# Stilometrische Untersuchung von Figurenreden in realistischen Erzähltexten

**Weimer, Lukas**
lukas.weimer@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

## Einführung

Das Poster stellt ein Korpus deutschsprachiger Erzählungen des 19. Jahrhunderts vor, in dem Figurenreden und ihre jeweiligen Sprecher annotiert und extrahiert wurden. Sie dienen als Basis für stilistische Auswertungen mit dem etablierten Abstandsmaß Delta. Es stellt sich die Frage, ob sich der Autorenstil in den jeweiligen Figurenreden niederschlägt, sich also Figuren desselben Autors zusammengruppieren, oder ob Figurentypen dominanter sind, sich gleiche Figurentypen also werkübergreifend stilistisch ähneln. Erste Ergebnisse hiervon werden als Grafiken präsentiert.

## Verwandte Forschung

Stilometrische Verfahren gehen v.a. auf John Burrows zurück. Sein entwickeltes Abstandsmaß *Delta* (Burrows 2002) gilt als Standardverfahren in der Stilometrie und es existieren zahlreiche Studien und Verbesserungsvorschläge (z.B. Smith/Aldridge 2011, Büttner et al. 2017). Für die einfache informatische Anwendung wurde es durch das R-Package *stylo* (Eder/Rybicki/Kestemont 2016) zugänglich gemacht. Die ersten quantitativen Untersuchungen des Figurenstils liefert ebenfalls erstmals Burrows (1987) in der anglistischen Literatur. Allerdings führt die unterschiedlich große Menge an Reden pro Figur zu disparatem Analysematerial. Um das Problem unterschiedlich langer Texte zu umgehen, nutzt Hoover (2017) Textauszüge bzw. zufällige Textanordnung in seiner Studie zur intratextuellen Stilvariation. Stilometrische Analysen erfreuen sich auch in der heutigen Forschung noch hoher Beliebtheit (so z.B. Bonch-Osmolovskaya/Skorinkin 2019, auf Dramentexte Galleron 2019).

## Korpus: Annotation und Datenaufbereitung

Das Korpus setzt sich aus acht realistischen Erzähltexten zwischen 1848 und 1871 zusammen, da dieser Zeitraum allgemein als Kernzeit des Realismus anerkannt ist (Aust 2006, Plumpe 2007). Um Vergleiche zu ermöglichen, enthält das Korpus zusätzlich drei Erzähltexte von vor 1848. Die Korpusauswahl beruht auf einem mehrschrittigen Prozess: Mit der Längenbegrenzung von 8.000-20.000 Wörtern wurde darauf geachtet, dass die Erzählungen einerseits lang genug sind, um stilometrische Verfahren anwenden zu können und andererseits kurz genug, um die manuelle Annotation in einem angemessenen zeitlichen Rahmen durchzuführen. Außerdem wurde darauf geachtet, sowohl kanonisierte als auch gänzlich unbekannte Texte zu integrieren, weibliche Autoren ins Korpus aufzunehmen und die Erstpublikationsorgane zu variieren. Wie in der damaligen Zeit üblich, wurde ein Großteil der Erzählungen in Zeitschriften, Almanachen oder Taschenbüchern veröffentlicht. Diese waren auf ganz verschiedene Leserschichten ausgerichtet, so dass eine Variation hier alle Stilniveaus erfassen sollte. Die Korpustexte sind die folgenden elf Erzählungen:

| Titel | Autor | Jahr | Wortanzahl |
|---|---|---|---|
| Der Gefangene | Malsburg, Otto von der | 1822 | 9.108 |
| Die Doppelgängerin | Ungern-Sternberg, Alexander von | 1834 | 8.094 |
| Die Judenbuche | Droste-Hülshoff, Annette von | 1842 | 16.191 |
| Der arme Spielmann | Grillparzer, Franz | 1848 | 15.132 |
| Das Erdbeerimareili | Gotthelf, Jeremias | 1850 | 15.720 |
| Bergmilch | Stifter, Adalbert | 1853 | 9.727 |
| Phosphorus Hollunder | François, Louise von | 1857 | 14.082 |
| Die schwarze Galeere | Raabe, Wilhelm | 1861 | 15.585 |
| Die zwölf Apostel | Marlitt, Eugenie | 1865 | 20.288 |
| Eine Malerarbeit | Storm, Theodor | 1867 | 9.392 |
| Der Leuchtturm von Livorno | Eckstein, Ernst | 1871 | 8.992 |

