Methodenbeitrag: Stilometrie

Jan Horstmann [1]

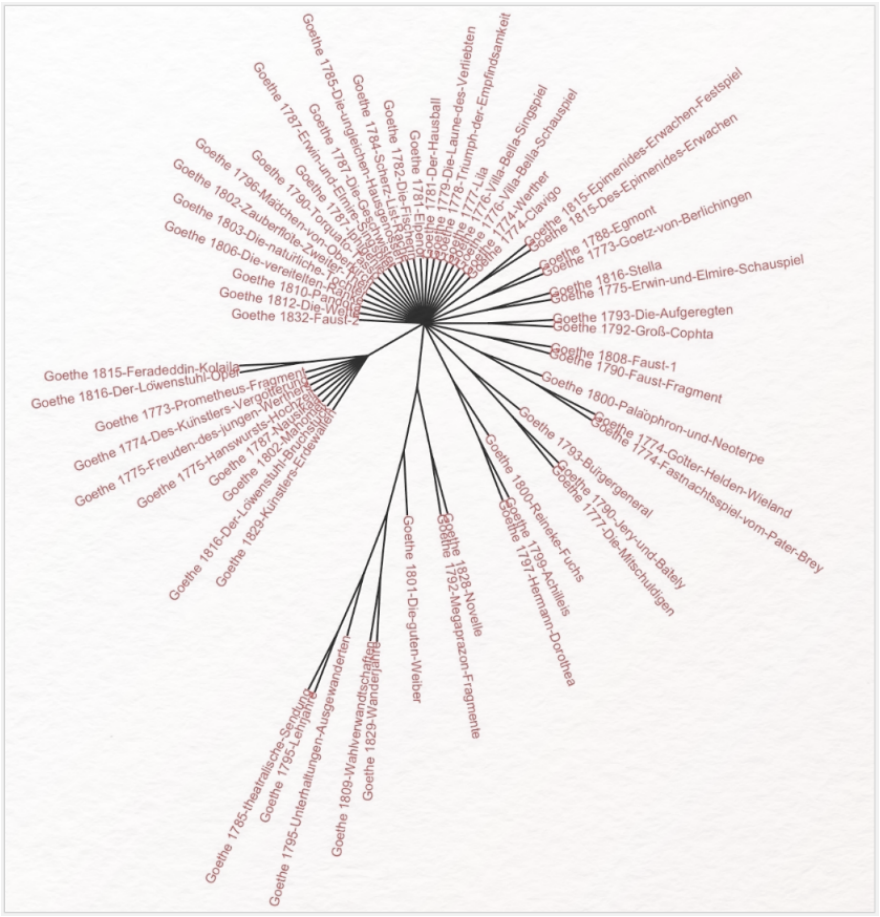
1. Universität Münster

forTEXT

Thema:	Stilometrie	DOI:	10.48694/fortext.3769
Jahrgang:	1	Ausgabe:	1
Erscheinungsdatum:	2024-02-26	Erstveröffentlichung:	2018-09-06 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Definition



Bootstrap Consensus Tree des dramatischen und epischen Œuvres Johann Wolfgang von Goethes

In der digitalen Stilometrie werden Texte oder Textpassagen auf Grundlage statistischer Verteilungen (i. d. R. der häufigsten Wörter) stilistisch miteinander verglichen. So lässt sich beispielsweise die stilistische Entwicklung oder Differenzierung eines literarischen Textes, eines Œuvres, oder gar einer ganzen Epoche quantitativ nachvollziehen. Insbesondere werden stilometrische Methoden bei Autorschaftsattributionen, Genreklassifikationen, Epochendifferenzierungen oder auch in der forensischen Linguistik eingesetzt.

