzelnen Serien dominiert werden. Außerdem haben wir ein Vergleichskorpus ‘Hochliteratur’ mit 500 Romanen von Autoren erstellt, die einen literarischen Preis gewonnen haben oder dafür vorgeschlagen wurden.

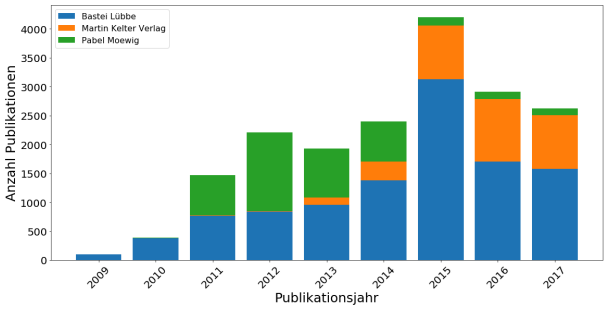


*Abbildung 1. Anzahl der Publikationen nach Verlag und Erscheinungsjahr*

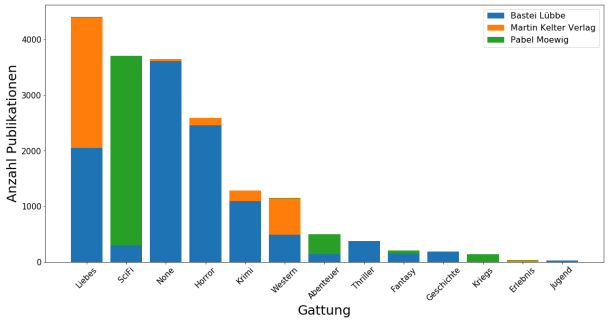


*Abbildung 2. Anzahl der Publikationen nach Gattung und Verlag*

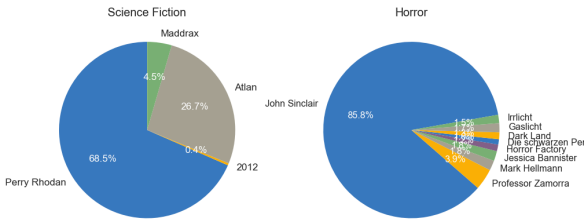


*Abbildung 3. Dominanz großer Serien*

# Methoden

Gattungen erkennen. Um einen Eindruck der Kohärenz der Gattungen aus inhaltlicher Perspektive zu erhalten, wird eine Document-Term-Matrix mit den 8000 häufigsten Substantiven verwendet und eine Dimensionsreduktion mit umap (McInnes 2018) durchgeführt, um eine zweidimensionale Darstellung zu ermöglichen. Um die Leistungsfähigkeit der Substantive für die Klassifikation der Texte nach Gattungen zu prüfen, wurde überwachtes Lernen mit durch Logistic Regression durchgeführt. Außerdem werden die Texte aufgrund eines stilistischen Maßes gruppiert: Kosinus Delta (Evert et. al 2017) der 2000 häufigsten Worte, deren Leistungsfähigkeit wird ebenfalls durch eine Klassifikation getestet.

Themen und Topoi. Wir verwenden zwei Verfahren, um Themen, Settings, Gegenstände, Figuren, rekurrente Formulierungen usw. der Gattungen zu identifizieren: Topic Modeling und Zeta. Ein Model (100 Topics) (Blei, Jordan, Ng 2002) wird über das gesamte Korpus mit Mallet 2.0.8 (McCallum 2002) erstellt. Zeta wird verwendet, um zu ermitteln, welche Worte in einer Gruppe von Texten im Vergleich mit einer zweiten bevorzugt werden (Burrows 2007, Craig, Kinney 2009). Für unsere Untersuchung wurden jeweils für jede Gattung 200 Texte dieser Gattung und 200 Texte aus allen anderen Gattungen zufällig gezogen und Eder's Zeta mit Stylo berechnet (Eder, Kestemont, Rybicki 2016); Parameter nach Empfehlungen in (Schöch et al. 2018).

Gattungskomplexität und Kontrast zu Hochliteratur . Zur Prüfung der Hypothese, dass Schemaliteratur sprachlich weniger komplex sei als Hochliteratur, wurde Vokabular und Syntax untersucht: Die lexikalische Komplexität wurde durch ein standardisiertes type-token-ratio (sttr, Fenstergröße 10.000) ( Kubát, Mili#ka 2013) sowie die Wortlänge ermittelt . Die syntaktische Komplexität wird zum einen durch die durchschnittliche Satzlänge und zum anderen durch die Variabilität von POS-Tags untersucht.

