ZfdG

Zeitschrift für digitale Geisteswissenschaften

Artikel aus:
Zeitschrift für digitale Geisteswissenschaften

»Delta« in der stilometrischen Autorschaftsattribution

Autor/in:
Andreas Büttner

Kontakt: andreas.buettner@uni-wuerzburg.de
Institution: Universität Würzburg, Institut für Philosophie
GND: 1139662686

Autor/in:
Friedrich Michael Dimpel

Kontakt: mail@dimpel.de
Institution: FAU Erlangen-Nürnberg, Department Germanistik und Komparatistik
GND: 1111656460

Autor/in:
Stefan Evert

Kontakt: stefan.evert@fau.de
Institution: FAU Erlangen-Nürnberg, Department Germanistik und Komparatistik
GND: 1139663577

Autor/in:
Fotis Jannidis

Kontakt: fotis.jannidis@uni-wuerzburg.de
Institution: Universität Würzburg, Institut für Deutsche Philologie
GND: 114523525 ORCID: 0000-0001-6944-6113

Autor/in:
Steffen Pielström

Kontakt: pielstroem@biozentrum.uni-wuerzburg.de
Institution: Universität Würzburg, Institut für Deutsche Philologie
GND: 1038678129

Autor/in:
Thomas Proisl

Kontakt: thomas.proisl@fau.de
Institution: FAU Erlangen-Nürnberg, Department Germanistik und Komparatistik
GND: 1139667327

Autor/in:
Isabella Reger

Kontakt: isabella.reger@uni-wuerzburg.de
Institution: Universität Würzburg, Institut für Deutsche Philologie
GND: 1139666193

Autor/in:
Christof Schöch

Kontakt: schoech@uni-trier.de
Institution: Universität Trier, Fachbereich II / Computerlinguistik und Digital Humanities
GND: 135594480 ORCID: 0000-0002-4557-2753

Autor/in:
Thorsten Vitt

Kontakt: thorsten.vitt@uni-wuerzburg.de
Institution: Universität Würzburg, Institut für Deutsche Philologie
GND: 1139665731

DOI des Artikels:
10.17175/2017_006

Nachweis im OPAC der Herzog August Bibliothek:
897375815

Erstveröffentlichung:
20.12.2017

Letzte Überprüfung aller Verweise:
18.12.2017

GND-Verschlagwortung:
Autorschaft | Literaturwissenschaft | Literarischer Stil |

Zitierweise:
Andreas Büttner, Friedrich Michael Dimpel, Stefan Evert, Fotis Jannidis, Steffen Pielström, Thomas Proisl, Isabella Reger: »Delta« in der stilometrischen Autorschaftsattribution. In: Zeitschrift für digitale Geisteswissenschaften. 2017. PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/2017_006.

Andreas Büttner, Friedrich Michael Dimpel, Stefan Evert, Fotis Jannidis, Steffen Pielström, Thomas Proisl, Isabella Reger

»Delta« in der stilometrischen Autorschaftsattribution

Abstracts

Der Artikel stellt aktuelle stilometrische Studien im Delta-Kontext vor. (1) Diskutiert wird, warum die Verwendung des Kosinus-Abstands zu einer Verbesserung der Erfolgsquote führt; durch Experimente zur Vektornormalisierung gelingt es, die Funktionsweise von Delta besser zu verstehen. (2) Anhand von mittelhochdeutschen Texten wird gezeigt, dass auch metrische Eigenschaften zur Autorschaftsattribution eingesetzt werden können. Zudem wird untersucht, inwieweit die mittelalterliche, nicht-normierte Schreibung die Erfolgsquote von Delta beeinflusst. (3) Am Beispiel von arabisch-lateinischen Übersetzungen wird geprüft, inwieweit eine selektive Merkmalseliminierung dazu beitragen kann, das Übersetzersignal vom Genresignal zu isolieren.

In this article, we present current stylometric studies on Delta. (1) We discuss why the use of cosine similarity improves the rate of success; our experiments on vector normalization lead to a better understanding of how Delta works. (2) Based on a corpus of Middle High German texts, we show that metrical properties can also be used for authorship attribution. The degree to which Delta is influenced by non-normalized medieval spellings is also investigated. (3) Using a corpus of Arabic-Latin translations, we explore how selective feature elimination can be used to separate the translator signal from the genre signal.

1. Einleitung

Wie ermittelt man, von welchem Autor ein anonym oder unter Pseudonym erschienenes literarisches Werk stammt? Mit Hilfe von stilometrischen Verfahren ist es möglich, Antworten zu Autorschaftsfragen zu formulieren, wenn ausreichend viele und ausreichend lange Vergleichstexte vorliegen. Überblicksbeiträge von Patrick Juola und Efstathios Stamatatos belegen die Vielfältigkeit der Bestrebungen, stilometrische Verfahren für die Autorschaftsattribution einzusetzen und weiterzuentwickeln.[1] Die Stilometrie hat eine lange Tradition in den digitalen Geisteswissenschaften: Mit der Analyse der *Federalist Papers* durch Mosteller und Wallace konnten schon Anfang der 1960er Jahre Erfolge verzeichnet werden.[2] Die frühen Versuche von Wickmann, einen klärenden Beitrag zur Verfasserfrage der *Nachtwachen von Bonaventura* zu leisten,[3] wurden hingegen durch das Auffinden einer handschriftlichen Notiz von Klingemann in Frage gestellt.[4]

In den letzten Jahren standen neben Beiträgen zu theoretischen Fragestellungen weniger Attributionsversuche bei konkreten Autorschaftsfragen im Zentrum der stilometrischen

[1] Juola 2006; Stamatatos 2009.
[2] Mosteller / Wallace 1963.
[3] Wickmann 1969; Wickmann 1974.
[4] Haag 1987. Die *Nachtwachen*-Ausgabe von Jost Schillemeit aus dem Jahr 2012 nennt nunmehr Klingemann als Autor (Klingemann 2012).

Forschung, sondern eher eine Kalibrierung der Werkzeuge: In Validierungsstudien wurde geprüft, mit welchen digitalen Techniken es in Experimenten mit Texten bekannter Autorschaft am besten gelingt, eine hohe Anzahl an Werken dem richtigen Autor zuzuordnen. Während Wickmann 1989 Probleme mit der Verfügbarkeit von Textmaterial konstatierte – »Die Texte sind meist zu kurz, und zum Vergleich stehen meist zu wenige Texte zur Verfügung«[5] –, hat es mittlerweile bei der Verfügbarkeit digitaler Texte ebenso sehr Fortschritte gegeben wie bei der Entwicklung der stilometrischen Techniken: Der unter dem Pseudonym Robert Galbraith erschienene Kriminalroman *The Cuckoo's Calling* konnte via Stilometrie Joanne K. Rowling zugeordnet werden; Rowling hat sich daraufhin zu ihrer Autorschaft bekannt.[6]

Während sich heute eine Autorschaft wie die von Rowling auf die Verkaufszahlen auswirkt, geht es bei älteren Werken auch darum, in welchem geistesgeschichtlichen und kulturhistorischen Kontext ein Werk wahrgenommen wird. Autorschaft kann an Spuren im Text dingfest gemacht werden, und das hat auch Rückwirkungen auf die Diskussion um den Tod bzw. die Rückkehr des Autors:[7] Wäre ein Autortext tatsächlich nur eine Ansammlung von verschiedenen Schreibweisen, »von denen keine einzige originell ist« und nur »ein Gewebe von Zitaten«,[8] dürfte es schwerfallen, die Erkennungsquoten von aktuellen stilometrischen Techniken zu erklären. Bei vormodernen Texten wurde unter dem Stichwort *New Philology* die Relevanz von Autorschaft angesichts der Varianz der mittelalterlichen Überlieferungslage und angesichts der Abschreibereinflüsse in Frage gestellt.[9] In dem Maße, wie sich Techniken zur Autorschaftsattribution trotz verschiedener Normalisierungstechniken oder gar gegenüber verschiedenen handschriftnahen Textfassungen als robust erweisen, wäre auch hier die Rolle des Autors wieder stärker zu akzentuieren.

Ein jüngerer Meilenstein der stilometrischen Autorschaftsattribution ist ohne Zweifel das von John Burrows vorgeschlagene »Delta«-Maß zur Bestimmung der stilistischen Ähnlichkeit zwischen Texten.[10] Die beeindruckend gute Leistung von Delta in verschiedenen Sprachen und Gattungen sollte allerdings nicht darüber hinwegtäuschen, dass die theoretischen Hintergründe weitgehend unverstanden geblieben sind.[11] Anders ausgedrückt: Wir wissen, dass Delta funktioniert, aber nicht, warum es funktioniert. In diesem Beitrag diskutieren wir verschiedene Entwicklungen in der stilometrischen Autorschaftsattribution mit Delta und seinen Varianten. Zugleich wird eine Reihe von Studien vorgestellt, die zu aktuellen internationalen Entwicklungen bei der Verwendung von Distanzmaßen wie Delta für die stilometrische Autorschaftsattribution beitragen.