2. Anwendungsbeispiel

Sie beschäftigen sich mit einem unter Pseudonym veröffentlichten literarischen Text und möchten herausbekommen, wer die Verfasserin oder der Verfasser gewesen ist, um eine kontextsensitive literaturwissenschaftliche

Analyse des Textes auf dieser Grundlage rechtfertigen zu können. Die Themenwahl, Figurenzeichnung, Plotentwicklung, das Setting oder auch der Stil erinnern Sie an andere Texte und Ihnen fallen drei Autorinnen ein (oder Sie ziehen dafür philologische Forschungsbeiträge zu Rate), die diesen Text potentiell geschrieben haben könnten. Sie stellen daher eine digitale Textsammlung (vgl. Korpus) aller Texte dieser Autorinnen zusammen, derer Sie habhaft werden können, und reichern diese Sammlung außerdem mit Texten weiterer vergleichbarer Autor*innen an, um die Möglichkeit fehlerhafter Ergebnisse zu minimieren. Eine digitale stilometrische Analyse wird Ihnen mit großer Zuverlässigkeit anzeigen, wem die Autorschaft des Textes am ehesten zugeschrieben werden kann.

3. Literaturwissenschaftliche Tradition

In mehreren Formen der literaturwissenschaftlichen Stilistik lassen sich Traditionslinien der digitalen Stilometrie ausmachen.

Die Stilanalyse bzw. Stilistik (damals noch mehr als normative Stilistik verstanden) löst als Textanalysemethode im späten 18. Jahrhundert die Rhetorik ab. Novalis erkannte 1798/99 in ihr „ungemein viel Aehnlichkeit mit der Declamationslehre - oder der Redekunst im strengern Sinne“ (Czapla 2003, 515). Nach Schaffrick und Willand (2014, 29) richtet sich die Stilometrie „grundsätzlich an einer vergleichenden Fragestellung aus und stellt den Stil verschiedener Epochen, Werke, Gattungen oder [...] Autoren gegenüber“. Manuell wurde dies bereits im 19. Jahrhundert vollzogen (Holmes 1998, 112; Kelih 2008, 31–44; Tuldava 2005, 370f.).
Die Stilanalyse als angewandte Stilforschung vor dem Hintergrund verschiedener Stiltheorien (Plummer 2007, 2007) wird von Meyer (2007, 70) auch als „Schlüsselqualifikation literaturwissenschaftlicher Arbeit“ bezeichnet, da „Basiskenntnisse im Bereich der Stilanalyse eine Voraussetzung für jede professionelle Beschäftigung mit literarischen wie auch nichtliterarischen Texten“ seien. Diese Schlüsselqualifikation findet in Form der digitalen Stilometrie ihre Tradition vor allem in quantitativ-stilistischen, aber auch in formalistisch und strukturalistisch ausgerichteten Stilanalysen. In der formalistischen Stilanalyse Michael Riffaterres (Czapla 2003, 517), auch „Funktionalstilistik“, genannt wird Stil als Normabweichung interpretiert (Meyer 2007, 74). In der strukturalistischen Stilanalyse wird hingegen die Regelhaftigkeit stilistischer Äquivalenz- und Oppositionsbeziehungen und damit die Erfüllung einer stilistischen Norm untersucht. Die strukturalistische Stilanalyse interessiert sich für „positive Merkmale einer poetischen Sprache, für artifizielle Gleichförmigkeiten (Isomorphien)“ (Meyer 2007, 75). Sowohl die formalistische als auch die strukturalistische Stilanalyse können somit als Vorgänger der digitalen Stilometrie betrachtet werden, da sie sowohl Abweichungen als auch Isomorphien exploriert.
Ganz besonders jedoch knüpft die digitale Stilometrie an die Vorstöße der statistischen Stilanalyse an, die sich mit „Wortwiederholungshäufigkeiten, Wortverteilungshäufigkeiten, Strukturanalogien, Bildung semantischer Felder etc.“ (Meyer 2007, 70) beschäftigt. Auch deshalb bezeichnet beispielsweise Czapla (2003, 515) die Stilistik als „Bindeglied zwischen Sprach- und Literaturwissenschaft“.
In der statistischen Stilanalyse begreift Doležel (1971, 253) Stil als „Wahrscheinlichkeitsbegriff“, d. h. unter der gegebenen Bedingung X (z. B. dass ein Text von einer bestimmten Autorin stammt) kommt die stilistische Erscheinung A (z. B. eine bestimmte Satzlänge) nur zu einer gewissen Wahrscheinlichkeit vor. Diese Art und Weise der Skizzierung von Stil arbeitet der späteren digitalen Modellierung bereits sehr entgegen und trägt der Tatsache Rechnung, dass eine Autorin, die für ihre kurzen Sätze bekannt ist, in ihren Texten auch längere Sätze schreiben kann und wird. Doležel (1971, 264) konstatiert außerdem: „[J]eder Text kann in einem multidimensionalen Raum beschrieben werden, in dem die Werte [der messbaren Textcharakteristiken] die individuellen Faktoren bilden“ und nimmt damit eine Grundidee der digitalen Stilometrie vorweg.