# Ergebnisse

Gattungen erkennen. In früheren Arbeiten mit Texten des 19. Jahrhunderts konnten wir stilometrisch (most frequent words) einige Gattungen sehr gut (z.B. Abenteuerroman), andere nur schlecht unterscheiden (z.B. Bildungsroman), während die thematischen Unterschiede (topic models) schlechtere Klassifikationsergebnisse erbrachten (Hettinger et al. 2016). Die Heftroman-Gattungen bilden sowohl inhaltlich (Abbildung 4) als auch stilistisch (Abbildung 5) recht deutlich abgegrenzte Einheiten. Diesen visuellen Eindruck bestätigen die Ergebnisse der Klassifikation mit den 8000 häufigsten Substantiven mit einem 0.90 F1 (macro) und 0.94 F1 (micro). Eine Klassifikation auf Basis der Distanzmatrix erreicht leicht schlechtere Werte von 0.78 F1 (macro) und 0.91 F1 (micro).

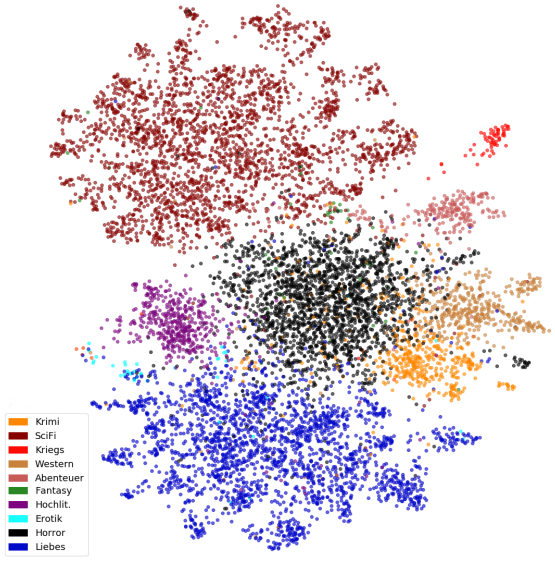


*Abbildung 4. Clustering der Texte auf Basis der 8000 most frequent nouns*

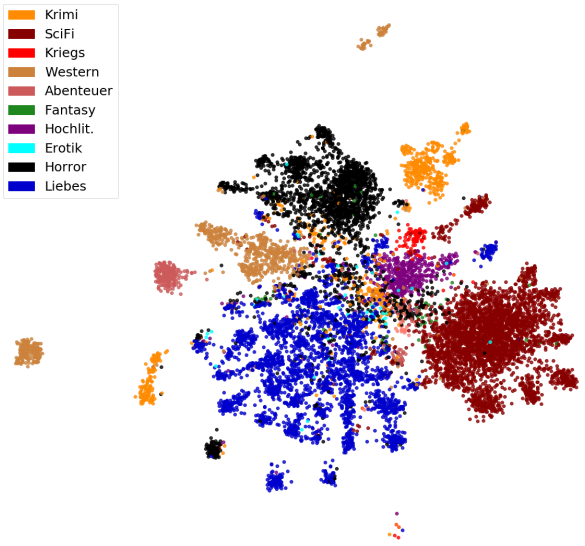


*Abbildung 5. Transformation der Distanzmatrix (cosine delta, 2000 mfw)*

Themen und Topoi. Das Topic Model eignet sich nur begrenzt zur inhaltlichen Erschließung des Korpus, da die ermittelten Topics zum größten Teil nicht interpretierbar sind. Einige wenige funktionieren sehr gut, z.B. in Abbildung 6 die wichtigsten Topics für die Gattung ‘Western’. Sie enthalten zentrale Entitäten (Pferde, Wagen), vor allem aber verdeutlichen sie, dass Kampfhandlungen wichtig für die Gattung sind.


*Abbildung 6. Häufigste Topics der Gattung “Western”*

Allerdings sind zahlreiche Topics zwar statistisch diskriminativ, aber aus literaturwissenschaftlicher Perspektive undankbar. So ist es offensichtlich, dass das Topic in Abb. 7 Kommunikationswörter als typisch für den Liebesroman aufführt. Aber die Worte des Topic in Abb. 8, fast ebenso diskriminativ für Fantasy und SciFi, überschneiden sich damit in vielen Punkten.