*Tabelle 1: Im Korpus enthaltene Erzählungen.*

Da einige der Texte noch nicht erschlossen waren, wurden sie vor der Annotation OCR-korrigiert. Für die Annotation wurde der im Zuge des Redewiedergabe-Projekts (Brunner et al. 2018) entstandene STWR-View des Annotationstools ATHEN (Krug et al. 2018) verwendet. Bei der Annotation wurden sämtliche direkten Figurenreden manuell annotiert und ihrem jeweiligen Sprecher zugeordnet (zur automatischen Zuordnung von Sprechern: Krug et al. 2016). So konnte die gesamte direkte Redemenge einzelner Figuren extrahiert werden. In direkte Reden einer Figur A eingelagerte Reden einer Figur B wurden dabei nur der Figur B als zugehörig annotiert. Auf diese Weise wurde sichergestellt, dass Figuren ausschließlich ihre eigenen Reden zugeordnet wurden (diese Problematik ist besonders relevant bei Binnenerzählungen). Zusätzlich wurden ausschließlich Figuren in die Auswertung integriert, deren gesamte Redemenge 200 Wörter übersteigt, um stilometrische Verfahren wirksam anwenden zu können. Diese Grenze ist für stilometrische Verfahren noch immer vergleichsweise niedrig. Eder (2015) hat evaluiert, dass korpusabhängig mindestens 2500-5000 Wortformen nötig sind, damit Auswertungen mit Delta zu guten Ergebnissen

führen. Aufgrund des Korpus dieser Studie kann dieser Mindestwert allerdings nicht eingehalten werden.

# Auswertung

Die folgenden Grafiken zeigen den Output des R-package *stylo* (Eder/Rybicki/Kestemont 2016) erstens der 100 häufigsten gesprochenen Wörter der Figuren und zweitens der 1000 häufigsten. Es wurde kein Sampling durchgeführt und ebenfalls kein Culling, ein Feature musste folglich nicht in einer bestimmten Anzahl Texte vorhanden sein, um in die Auswertung einbezogen zu werden. Als Abstandsmaß wurde klassisch Burrows‘ Delta gewählt, die Outputgrafiken sind Cluster-Analysen, die stilistisch ähnliche Figuren zueinander gliedern. Zu beachten ist die oben erwähnte Mindestmenge von 200 Wörtern pro Figur. Das führt dazu, dass bei der Analyse der 1000 häufigsten Wörtern von einigen Figuren alle gesprochenen Wörter in der Auswertung enthalten sind.

Auswertung mit 100 häufigsten Wörtern:

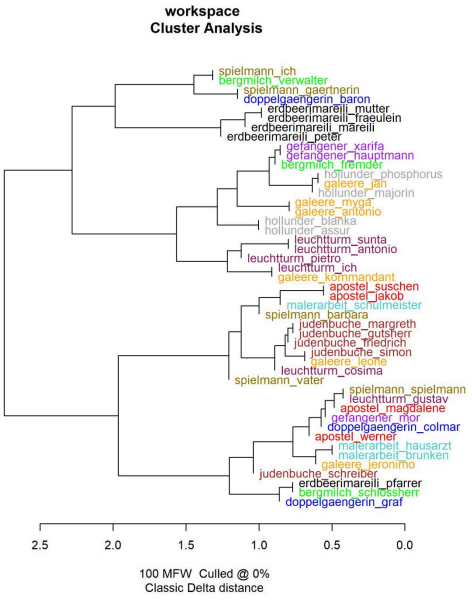


*Abbildung 1: Auswertung mit 100 häufigsten Wörtern.*

Auswertung mit 1000 häufigsten Wörtern:

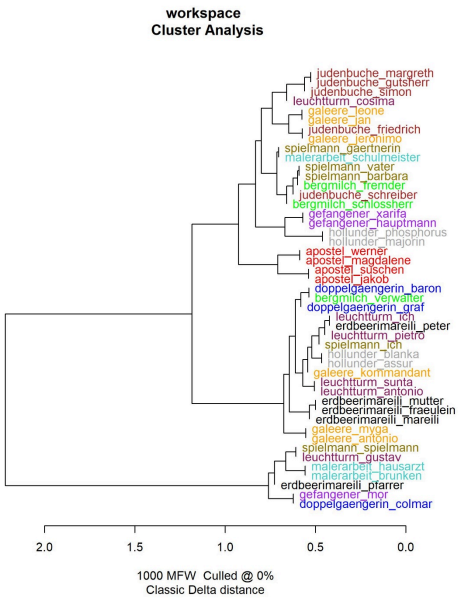


*Abbildung 2: Auswertung mit 1000 häufigsten Wörtern.*

In beiden Auswertungen ist zu erkennen, dass sich häufig Figuren desselben Autors zueinander gliedern. Besonders beim Mundartdichter Gotthelf (*Erdbeerimareili*) ist das sehr verständlich. Dennoch gibt es Abweichungen. Besonders bei den 1000 häufigsten Wörtern gruppieren sich auf dem untersten Ast die Figuren mit der größten Redemenge zusammen. Dies sind häufig Binnenerzähler, die in ihrem Redestil häufig schnell die Funktion und den Stil von Erzählerrede einnehmen (Bockwinkel 2016). Um zu untersuchen, ob das Clustering nur der insgesamt größeren Redemenge dieser Figuren geschuldet ist, wurde in mehreren Analysen Sampling durchgeführt. Die hier nicht abgebildeten Auswertungen bestätigen das Ergebnis, wenngleich der Abstand der Binnenerzähler zu den übrigen Figuren geringer wird. Außerdem nimmt Colmar aus der *Doppelgängerin* einen größeren Abstand zu den übrigen Binnenerzählern ein. Auch das ist nachvollziehbar, da Colmar im Gegensatz zu ihnen nur über einen kleinen Teil der Erzählung als Binnenerzähler fungiert und sonst wie eine „normale“ Figur agiert. Figurentypen gliedern sich in dieser ersten Vorstudie dagegen nicht zusammen. Figurenpaare, die sich in gegenseitiger Liebe befinden (wie Magdalene-Werner, Xarifa-Mor, Myga-Jan, Cosima-Antonio) gruppieren sich nur teilweise als Paar und gar nicht als Figurengruppe. Weitere Schlussfolgerungen, dass sich beispielsweise gleiches Geschlecht, Figuren aus ähnlichen Subgenres (Abenteuer/Liebe) oder Erzählungen aus einer bestimmten Epoche gruppieren, können in dieser ersten Vorstudie ebenfalls noch nicht gezogen werden. Gleichfalls kann diese Studie aber auch noch nicht als Beweis fungieren, dass sich deren Stil nicht ähnelt.

# Ausblick

Im weiteren Verlauf der Arbeit müssen die Maße verfeinert und sollen andere Abstandsmaße getestet, Variablen geändert und Ergebnisse evaluiert werden. Die Problematik der Kürze der Texte könnte durch eine Optimierung des Verfahrens verringert werden. So könnten eine Kombination aus Wortform- *Grammar-Tags* und besonders gut zur Autorschaftsattribution geeigneter Wörter Verbesserungen bringen (Dimpel 2019). Eine Integration von Gedanken- und Schriftzitaten ist ebenfalls denkbar. Interessant wäre auch die Berücksichtigung von indirekter Rede, da hier ebenfalls die Figurenstimme stark ist. In einem Schritt weg von der Stilometrie sollen in späteren Tests darüber hinaus Topic Modeling und Sentimentanalyse durchgeführt werden, um die Figurenreden auch auf diesen Ebenen zu vergleichen.

