Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften
–
Bericht über den Stand der Forschung

Sina Bock, Keli Du, Michael Huber, Stefan Pernes, Steffen Pielström

(Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte, Julius-Maximilians-Universität Würzburg)

DARIAH-DE Working Papers Nr. 18

Bock, S.; Du, K.; Huber, M.; Pernes, S., Pielström, S.: Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften : Bericht über den Stand der Forschung
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2016
(DARIAH-DE working papers 18)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2016-4
URN: http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2016-4-0

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch

Abstract: Der Beitrag beschreibt den Forschungsstand zum Einsatz quantitativer Verfahren der Textanalyse in den Geisteswissenschaften. Dabei werden für zentrale Verfahren der stilistischen Analyse und thematischen Erschließung von Textbeständen jeweils die grundlegenden Konzepte und Intuitionen zu ihrer Wirkungsweise, ihre Anfänge und der heutige Entwicklungsstand beschrieben.

Keywords: Textanalyse, Quantitative Methoden, Autorenschaftsattribution, Distributionelle Semantik

Text Analysis, Quantitative Methods, Authorship Attribution, Distributional Semantics

Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften

Bericht über den Stand der Forschung

Sina Bock Keli Du Michael Huber Stefan Pernes
Steffen Pielström

Lehrstuhl für Computerphilologie und Neuere Deutsche Literaturgeschichte,
Julius-Maximilians-Universität Würzburg

DARIAH-DE
Working Papers

Sina Bock, Keli Du, Michael Huber, Stefan Pernes, Steffen Pielström: „Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften". *DARIAH-DE Working Papers* Nr. 18. Göttingen: DARIAH-DE, 2016. URN: urn:nbn:de:gbv:7-dariah-2016-4-0.

Die *DARIAH-DE Working Papers* werden von Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch herausgegeben.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Dieser Beitrag ist ursprünglich im Oktober 2015 als Report R 5.2.3 im Rahmen von DARIAH-DE (BMBF, Förderkennzeichen 01UG1110A-N) entstanden.

Zusammenfassung

Der Beitrag beschreibt den Forschungsstand zum Einsatz quantitativer Verfahren der Textanalyse in den Geisteswissenschaften. Dabei werden für zentrale Verfahren der stilistischen Analyse und thematischen Erschließung von Textbeständen jeweils die grundlegenden Konzepte und Intuitionen zu ihrer Wirkungsweise, ihre Anfänge und der heutige Entwicklungsstand beschrieben.

Schlagwörter

Textanalyse, Quantitative Methoden, Autorenschaftsattribution, Distributionelle Semantik

Keywords

Text Analysis, Quantitative Methods, Authorship Attribution, Distributional Semantics

Inhaltsverzeichnis

1 Quantitative Textanalyse in den Geisteswissenschaften

Die zunehmende digitale Verfügbarkeit von Textquellen für die philologische Forschung führt dazu, dass das zur Bearbeitung bestimmter Forschungsfragen idealerweise zu berücksichtigende Material sich oftmals durch seine schiere Menge der Möglichkeit einer intensiven Lektüre entzieht (Crane 2006). Neben dem im 20. Jahrhundert als literarisches Analyseverfahren etablierten sogenannten *Close Reading*, das eine detaillierte und vollständige Analyse der untersuchten Texte voraussetzt (Wenzel 2004), gewinnt damit das auf computergestützen Verfahren basierende *Distant Reading* immer mehr an Bedeutung (Moretti 2000). So können in einem digitalen Literaturkorpus anhand quantitativer Verfahren sowohl sprachlich-stilistische als auch thematisch-inhaltliche Aspekte untersucht werden. Der erste Abschnitt des Working Papers widmet sich dem Bereich der Stilometrie mit einem Fokus auf Methoden der Autorenschaftszuschreibung. Dabei werden grundlegende Konzepte wie das des Abstandsmaßes erklärt und geeignete maschinelle Lernverfahren besprochen. Im darauf folgenden zweiten Abschnitt wird auf den Bereich der automatischen inhaltlichen Erschließung eingegangen – ausgehend von der Modellierung einheitlicher semantischer Räume bis hin zur Berechnung darin enthaltener Themenkomplexe, besser bekannt als *Topic Modeling*.

2 Stilanalyse

Eine Eigenschaft der menschlichen Wahrnehmung ist es, die charakteristischen Merkmale einer Person oder einer Sache zu erfassen und einzuordnen – dabei ist Stil allgegenwärtig (Argamon, Burns und Dubnov 2010). So wird ein literarisches Werk nicht nur durch den individuellen Stil der Autorinnen und Autoren geprägt, sondern lässt sich auch mithilfe markanter Merkmale Gattungen und Epochen zuordnen. Methoden computergestützter Stilometrie erlauben es, stilistische Unterschiede zu quantifizieren und zu visualisieren. Somit lässt sich der Stil verschiedener Autorinnen und Autoren vergleichen, anonyme oder undatierte Texte können einer Autorin/einem Autor oder einer Epoche zugeordnet oder spezifische Eigenschaften innerhalb einer Gattung herausgestellt werden. Hierfür stehen heutzutage verschiedene, frei verfügbare *Software Tools*, wie z.B. *Voyant Tools* (Sinclair und Rockwell 2016), *Stylo* (Eder, Kestemont und Rybicki 2013) und *PyDelta* (Jannidis 2014) zur Verfügung. Klassische Methoden in diesem Bereich sind die *Principal Component Analysis* und die Quantifizierung stilistischer Unterschiede durch Textabstandsmaße, neuere, teilweise darauf aufbauende stilometrische Verfahren bedienen sich Techniken aus dem Bereich der *Clustering*-Verfahren und des überwachten maschinellen Lernens.

2.1 Principal Component Analysis

Will man die Unterschiedlichkeit zweier Texte modellieren, so kann man jeden Text als Datenpunkt in einem mehrdimensionalen Koordinatensystem betrachten. Die Achsen dieses Koordinatensystems repräsentieren messbare Eigenschaften der Texte, sogenannte **features**. In der Autorenschaftsattribution sind das meistens die relativen Häufigkeiten der am häufigsten verwendeten Wörter, also überwiegend von Funktionswörtern wie „und", „der" und „die". Je nach Fragestellung kann auch die Verwendung anderer *features* sinnvoll sein, z.B. die Häufigkeiten von Wortgruppen, grammatischen Konstruktionen

oder selteneren Inhaltswörtern. In jedem Fall geht die Zahl der berücksichtigten *features*, oft sind es die Frequenzen von mindestens 50-100 der häufigsten Wörter, über die drei visuell darstellbaren Dimensionen hinaus. Es gibt aber eine Reihe von Analysetechniken für derartige hochdimensionale Datensätze.