Der erste Abschnitt dieses Beitrags bietet einen Überblick über den Forschungsstand und analysiert, warum die Veränderung von Delta durch Verwendung des Kosinus-Abstands

[5] Wickmann 1989, S. 534.
[6] Vgl. Juola 2013.
[7] So schon Burrows 1999 und Craig 2000. Zur Autorschaftsdiskussion vgl. exemplarisch Barthes 2000; Jannidis et al. 1999. Zum Verhältnis von Autor- und Stilbegriff Jannidis 2014.
[8] Barthes 2000, S. 190.
[9] Vgl. Strohschneider 1997.
[10] Burrows 2002.
[11] Vgl. Argamon 2008.

zwischen den Vektoren eine so deutliche Verbesserung der Ergebnisse erbracht hat.[12] Am Beispiel einer Sammlung deutscher Romane aus dem 19. und 20. Jahrhundert wird gezeigt, wie sich verschiedene Strategien der Normalisierung oder anderweitigen Transformation des Merkmalsvektors (hier: Wortformen und ihre Häufigkeiten) auf die Attributionsqualität auswirken und inwiefern dies Einblick darin gewährt, wie sich Informationen über Autorschaft im Merkmalsvektor manifestieren – was auch einen Aspekt der Leistungsfähigkeit des klassischen Delta erklärt.

Ein zweiter Abschnitt beleuchtet den Einsatz unterschiedlicher Merkmalstypen für die Ähnlichkeitsbestimmung von Texten mit Delta. Anhand einer Sammlung von mittelhochdeutschen Texten wird gezeigt, dass nicht nur die äußerst häufigen Funktionswörter, sondern auch metrische Eigenschaften für die Autorschaftsattribution eingesetzt werden können. Zugleich wird das Problem der variablen, nicht-normierten Schreibweise von Wörtern diskutiert, das immer dann auftritt, wenn Texte älterer Sprachstufen stilometrisch analysiert werden.

Der letzte Abschnitt behandelt am Beispiel der Übersetzerattribution bei arabisch-lateinischen Übersetzungen philosophischer Texte die Manipulation des Merkmalsvektors nicht durch verschiedene Normalisierungsstrategien, sondern durch gezielte, selektive Merkmalseliminierung. Das Verfahren verbessert nicht nur die Attributionsqualität, sondern erlaubt auch die Isolierung des Autorsignals einerseits, des disziplinenbezogenen Signals andererseits und gibt einen Einblick, welche Einzelmerkmale für das Autorsignal statistisch gesehen entscheidend sind.

Die drei Beispiele demonstrieren auf diese Weise verschiedene aktuelle Entwicklungen in der stilometrischen Autorschaftsattribution mit Delta und seinen Varianten und zeigen, wie bei der Anwendung stilometrischer Distanzmaße auf ganz unterschiedliche Gegenstandsbereiche ähnliche methodische Fragen zu berücksichtigen sind.[13]

2. Burrows Delta verstehen

2.1 Überblick zum Forschungsstand

Burrows Delta ist einer der erfolgreichsten Algorithmen der Stilometrie (Computational Stylistics).[14] In einer ganzen Reihe von Studien wurde seine Brauchbarkeit nachgewiesen.[15] Zur Berechnung von Delta werden in einem ersten Schritt alle in einer Textsammlung vorkommenden Wörter (genauer: alle unterschiedlichen Wortformen, also Types) nach ihrer Gesamthäufigkeit sortiert; ihre relativen Häufigkeiten werden in den einzelnen Dokumenten

[12] Vgl. Smith / Aldridge 2011; Jannidis et al. 2015.
[13] Dieser Beitrag beruht auf den Vorträgen der Sektion »›Delta‹ in der stilometrischen Autorschaftsattribution«, die auf der DHd-Tagung 2016 in Leipzig gehalten wurden. Für die Schriftfassung wurden Anregungen aus den Diskussionen aufgegriffen; die Vorträge wurden überarbeitet und ergänzt.
[14] Burrows 2002.
[15] Z.B. Hoover 2004a; Rybicki / Eder 2011.

berechnet, um Textlängenunterschiede auszugleichen. Wir bezeichnen die relative Häufigkeit des *i*-ten Wortes (gemäß der Gesamtreihenfolge) in einem Dokument *D* mit $f_i(D)$. Im zweiten Schritt werden alle Werte durch eine z-Transformation standardisiert (Abbildung 1), wobei μ_i der Mittelwert über die relativen Häufigkeiten des Wortes *i* in allen Dokumenten und σ_i die zugehörige Standardabweichung bezeichnet.

$$z_i(D) = \frac{f_i(D) - \mu_i}{\sigma_i}$$

Abb. 1: Z-Transformation [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Durch diese Standardisierung tragen alle Worte in gleichem Maße zum Differenzprofil, das im dritten Schritt generiert wird, bei. Dazu wird für jedes Wort die Differenz zwischen seinen z-Scores in zwei Texten D_1 und D_2 ermittelt. Nun kann der Abstand dieser beiden Texte voneinander berechnet werden, indem die Absolutbeträge der Differenzen für alle ausgewählten Wörter aufaddiert werden (Abbildung 2): *m* steht für die Anzahl der häufigsten Wörter (MFW), die für die Untersuchung herangezogen werden.

$$\Delta_{Bur} = \sum_{i=1}^{m} \left| z_i(D_1) - z_i(D_2) \right|$$

Abb. 2: Delta. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Je kleiner der Abstand zwischen zwei Texten, also die Summe der Absolutdifferenzen, desto ähnlicher – so die gängige Interpretation – sind sich die Texte stilistisch und desto höher ist die Wahrscheinlichkeit, dass sie vom selben Autor verfasst wurden.

Trotz seiner Einfachheit und seiner praktischen Nützlichkeit mangelt es bislang allerdings an einer Erklärung für die Funktionsweise des Algorithmus. Argamon zeigt, dass der dritte Schritt in der Berechnung von Burrows' Delta sich als der sogenannte *Manhattan*-Abstand zwischen zwei Punkten in einem mehrdimensionalen Raum verstehen lässt, wobei jede

Dimension der Häufigkeit (bzw. dem z-Score) eines bestimmten Wortes entspricht.[16] Er schlägt u. a. vor, stattdessen den euklidischen Abstand zu verwenden, also die Länge der direkten Verbindungslinie zwischen den beiden Punkten, weil dieser »possibly more natural«[17] sei und zudem eine wahrscheinlichkeitstheoretische Interpretation der standardisierten *z*-Werte erlaubt. Diese Variante wird als Δ_Q (für »Quadratic Delta«) bezeichnet, da hier die Quadrate der Differenzen aufsummiert werden. Bei einer empirischen Prüfung zeigte sich, dass keiner von Argamons Vorschlägen eine Verbesserung bringt.[18]

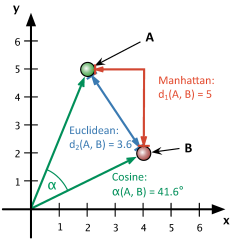

Abb. 3: Veranschaulichung des Abstands zwischen zwei Texten A und B, wenn nur die zwei häufigsten Wörter berücksichtigt werden (also $m = 2$). Burrows Delta verwendet den Manhattan-Abstand. Argamons Vorschlag, den euklidischen Abstand zu verwenden (von ihm als Quadratic Delta bezeichnet), verschlechterte die Clustering-Ergebnisse, während der Vorschlag von Smith / Aldridge, den Kosinus-Abstand bzw. Winkel zwischen den Vektoren zu verwenden, eine deutliche Verbesserung erbrachte. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Smith / Aldridge schlagen vor, wie im Information Retrieval üblich, den Kosinus des Winkels zwischen den Dokumentvektoren als Ähnlichkeitsmaß zu verwenden bzw. äquivalent den Winkel selbst als Abstandsmaß zu wählen.[19] Diese Kosinus-Variante von Delta (Δ_{Cos}) übertrifft Burrows Delta (Δ_{Bur}) fast immer an Leistungsfähigkeit und weist, im Gegensatz zu den anderen Varianten, auch bei der Verwendung sehr vieler MFWs keine Verschlechterung auf.[20] Es stellt sich die Frage, warum Δ_{Cos} besser ist als Δ_{Bur} und ob auf diese Weise die überraschende Leistungsfähigkeit von Delta-Maßen erklärt werden kann.

Entscheidend für unsere weitere Analyse war die Erkenntnis, dass man die Verwendung des Kosinus-Abstands als eine Vektor-Normalisierung verstehen kann, da für die Berechnung des Winkels – anders als bei Manhattan- und euklidischem Abstand – die Länge der Vektoren keine Rolle spielt (vgl. Abbildung 3). Mathematisch lässt sich leicht nachweisen, dass Δ_Q und Δ_{Cos} nach einer expliziten Vektor-Normalisierung äquivalent zueinander sind. Experimente haben gezeigt, dass durch eine Vektor-Normalisierung auch die Ergebnisse anderer Varianten von Delta-Maßen erheblich verbessert und Leistungsunterschiede zwischen ihnen weitgehend neutralisiert werden.[21]

[16] Argamon 2008.
[17] Argamon 2008, S. 134.
[18] Jannidis et al. 2015.
[19] Smith / Aldridge 2011; vgl. auch Baeza-Yates / Ribeiro-Neto 1999, S. 27.
[20] Jannidis et al. 2015.
[21] Evert et al. 2015.