4. Diskussion

Gern und häufig wird eine der bislang populärsten Verwendungen computergestützter stilometrischer Verfahren herangezogen, wenn es darum geht, die Wirkkraft dieser digitalen Methode zu veranschaulichen: Der unter dem Pseudonym „Robert Gailbraith“ 2013 veröffentlichte Roman *The Cuckoo's Calling* wurde mithilfe einer stilometrischen Analyse der Bestsellerautorin J. K. Rowling zugeschrieben, die sich daraufhin zur Autorschaft bekannte und dadurch nicht nur die Verkaufszahlen des Buches in die Höhe schnellen ließ, sondern nebenher auch noch die Methode selbst berühmt machte (Juola 2015). Der Algorithmus, der hinter diesem Verfahren steckt, heißt Burrows' Delta und ist der am häufigsten angewendete in der computergestützten Stilometrie.
Die sog. Delta-Messung wird auch nach ihrem Erfinder „Burrows' Delta“ genannt (Burrows 2002). Bei dieser Methode wird Stil jedoch anders gedacht als in vielen der traditionelleren Stilanalysen: Statt semantischer Inhaltswörter werden hier besonders Funktionswörter betrachtet, oder um genau zu sein: die häufigsten Wörter (noch genauer: Tokens (vgl. Type/Token)) eines Textes oder einer Textsammlung. Diese *most frequent words* (MFW) werden von Autor*innen „kaum bewusst manipuliert“ (Jannidis und Lauer 2014, 180) und bieten aufgrund ihres schlicht häufigeren Vorkommens eine verlässlichere Datenbasis für eine automatische Vergleichsanalyse als seltene Wörter.
Eine Charakterisierung des Stils einer bestimmten Autorin à la „XY benutzt besonders viele Neologismen“

wird es in der digitalen Stilometrie daher nicht geben. Gerade im Zusammenhang mit der Untersuchung des Phänomens Autorenstil bedarf es jedoch numerischer Merkmale, die unabhängig von Textsorte, Thema und dem Verstreichen von Zeit wiederkehren (Jannidis und Lauer 2014, 178). In der vergleichenden Stilanalyse Thomas und Heinrich Manns und Hermann Hesses untersucht Grimm (1991) daher beispielsweise die Vorkommnisse von syntaktischen, lexikalischen und morphologischen Mitteln, aber auch von Semikola oder Auslassungen durch drei Punkte. Da viele Autor*innen im Laufe ihres Lebens ihren Stil verändern (oder der Stil sich unbewusst verändert), ist es nach wie vor nicht gelungen, einen genuinen Autorenstil, der sich über das jeweilige Œuvre als Ganzes erstreckt, statistisch dingfest zu machen. Jannidis und Lauer (2014) unterscheidet in diesem Zusammenhang einen starken und einen schwachen Begriff von Autorstil, wobei der starke Begriff sämtliche Texte eines Autors/einer Autorin meint, der schwache jedoch nur einige Texte. Die Konzentration auf den schwachen Begriff von Autorenstil macht es möglich, ein Œuvre als Ganzes zu begreifen und dennoch seine dynamische stilistische Entwicklung zu beschreiben.