Eines der ersten Verfahren, das in der quantitatitven Textanalyse eingesetzt wurde, ist die *Principal Component Analysis* (PCA), die – lange vor der quantitativen Textanalyse – von Karl Pearson (Pearson 1901) und von Harold Hotelling (Hotelling 1933) entwickelt wurde. Ziel der PCA ist, in einem hochdimensionalen Datensatz eine Betrachtungsebene zu finden, in der sich möglichst viel von der Varianz der Daten visuell erfassen lässt.

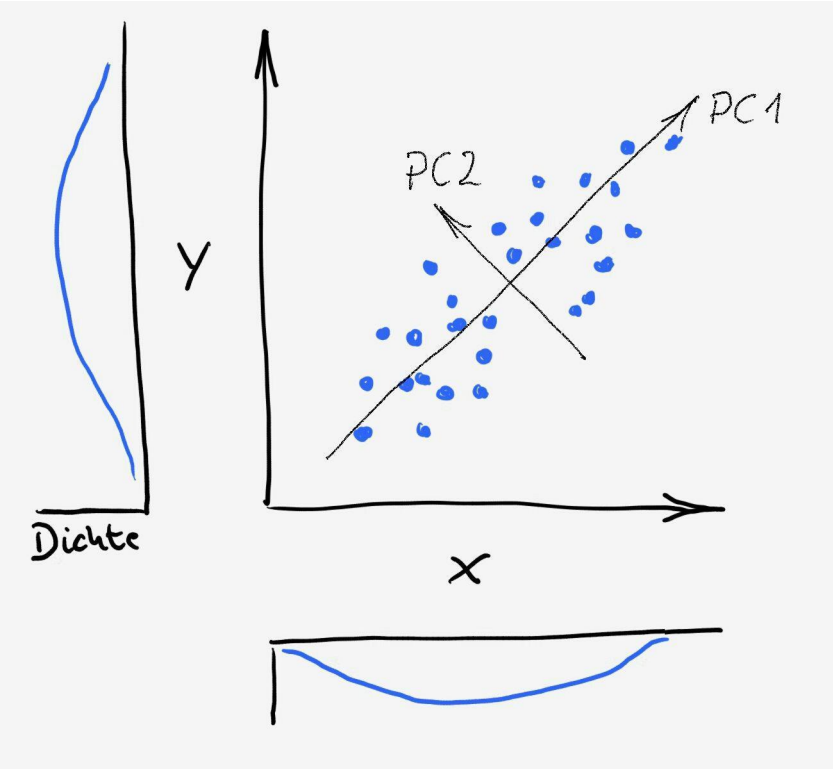

Abbildung 1: Vereinfachte Darstellung einer PCA auf nur zwei Dimensionen. Bei gleichzeitiger Betrachtung aller (zwei) Dimensionen sind hier zwei unterscheidbare Gruppen zu erkennen. Reduziert auf eine einzige Dimension, X oder Y, zeigt sich in den Daten aber keine bimodale Verteilung; die Gruppen lassen sich nicht mehr unterscheiden. Ebenso kann es in einem Datensatz mit 100 oder mehr Dimensionen schwierig werden, jene Dimensionen (oder Kombinationen von Dimensionen) auszumachen, in denen Unterschiede deutlich werden. Die Achsen der beiden Principal Components, die sich für diesen Datensatz berechnen lassen, sind hingegen an die Varianzverteilung der Datenpunkte angepasst. Aus: DARIAH-DE Report 5.2.3: Stand der Forschung in der Textanalyse (Baumgardt u. a. 2015, 5).

Hierfür werden die Dimensionen der Daten mit Hilfe der sogenannten **Singulärwertzerlegung** in ein neues Set von Variablen, die **Principal Components**, transformiert. Diese *Principal Components* kann man als Achsen eines alternativen Koordinatensystems verstehen, in dem die selben Datenpunkte in der selben Anordnung aufgetragen sind. Die erste Achse dieses neuen Bezugsystems (PC1) führt exakt durch die Datenpunkte in Richtung ihrer größten Ausdehnung, sie beschreibt also die größte Varianz

der Daten (Abb. 1), die weiteren Achsen (PC2 bis PCn) repräsentieren die anderen neuen, orthogonal zur PC1 verlaufenden Achsen in Reihenfolge der Varianz, die der Datensatz in diesen Dimensionen jeweils hat. Folglich kann diese Technik eingesetzt werden, um aus einem Datensatz mit beliebig vielen Dimensionen eine zweidimensionale Darstellung (mit PC1 und PC2 als X- bzw. Y-Achse) zu erzeugen, die exakt diejenige Betrachtungsbene zeigt, in der der größte Teil der Datenvarianz zu sehen ist und Unterschiede zwischen Gruppen vermutlich am besten herausgestellt werden (Abb. 2, Smith (2002)).

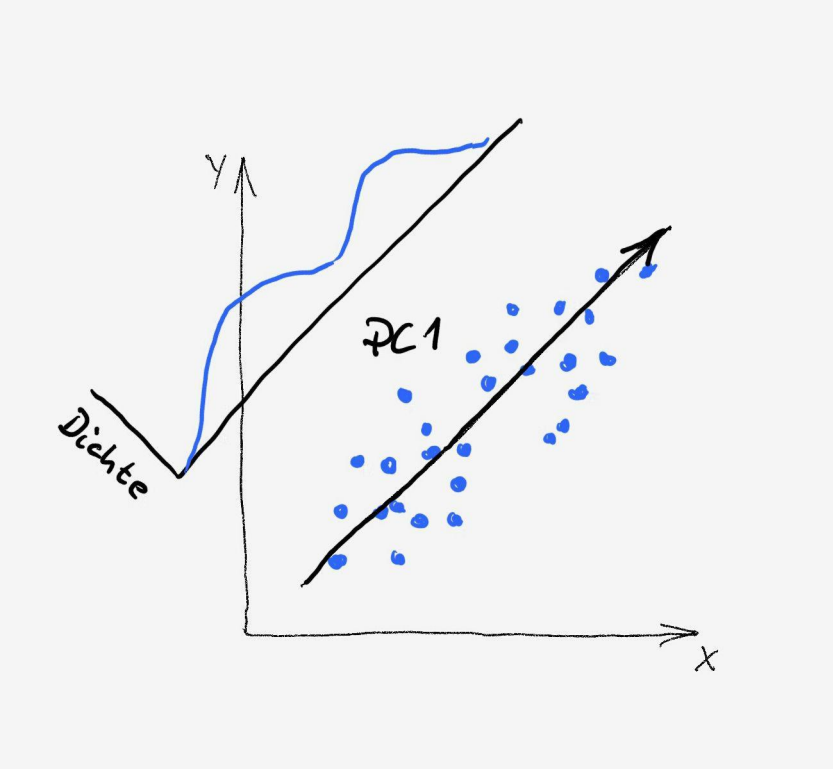

Abbildung 2: Entlang der neu berechneten Achse PC1 verläuft die Dichtekurve bimodal. Nun wird der Unterschied zwischen den beiden Gruppen schon in einer einzigen Dimension sichtbar. Aus: DARIAH-DE Report 5.2.3: Stand der Forschung in der Textanalyse (Baumgardt u. a. 2015, 6).