Daraus wurden zwei Hypothesen abgeleitet:

- (H1, Ausreißer-Hypothese) Verantwortlich für die Leistungsunterschiede sind vor allem einzelne Extremwerte (»Ausreißer«), d. #h. besonders große (positive oder negative) z-Scores, die nicht für Autoren, sondern nur für einzelne Texte spezifisch sind. Da das euklidische Abstandsmaß besonders stark von solchen Ausreißern beeinflusst wird, stellen sie eine naheliegende Erklärung für das schlechte Abschneiden von Argamons Δ_Q dar. Der positive Effekt der Vektor-Normalisierung wäre dann so zu deuten, dass durch die Vereinheitlichung der Vektorlängen der Betrag der z-Scores von textspezifischen Ausreißern deutlich reduziert wird.
- (H2, Schlüsselprofil-Hypothese) Das charakteristische stilistische Profil eines Autors spiegelt sich eher in der qualitativen Kombination bestimmter Wortpräferenzen wider, also im grundsätzlichen Muster von über- bzw. unterdurchschnittlich häufigem Gebrauch der Wörter, als in der Amplitude dieser Abweichungen. Ein Textabstandsmaß ist vor allem dann erfolgreich, wenn es strukturelle Unterschiede der Vorlieben eines Autors erfasst, ohne sich davon beeinflussen zu lassen, wie stark das Autorenprofil in einem bestimmten Text ausgeprägt ist. Diese Hypothese erklärt unmittelbar, warum die Vektor-Normalisierung zu einer so eindrucksvollen Verbesserung führt: durch sie wird die Amplitude des Autorenprofils in verschiedenen Texten vereinheitlicht.

2.2 Neue Erkenntnisse

2.2.1 Korpora

Für die hier präsentierten Untersuchungen verwenden wir drei vergleichbar aufgebaute Korpora in Deutsch, Englisch und Französisch. Jedes Korpus enthält je 3 Romane von 25 verschiedenen Autoren, insgesamt also jeweils 75 Texte. Die deutschen Romane aus dem 19. und dem Anfang des 20. Jahrhunderts stammen aus der Digitalen Bibliothek von TextGrid.[22] Die englischen Texte aus den Jahren 1838 bis 1921 sind dem Project Gutenberg[23] entnommen, und die französischen Romane von Ebooks libres et gratuits[24] umfassen den Zeitraum von 1827 bis 1934. Im folgenden Abschnitt stellen wir aus Platzgründen nur unsere Beobachtungen für das deutsche Romankorpus vor. Die Ergebnisse mit Texten in den beiden anderen Sprachen bestätigen – mit kleinen Abweichungen – unseren Befund.[25]

2.2.2 Experimente

[22] https://textgrid.de/Digitale-Bibliothek.
[23] http://www.gutenberg.org/.
[24] https://www.ebooksgratuits.com/.
[25] Für die mit der englischen Textsammlung ermittelten Ergebnisse, vgl. Evert et al. 2017.

Um die Rolle von Ausreißern, und damit die Plausibilität von H1 näher zu untersuchen, ergänzen wir Δ_{Bur} und Δ_Q um weitere Delta-Varianten, die auf dem allgemeinen Minkowski-Abstand basieren (Abbildung 4).

$$\Delta_p = \left(\sum_{i=1}^{m} \left| z_i(D_1) - z_i(D_2) \right|^p \right)^{1/p} \text{ für } p \geq 1.$$

Abb. 4: Delta auf Basis des Minkowski-Abstands [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Wir bezeichnen diese Abstandsmaße allgemein als L_p-Delta. Der Spezialfall p = 1 entspricht dem Manhattan-Abstand (also L_1-Delta = Δ_{Bur}), der Spezialfall p = 2 dem euklidischen Abstand (also L_2-Delta = Δ_Q). Je größer p gewählt wird, desto stärker wird L_p-Delta von einzelnen Ausreißerwerten beeinflusst.

Abbildung 4 vergleicht vier unterschiedliche L_p-Abstandsmaße (für p = 1, 1.4 ≈2, 2, 4) mit Δ_{Cos}. Wir übernehmen dabei den methodologischen Ansatz von Evert et al.:[26] die 75 Texte werden auf Basis der jeweiligen Delta-Abstände automatisch in 25 Cluster gruppiert; anschließend wird die Güte der Autorschaftszuschreibung mit Hilfe des *adjusted Rand index* (ARI) bestimmt. Ein ARI-Wert von 100% entspricht dabei einer perfekten Erkennung der Autoren, ein Wert von 0% einem rein zufälligen Clustering. Offensichtlich nimmt die Qualität von L_p-Delta mit zunehmendem p ab; zudem lässt die Robustheit der Maße gegenüber der Anzahl von MFWs erheblich nach.

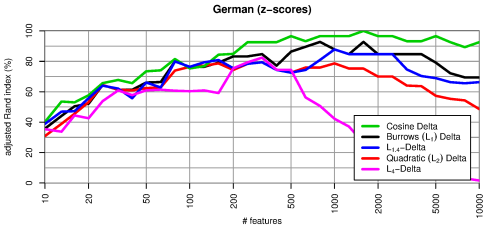

Abb. 5: Clustering-Qualität verschiedener Delta-Maße in Abhängigkeit der Anzahl der MFWs, die als Merkmale verwendet werden. Wie bereits von Jannidis et al. 2015 und Evert et al. 2015 festgestellt wurde, liefert Δ_{Bur} (L_1) durchgängig bessere Ergebnisse als Argamons Δ_Q (L_2) (vgl. Jannidis et al. 2015; Evert et al. 2015). Δ_Q erweist sich als besonders anfällig gegenüber einer zu großen Anzahl von MFWs. Δ_{Cos} ist in dieser Hinsicht robuster als alle anderen Delta-Varianten und erreicht über einen weiten Wertebereich eine nahezu

[26] Evert et al. 2015, S. 82.

perfekte Autorschaftszuschreibung (ARI > 90%). [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Eine Vektor-Normalisierung verbessert die Qualität aller Delta-Maße erheblich (Abbildung 4). Argamons Δ_Q ist in diesem Fall, wie bereits erwähnt, äquivalent zu Δ_{Cos}: die rote Kurve wird von der grünen vollständig überdeckt. Aber auch andere Delta-Maße (Δ_{Bur}, $L_{1.4}$-Delta) erzielen praktisch dieselbe Qualität wie Δ_{Cos}. Einzig das für Ausreißer besonders anfällige L_4-Delta fällt noch deutlich gegenüber den anderen Maßen ab. Diese Ergebnisse scheinen zunächst H1 zu bestätigen.

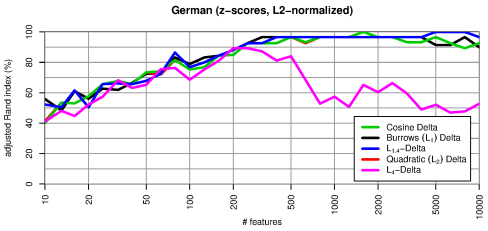

Abb. 6: Clustering-Qualität verschiedener Delta-Maße mit Längen-Normalisierung der Vektoren. In diesem Experiment wurde die euklidische Länge der Vektoren vor Anwendung der Abstandsmaße auf den Standardwert 1 vereinheitlicht. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Ein anderer Ansatz zur Abmilderung von Ausreißern besteht darin, besonders große und kleine z-Scores ›abzuschneiden‹. Wir setzen dazu alle $|z| > 2$ (ein übliches Ausreißerkriterium) je nach Vorzeichen auf den Wert +2 oder –2. Abbildung 7 zeigt, wie sich unterschiedliche Maßnahmen auf die Verteilung der Merkmalswerte auswirken. Die Vektor-Normalisierung (links unten) führt nur zu minimalen Änderungen und reduziert die Anzahl von Ausreißern praktisch nicht. Das Abschneiden extremer *z*-Werte wirkt sich nur auf eine beschränkte Anzahl überdurchschnittlich häufiger Wörter aus (rechts oben). Wie in Abbildung 5 zu sehen ist, wird durch diese Maßnahme ebenfalls die Qualität aller L_p -Deltamaße deutlich verbessert. Der positive Effekt fällt aber merklich geringer aus als bei einer Vektor-Normalisierung.

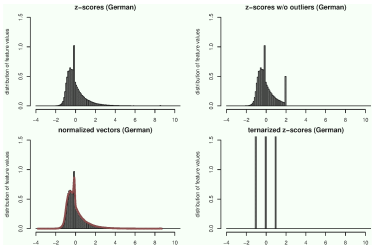

Abb. 7: Verteilung von Merkmalswerten über alle 75 Texte bei Vektoren mit *m* = 5000 MFWs. Gezeigt wird die Verteilung der ursprünglichen *z*-Werte (links oben), die Verteilung nach einer Längen-Normalisierung (links unten), die Verteilung beim Abschneiden von Ausreißern mit $|z| > 2$ (rechts oben) sowie eine ternäre Quantisierung in Werte –1, 0 und +1 (rechts unten). Im linken unteren Bild gibt die rote Kurve die Verteilung

der *z*-Werte ohne Vektor-Normalisierung wieder; im direkten Vergleich ist deutlich zu erkennen, dass die Normalisierung nur einen minimalen Einfluss hat und Ausreißer kaum reduziert. Grenzwerte für die ternäre Quantisierung sind $z < -0.43$ (–1), $-0.43 \leq z \leq 0.43$ (0) und $z > 0.43$ (+1). Diese Grenzwerte sind so gewählt, dass bei einer idealen Normalverteilung jeweils ein Drittel aller Merkmalswerte in die Klassen –1, 0 und +1 eingeteilt würde. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

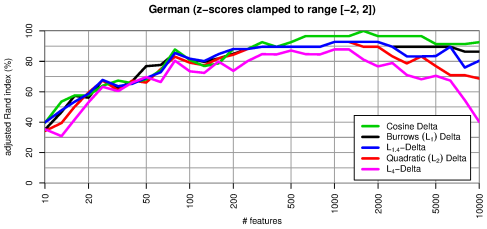

Abb. 8: Clustering-Qualität nach »Abschneiden« von Ausreißern, bei dem Merkmalswerte $|z| > 2$ je nach Vorzeichen durch die festen Werte –2 bzw. +2 ersetzt wurden. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Insgesamt erweist sich Hypothese H1 somit als nicht haltbar. H2 wird durch das gute Ergebnis der Vektor-Normalisierung unterstützt, kann aber nicht unmittelbar erklären, warum auch das Abschneiden von Ausreißern zu einer deutlichen Verbesserung führt. Um diese Hypothese weiter zu untersuchen, wurden reine ›Schlüsselprofil‹-Vektoren erstellt, die nur noch zwischen überdurchschnittlicher (+1), unauffälliger (0) und unterdurchschnittlicher (–1) Häufigkeit der Wörter unterscheiden (vgl. Abbildung 7, rechts unten).