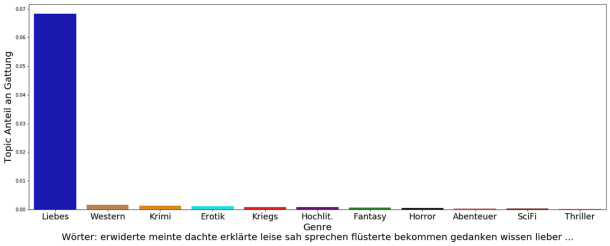


*Abbildung 7. Kommunikationswörter des Gattung Liebesroman*

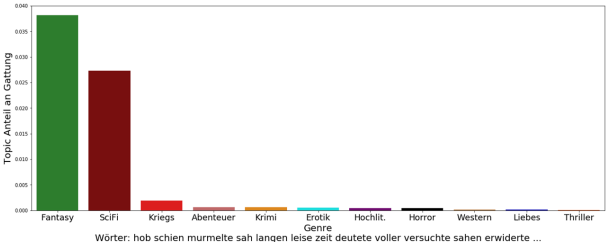


*Abbildung 8. Distinktives Topic für Fantasy und SciFi*

Wie das Clustering der Texte aufgrund der Topics in Abbildung 9 zeigt, können diese als Proxies für die Verteilungen von Worten in Romanen dienen, aber sie erschließen die Texte - anders als das in vielen vergleichbaren Untersuchungen der Fall ist - nur in wenigen Fällen inhaltlich.

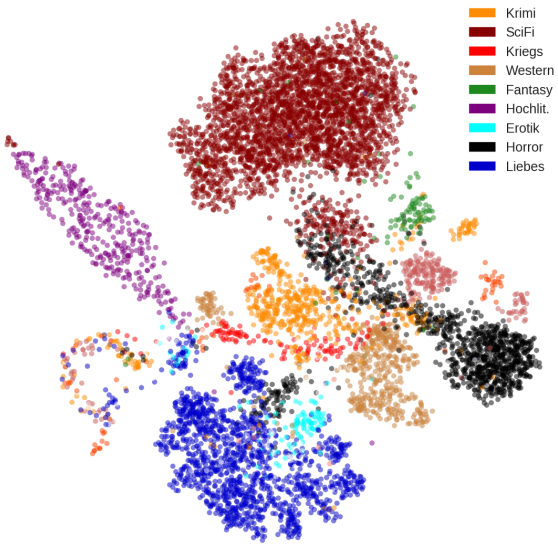


*Abbildung 9. Transformation der Verteilung der Topics pro Dokumenten mit umap*

Die Wörter, die aufgrund von Zeta, von der Gattung bevorzugt werden, leisten dagegen die inhaltliche Erschließung der Gattungen ausgesprochen gut (siehe die Beispiele in Abbildung 10 und 11): Vor allem

Figuren (z.B. Seeleute, Fürsten, Oberarzt, usw.) sowie Objekte und Settings (z.B. Bordwand, BH, Bergwald) entsprechen den Erwartungen, lassen aber zugleich einen Einblick in die Spezifika der Genres zu, z.B. dass es sich im Falle der Science Fiction um eine langandauernde 'space opera' handelt oder dass der Handlungsort der Liebesromane häufig ein Oberschichtmilieu ist. Nur im Fall der Hochliteratur ist die Divergenz der sprachlichen Register / der Begriffe auffällig groß. Dieser Effekt schlägt sich in Abbildung 12 nieder, es ist zu beobachten, dass die Kohärenz, hier interpretiert als Maß für die Geschlossenheit einer Gruppe von Worten, der Zeta Wörter für Hochliteratur in etwa der einer Zufallsstichprobe entspricht. Wie gut die Zeta-Wörter die Gattungen repräsentieren, wird daran deutlich, dass eine Klassifikation (svm, 150 Zetawörter pro Gattung, 600 Texte aus jeder Gattung mit Oversampling, wo notwendig) einen F1 (micro) score von 0.90 erreicht.