Dieses rechnerisch aufwändige Verfahren fand mit Aufkommen des Computers zunehmend mehr Berücksichtigung in unterschiedlichen Bereichen wie beispielsweise der Biologie, der Meterologie oder bei Bildkompressionsverfahren. Im Bereich der Textanalyse setzten (Mosteller und Wallace 1964) die Methode zur Untersuchung der *Federalist Papers* erstmals im Zusammenhang mit Autorschaftsattributionen ein.

Vor allem wenn es um die Zuordnung eines einzelnen Textes unbekannter Herkunft zu einem von zwei Autorinnen beziehungsweise Autoren geht, für die jeweils mehrere sicher zugeordnete Vergleichstexte vorliegen, ist die PCA oftmals gut geeignet, die stilistische Ähnlichkeit zu einer der beiden Textgruppen visuell herauszustellen (Burrows (1989), Binongo und Smith (1999), Binongo (2003)). Aber auch zur Analyse der zeitlichen Entwicklung von Schreibstilen (Brainerd 1980), oder der stilistischen Unterschiede zwischen Dialogen und narrativen Textpassagen (Burrows 1987a), kann die PCA eingesetzt werden.

2.2 Die Quantifizierung stilistischer Unterschiede

Die Analyse stilistischer Unterschiede lässt sich noch weiter operationalisieren, indem man diese auch tatsächlich quantifiziert. Die aus der PCA bekannte Form der Modellierung von Texten als Datenpunkte in einem hochdimensionalen Koordinatensystem bietet hierbei die Möglichkeit, Abstände zwischen diesen Punkten direkt zu berechnen und als Maß für die stilistische Verschiedenheit zweier Texte zu verwenden. Als **Textabstandsmaße** bieten sich grundsätzlich die sogenannte **Manhattan**-Distanz, d.h. die Summe der Abstände in den einzelnen Dimensionen [1], und die **Euklidische**-Distanz [2] an. Das erste Verfahren dieser Art, das in der Textanalyse erfolgreich war und bis heute in vielen Bereichen sehr erfolgreich eingesetzt wird, wurde von (Burrows 2002) vorgestellt.

Nachdem Burrows in seiner ersten Studie über die Figurenreden in Jane Austens Romanen mit statistischen Analysen erfolgreich hatte nachweisen können, dass die Sprechakte der Figuren sich systematisch unterscheiden (Burrows 1987b) untersuchte er diese Methode im Zusammenhang mit Fragen bei Autorenschaftsattributionen und Epochenzugehörigkeit. In so genannten „geschlossenen" Spielen, also bei einer kleinen Anzahl potentieller Kandidaten, die als Autorinnen und Autoren für einen anonymen Text in Frage kommen, können Autorenschaftsattributionen mit Hilfe der PCA sehr zuverlässig vorgenommen werden. Bei einem „offenen" Spiel, in dem die möglichen Autorinnen und Autoren kaum eingegrenzt werden können, reichen diese Methoden jedoch nicht mehr aus. In diesem Rahmen muss eine PCA für jede Kandidatin/jeden Kandidaten, der dem Spiel hinzugefügt wird, erneut durchgeführt werden. Burrows Methode zielt darauf ab mit Hilfe statistischer Merkmale/Variablen aus einem offenen Spiel ein geschlossenes Spiel zu machen. 2002 stellte er dann ein Verfahren vor, das er Delta nannte. In seiner Studie verwendet er ein Referenzkorpus mit Texten von 25 Poetinnen und Poeten des 17. Jahrhunderts. Zunächst wird dabei eine Liste der 150 häufigsten Wörter des Korpus erzeugt (hierbei handelt es sich meist um Funktionswörter wie Artikel, Hilfsverben, Personalpronomen, Präpositionen, Konjunktionen). Da die Frequenzen in Worthäufigkeitslisten sehr schnell abfallen, alle Wörter aber gleich gewichtet werden sollten, damit die Analysen nicht von den häufigsten Wörtern dominiert werden, **standardisierte** er die Wortfrequenzen [3] und summierte die Beträge der Wertdifferenzen aller Dimensionen. **Burrows Delta** ist somit also ein Maß, das ausdrückt wie sehr sich die standardisierten, relativen Wortfrequenzen zweier Texte voneinander unterscheiden. Bei einem Text unbekannter Herkunft erlaubt Burrows Delta zu berechnen, welchen anderen Texten dieser am ähnlichsten ist. Burrows konnte mit diesem Verfahren bei einer Textlänge von über 2000 Wörtern 19 von 20 Gedichten der richtigen Autorin beziehungsweise dem richtigen Autor zuordnen. Bei einer geringeren Textlänge befand sich der/die entsprechende Autor/in in 85% aller Fälle noch unter den ersten fünf Kandidaten, so dass mit Delta aus einem offenen Spiel ein geschlossenen Spiel gemacht werden konnte. Im Laufe der Zeit wurde Burrows Delta getestet und weiterentwickelt. So zeigte (Hoover 2004a) dass Burrows Delta für Autorschaftsattributionen besonders gute Ergebnisse erzielt, wenn die 150 (oder mehr) häufigsten Wörtern berücksichtigt werden. Seine Untersuchungen zeigen außerdem, dass sich die Ergebnisse weiter optimieren lassen, wenn Personalpronomina und Wörter mit einer Frequenz von über 70% nicht in die Analyse miteinbezogen werden. Eine Auflösung von Kontraktionen in den von ihm untersuchten englischsprachigen Texten führt jedoch

[1] „Manhattan"-Distanz in Anspielung auf die Strecke, die man in einem rechtwinkligen Straßennetz zurücklegen muss, um von einem Punkt zu einem anderen zu gelangen.

[2] Hierbei handelt es sich um die kürzeste Verbindunglinie zwischen zwei Punkten in einem mehrdimensionalen Raum.