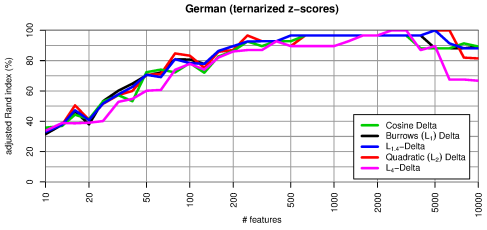

Abb. 9: Clustering-Qualität bei ternärer Quantisierung der Vektoren in überdurchschnittliche (+1, bei $z > 0.43$), unauffällige (0, bei $-0.43 < z < 0.43$) und unterdurchschnittliche (–1, bei $z < -0.43$) Häufigkeit der Wörter. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abbildung 9 zeigt, dass solche Profil-Vektoren hervorragende Ergebnisse erzielen, die der Vektor-Normalisierung praktisch ebenbürtig sind. Selbst das besonders anfällige L_4-Deltamaß erzielt eine weitgehend robuste Clustering-Qualität von über 90%. Wir interpretieren diese Beobachtung als eine deutliche Bestätigung der Hypothese H2.

2.3 Diskussion und Ausblick

H1, die Ausreißerhypothese, konnte widerlegt werden, da die Vektor-Normalisierung die Anzahl von Extremwerten kaum verringert und dennoch die Qualität aller L_p -Maße deutlich verbessert wird. H2, die Schlüsselprofil-Hypothese, konnte dagegen bestätigt werden. Die ternäre Quantisierung der Vektoren zeigt deutlich, dass nicht das Maß der Abweichung bzw. die Größe der Amplitude wichtig ist, sondern das Profil positiver und negativer Abweichungen über die MFWs hinweg. Auffällig ist das unterschiedliche Verhalten der Maße, wenn mehr als 2.000 MFWs verwendet werden. Fast alle Varianten zeigen bei sehr vielen Features eine Verschlechterung, aber sie unterscheiden sich darin, wann dieser Verfall einsetzt. Wir vermuten, dass das Vokabular in diesem Bereich weniger spezifisch für den Autor als für Themen und Inhalte ist. Die Klärung dieser Fragen wird zusätzliche Experimente erfordern.

3. Burrows' Delta im Mittelalter: Wilde Graphien und metrische Analysedaten

Während im vorstehenden Abschnitt bestätigt wurde, dass Delta für germanische Sprachen ausgezeichnet funktioniert,[27] ist bei älteren Sprachstufen wie dem Mittelhochdeutschen das Problem der nicht-normierten Schreibung zu bedenken: Das Wort »und« kann als »unde«, »unt« oder »vnt« verschriftet sein. Ein Teil dieser Varianz wird zwar in normalisierten Ausgaben ausgeglichen, jedoch nicht vollständig. Viehhauser diskutiert in einer ersten Delta-Studie zum Mittelhochdeutschen diese Probleme:[28] Wolfram von Eschenbach benutzt zum Wort »kommen« die Präteritalform »kom«, Hartmann von Aue verwendet »kam«, eine Form, die eher in den südwestdeutschen Raum gehört: Hier gehen also auch dialektale Aspekte ein. Die Bedingungen für den Einsatz von Delta auf der Basis der most-frequent-words (MFWs) erscheinen auf den ersten Blick also als denkbar ungünstig; Viehhauser ist skeptisch, inwieweit Autor, Herausgeber, Schreibereinflüsse oder Dialekt erfasst werden, auch wenn seine Ergebnisse zeigen, dass Delta Texte von gleichen Autoren korrekt sortiert.

Normalisierte Texte sind besser für Autorschaftsstudien geeignet, da hier die Zufälligkeiten von Schreibergraphien reduziert sind; Längenzeichen stellen dort meist weitere lexikalische Informationen zur Verfügung – etwa zur Differenzierung von »sin« (»Sinn«) versus »sîn« (»sein«; allerdings ohne Disambiguierung von »sîn« als verbum substantivum oder Pronomen). In diplomatischen Transkriptionen sind dagegen etwa »u-e« Superskripte und andere diakritische Zeichen enthalten; die gleiche Flexionsform des gleichen Wortes kann in verschiedenen Graphien erscheinen.

Anlass zu vorsichtigem Optimismus bietet allerdings eine Studie von Maciej Eder, die den Einfluss von Noise (wie OCR-Fehler oder Schreibervarianten) analysiert – mit dem Ergebnis etwa für das Neuhochdeutsche, dass ein zufälliger Austausch von 12% aller Buchstaben in jedem Text bei 100–400 MFWs die Ergebnisse kaum beeinträchtigt; bei einer mäßig großen zufälligen Störung der MFW-Frequenzen verschlechtert sich die Quote der korrekten

[27] Vgl. auch Hoover 2004b, S. 477–495; Argamon 2008, S. 131–147; Eder / Rybicki 2011, S. 315–321; Eder 2013, S. 603–614; Jannidis / Lauer 2014; zu SVM vgl. Eder 2015, S. 9f.
[28] Viehhauser 2015.

Attributionen bei 200–400 MFWs ebenfalls kaum. Ersetzt man im Autortext Passagen durch zufällig gewählte Passagen anderer Autoren, ergibt sich bei der Quote lediglich ein »gentle decrease of performance«; im Lateinischen bleibt die Quote gut, selbst wenn 40% des Originalvokabulars ausgetauscht werden.[29]

Während die 17 Texte, die Viehhauser analysiert, in normalisierten Ausgaben vorliegen, werden hier zunächst 37 heterogene Texte von sieben Autoren getestet – sowie drei Texte mit fraglicher Autorzuschreibung zu Konrad von Würzburg.[30] Ein Teil ist normalisiert (Hartmann, Wolfram, Gottfried, Ulrich, Wirnt, Konrad), andere liegen zum Teil in diplomatischen Transkriptionen vor: Bei Rudolf von Ems sind *Gerhard*, *Alexander* und *Barlaam* normalisiert, nicht normalisiert sind *Willehalm* und *Weltchronik* (hier etwa »ubir« statt »über«). Beim Stricker ist lediglich der *Pfaffe Amis* normalisiert verfügbar.[31]

Per Skript werden automatisch Längenzeichen eliminiert, damit nicht Texte mit und ohne Längenzeichen auseinander sortiert werden. Tustep-Kodierungen etwa für Superskripte werden in konventionelle Buchstaben transformiert. Dennoch bleiben große Unterschiede: Die Genitivform zu »Gott« lautet teils »gotes«, teils »gotis«, so dass eigentlich eine primäre Sortierung entlang der Unterscheidung normalisiert–nicht-normalisiert zu erwarten wäre. Das Ergebnis ist jedoch frappierend: Auf der Basis von 200 MFWs[32] gelingt stylo-R ohne Pronomina und bei Culling=50% eine fehlerfreie Sortierung nach Autorschaft; Delta ordnet Rudolf zu Rudolf – ob normalisiert oder nicht (Abbildung 10).

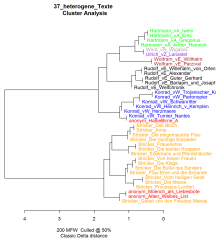

Abb. 10: Clusteranalyse [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

3.1 Validierungstests

3.1.1 Vektorlänge

[29] Eder 2013, S. 612f. Interessant ist die Überlegung: »Luckily enough, stylometric investigations are usually based on texts written in the same period, thus even if the impact of the aforementioned noise is substantial, it affects all the texts in question to a similar degree.«
[30] Hier mit dem R-Paket stylo (Eder et al. 2016).
[31] Mittelhochdeutsche Texte liegen auf vielfältige Art mehr oder weniger normalisiert vor. Da es bei Delta um Abweichungen von Wortfrequenzen geht, werden Texte, die sich relativ eng an Lachmanns Mittelhochdeutsch orientieren, als normalisiert betrachtet; davon abweichende Texte werden hier als nicht-normalisiert rubriziert, selbst wenn sie (etwa in Relation zu einem diplomatischen Abdruck) einer mäßigen Normalisierung unterzogen wurden, da sie eine größere Formenvielfalt aufweisen.
[32] Mit diesem Parameter arbeiten auch Viehhauser 2016 und Eder 2015, S. 169.