| Abenteuer | Adels | Arzt | Erotik | Horror | Hochlit |
|---|---|---|---|---|---|
| aye | schlosspark | infusion | orgasmus | geisterjäger | klo |
| spanier | fürsten | zillertal | t-shirt | rover | it |
| achtern | gräfin | fee | slip | silberkugeln | me |
| bordwand | baron | sprechstunde | nippel | silberkugel | hitler |
| ausguck | fürstin | innsbruck | reißverschluss | beretta | texte |
| außenbords | schlossbewohner | röntgen | po | geweihten | for |
| kahn | durchlaucht | grünwald | brustwarze | dämons | weltkrieg |
| gesegelt | hinzusetzte | spatz | schamlippen | zombie | on |
| degen | fürst | oberarzt | klasse | vampiren | andauernd |
| holzbein | teenagern | facharzt | strähnen | untoten | wischt |
| pullen | privaträume | hausarzt | warmer | blutsauger | juden |
| backbordseite | schlosshof | gebärmutter | penis | friedhofs | russland |
| mast | fürstenpaar | operationssaal | duschen | abbé | cola |
| seemann | boxer | pflegerin | nässe | augenhöhlen | wörter |
| wikinger | cousin | assistenzarzt | angefühlt | dämonischen | präsidenten |
| backbord | standesgemäß | bergdoktor | lecken | luzifer | what |
| großmast | bediensteten | notarztwagen | unterleib | dämonenpeitsche | christus |

*Abbildung 10. Genretypische Wörter (Zeta)*

| Heimat | Kriegs | Krimi | Liebes | SciFi | Western |
|---|---|---|---|---|---|
| madl | leutnant | streifenwagen | dienerschaft | galaxis | saloon |
| bissel | mg | ford | gnädigen | raumschiff | hufschlag |
| brotzeit | munition | field | diwan | planet | texas |
| tonis | russen | officer | gnädiges | universums | winchester |
| bös | russischen | schalldämpfer | anerbieten | schleuse | cowboys |
| obstler | deutscher | handschellen | unbeschreiblichen | weltraum | cowboy |
| förster | einschläge | dienstwaffe | vornehmer | hangar | camp |
| ausschaut | flugzeuge | detective | gottlob | schutzschirm | well |
| gell | oberst | brooklyn | destille | raumschiffs | reitern |
| leut | funker | notebook | teetisch | raumfahrer | county |
| trenker | meldet | inspektoren | liebenswürdigkeit | jahrtausenden | kansas |
| bergwald | flanke | ermittlung | mancherlei | terra | creek |
| bursch | lastwagen | mafia | reizendes | lichtjahre | banditen |
| bergführer | feindes | plaza | umzukleiden | humanoiden | mountains |
| feschen | pistole | ganoven | umkleiden | geortet | karabiner |
| gerad | ne | bewußte | namenlos | unsterblichen | arizona |
| bergtour | flieger | datenbanken | frohen | projektion | indianern |

*Abbildung 11. Genretypische Wörter (Zeta)*

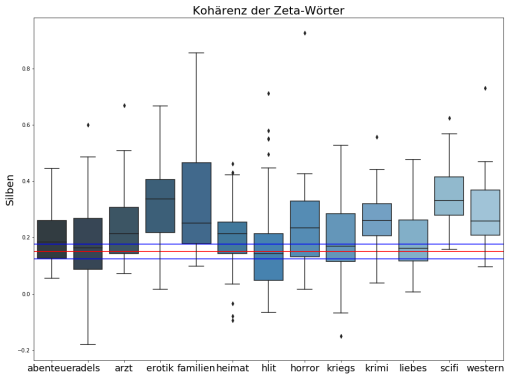


*Abbildung 12. Kohärenz der Zetawörter mittels word embedding. Rote/blaue Linien: Wert für zufällig gewählte Worte. Rot: Mittelwert, Blau +/- eine Standardabweichung*

Gattungskomplexität und Kontrast zu Hochliteratur . Die Annahme, dass Heftromane weniger komplex seien als Hochliteratur ist schon Teil des Namens: Schemaliteratur. Zugleich ist es offensichtlich, dass literarische Komplexität unterschiedliche Aspekte betreffen kann. Die hier untersuchten Aspekte sind teilweise von der ideologiekritischen Trivialliteraturforschung bereits an kleinen Samples untersucht worden, die im Fall der Heftromane zu dem Ergebnis kommt, dass der Wortschatz kleiner sei, die durchschnittliche Satzlänge kürzer und die Komplexität der Sätze geringer (Nusser 1982: 88f.) Zwar wurde die ideologiekritische Forschung zur populären Literatur in den letzten Jahren kritisiert, die quantitativen Forschungen sind jedoch nicht durch neue ersetzt worden.