[3] Ein Vorgang, der auch als z-Transformation bezeichnet wird.

meist zu schlechteren Ergebnissen. In Hoovers Referenzkorpus konnten unter Berücksichtigung der 100 bis 300 häufigsten Wörter 18 von 20 Autorinnen und Autoren treffend zugeordnet werden. (Hoover 2004b) entwickelte und testete auch mehrere alternative Varianten von Delta, ohne aber eine wesentliche Verbesserung erreichen zu können. (Argamon 2007) wies in einer umfassenden mathematischen Analyse des Delta-Verfahrens auf mehrere Annahmen hin, die diese Methode implizit vorraussetzt und zeigte damit eine Reihe von Unzulänglichkeiten auf. Zunächst stellte er heraus, daß bei der Berechnung Burrows Delta eine Standardisierung oder z-Transformation mit der Manhattan-Distanz kombiniert wird. Während erstere normalerweise für normalverteilte Daten eingesetzt wird, eignet sich letztere eher für Daten, die einer Laplace-Verteilung folgen. Argamon schlägt daher vor, vorrausgesetzt die Wortfrequenzen sind tatsächlich normalverteilt, statt der Manhattan-Distanz die Euklidische Distanz einzusetzen, ein Verfahren, das als **Argamons Delta** Eingang in einige gängige Stilometrie-Tools gefunden hat. Desweiteren weist Argamon darauf hin, dass die Wortfrequenzen eigentlich nicht alle voneinander unabhängig sind, d.h. viele Wörter miteinander korrelieren. Er schlägt vor, diese Korrelationen auf Basis einer Singulärwertzerlegung aus der *feature*-Matrix herauszurechnen, um so die statistische Unabhängigkeit der *features* herzustellen. Empirische Untersuchungen zeigen allerdings, daß Argamons Varianten der Delta-Methode in der Praxis keine Verbesserung der Erfolgsquote bei der Autorenschaftsattribution bringen (Jannidis u. a. 2015). Rybicki und Eder entwickelten eine Variante, die speziell an die Bedürfnisse stark flektierter Sprachen wie Polnisch und Latein angepasst ist (Rybicki und Eder 2011). Im Vergleich zu einer weitgehend unflektierten Sprache, wie dem Englischen, ist bei Sprachen mit größerer morphologischer Formenvielfalt zu erwarten, daß die relative Häufigkeit der häufigen Wörter insgesamt weniger groß ist. Beim sogenannten **Eders Delta** werden die *features* nach ihrem Rang in der Liste der häufigsten Wörter gewichtet, um diesen Unterschied zu kompensieren. Smith und Aldrige schlugen, in Anlehnung an etablierte Verfahren aus dem *Information Retrieval*, anstelle des bei Burrows verwendeten Manhattan-Distanzmaßes die Verwendung des Cosinus-Maßes vor. Die einzelnen, durch eine Reihe von numerischen *feature*-Werten repräsentierten Texte werden hierbei nicht als Datenpunkte in einem Koordinatensystem aufgefasst, sondern als Vektoren. Der Cosinus des Winkels zwischen diesen Vektoren dient hierbei als Maß für die Unterschiedlichkeit (Abb. 3). Empirische Tests konnten zeigen, dass **Cosinus Delta** für die Autorenschaftszuschreibung tatsächlich wesentlich bessere Ergebnisse ermöglicht als andere Varianten, und auch bei der Verwendung von mehr als 2000 Wörtern als *features* immer noch verlässlich ist, während die Performanz andere Maße in diesem Bereich beginnt wieder abzunehmen (Jannidis u. a. 2015). Ein wesentlicher Grund dafür liegt vermutlich darin, dass in diesem Abschnitt der Wortliste zunehmend Worte auftreten, die nur in einzelnen Texten in hoher Frequenz vorkommen. Solche einzelnen Worte können die Abstände zwischen Texten, die von der selben Autorin beziehungsweise dem selben Autor stammen, bei anderen Delta-Verfahren sehr groß werden lassen. Sie haben aber einen geringeren Effekt auf die Cosinus-Distanz, die Wirkung der Ausreißer wird hier in ähnlicher Weise gedämpft wie bei einer **Vektor-Normalisierung** (Jannidis u. a. 2015).

2.3 Clusteranalyse und überwachtes maschinelles Lernen

Bei der Untersuchung einer ganzen Gruppe von Texten kann die oben beschriebene Quantifizierung der Distanzen der einzelnen Texte zueinander die Grundlage für eine sogenannte Clusteranalyse bilden. Ziel der Clusteranalyse ist es, aus einer vollständigen quantifizierenden Beschreibung aller

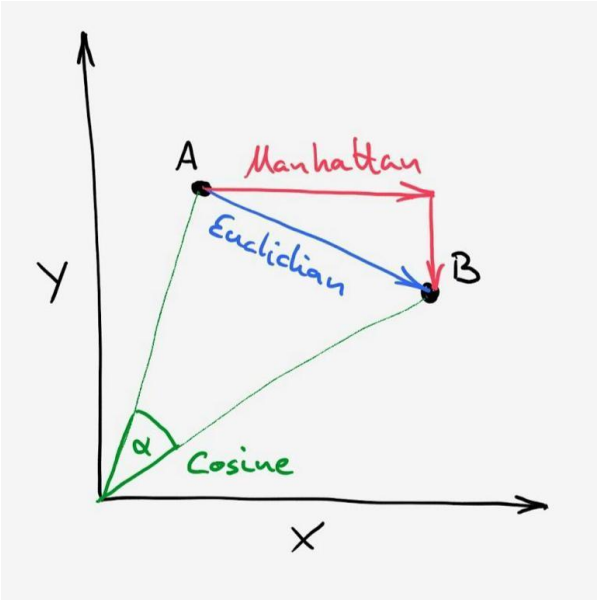

Abbildung 3: Der Abstand zweier Punkte A und B in einem Koordinatensystem: Manhattan-, Euklidische und Cosinus-Distanz. Aus: (Jannidis u. a. 2015).

Zweierbeziehungen in einer Gruppe von Objekten Untergruppen, sogennante **Cluster**, zu identifizieren, in denen die Elemente eines Clusters sich einander möglichst ähnlich sind, während sich die Elemente aus verschiedenen Clustern möglichst unähnlich sind (Heyer, Quasthoff und Wittig 2008). Solche Clusteringverfahren, die schon lange in den Naturwissenschaften eingesetzt werden, um Phänomene anhand messbarer Kriterien zu gruppieren (z.B. in der biologischen Taxonomie oder bei der Klassifikation von Lebensräumen), werden heutzutage auch zur Gruppierung von Texten eingesetzt und sind in den gängigen stilometrischen *Software Tools* integriert. Ein weit verbreitetes Vorgehen hierbei ist, auf Basis der Delta-Abstände zwischen den Texten ein hierarchisches Clusteringverfahren durchzuführen und das Ergebnis in einem sogenannten Baumdiagramm oder Dendrogramm zu visualisieren. Solche Grafiken können dann interpretiert werden: Stilistisch ähnliche Autorinnen und Autoren finden sich zum Beispiel auf benachbarten Ästen des Dendrogramms und ein Text unbekannter Herkunft sollte sich zwischen den anderen Texten des tatsächlichen Urhebers, oder zumindest in deren Nähe wiederfinden.