Dieser Befund ist Anlass für eine Serie an automatisierten Tests: Bei welchem Vektor und ab welcher Textlänge liefert Delta zuverlässige Ergebnisse? Wie wirkt sich das Einbringen von Noise aus?[33]

Per Perlskript wird ein Delta-Test implementiert, der verschiedene ›Ratetexte‹ in einem ›Ratekorpus‹ mit bekannter Autorschaft gegen ein Trainingskorpus mit ebenfalls bekannter Autorschaft prüft. Ermittelt wird in einem ersten Validierungslauf jeweils der Prozentsatz der richtig erkannten Autoren für jeweils eine Vektorlänge; die Vektorlänge wird dabei in 100er Schritten bis auf 2.500 MFWs erhöht. Gegen ein normalisiertes Trainingskorpus mit 18 Texten werden 19 Ratetexte getestet, zudem werden 17 nicht-normalisierte Texte im Ratekorpus gegen ein gemischtes (normalisiert / nicht-normalisiert) Trainingskorpus mit 21 Texten getestet.

Es zeigt sich (Abbildung 11): Bei den normalisierten Ratetexten (blaue Linie) ist die Erkennungsquote in einem relativ breiten Bereich sehr gut, nämlich von 200–900 MFWs. Bei den nicht-normalisierten Texten (orange Linie) ist die Quote nur für einen kleineren Bereich sehr gut (200, 400 und 500 MFWs). Dass auch bei den nicht-normalisierten Texten bis 500 MFWs noch derart gute Ergebnisse erzielt werden, schürt den Verdacht, dass bei den nicht-normalisierten Texten die Zufälligkeiten der Graphie spezifischer sind als das Autorsignal. Dass die Quote bei den nicht-normalisierten Texten bei längeren Wortlisten schlechter wird, kann durch einen Blick in die Frequenzlisten erklärt werden: Hier sind die ganz häufigen 70 Wörter regelmäßig bei den meisten Texten vertreten. Von den weniger häufigen Wörtern kommen einige jedoch in nicht-normalisierten Texten nicht vor, so dass hier mit zunehmender Länge der Wortliste zunehmend Nullwerte verzeichnet sind.

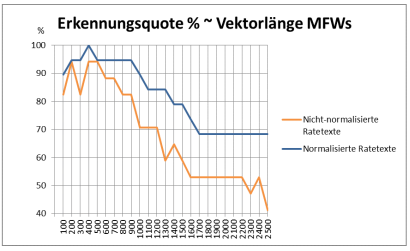

Abb. 11: Vektorlänge [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Interessante Fehlattributionen (etwa bei Konrads *Herzmäre* – ein relativ kurzer Text mit knapp 3000 Wörtern) machen weitere Validierungsläufe nötig. Während Burrows davon ausgeht, dass Delta ab 1.500 Wörtern anwendbar sei, zeigt Eder, dass Delta im Englischen ab 5.000 Wörter sehr gute und unter 3.000 Wörter teils desaströse Ergebnisse liefert; nur im Lateinischen werden ab 2.500 Wörter gute Ergebnisse erreicht.[34]

[33] Testdesign in Anlehnung an Eder 2015, S. 168–170.
[34] Burrows 2002, S. 276; Eder 2015, S. 170–173.

3.1.2 Korrelation Vektorlänge und Textlänge in konventionellen Segmentierungen

Hier wird die Textlänge linear begrenzt, die Texte werden nach 1.000, 2.000 Wörtern usw. abgeschnitten. Das Korpus ist kleiner als zuvor, da zu kurze Texte herausgenommen werden. Bei den normalisierten Texten befinden sich 16 Texte im Trainingskorpus und 15 im Ratekorpus. Bei den nicht-normalisierten Texten sind 20 Texte im Trainingskorpus und 10 im Ratekorpus. Das kleine Ratekorpus erlaubt bei den nicht-normalisierten Texten keine repräsentativen Aussagen. Insgesamt werden dabei 13.200 Delta-Werte berechnet.[35]

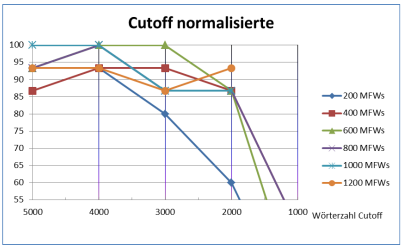

Abb. 12: Cutoff normalisierte Texte [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

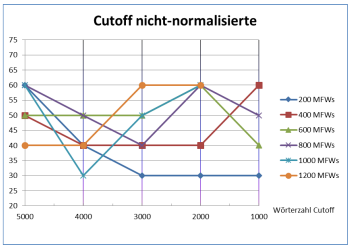

Abb. 13: Cutoff nicht-normalisierte Texte [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Fast alle Ergebnisse sind schlechter als in Testreihe A) mit ungekürzten Texten. Bei den normalisierten Texten sind die Ergebnisse ab 4.000 Wörter Textlänge weitgehend gut bis sehr gut, nur bei 400 MFWs (rot) zeigt sich ein Einbruch (Abbildung 12). Bei kürzeren Vektoren verschlechtern sich die Ergebnisse deutlich. Bei den nicht-normalisierten Texten sind die Ergebnisse desaströs (Abbildung 13). Zu prüfen wäre an anderer Stelle, ob die Schreiber (oder Editoren bei mäßiger Normalisierung) nach und nach eine klarere Präferenz zu bestimmten

[35] Normalisierte Texte: Im Ratekorpus befinden sich Texte von Berthold von Regensburg (3), Meister Eckhart, Hartmann von Aue (3), Jansen Enikel (1), Konrad von Würzburg (3), Rudolf von Ems (2), Wolfram von Eschenbach (2). Im Trainingskorpus sind neben Texten von Autoren, die im Ratekorpus enthalten sind, zusätzlich als Diskriminatoren noch das *Nibelungenlied* sowie Texte von David von Augsburg, Gottfried von Straßburg, Heinrich von Veldeke, Heinrich von Neustadt, Konrad von Heimesfurt, Stricker, Ulrich von Zatzikhoven und Wirnt von Grafenberg. Nicht-normalisierte Texte: Im Ratekorpus befinden sich Texte von David von Augsburg, Hartmann von Aue, Heinrich von Neustadt, Konrad von Heimesfurt, Pleier, Rudolf von Ems (2), Stricker (2), Ulrich von Türheim. Im Trainingskorpus sind neben Texten von Autoren, die im Ratekorpus enthalten sind, zusätzlich das *Nibelungenlied* sowie Texte von Berthold von Regensburg, Meister Eckhart, Gottfried von Straßburg, Heinrich von Veldeke, Jansen Enikel, Konrad von Würzburg, Lamprecht von Regensburg, Ulrich von Liechtenstein, Ulrich von Zatzikhoven, Wirnt von Grafenberg und Wolfram von Eschenbach.

Graphien entwickeln, so dass gerade am Textanfang eine besonders große Heterogenität herrscht.

3.1.3 Korrelation Vektorlänge und Textlänge bei randomisierter Wortauswahl (bag-of-words)

In diesem Szenario wird getestet, welche Erkennungsquoten sich ergeben, wenn anstelle einer linearen Kürzung eine zufällige Auswahl der Token (*bag-of-words*) vorgenommen wird.[36] Weil jede *bag-of-words* neu per Zufall zusammengestellt wird, schwankt die Erkennungsquote etwas, wenn man das Programm mit gleichem Korpus mehrmals laufen lässt – je nach ›Laune‹ des Zufallsgenerators. Um dieses aleatorische Element etwas zu zügeln, wird jeder Test pro *bag-of-words*-Länge und pro Vektor 25 Mal durchgeführt und dann der Mittelwert dieser 25 Erkennungsquoten im Diagramm verwendet. Abbildungen 14 und 15 zeigen die Ergebnisse mit dem Korpus aus Testreihe 3.1.2 mit insgesamt 220.000 Delta-Berechnungen:

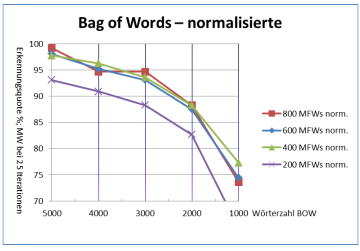

Abb. 14: *Bag-of-words* normalisierte Texte [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

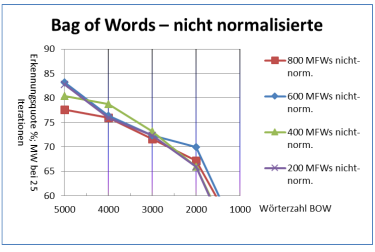

Abb. 15: *Bag-of-words* nicht-normalisierte Texte [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Bei den normalisierten Texten ist die Quote bei einer Textlänge von 5.000 sehr gut bei 400–800 MFWs, bei 200 etwas schlechter (Abbildung 14). Bei 3.000–4.000 bleibt die Quote gut oder fast gut. Bei den nicht-normalisierten Texten werden nur mäßige Quoten erreicht (Abbildung 15); diese Ergebnisse sind eher ernüchternd.[37]

[36] In Anlehnung an Eder 2015, S. 170.
[37] Vgl. jedoch Eder 2013, S. 609f. (hier bei Noise=0%): Für polnische oder lateinische Texte ermittelt Eder Quoten von etwas unter 80% oder 90%.