„Welche Stilprinzipien jeweils für eine Epoche, Textsorte oder Autor-Individualität von tendenzieller Geltung sind und was dabei jeweils als sprachliche Angemessenheit, Ornamentik, Eleganz o. ä. postuliert wird, ist von zeitgebundenen Paradigmen abhängig“ (Michel 2003, 520). Die digitale Stilometrie bringt mit ihrer Konzentration auf die *most frequent words* in den Kanon dieser Stilprinzipien eine neue Perspektive ein, die einen großen Erkenntniszuwachs bringt, die anderen Aspekte jedoch nicht ablösen kann: Über Ornamentik, Eleganz etc. von Texten vermag sie keine Aussage zu treffen.

Die digitale Stilometrie ist zudem rein textimmanent und kann extratextuelle Zeichen wie z. B. die Beschaffenheit des Manuskripts mit all seinen stilistischen Zusatzinformationen, die in der Druckfassung verloren gegangen sind (Meyer 2007, 78), nicht mit deuten. Dies ist freilich ein generelles Problem vieler digitaler Methoden und betrifft insbesondere die Möglichkeiten der Textdigitalisierung (Horstmann 2024b) und die digitale Manuskriptanalyse (Horstmann 2024c).

Die Konzentration auf ein bestimmtes Stilmerkmal beim Vergleich unterschiedlicher Texte, Gattungen oder Œuvres macht die Stilanalyse jedoch handhabbar, oder wie Jannidis und Lauer (2014, 172) zusammenfasst: „Deutlich moderater wäre ein Test, dem es gelingt, aufgrund bestimmter Merkmale und Merkmalskombinationen die Texte eines Autors von den Texten anderer Autoren zu unterscheiden, also ein individualisierendes oder unterscheidendes Verfahren“. Man sollte die vergleichsweise unbewusste Verwendung von häufigen Funktionswörtern in Texten jedoch nicht mit einer Art ‚stilistischem Fingerabdruck‘ von Autor*innen gleichsetzen, denn einzigartige Merkmale oder Merkmalsbündel, die Stil individuell definitiv bestimmen, gibt es nicht. Stilometrische Verfahren liefern stattdessen *wahrscheinliche* Autorschaftszuschreibungen und diese Wahrscheinlichkeit kann höher oder niedriger sein (Jannidis und Lauer 2014, 183).

Neben der Autorschaftsattribution kann die stilometrische Analyse nach Burrows in weiteren Zusammenhängen fruchtbar gemacht werden: So lässt sie Rückschlüsse auf Textsortenzugehörigkeiten, Œuvreperiodisierungen, Übersetzungsstilistiken, Epochenzusammenhänge, Genderzugehörigkeiten etc. zu. Es hat sich zudem gezeigt, dass die Wahl der Menge der häufigsten Wörter sprachabhängig ist (Rybicki und Eder 2011), da „bei Sprachen mit größerer morphologischer Formenvielfalt zu erwarten [ist], daß die relative Häufigkeit der häufigen Wörter insgesamt weniger groß ist“ (Bock u. a. 2016, 9). Rybicki und Eder (2011) kommen in ihren Experimenten daher zu dem Ergebnis, dass Burrows’ Delta am besten mit englischen oder deutschen Prosatexten funktioniert.

Die quantitative digitale Stilometrie ermöglicht einen neuen Blick auf Stile in Textsammlungen. Für gute Ergebnisse ist eine ausreichende (möglichst große) Datengrundlage jedoch unumgänglich. Ist diese gegeben, spricht die Qualität der Ergebnisse für sich: Sie sind mit so hoher Wahrscheinlichkeit korrekt, dass die Verfahren sogar in der forensischen Linguistik Anwendung finden und als vor Gericht haltbar gelten können (Fobbe 2011, 109f.). Dieser quantitative Blick (vgl. Distant Reading) ist jedoch nur eine Perspektive auf Stil und kann und will (insbesondere bei kürzeren Texten) die qualitative Stilanalyse (vgl. Close Reading) nicht ersetzen (Tuldava 2005, 369).