Für eine Klassifikationsaufgabe, wie die Zuschreibung eines einzelnen Textes unbekannter Urheberschaft zu einem von mehreren möglichen Autorinnen und Autoren, kann aber auch ein Klassifikationsalgorithmus eingesetzt werden. Hierfür steht eine Reihe von Techniken aus dem Bereich des überwachten maschinellen Lernens (ML) zur Verfügung. ML, unter anderem definiert als „…der Wissenserwerb eines künstlichen Systems." (Reichel 2008), kann als Begriff in diesem Zusammenhang recht irreführend sein. Einige der bereits weiter oben beschriebenen Analysetechniken werden heute in maschinellen Lernverfahren eingesetzt und darum oft dem ML zugerechnet. Sowohl die PCA als auch die verschiedenen Formen der Clusteranalyse werden in gängigen Lehrbüchern als **unüberwachte** maschinelle Lernverfahren aufgeführt. „Unüberwacht" darum, weil sie helfen sollen, Muster und Strukturen zu erkennen, ohne dass irgendeine Form von Vorwissen bei der Analyse berücksichtigt wird: Die Autorinnen und Autoren der Texte bekannter Urheberschaft spielen bei der Durchführung von PCA und Clusteranalyse keine Rolle,

lediglich bei der anschließenden Interpretation. Im Gegensatz dazu nutzen **überwachte** Verfahren vorhandene Metadaten. Sie werden vor allem für Klassifikationsaufgaben eingesetzt und benötigen stets ein sogenanntes Test-Set, d.h. ein Set, anhand dessen sie Eigenschaften der einzelnen Klassen erlernen und neue Texte diesen Klassen zuordnen können. Verfahren des ML werden in vielen alltäglichen Bereichen eingesetzt wie beispielsweise beim Filtern von Spam-Nachrichten. Hier wird ein Algorithmus regelbasiert, also mittels bestimmter Schlüsselwörter und/oder bereits von der Benutzerin/dem Benutzer als Spam eingestufter E-Mail-Adressen so trainiert, dass er auch künftige unerwünschte E-Mail-Nachrichten mit hoher Wahrscheinlichkeit erkennen und aussortieren kann. Ein weiteres Anwendungsgebiet des ML ist der Bereich der Spracherkennung. Auch für die Klassifikation literarischer Texte stellt z.B. die Software *Stylo* (Eder, Kestemont und Rybicki 2013) eine Reihe etablierter Lernalgorithmen zur Verfügung, darunter die *Naive-Bayes*-Klassifikation, der *k-Nearest-Neighbour*-Algorithmus und *Support Vector Machines*. Als zuverlässigster Algorithmus für die Autorenschaftsattribution hat sich in empirischen Untersuchungen das ebenfalls in *Stylo* implementierte *Nearest-Shrunken-Centroids*-Verfahren erwiesen (Eder 2015).

3 Inhaltsanalyse

Aufgrund technischer Voraussetzungen (zu diesen zählen etwa geeignete statistische Modelle, performante Soft- und Hardware sowie ausreichend große Datenbestände) beginnt die automatische inhaltliche Erschließung von Textbeständen erst mit Ende des 20. Jahrhunderts eine Reife zu erlangen, die eine breite Anwendung in unterschiedlichen Forschungsbereichen möglich macht. Im Kontext philologischer Fragestellungen können solche Verfahren neben einer Strukturerschließung und der verbesserten Auffindbarkeit von Inhalten auch dazu beitragen, theoretische Konstrukte auf einer Makroebene – z.B. literarischer Gattungen oder Epochen – zu modellieren und damit ein genaueres Verständnis der untersuchten Phänomene zu erlangen. Für die Verfahren, die im Folgenden vorgestellt werden, gilt, dass nicht auf zuvor manuell definierte Wissensbestände oder Annotationen in den Texten zurückgegriffen werden muss, sondern der Ablauf *ungesteuert* und anhand von *unstrukturiertem Text* vonstatten gehen kann.

3.1 Key Words in Context

Key Words in Context (KWIC), bekannt vor allem aus der Korpuslinguistik, gilt als ein klassisches Werkzeug, das es ermöglicht, größere Textmengen im Hinblick auf bestimmte Schlüsselworte zu erschließen. KWICs werden genutzt, um eine Abfolge von Zeichen, meist Wörter in ihrem syntaktischen Zusammenhang oder dem Kontext ihrer Äußerung darzustellen, wobei die gesuchte Zeichenfolge zentriert untereinander dargestellt wird. Solche Abfolgen von Zeichen beziehungsweise Wörtern werden auch als *N-Gramme* bezeichnet.

Das Konzept der N-Gramme ist maßgeblich durch den Google N-Gram Viewer (Michel u. a. 2011) bekannt. Hierbei wird eine Suchanfrage in Form eines Wortes oder einer Wortfolge an das Google-Books-Korpus gestellt. Das Ergebnis ist eine Kurve, welche die Häufigkeit des Auftretens der N-Gramme in ihrem chronologischen Verlauf zeigt. Bei der Erschließung von Textkorpora anhand von N-Grammen

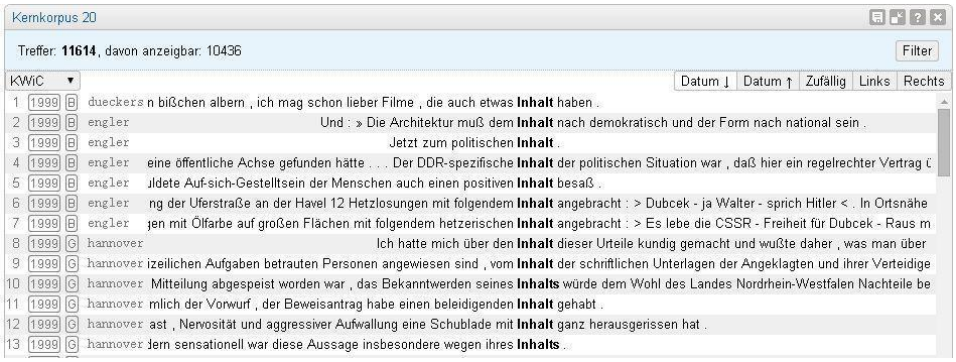

Abbildung 4: Key Words in Context *im Digitalen Wörterbuch der deutschen Sprache (Klein und Geyken 2010).*

werden häufig auftretende Muster innerhalb der Satzgrenze zu Tage gefördert, es besteht jedoch auch die Möglichkeit, inhaltliche Muster auf der Ebene ganzer Texte bzw. Textteile zu erheben.