# Bibliographie

**Aust, Hugo** (2006): *Realismus*. Lehrbuch Germanistik, Stuttgart: Metzler.

**Bockwinkel, Peggy** (2018): "Wie anders ist Figurenrede? Die Rolle der direkten Rede in quantitativen Erzähltextanalysen", in: Bockwinkel, Peggy / Nickel, Beatrice / Viehhauser, Gabriel (eds.): *Digital Humanities. Perspektiven der Praxis.* Berlin: Frank&Timme 117-148.

**Bonch-Osmolovskaya, Anastasia / Skorinkin, Daniil** (2019): "The Complexity of Character-building: Speech, Portraits, Interactions in Leo Tolstoy's 'War and Peace'", in: *Conference Abstracts of DH2019 Utrecht*.

**Brunner, Annelen / Engelberg, Stefan / Jannidis, Fotis / Tu, Ngoc Duyen Tanja / Weimer, Lukas** (2018): "Projektvorstellung – Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse", in: *Konferenzabstracts der DHd2018 Köln* 458-460.

**Burrows, John** (1987): *Computation into Criticism*. A Study of Jane Austen's Novels and an Experiment in Method, Oxford: Clarendon Press.

**Burrows, John** (2002): "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing* 17/3: 267-287.

**Büttner, Andreas / Dimpel, Michael / Evert, Stefan / Jannidis, Fotis / Proisl, Thomas / Reger, Isabella / Schöch, Christof / Vitt, Thorsten** (2017): "'Delta' in der stilometrischen Autorschaftsattribution", in: *Zeitschrift für digitale Geisteswissenschaft* 2.

**Dimpel, Friedrich Michael / Zeppezauer-Wachauer, Katharina / Schlager, Daniel** (2019): "Der Streit um die Birne. Autorschafts-Attributionstest mit Burrows' *Delta* und dessen Optimierung für Kurztexte am Beispiel der ‚Halben Birne' des Konrad von Würzburg", in: *Das Mittelalter* 24/1: 71-90.

**Eder, Maciej** (2015): "Does Size Matter? Authorship Attribution, Small Samples, Big Problem", in: *Digital Scholarship in the Humanities* 30/2: 167-182.

**Eder, Maciej / Rybicki, Jan / Kestemont, Mike** (2016): "Stylometry with R: A Package for Computational Text Analysis", in: *The R Journal* 8/1: 107-121.

**Galleron, Ioana** (2019): "Stylometric Analyses of Character Speeches in French Plays", in: *Conference Abstracts of DH2019 Utrecht*.

**Krug, Markus / Jannidis, Fotis / Reger, Isabella / Macharowsky, Luisa / Weimer, Lukas / Puppe, Frank** (2016): "Attribuierung direkter Reden in deutschen Romanen des 18.-20. Jahrhunderts. Methoden zur Bestimmung des Sprechers und des Angesprochenen", in: *Konferenzabstracts der DHd2016 Leipzig* 181-186.

**Krug, Markus / Tu, Ngoc Duyen Tanja / Weimer, Lukas / Reger, Isabella / Konle, Leonard / Jannidis, Fotis / Puppe, Frank** (2018): "Annotation and beyond – Using ATHEN Annotation and Text Highlighting Environment", in: *Konferenzabstracts der DHd2018 Köln* 19-21.

**Plumpe, Gerhard** (2007): "Realismus", in: Müller, Jan-Dirk / Braungart, Georg / Fricke, Harald / Grubmüller, Klaus / Vollhardt, Friedrich / Weimar, Klaus (eds.): *Reallexikon der deutschen Literaturwissenschaft* 3. Berlin / New York: De Gruyter 221-224.

**Smith, Peter W. H. / Aldridge, W.** (2011): "Improving Authorship Attribution: Optimizing Burrows' Delta Method", in: *Journal of Quantitative Linguistics* 18/1: 63-88.