5. Technische Grundlagen

Die stilometrische Analyse mit Burrows’ Delta betrachtet die häufigsten Wörter der Textsammlung. In Burrows’ eigenem Beispiel sind das die 30 häufigsten Wörter und diese MFW (*most frequent words*) bezeichnet er als „markers of potentially equal power“ (Burrows 2002, 271) für Stildifferenzen. Nun ist es bei nach Häufigkeiten angeordneten Wortlisten so, dass die Zahlenwerte sehr schnell abfallen, dass also beispielsweise die ersten 3 bis 5 Wörter sehr viel häufiger vorkommen als die Wörter 6-10 usw. Damit die häufigsten dieser häufigen Wörter das Ergebnis der stilometrischen Analyse nicht allein dominieren, sondern alle 30 häufigsten Wörter in der Berechnung ein gleiches Gewicht bekommen, wird nicht mit Rohwerten, sondern mit dem sog. *z-score* gerechnet. Im *z-score* wird vom Zahlenwert der häufigsten Wörter der Mittelwert abgezogen und das Ergebnis wird durch die Standardabweichung (d. h. die durchschnittliche Streuung um den Mittelwert) geteilt.

Um schließlich das Delta zu berechnen, werden die 30 (bzw. *n*) *z-scores* des einen Textes oder der einen Textsammlung von den 30 *z-scores* des anderen Textes oder der anderen Textsammlung abgezogen. Diese Differenzen werden aufaddiert und das Ergebnis wird durch die gewählte Menge der häufigsten Wörter (in diesem Beispiel 30) geteilt. Bei dieser rechnerischen Komplexitätsreduktion kommt ein numerischer Wert heraus; Delta kann

also in einer Zahl angegeben werden und wird auch als Distanzmaß bezeichnet (Schöch 2017, 292). Kleinere Deltawerte (im Verhältnis zu allen in einem Analysedurchgang jeweils errechneten Deltawerten) stehen laut dieser Berechnung für größere stilistische Nähe bzw. für kleinere stilistische Differenz. Eine etwas ausführlichere Erläuterung dieser Rechnung findet sich bei Jannidis und Lauer (2014, 183–185). Eine Vereinfachung der Formel (bei welcher der Mittelwert herausgekürzt und somit der Bezug auf einen normgebenden Haupttext geschmälert wird) bietet Argamon (2007).

Bei der quantitativen Berechnung stilistischer Unterschiedlichkeit entstehen schnell Werte, die durch eine Vielzahl von Faktoren bestimmt werden. Wenn in einer Textsammlung von nur 10 Texten beispielsweise die jeweils 30 häufigsten Wörter stilometrisch analysiert werden sollen (Rybicki und Eder 2011), lässt sich jeder einzelne Text durch einen Vektor mit 30 Zahlen repräsentieren: die Frequenzen der jeweils häufigsten 30 Wörter. Dadurch nimmt jeder Text einen spezifischen Punkt in einem 30-dimensionalen Raum ein.

Um dieses hochdimensionale Raumkonstrukt greifbar zu machen, werden die 30 Dimensionen auf zwei Achsen reduziert. Dazu braucht man die sog. *principal component analysis* (PCA) (vgl. PCA). Anstatt für jedes Wort eine neue Dimension zu eröffnen, bildet dabei die Varianz selbst die Grundlage für die neu definierten Dimensionen: x-Achse = PC1 und y-Achse = PC2 (bei Bedarf noch eine z-Achse = PC3). Entlang der Achse Principal Component 1 werden somit die Texte in Bezug auf die größte Varianz aufgefächert, entlang der Achse Principal Component 2 in Bezug auf die zweitgrößte Varianz usw. Damit hat man in der Regel schon eine so große Menge an Daten abgedeckt, dass weitere Principal Components vernachlässigt werden können, ohne dass der Informationsverlust zu groß wäre.