Wir nehmen das Verhältnis von Types zu Tokens in einem Text als Maß für die Variabilität der Sprache und als Größe des Wortschatzes. Wie die Boxplots in Abbildung 13 zeigen, unterscheidet sich Hochliteratur ('hlit') in diesem Punkt keineswegs grundlegend vom Heftroman. t-Test.

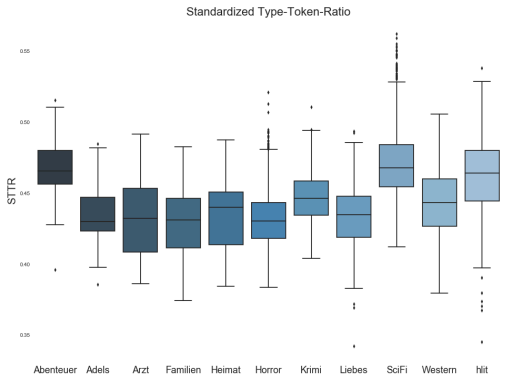


*Abbildung 13. Standardized Type-Token-Ratio*

Auch bei der durchschnittlichen Länge der Worte, häufig verwendet für Maße der Leseschwierigkeit von Texten, ist die Varianz innerhalb der Heftromane größer der Unterschied zwischen den Heftromanen und der Hochliteratur (siehe Abbildung 14).

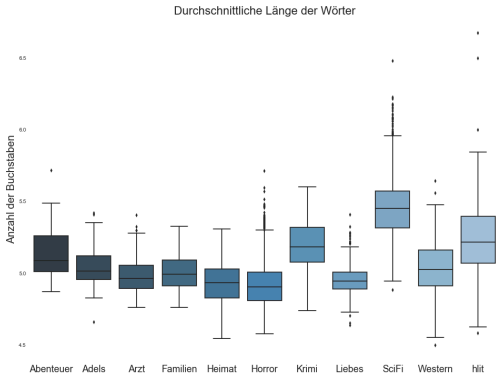


*Abbildung 14. Durchschnittliche Wortlänge in Buchstaben*

Die Messung der Satzlänge bestätigt die Untersuchungen aus den 1970er Jahren (Abbildung 15): Die Sätze sind im Durchschnitt über alle Gattungen hinweg kürzer, auch die Varianz der Satzlänge ist im Bereich der Hochliteratur deutlich größer. Allerdings widerspricht die Messung der Part-of-Speech Trigramme der Annahme, dass die Satzbaupläne ebenfalls schematischer sind (Abbildung 16).