3.2 Ansätze zur Modellierung thematischer Felder

Topic Modeling bezeichnet eine Gruppe von Verfahren, die es ermöglichen, anhand einer statistischen Analyse des lexikalischen Inventars Rückschlüsse auf die zugrunde liegende thematische Struktur einer Sammlung von Texten zu ziehen (Blei 2012). Ausgangspunkt ist meist eine *Term-Document* Matrix – jede Zeile in dieser Matrix steht für ein Wort bzw. eine Wortform und jede Spalte steht für eine zusammenhängende Textpassage, wie z.B. ein Absatz, Kapitel oder ein Dokument. Die Zellen enthalten die Worthäufigkeit für jedes Dokument. Eine solche Darstellung wird auch als **bag-of-words** Modell bezeichnet – eine Vereinfachung, die nicht Syntax und Wortfolge, sondern nur die Häufigkeit des Auftretens von Wörtern abbildet. Für die Bearbeitung einiger korpusanalytischer Fragestellungen ist diese Matrix bereits ausreichend. Um jedoch anhand von statistischen Methoden die Semantik, also den Bedeutungsgehalt eines Textes, besser erschließen zu können, wurde eine Reihe von Transformationen entwickelt, die es ermöglichen, Aussagen über die inhaltlichen Zusammenhänge von Wörtern und Dokumenten zu treffen. Da es sich dabei um einen rein statistischen Ansatz handelt, funktionieren die Verfahren sprachunabhängig und können sogar eingesetzt werden, um thematische Strukturen in mehrsprachigen Korpora zu verfolgen. Ein wichtiger Einsatzbereich ist auch die Klassifikation und das Auffinden verwandter Dokumente (*Information Retrieval*). Desweiteren wurden Parallelen zum menschlichen Konzepterwerb festgestellt. Zurzeit ist *Latent Dirichlet Allocation* (LDA) ein weit verbreitetes Verfahren im Bereich des Topic Modeling. Als bekannte Vorgänger der LDA sollen im Folgenden auch zwei weitere Methoden skizziert werden: *Latent Semantic Analysis* (LSA) und *Probabilistic Latent Semantic Analysis* (pLSA). Synonym für beide wird auch die Bezeichnung *Latent Semantic Indexing* (LSI) verwendet. Beide Methoden zielen darauf ab, inhaltlich zusammenhängende Worte zu gruppieren und dabei auf die Bedeutung dieser Gruppierungen (beziehungsweise Dokumente, Passagen,...) zu schließen.

3.3 Latent Semantic Analysis

Ausgehend von einem ausreichend großen Korpus kann mit Hilfe der **Latent Semantic Analysis** (Landauer und Dumais 2008) ein *semantischer Raum* berechnet werden, der Aussagen darüber zulässt, wie ähnlich die im Korpus enthaltenen Worte und Dokumente sind. Grundlage ist eine *Term-Document* Matrix und – wie für einige andere Verfahren auch (siehe: PCA) – die Singulärwertzerlegung (*Singular Value Decomposition*, SVD) der Matrix. Die SVD ermöglicht es, eine wechselseitige Einschränkung zu berücksichtigen, nämlich die Berechnung der Wortbedeutungen als durchschnittlicher Effekt auf die Bedeutung von Dokumenten und umgekehrt, die Bedeutung von Dokumenten als durchschnittlicher Effekt auf die Bedeutung von Wörtern. Das anschließende Ergebnis dieser Operation sind zwei Matrizen, die Informationen zu den im Modell enthaltenen Wörtern, respektive Dokumenten, enthalten sowie eine dritte Matrix, die den *Eigenwerten* der Ursprungsmatrix entspricht. Im Fall der LSA gilt weiters, dass eine reduzierende Form der SVD angewandt wird, die eine bestmögliche *k*-dimensionale Approximation an die Ursprungsmatrix darstellt. Im Anschluss ist es aufgrund der einheitlichen Vektorform der berechneten Wort- und Dokument-Matrizen möglich, die Ähnlichkeiten von Wort-Wort, Dokument-Dokument, und Wort-Dokument, anhand eines Abstandsmaßes für Vektoren, wie z.B. der *Kosinus-Ähnlichkeit*, zu berechnen. Landauer und Dumais weisen darauf hin, dass die relative Nähe der Worte zueinander, die aus dem Modell abgelesen werden kann, nicht einer einfachen *Kookkurrenz* der Worte entsprechen, sondern einer „Ähnlichkeit der Effekte, die jene Worte auf die Textpassagen haben, in denen sie vorkommen" (Landauer und Dumais 2008).

3.4 Probabilistic Latent Semantic Analysis

Obwohl mit LSA erstmals ein Verfahren zur Berechnung von sogenannten *latenten semantischen Räumen* entwickelt wurde, handelt es sich dabei noch nicht um ein Topic Model im engeren Sinn. Das Konzept des *Topics* wird in der Regel als *latente Variable* umgesetzt und repräsentiert dabei ein Teilvokabular des Korpus, aus dem sich Dokumente zusammensetzen können. Im Gegensatz zu Wörtern und Dokumenten handelt sich dabei um eine unbeobachtbare Größe, die erst durch die Berechnung des Modells entsteht. LSA kommt ohne solche latenten Variablen aus, was in diesem Fall bedeutet, dass sich jedes Dokument aus einem einzigen Vokabular – nämlich auf Grundlage des gesamten Korpus – zusammensetzt.

Im Rahmen einer Erweiterung des Verfahrens, nämlich der **Probabilistic Latent Semantic Analysis** (pLSA) (Hoffmann 1999), wird es möglich, Dokumente als Zusammensetzung mehrerer Topics zu beschreiben. Dazu wird für jedes Dokument eine Wahrscheinlichkeitsverteilung über eine vorher festgelegte Anzahl von Topics (latente Variable) angenommen. Anhand dieser Verteilung werden schließlich die Worte eines Dokuments „gezogen". Es handelt sich dabei um eine realistischere Annäherung an die inhaltliche Zusammensetzung eines Korpus beziehungsweise eines Textes. Ein weiterer Unterschied ist, dass keine Heuristik („die ersten *k* Singulärwerte") zur Dekomposition der Matrix eingesetzt, sondern anhand einer *Expectation-Maximization* eine optimale Approximation an die Ursprungsdaten erreicht wird. Obwohl das Verfahren dadurch robuster gegen *Overfitting* wird (eine übermäßig genaue Modellierung der Trainingsdaten, die sich nicht ausreichend auf neue, bisher unbekannte Fälle generalisieren lässt), können die damit generierten Modelle nicht ohne Qualitätsverlust um neue Dokumente erweitert werden. Gerade bei großen Korpora kann das rechnerisch aufwändig und unpraktikabel sein. Es handelt

sich dabei um eine Eigenschaft, die in erster Linie bei *Document Classification* und *Retrieval*-Aufgaben zum Nachteil gerät – ein Nachteil der jedoch einen wesentlichen Grund für die Entwicklung der **Latent Dirichlet Allocation** (LDA) darstellt (Blei 2012).