Die Reduktion auf zwei Dimensionen bietet den großen Vorteil, dass sämtliche Texte einer stilometrischen Analyse innerhalb eines Koordinatensystems wie auf einer Karte visualisiert werden können, in dem räumliche Nähe (da Delta ein Distanzmaß ist) auch für stilistische Nähe steht. (Zu den Chancen und Risiken von unterschiedlichen Datenvisualisierungen vgl. Textvisualisierung (Horstmann und Stange 2024)). Statt die Vollständigkeit der gesamten Daten abzubilden, liegt der Fokus der PCA-Analyse also auf Aspekten, die für die stilistische Varianz von besonderer Bedeutung sind (Bock u. a. 2016; Tweedie 2005, 388).

Bei dieser und auch den im Folgenden erwähnten Verfahren und Möglichkeiten der Ergebnisvisualisierung bestimmen immer Sie selbst die jeweiligen Parameter, die in die Analyse einbezogen werden, und auch die Ergebnisse müssen von Ihnen vor dem Hintergrund Ihres literaturwissenschaftlichen Fachwissens gedeutet werden. Diese Aufgaben kann die digitale Methode nicht übernehmen. Eine häufiger verwendete Möglichkeit, stilistische Nähe von Texten zu visualisieren, ist die Darstellung in einem Baumdiagramm oder Dendrogramm, bei dem i. d. R. die vertikale Anordnung von Textgruppen in sog. Clustern auf den gleichen Ästen des Diagramms die stilistische Nähe dieser Texte zueinander anzeigt.

Die in diesem Beitrag aufgeführte Visualisierung (s. o.) nennt sich *Bootstrap Consensus Tree* und wurde mit Stylo (Horstmann 2024a) (s. nächsten Absatz) erstellt. Sie zeigt die stilistische Varianz von Goethes dramatischem und epischem Œuvre. Die Visualisierung beruht (im Gegensatz zum Dendrogramm) auf wiederholten und vergleichenden Durchläufen der Berechnung mit unterschiedlichen Parametern (bspw. jeweils unterschiedlich viele MFW) und zeigt stilistisch ähnliche Texte an den gleichen Ästen an. Dabei werden nur diejenigen Ähnlichkeiten angezeigt, die in einem vorher festgelegten prozentualen Anteil der Analysedurchläufe auftreten. Die Entfernung vom Zentrum hat dabei keine semantische Aussagekraft, sondern ist lediglich der visuellen Darstellbarkeit geschuldet.

Auch für Einsteiger*innen nach kurzer Einführung relativ leicht zugänglich und nutzbar ist das sog. „Stylo“-Package (Eder, Rybicki und Kestemont 2016), das in der Java-basierten statistischen Programmierumgebung R kostenlos angewendet werden kann und bei Bedarf sogar über eine grafische interaktive Benutzeroberfläche (vgl. GUI) verfügt. Die grundlegenden Befehle werden hier über speziell dafür eingerichtete Buttons ausgeführt und Skript- oder Codekenntnisse sind (bis auf das Laden und Starten des Programms) nicht zwangsläufig vonnöten.

Externe und weiterführende Links

- R Project: https://web.archive.org/save/https://www.r-project.org/ (Letzter Zugriff: 20.02.2024)

Bibliographie

Argamon, S. 2007. Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing* 23, Nr. 2 (1. Oktober): 131–147. doi: 10.1093/llc/fqn003, https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqn003 (zugegriffen: 13. Dezember 2019).

Bock, Sina, Keli Du, Michael Huber, Stefan Pernes und Steffen Pielström. 2016. Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften: Bericht über den Stand der Forschung. DARIAH Working Papers Nr. 18: 4–9. http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2016-18.pdf.

Burrows, John. 2002. Delta: A Measure for Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17, Nr. 3: 267–287.

Czapla, Ralf Georg. 2003. Stilistik. In: *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, 3: P-Z:515–518. Berlin, New York: de Gruyter.

Doležel, Lubomír. 1971. Ein Begriffsrahmen für die statistische Stilanalyse. In: *Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven*, hg. von Jens Ihwe, 1: Grundlagen und Voraussetzungen:253–273. Frankfurt am Main: Athenäum.