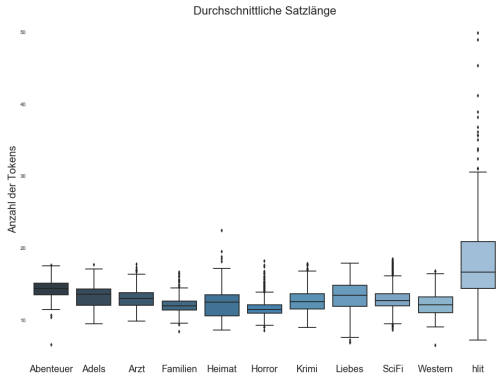


*Abbildung 15. Durchschnittliche Satzlänge in Token*

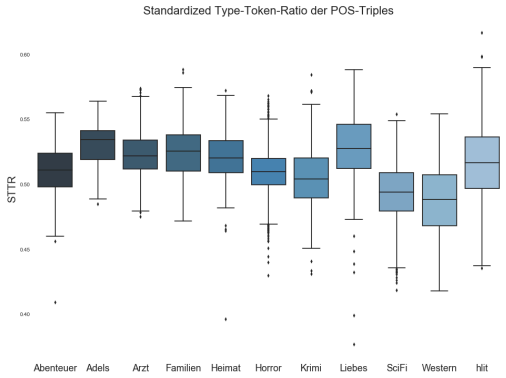


*Abbildung 16. Schematisierung von Satzbauplänen anhand von POS-Triples*

# Diskussion

Die Gattungen der Heftromane sind - so der vorläufige Befund - deutlich umrissen und lassen sich durch die Zetawörter auch inhaltlich gut erschließen. Zwei Thesen zum Verhältnis von Hochliteratur zum Heftroman können dagegen als widerlegt gelten: Das Gebiet der Heftromane ist keineswegs besonders “homogen” (Domagalski 1980), vielmehr ist die Binnenvarianz sehr deutlich. Zum anderen ist die Sprache der Heftromane nicht eindeutig schlichter (Nusser 1982). Auch hier ist die Varianz innerhalb der Gattungen auffällig; insbesondere die Science-Fiction Romane weichen deutlich ab. Diese Arbeit markiert erst den Anfang unserer Untersuchungen. In den nächsten Schritten werden Figurenkonstellation, Analyse von Erzähler- und Figurenrede sowie Sentiment untersucht werden.

# Fußnoten

1. Strategische Prioritäten 2017 – 2020, urn:nbn:de:101-2017021403

# Bibliographie

**Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003)**: *“Latent dirichlet allocation”*. Journal of machine Learning research, 3 (Jan), 993-1022.

**Burrows, J.**: *»All the Way Through. Testing for Authorship in Different Frequency Strata«*, in: Literary and Linguistic Computing 22.1 (2007), S.27–47.

**Craig, H., Kinney, A. (2009)**: *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge.

**Eder, M., Rybicki, J., and Kestemont, M. (2016)**: *Stylometry with R: a package for computational text analysis*. R Journal, 8(1): 107-121.

**Foltin, H. F. (1965)**: *Die minderwertige Prosaliteratur. Einteilung und Bezeichnung*. In: DVjs 39 H. 2, 288–323.

**Domagalski, P. (1980)**: *Trivialliteratur. Geschichte. Produktion, Rezeption.* Freiburg im Breisgau.

**Hügel, H.-O. (2007)**: *Lob des Mainstreams. Zu Theorie und Geschichte von Unterhaltung und Populärer Kultur*. Köln: Halem.

**Evert, S. Proisl, T., Reger, I., Pielström, S., Schöch, C. Vitt, T. (2017)**: *Understanding and explaining Delta measures for authorship attribution*. In: Digital Scholarship Humanities 32, 2,1, p. ii4-ii16.

**Hettinger, L., Jannidis, F., Reger, I., Hotho, A. (2016)**: *Classification of Literary Subgenres*. Abstracts DHd-Tagung 2016, Leipzig 2016.

**Kelleter, F. (2012)**: *Populäre Serialität. Eine Einführung*. In: ders.: (Hg.): Populäre Serialität: Narration – Evolution – Distinktion. Zum seriellen Erzählen seit dem 19. Jahrhundert. Bielefeld: transcript, S. 11-46.

**Kubát, M. & Mili#ka, J. (2013)**: *Vocabulary richness measure in genres*. Journal of Quantitative Linguistics, 20(4), 339–349.

**McCallum, A.**: *"MALLET: A Machine Learning for Language Toolkit."* http://mallet.cs.umass.edu . 2002.

**McInnes, L. Healy, J. (2018)**: *"UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction"*, ArXiv e-prints 1802.03426

**Nast, M. (2017)**: *"Perry Rhodan" lesen. Zur Serialität der Lektürepraktiken einer Heftromanserie*. Bielefeld: transcript.

**Nusser, P.**: *Romane für die Unterschicht. Stuttgart 1982.*

**Nutz, W.; Schlögell, V. (1991)**: *Die Heftroman-Leserinnen und -Leser in Deutschland. Beiträge zur Erfassung populärkulturaler Phänomene*. In: Communications 16 (2). DOI: 10.1515/ comm.1991.16.2.133.

**Schöch, C. (2018)**: *Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie*. In **Bernhart, T., et al. (eds.)**: *Quantitative Ansätze in der Literatur- und Geisteswissenschaften*. Berlin: de Gruyter. 77-94.

**Schöch, C. Schlör, D., Zehe, A., Gebhard, H, Becker, M., Hotho, A.**: *Burrows' Zeta: Exploring and Evaluating Variants and Parameters*. Abstracts DH 2018.

*Stiftung Lesen. Lesen in Deutschland 2008.* https:// www.stiftunglesen.de/download.php? type=documentpdf&id=11

**Stockinger, C. (2018)**: *Das Groschenheft*. In: **Carlos Spoerhase und Steffen Martus (Hg.)**: *Gelesene Literatur. Populäre Lektüre im Zeichen des Medienwandels*. München: text + kritik. (Im Druck)