3.1.4 Auswirkung bei der Eliminierung von Pronomina

Unter ansonsten gleichen Bedingungen wie unter 3.1.3 wird darauf verzichtet, Pronomina aus der Wortliste zu tilgen (Abbildung 16 und Abbildung 17):

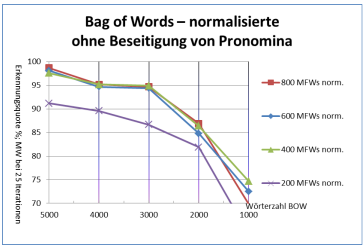

Abb. 16: *Bag-of-words*, normalisierte Texte, ohne Beseitigung der Pronomina [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

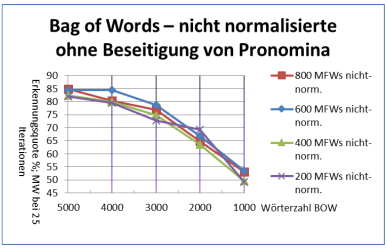

Abb. 17: *Bag-of-words*, nicht-normalisierte Texte, ohne Beseitigung der Pronomina [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Hier ein Vergleich der Testreihen 3.1.2–3.1.4:

Mittelwerte: 200-800MFWs	normalisierte
Textlänge	5000
Bag-of-words ohne Pronomina	**97,1%**
Bag-of-words mit Pronomina	96,4%
Cutoff	91,7%

Verglichen werden in der Tabelle die drei Testreihen über die Mittelwerte von 200–800 MFWs. Bei den normalisierten Texten sind die Quoten minimal schlechter als ohne Pronomina. Das *cutoff*-Verfahren (lineare Kürzung) liefert deutlich schlechtere Erkennungsquoten. Als Anwendungsempfehlung für normalisierte Texte ergibt sich: Es sollten Vektoren von 400–800 Wörter mittels *bag-of-words* getestet werden, Texte sollten nicht kürzer als 3.000 Wörter, am besten länger als 5.000 Wörter sein.

Bei den nicht-normalisierten Texten erhält man mit Pronomina bessere Quoten als ohne. Bei nicht-normalisierten Texten sollten also die Pronomina beibehalten werden. Das ist

plausibel, da sich mit Pronomina die Problematik der Nullwerte weniger gravierend auswirkt, die gerade bei der handschriftlichen Varianz im Spiel ist. Allerdings ist die Datenbasis für nicht-normalisierte Texte bedenklich klein: Im Ratekorpus befinden sich nur die zehn digital verfügbaren Texte von acht Autoren. Diese Tests sind also nicht repräsentativ.

3.1.5 Auswirkungen beim Hinzufügen von Noise

Im nächsten Schritt werden den Ratetexten aus einer Noise-Datei Fehler hinzugefügt. Die Noise-Datei enthält Wortformen (Types) aus zwei mhd. Texten, die nicht im Untersuchungskorpus enthalten sind: *Wilhelm von Österreich* von Johann von Würzburg und *Augsburger Sachsenspiegel*. Damit wird die Integration von anderen mittelhochdeutschen Wortformen simuliert (insgesamt 18.292 Wortformen). Zudem wird der Wortschatz des altfranzösischen Karrenritterromans von Chrétien de Troyes in die Noise-Datei aufgenommen, um fremde Wortformen zu simulieren, die etwa bei Schreiberfehlern entstehen können. Die mittelhochdeutschen Noise-Wortformen können zwar im Untersuchungskorpus vorkommen; da jedoch eine große Zahl an Noise-Wortformen verwendet wird, ist es sehr wahrscheinlich, dass sie nur zu einem sehr geringen Prozentsatz in der MFWs-Liste ohne Noise vorhanden wären.

Getestet werden *bag-of-words* mit 5.000 Wörtern bei ansonsten gleichem Testdesign. Aus dieser Datei werden prozentual aufsteigend randomisiert Wörter der *bag-of-words* des Ratetextes ausgetauscht. Aus Laufzeitgründen erfolgt die Mittelwertbildung nur aus 10 Werten pro *bag-of-words*-Berechnung; mit Noise ist ohnehin ein aleatorisches Element Gegenstand des Tests. Für jede *bag-of-words*-Berechnung wird Noise erneut randomisiert hinzugefügt: In jeder *bag-of-words* sind also verschiedene Noise-Wörter vorhanden. Aufgrund der doppelten Aleatorik (bei Erstellung der *bag-of-words* und beim Einbringen von Noise) verläuft die Kurve nicht konstant linear. Durchgeführt werden in drei Testläufen 2.329.600 Delta-Berechnungen.

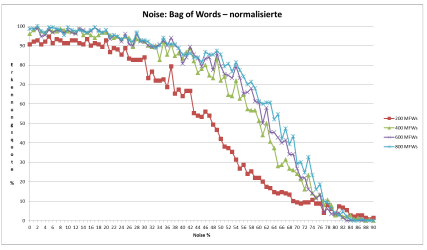

Abb. 18: Noise bei normalisierten Texten [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Bei Vektoren von 400–800 bleiben die Ergebnisse bei normalisierten Texten erstaunlich lange stabil (Abbildung 18); das Einbringen von Noise hat kaum Einfluss auf die Erkennungsquote, solange nicht mehr als 17% des Vokabulars ausgetauscht werden. Nach 17% Noise wird es geringfügig schlechter, die Quoten liegen für Vektorlänge 400–600 jedoch noch

über 90%. Selbst wenn 38% des Vokabulars eines Textes ausgetauscht wird, kann Delta bei Vektorlängen von 600 und 800 noch in über 90% der Fälle die richtigen Autoren ermitteln.

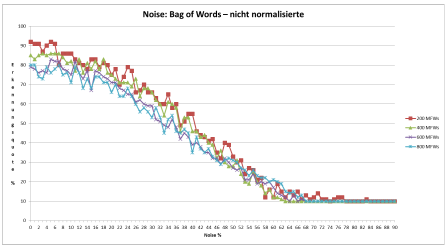

Abb. 19: Noise bei nicht-normalisierten Texten [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Die nicht-normalisierten Texte sind deutlich weniger immun gegen Noise als die normalisierten Texte: Bis ca. 7% bleiben die Quoten relativ stabil, danach sinkt die Quote nahezu linear (Abbildung 19).

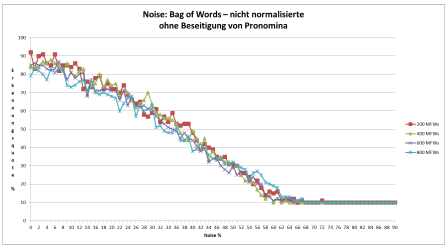

Abb. 20: Noise bei nicht-normalisierten Texten ohne Beseitigung von Pronomina [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Da sich bei der Testreihe 3.1.4 gezeigt hat, dass bei nicht-normalisierten Texten bessere Quoten unter Beibehaltung der Pronomina erzielt werden, wird dieser Test nochmals mit den Pronomina durchgeführt: Die Quoten mit Pronomina (Abbildung 20) liegen zwar knapp 2% über den Quoten ohne Pronomina (Abbildung 19), ansonsten bleibt der Befund jedoch gleich.

Die Stabilität der Erkennungsquoten gibt Grund zum Optimismus für eine Anwendbarkeit bei normalisierten mittelhochdeutschen Texten. Am besten geeignet ist der Vektorbereich von 400–800 MFWs bei langen Texten mittels *bag-of-words*. Wenn man zu dem ohnehin mutmaßlich vorhandenen Rauschen in der Überlieferungskette noch gut 17% an weiteren Fehlern hinzufügt, bleibt Delta davon unbeirrt. Auch wenn die Ergebnisse für nicht-normalisierte Texte zurückfallen, ist es angesichts der wilden mittelhochdeutschen Graphien doch überraschend, dass die Delta-Performanz selbst hier einigermaßen robust bleibt.

Allerdings muss die Vorläufigkeit betont werden: Während im ersten Abschnitt Validierungsstudien mit 75 Texten möglich waren, ist es um die digitale Verfügbarkeit von längeren mittelhochdeutschen Texten, von denen mindestens zwei Texte dem gleichen

Autor zugeschrieben sind, derzeit noch deutlich schlechter bestellt. Die Aussagekraft der vorliegenden Studien wird daher durch die Korpusgröße limitiert: Insbesondere bei den nicht-normalisierten Texten konnte nur ein bedenklich schmales Korpus verwendet werden.

Auch wenn es vorteilhaft wäre, bei der Korpuserstellung nur Texte einer Gattung und insbesondere nur Vers- oder Prosatexte zu verwenden,[38] würde eine solche Begrenzung die Korpusgröße deutlich reduzieren. Daher gehen in diese Studien mittelhochdeutsche Texte aus den verschiedensten Gattungen und sowohl Vers- als auch Prosatexte ein; das Korpus besitzt also keine klaren Konturen. Ob sich dieses Problem der Korpuserstellung künftig umgehen lässt, ist jedoch nicht nur eine Frage der digitalen Verfügbarkeit: Bei Autoren mit mehreren ausreichend langen Texten verteilt sich das Œuvre oft auf verschiedene Gattungen – etwa bei Wolfram von Eschenbach.