Eder, Maciej, Jan Rybicki und Mike Kestemont. 2016. Stylometry with R: A Package for Computational Text Analysis. *The R Journal* 8, Nr. 1: 107–121. https://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf (zugegriffen: 12. November 2018).

Fobbe, Eilika. 2011. *Forensische Linguistik. Eine Einführung*. Tübingen: Narr.

Grimm, Christian. 1991. *Zum Mythos Individualstil. Mikrostilistische Untersuchungen zu Thomas Mann*. Würzburg: Königshausen & Neumann.

Holmes, D. I. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13, Nr. 3 (1. September): 111–117. doi: 10.1093/llc/13.3.111, https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/13.3.111 (zugegriffen: 13. Dezember 2019).

Horstmann, Jan. 2024a. Toolbeitrag: Stylo. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 1. Stilometrie (26. Februar). doi: 10.48694/fortext.3770, https://fortext.net/tools/tools/stylo.

———. 2024c. Methodenbeitrag: Digitale Manuskriptanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3744, https://fortext.net/routinen/methoden/digitale-manuskriptanalyse.

———. 2024b. Methodenbeitrag: Möglichkeiten der Textdigitalisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3741, https://fortext.net/routinen/methoden/moeglichkeiten-der-textdigitalisierung.

Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, https://fortext.net/routinen/methoden/textvisualisierung.

Jannidis, Fotis und Gerhard Lauer. 2014. Burrows's Delta and Its Use in German Literary History. In: *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, hg. von Matt Erlin und Lynne Tatlock, 29–54. Rochester, New York: Camden House.

Juola, Patrick. 2006. Authorship Attribution. *Foundations and Trends in Information Retrieval* 1, Nr. 3: 233–334. doi: 10.1561/1500000005,.

———. 2015. The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in The Humanities* 30, Nr. 1: 100–113. doi: 10.1093/llc/fqv040, (zugegriffen: 3. September 2018).

Kelih, Emmerich. 2008. *Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft*. Hamburg: Kovač.

Meyer, Urs. 2007. Stilanalyse. In: *Handbuch Literaturwissenschaft*, hg. von Thomas Anz, 2: Methoden und Theorien:70–80. Stuttgart, Weimar: Metzler.

Michel, Georg. 2003. Stilprinzip. In: *Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte*, III: P-Z:518–521. Berlin, New York: de Gruyter.

Plummer, Patricia. 2007. Stilistik. In: *Metzler Lexikon Literatur. Begriffe und Definitionen*, hg. von Dieter Burdorf, Christoph Fasbender, und Burkhard Moenninghoff, 734. Stuttgart, Weimar: Metzler.

Rybicki, Jan und Maciej Eder. 2011. Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? *Literary and Linguistic Computing* 26, Nr. 3: 315–321.

Schaffrick, Matthias und Marcus Willand, Hrsg. 2014. *Theorien und Praktiken der Autorschaft*. Berlin, Boston: de Gruyter.

Schöch, Christof. 2017. Quantitative Analyse. In: *Digital Humanities. Eine Einführung*, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein, 279–298. Stuttgart: Metzler.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60, Nr. 3: 538–556. doi: 10.1002/asi.21001, https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001 (zugegriffen: 13. Dezember 2019).

Tuldava, Juhan. 2005. Stylistics, author identification. In: *Quantitative Linguistik: ein internationales Handbuch*, hg. von Reinhard Köhler, Gerold Ungeheuer, und Herbert Ernst Wiegand, 368–387. Berlin, New York: de Gruyter.

Tweedie, Fiona J. 2005. Statistical Models in Stylistics and Forensic Linguistics. In: *Quantitative Linguistik: ein internationales Handbuch*, hg. von Reinhard Köhler, Gerold Ungeheuer, und Herbert Ernst Wiegand, 387–397. Berlin, New York: de Gruyter.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

PCA PCA steht für *Principal Component Analysis* . Die Hauptkomponentenanalyse ist ein komplexes, statistisches Verfahren zur Reduktion und Veranschaulichung umfangreicher Datensätze.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein

konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.