3.5 Latent Dirichlet Allocation

Vom Grundprinzip her unterscheiden sich pLSA und LDA als *probabilistische Topic Models* nicht wesentlich voneinander – in beiden Fällen wird von beobachtbaren Variablen – den Wörtern und Dokumenten – auf die nicht-beobachtbare, latente Variable der Topics geschlossen. Folgende Punkte fassen die Beziehung zwischen Topics, Dokumenten und Wörtern noch einmal zusammen:

- Ein **Dokument** setzt sich aus mehreren Topics in jeweils unterschiedlicher Gewichtung zusammen.
- Ein **Topic** ist eine Wahrscheinlichkeitsverteilung über das gesamte Vokabular des Korpus. Die Wahrscheinlichkeit, dass ein Wort zu dem Thema gehört, ist unterschiedlich. Zum Beispiel gehört das Wort „Sprache" mit hoher Wahrscheinlichkeit zu dem Thema „Sprachwissenschaft". Im Vergleich dazu gehört das Wort „Brezel" mit geringer Wahrscheinlichkeit zu dem gleichen Thema. Diese „kleinere Wahrscheinlichkeit" kann sehr nah an 0% liegen, aber sie ist keinesfalls gleich 0%.
- Ein **Wort** kann mit hohen Wahrscheinlichkeiten mehreren Topics angehören. Zum Beispiel ist das Wort „Spiel" ein wichtiger Begriff nicht nur für das Thema „Sport", sondern auch für das Thema „Theater".

Probabilistische Topic Models können am einfachsten als generativer Prozess beschrieben werden. Es wird ein zufälliger Prozess angenommen, der zur Entstehung der Dokumente geführt hat: Zu Beginn wird eine zufällige Zusammensetzung von Topics für das gesamte Korpus generiert. Anschließend wird für jedes Wort in jedem Dokument zufällig ein Topic aus der Topic-Verteilung ausgewählt. Auf Grundlage dieses Topics wird schließlich zufällig ein Wort aus dem Gesamtvokabular des Korpus ausgewählt. So werden für jedes Dokument die Worte in einem zweistufigen Prozess generiert.

In der praktischen Anwendung eines Topic Models auf ein bestehendes Korpus handelt es sich jedoch nicht um einen generativen Prozess, sondern um eine Umkehrung des eben beschriebenen Vorgangs. Während die Topics und ihre Verteilung für den generativen Prozess den Ausgangspunkt darstellen, sind sie in der Anwendung auf ein bestehendes Korpus die unbekannte, latente Variable. In diesem Fall sind die Dokumente bekannt und beobachtbar, die Topics, ihre Verteilungen für jedes Dokument sowie die Zuweisung zu den Wörtern sind jedoch unbekannt und müssen aus den beobachtbaren Informationen gewonnen werden. Die Frage lautet also: Wie sieht die versteckte Topic-Struktur aus, die am wahrscheinlichsten das zu untersuchende Korpus erzeugt hat? Dieser Vorgang beschreibt die zentrale technische Problemstellung probabilistischer Topic Models, denn hierbei ist es notwendig, alle denkbaren Topic-Zusammensetzungen zu berücksichtigen, um zu einem möglichst optimalen Ergebnis zu gelangen (auch: *Parameter Estimation*). Hierin besteht die mathematische Komplexität solcher Topic Models und gleichzeitig auch der Vorteil von LDA gegenüber pLSA: Da die Anzahl aller möglichen Topic-Zusammensetzungen exponentiell groß und nicht direkt berechenbar ist (da sich die Topic Struktur auch auf jedes einzelne Wort auswirkt), muss der Vorgang approximiert werden. Solche aufwändigen Berechnungen möglichst effizient zu gestalten ist ein aktiver Forschungsbereich, wobei sich besonders

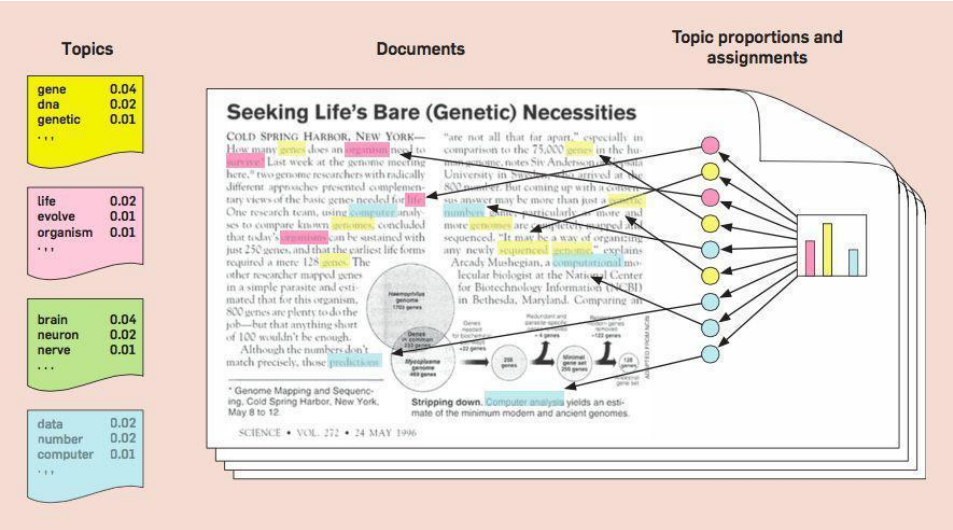

Abbildung 5: Intuition zu Topic Modeling: Aus der Gesamtmenge von Topics (links) wird für jedes Dokument eine Topic-Zusammensetzung gewählt (Histogramm-Skizze rechts). Anschließend wird für jedes Wort eine Topic-Zuordnung getroffen (farbige Kreise) und anhand dieses Topics ein Wort gewählt (Blei 2012, 78).

Ansätze der Bayes'schen Statistik dafür zu eignen scheinen. Die für LDA am häufigsten eingesetzte Methode ist das *Gibbs Sampling*, ein Verfahren der statistischen Physik, das in der Lage ist, sich über ein ausgeklügeltes Ziehen von Stichproben an hochdimensionale, multivariate Verteilungen anzunähern und dabei auch auf unbeobachtbare Variablen, die in dem multivariaten Modell enthalten sind, zu schließen (Blei 2012).

3.6 Topic Modeling Varianten und Erweiterungen

Der Anwendungsbereich von Topic Modeling Verfahren liegt vor allem in der Analyse von großen Korpora, wobei es älteren, weniger spezialisierten statistischen Methoden überlegen ist (Evans 2014). Topic Modeling ermöglicht einen Überblick über die im Korpus auftretenden Themenfelder und kann somit also eine Art Inhaltsanalyse für große Korpora leisten. Darüber hinaus können Modelle auch für struktuiertere Aufgaben, beispielsweise der Evaluation von Forschungsergebnissen, erstellt werden (Heuser und Le-Khac 2012). Ein weiterer Anwendungsbereich ist das *Clustering* von Texten. Hierbei werden die produzierten Themen für eine Klassifizierung von Texten benutzt, wie zum Beispiel Genre oder Epoche (Underwood und Goldstone 2012). Zum Forschungsstand im Bereich des Topic Modeling hat David Mimno eine ausführliche Literaturliste erstellt (Mimno). Die Liste enthält über 100 Literaturangaben, angeordnet in 15 Teilbereiche. Im Hinblick auf eine Anwendung in der Literaturwissenschaft und anderen Philologien sind inbesondere die folgenden zu nennen:

- *Bibliometrics*: Topic Modeling wird z.B. dafür eingesetzt, einflussreiche Texte im Korpus zu identifizieren (Gerrish und Blei 2010).
- *Cross-language*: Anwendung von Topic Modeling für die Erzeugung der multilingualen Topics (Jagarlamudi und Daumé 2010).