3.1.6 Alternative Noise-Konzepte

In der Diskussion auf der DHd-Tagung 2016 in Leipzig wurde angeregt, Noise nicht nur ausschließlich zufallsgeneriert einzubringen, sondern die Verteilung der Wortformen auf Frequenzklassen zu bedenken. Allerdings kann eine Verteilung von bspw. 500 Noise-Wörtern auf die Types mit den höchsten Frequenzen kaum eine nennenswerte Auswirkung auf die Delta-Ergebnisse haben, da so nur die Werte von wenigen sehr häufigen Types verfälscht würden, das Gros der positiven Z-Wert-Differenzen bliebe davon unbetroffen.[39] Versucht wird nun eine Verteilung von Noise-Wörter auf drei verschiedene Häufigkeitsbereiche: Noise wird A) auf die die Top-100-Types der Ratedatei-*bag-of-words* verteilt, B) auf Types mit ein oder zwei Token sowie C) auf mittelfrequente Types zwischen diesen Bereichen. Ausgetauscht werden abwechselnd einige bzw. alle Token zu einem Type in einem dieser drei Häufigkeitsbereiche.[40] Im Test I (Reihe Mod_I in Abbildung 21) werden sämtliche Token zum jeweiligen Type gegen je ein identisches Token aus der Noise-Datei ausgetauscht. Im Test II (Reihe Mod_II in Abbildung 21) wird eine zufällige Anzahl der Token zum jeweiligen Type gegen je eine identische Wortform aus der Noise-Datei ausgetauscht (alle ausgetauschten Token werden durch denselben Type ersetzt).[41] Die Modifikation I simuliert eine systematisch abweichende Graphie etwa eines Abschreibers (stets »vnt« statt »unt«), die Modifikation II simuliert eine weniger systematisch abweichende Graphie (manchmal »vnt« statt »unt«).

[38] Vgl. Schöch 2014, S. 140–147.
[39] Exemplarische Verteilung einer *bag-of-words* mit 5.000 Token: Die ersten 30 Types entsprechen bspw. ca. 1.500 Token, die ersten 100 Types entsprechen knapp 50% der Token. Würde man diese ersten 30 Types vollständig mit 1.500 Noise-Wörtern ersetzen, würden bei einem Vektor von 400 MFWs trotz 30% Noise 370 positive Z-Wert-Differenzen unbeeinflusst bleiben.
[40] In einer explorativen, nicht repräsentativen Ministudie wurde die Varianz anhand von 40 Parzivalversen und 13 Handschriften untersucht – für die Überlassung der Daten sei Michael Stolz herzlich gedankt. Von den 100 häufigsten Types im gesamten normalisierten *Parzival*-Text kommen in dem Handschriften-Sample 92 Types in irgendeiner Variante vor. Zu 26 dieser 92 Types gibt es im Sample graphische Varianten – ein Indikator dafür, dass eine Berücksichtigung der Top-100-Types im ternären Wechsel mit mittel- und niederfrequenten Types nicht völlig an der Überlieferungslage vorbei gehen könnte.
[41] Sobald in einem Häufigkeitsbereich alle Types mit Noise bedient sind, werden nur noch die anderen Häufigkeitsbereiche abgearbeitet. Bei Modifikation II wird nur ein zufälliger Anteil der Noise-Token zu einem Type durch Noise ersetzt. Der übrige, nicht ersetzte Anteil der Token wird bei hohen Noise-Prozentraten ebenfalls sukzessiv durch Noise ersetzt, sobald zu allen Types eine Noise-Ersetzung stattgefunden hat und solange noch weitere Noise-Wörter unterzubringen sind.

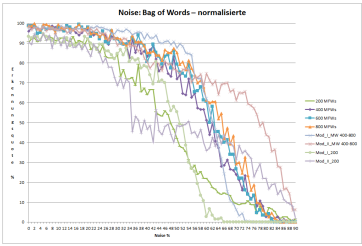

Abb. 21: Noise mit Modifikationen bei normalisierten Texten [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

In Abbildung 21 sind zunächst die Werte für 200–800 MFWs aus Grafik 15 nochmals dargestellt – also Werte, bei denen per Zufall und ohne Rücksicht auf die Frequenzklasse Wörter aus der *bag-of-words* gegen Noise-Wörter ausgetauscht wurden. Für die Modifikationen I und II werden die nahe beieinanderliegenden Werte für die Vektoren 400, 600 und 800 in ihrem Mittelwert wiedergegeben sowie – aufgrund der großen Abweichungen – jeweils die Werte für 200 MFWs bei beiden Modifikationen.

Für die Modifikation II liegen die Quoten von 400–800 bis ca. 50% Noise in einem ähnlichen Bereich wie für den per Zufall generierten Noise von 400–800. Ab 50% sind die Quoten sogar besser. Die hochfrequenten Types beanspruchen einen großen Teil des verfügbaren Noise-Materials, das dann nicht mehr für eine Verfälschung der mittel- und niederfrequenten Types zur Verfügung steht. Für die Modifikation I ist dieser Effekt bis ca. 55% spürbar, hier sind die Ergebnisse besser als bei zufälligem Noise. Erst bei Noise >55% erfolgt ein Einbruch gegenüber dem per Zufall generierten Noise.

Bei 200 MFWs bleiben beide Modifikationen in einer ähnlichen Größenordnung wie bei dem zufälligen Noise bei 200 MFWs. Während die Modifikation II nach 20% schlechter wird als der per Zufall generierte Noise bei 200 MFWs, bleibt Modifikation I hier etwas besser.

3.2 Noise-Reduktion: Metrik-Delta

Bei einem weiteren Versuch geht es darum, die Einflüsse von Schreibergraphie und Normalisierungsart zu reduzieren, indem nicht der Wortschatz, sondern abstraktere Daten verwendet werden; nach Hirst / Feiguina erzielen Tests auf der Basis von Part-of-Speech–Bigrammen gute Ergebnisse.[42] Eine Alternative dazu stellt die Verwendung von metrischen Analysedaten dar, wie sie das Modul »Automatische mittelhochdeutsche Metrik 2.0« bereitstellt.[43] Das Metrik-Modul gibt Kadenzen aus (etwa »weiblich klingend«). Die metrische Struktur wird mit »0« (unbetonte Silben) und »1« (betonte Silben) ausgegeben; der dritte *Parzival*-Vers (*gesmœhet unde gezieret*) hat das Muster »01010011«. Anstatt MFWs werden nun

[42] Hirst / Feiguina 2007.
[43] Vgl. Dimpel 2015, S. 1–26. Das Metrik-Tool (eine Weiterentwicklung aus Dimpel 2004) steht unter http://www.archiv.mediaevistik.germanistik.phil.uni-erlangen.de/cgi-bin/metrik/metrik2.pl. Die Fehlerquote liegt nun etwas unter 2%.

Metrikmuster und Kadenzinformationen als Features verwendet. Weil das Ausgangsmaterial weniger variationsreich ist, wird wie bei Hirst / Feiguina mit Bigrammen und Trigrammen gearbeitet. Abbildung 22 zeigt zunächst einen Stylo-R-Plot mit einigen Klassikern und einigen Mären.

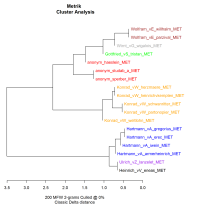

Abb. 22: Stylo-R-Clusteranalyse mit Metrik-Daten I [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abbildung 23 zeigt einen etwas größeren Test mit zwei Fehlattributionen.

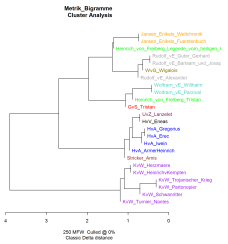

Abb. 23: Stylo-R-Clusteranalyse mit Metrik-Daten II [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Interessant sind die Fehler mit Blick auf Gattungen (Abbildung 23): Der grünfarbige *Tristan* von Heinrich von Freiberg liegt näher bei anderen Texten, die überwiegend ebenfalls der höfischen Epik zuzuordnen sind – eigentlich müsste Heinrichs *Tristan* mit Heinrichs *Legende vom heiligen Kreuz* clustern. Nicht hier im Plot, sondern im Validierungstest ist Rudolfs *Alexander* bei Jansen Enikels *Fürstenbuch* lokalisiert statt bei Rudolfs *Barlaam*. Bekanntlich kann das Autorsignal manchmal vom Gattungssignal überlagert werden. Die Sonderstellung von Rudolfs *Alexander* und von Heinrich von Freiberg wird hier ebenfalls visualisiert – bemerkenswert ist die Nähe von Gottfrieds *Tristan* zur *Tristan*-Fortsetzung des Freibergers.[44]

Validierungstests sind bislang nur mit einem kleineren Korpus möglich, da das Metrik-Modul Längenzeichen und Texte mit vierhebigen Reimpaarversen benötigt (Abbildung 24).

[44] Ähnlich bei anderen Parametern, allerdings wird – wie bereits von Viehhauser 2015 bei konventionellen MFWs beobachtet – Hartmanns *Armer Heinrich* nicht immer in Hartmanns Werk einsortiert, eventuell ein gattungsspezifischer Befund.

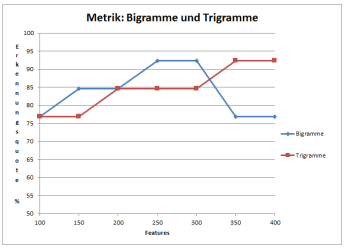

Abb. 24: Metrik-Daten Validierungstest [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Bei 13 Ratedateien und 11 Trainingsdateien ergibt sich bei einem Vektor mit 250–300 Features eine Erkennungsquote von 92,3%. Bei anderen Vektoren werden die Quoten etwas schlechter.