- *Evaluation*: Methoden zur Evaluation von Topic Models (Wallach u. a. 2009).
- *NLP*: Eine erweiterte Form von LDA kann auch für Natural Language Processing Aufgaben angewendet werden, wie z.B. für Part-of-Speech Tagging (Toutanova und Johnson 2007).
- *Networks*: Topic Models kann auch für die Klassifizierung von Netzwerken angewendet werden. Zum Beispiel die Entdeckung von Gruppen von Entitäten und ihre Attributten in einem Entity-Relationship Modell (Wang, Mohanty und McCallum 2005).
- *Non-parametric*: Erweiterungen, die in der Lage sind, optimale Parameter selbst zu finden. Dabei entstehen hierarchische Topics, ähnlich wie bei einem hierarchischen Clustering (Blei u. a. 2003).

Literatur

Argamon, Shlomo. 2007. Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing* 23, Nr. 2: 131—147.

Argamon, Shlomo, Kevin Burns und Shlomo Dubnov. 2010. *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*. Springer.

Baumgardt, Frederick, Sina Bock, Keli Du, Philip Dürholt, Michael Huber, Matt Munson, Stefan Pernes, Steffen Pielström und Michael Sünkel. 2015. Report 5.2.3: Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften: Bericht über den Stand der Forschung. DARIAH-DE.

Binongo, José. 2003. Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. *Chance* 16, Nr. 2: 9—17.

Binongo, José und M. W. A. Smith. 1999. The application of principal component analysis to stylometry. *Literary and Linguistic Computing* 14, Nr. 4: 445—466.

Blei, David M. 2012. Probabilistic topic models. *Communications of the ACM* 55, Nr. 4: 77–84.

Blei, David M., Thomas L. Griffiths, Michael I. Jordan und Joshua B. Tenenbaum. 2003. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In: *Advances in Neural Information Processing Systems*, 17–24.

Brainerd, Barron. 1980. The Chronology of Shakespeare's Plays: A Statistical Study. *Computers and the Humanities* 14, Nr. 4: 221–230.

Burrows, John F. 1987a. Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing* 2, Nr. 2: 61–70.

Burrows, John F. 1987b. *Computation into criticism: a study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

Burrows, John F. 1989. „An ocean where each kind. . .": Statistical analysis and some major determinants of literary style. *Computers and the Humanities* 23, Nr. 4-5: 309–321.

Burrows, John F. 2002. „Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary*

and Linguistic Computing 17, Nr. 3: 267–287.

Crane, Gregory R. 2006. What Do You Do with a Million Books? *D-Lib Magazine* 12, Nr. 3: 1–7.

Eder, Maciej. 2015. Taking stylometry to the limits: benchmark study on 5,281 texts from „Patrologia Latina". In: *Digital Humanities 2015 Conference Abstracts*.

Eder, Maciej, Mike Kestemont und Jan Rybicki. 2013. Stylometry with R: a suite of tools. In: *Digital Humanities 2013: Conference Abstracts*, 487–489.

Evans, Michael S. 2014. A computational approach to qualitative analysis in large textual datasets. *PloS ONE* 9, Nr. 2: 1–10.

Gerrish, Sean M. und David M. Blei. 2010. A language-based approach to measuring scholarly impact. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 375–382.

Heuser, Ryan und Long Le-Khac. 2012. A quantitative literary history of 2,958 nineteenth-century British novels: The semantic cohort method. *Stanford Literary Lab Pamphlets* 4: 1–66.

Heyer, Gerhard, Uwe Quasthoff und Thomas Wittig. 2008. Text mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. Bochum: W3LVerlag.

Hoffmann, Thomas. 1999. Probabilistic Latent Semantic Analysis. In: *Proceedings of Uncertainty in Artificial Intelligence, UAI'99*, 289–296.

Hoover, David L. 2004a. Delta Prime? *Literary and Linguistic Computing* 19, Nr. 4: 477–495.

Hoover, David L. 2004b. Testing Burrows's Delta. *Literary and Linguistic Computing* 19, Nr. 4: 453–475.

Hotelling, Harold. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24, Nr. 6: 417–441.

Jagarlamudi, Jagadeesh und Hal Daumé. 2010. Extracting Multilingual Topics from Unaligned Comparable Corpora. In: *Proceedings of the European Conference on Information Retrieval*, 444–456.

Jannidis, Fotis. 2014. The Pydelta Package. http://github.com/fotis007/pydelta.

Jannidis, Fotis, Stefan Evert, Thomas Proisl, Steffen Pielström, Christof Schöch und Thorsten Vitt. 2015. Towards a better understanding of Burrows's Delta in literary authorship attribution. In: *Proceedings of NAACL-HLT Fourth Workshop on Computational Linguistics for Literature*, 79–88.

Klein, Wolfgang und Alexander Geyken. 2010. Das digitale Wörterbuch der deutschen Sprache (DWDS). *Lexicographica* 26: 79–93.

Landauer, Thomas K. und Susan Dumais. 2008. Latent Semantic Analysis. *Scholarpedia* 3, Nr. 11: 4356.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, u. a. 2011. Quantitative analysis of culture using millions of digitized books. *Science*

331, Nr. 6014: 176–182.

Mimno, David. Topic Modeling Bibliography. http://mimno.infosci.cornell.edu/topics.html.

Moretti, Franco. 2000. Conjectures on world literature. *New Left Review* 1: 54–68.

Mosteller, Frederick und David Wallace. 1964. *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Pearson, Karl. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, Nr. 11: 559–572.

Reichel, Uwe. 2008. Maschinelles Lernen I. Course Materials, LMU München. http://www.phonetik.uni-muenchen.de/~reichelu/kurse/machine_learning/machine_learning_1.pdf.

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.

Sinclair, Stéfan und Geoffrey Rockwell. 2016. Voyant Tools. http://voyant-tools.org.

Smith, Lindsay I. 2002. A tutorial on principial component analysis. http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf.

Toutanova, Kristina und Mark Johnson. 2007. A Bayesian LDA-based model for semi-supervised partofspeech tagging. In: *Advances in Neural Information Processing Systems*, 1521–1528.

Underwood, Ted und Andrew Goldstone. 2012. What can topic models of PMLA teach us about the history of literary scholarship? http://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship.

Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov und David Mimno. 2009. Evaluation Methods for Topic Models. In: *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112.

Wang, Xuerui, Natasha Mohanty und Andrew McCallum. 2005. Group and Topic Discovery from Relations and their Attributes. In: *Advances in Neural Information Processing Systems*, 1449–1456.

Wenzel, Peter. 2004. New Criticism. In: *Grundbegriffe der Literaturtheorie*, hg. von A. Nünning. Stuttgart und Weimar: Metzler.