Auch auf Metrik-Basis stellen sich Ergebnisse ein, die in verschiedener Hinsicht erfreulich sind: (1) Bei Tests auf Grundlage von Metrik-Daten ist eine etwas geringere Abhängigkeit von Schreibergraphie und von Normalisierungsgewohnheiten gegeben.[45] (2) Zudem kann Autorschaft offenbar nicht nur mit dem vergleichsweise einfachen Parameter MFWs dargestellt werden: Nicht nur eine pure Wortstatistik führt zum Ziel, vielmehr erweisen sich auch Daten als fruchtbar, die aus einer Verbindung von philologischen mit digitalen Methoden resultieren. (3) Bei der metrischen Struktur handelt es sich um ein Stilmerkmal, das Autoren oft intentional kunstvoll gestalten. Während es als *communis opinio* gilt, dass vor allem die unbewussten Textmerkmale wie die Verteilung der MFWs für die Autorschaftsattribution relevant sind, da sie nicht so einfach zu fälschen sind, kann es, wie in dem vorliegenden Fall, durchaus sinnvoll und philologisch relevant sein, Autorschaft mit Hilfe eines mutmaßlich bewusst gestalteten Stilmerkmals zu unterscheiden. Autorschaft lässt sich zumindest hier auch über ein Merkmal erfassen, das dem bewussten künstlerischen Zugriff unterliegen kann.

4. Stilometrie interdisziplinär: Merkmalsselektion zur Differenzierung zwischen Übersetzer- und Fachvokabular

4.1 Übersetzungen und Übersetzer

Nicht nur mittelhochdeutsche, sondern auch lateinische Texte des Mittelalters lassen eine erstaunliche Bandbreite der Schreibweisen erkennen. Für die folgenden Untersuchungen soll diese Problematik jedoch zurückgestellt werden, da das hier vorliegende Textkorpus noch einige weitere Eigenschaften aufweist, die der gelingenden stilometrischen Analyse im Wege stehen. In diesem Abschnitt verwenden wir Deltamaße zur Identifikation von Übersetzern.

[45] Herausgeber haben zwar mitunter aus metrischen Gründen in den Text eingegriffen. Wenn ein Herausgeber aus metrischen Gründen lieber das Wort »unde« statt »unt« verwendet, dann geht in den Metrik-Delta ein ähnlicher Fehler wie in den konventionellen Delta-Test ein. Immerhin immunisieren Metrik-Daten gegen Graphie-Varianten wie »und« versus »unt«.

Textgrundlage ist eine Sammlung von im 12. Jahrhundert entstandenen arabisch-lateinischen Übersetzungen wissenschaftlicher Texte aus verschiedenen Disziplinen. Faktoren der Zusammensetzung des Textkorpus, die sich bekanntlich negativ auf die Qualität der Ergebnisse auswirken können, sind unter anderem zu kurze Texte, unterschiedliche Genres der Texte und eine Überlagerung von Autor- und Übersetzerstilen.[46] Gerade inhaltliche Unterschiede zwischen Texten stellen ein Hindernis bei der Erkennung der Autoren dar, das nur mit erheblichem technischen Aufwand überwunden werden kann.[47] Obwohl alle diese Faktoren zutreffen, waren erste Erfolge auf einem kleineren, nur aus philosophischen Texten bestehenden Teilkorpus Motivation genug, weiter nach Lösungen zu suchen.[48] Im Folgenden zeigen wir einen Weg auf, wie die aus den oben genannten Faktoren resultierenden Einschränkungen durch den Einsatz maschineller Lernverfahren kompensiert werden können. Gleichzeitig eröffnet sich dadurch eine Möglichkeit, unter den meistverwendeten Wörtern solche zu identifizieren, deren Häufigkeiten einerseits eher Rückschlüsse auf die Übersetzer oder andererseits eher Rückschlüsse auf die wissenschaftliche Disziplin der Texte zulassen.

4.2 Das Korpus

Die hier verwendete Textsammlung wurde mit dem philologischen Ziel angelegt, die Übersetzer zu identifizieren, die im 12. Jahrhundert eine Vielzahl von Texten aus dem Arabischen ins Lateinische übertragen und damit in den verschiedensten Disziplinen die weitere Entwicklung der europäischen Wissenschaften nachhaltig beeinflusst haben.[49] Es handelt sich dabei um Texte unterschiedlicher Autoren, darunter Aristoteles, Ptolemäus und al-Fārābī, aus den Bereichen Philosophie, Mathematik, Astronomie, Astrologie, Medizin, Geologie und Meteorologie, aber auch um religiöse, magische und alchemistische Traktate, wobei einzelne Texte nicht eindeutig einer Disziplin zugeordnet werden können. Elf der Übersetzer sind namentlich bekannt, fast die Hälfte der Texte ist jedoch nur anonym überliefert.

Für die Experimente wird ein Testkorpus so zusammengestellt, dass von jedem Übersetzer und aus jeder Disziplin mindestens drei Texte zur Verfügung stehen. Dieses besteht aus insgesamt 37 Texten von 6 Übersetzern, wobei die Texte aus 4 Disziplinen stammen (vgl. Abbildung 25). Das daraus resultierende Textkorpus ist nicht balanciert: Die Anzahl der Texte pro Übersetzer ist ungleich verteilt, die Länge der Texte liegt zwischen 500 und fast 200.000 Wörtern; insgesamt sind die Texte auch deutlich kürzer als diejenigen der oft verwendeten Romankorpora.[50]

[46] Eder 2015, passim; Schöch 2014, passim; Rybicki 2012, passim.
[47] Stamatatos et al. 2000, passim; Kestemont et al. 2012, passim.
[48] Hasse / Büttner 2017.
[49] Vgl. hierzu auch die Projekthomepage des Digital Humanities-Zentrums KALLIMACHOS der Universität Würzburg.
[50] Vgl. etwa Jannidis et al. 2015.

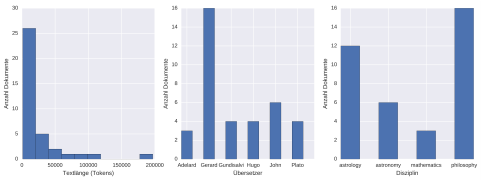

Abb. 25: Verteilung der Textlängen, Übersetzer und Disziplinen im verwendeten Teilkorpus [Andreas Büttner, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Weitere die Analyse erschwerende Faktoren sind Doppelübersetzungen desselben Originaltextes durch zwei Übersetzer und die – historisch nicht völlig eindeutig belegte – Zusammenarbeit einiger Übersetzer. Auf der anderen Seite sind die unterschiedlichen Disziplinen prinzipiell einfacher und eindeutiger unterscheidbar als literarische Gattungen in Romankorpora.

4.3 Methoden

4.3.1 Delta-Maße

Wie in Abschnitt 2.1 beschrieben, wurde ausgehend von Burrows' ursprünglichem Deltamaß eine ganze Reihe von Deltamaßen für die Autorschaftszuschreibung vorgeschlagen, die sich in der Regel nur im dritten Verarbeitungsschritt, der Berechnung der Textabstände, unterscheiden.[51] Für die folgenden Experimente verwenden wir Kosinus-Delta, das sich in verschiedenen Evaluationen als das robusteste Mitglied der Delta-Familie erwiesen hat.[52]

4.3.2 Rekursive Merkmalseliminierung

Rekursive Merkmalseliminierung (recursive feature elimination, RFE)[53] ist eine Methode zur Selektion einer möglichst kleinen Teilmenge von Merkmalen, mit der trotzdem möglichst optimale Ergebnisse mit einem überwachten maschinellen Lernverfahren erzielt werden können. Evert et al. experimentieren zur Autorschaftszuschreibung mit durch RFE ermittelten Wörtern als Alternative zu den üblichen *n* häufigsten Wörtern.[54]

[51] Burrows 2002; bspw. Hoover 2004b; Argamon 2008; Smith / Aldridge 2011; Eder et al. 2013.
[52] Vgl. bspw. Jannidis et al. 2015; Evert et al. 2015.
[53] Guyon et al. 2002.
[54] Evert et al. 2015.

Da RFE auf einem überwachten Lernverfahren (üblicherweise einer *Support Vector Machine*) basiert, müssen zumindest für eine Teilmenge der Dokumente die wahren Autoren bzw. Übersetzer bekannt sein. Das rekursive Verfahren trainiert zunächst den Klassifikator auf allen Merkmalen (d. h. im vorliegenden Fall den häufigsten Wörtern), wobei den einzelnen Merkmalen Gewichte zugeordnet werden. Anschließend werden die *k* Merkmale mit den niedrigsten absoluten Gewichten entfernt (*pruning*). Die Schritte Training und *pruning* werden nun auf den verbleibenden Merkmalen so lange wiederholt, bis die gewünschte Anzahl von Merkmalen übrigbleibt. Alternativ kann durch Kreuzvalidierung eine optimale Merkmalsmenge bestimmt werden.

In den folgenden Experimenten kombinieren wir beide Varianten und verkleinern die Merkmalsmenge (also die Menge der verwendeten Wörter) zunächst schrittweise auf die 500 besten Merkmale, um anschließend die optimale Teilmenge zu bestimmen.

4.4 Experimente

Zunächst führen wir mit dem Testkorpus einige Versuche zur Anpassung der stilometrischen Methoden durch. Als Maß der Qualität des Clusterings dient dabei wiederum der *Adjusted Rand Index (ARI)*, der zwischen –1 und 1 liegen kann. Ein vollständig korrektes Clustering erhält einen ARI von 1, eine zufällige Gruppierung der Elemente einen ARI um 0, und negative Werte weisen auf ein Clustering hin, das schlechter als zufällig ist. Wie in Abbildung 26 dargestellt, wird bei Verwendung von Kosinus-Delta der höchste ARI für das Clustering der Übersetzer bereits bei weniger als 500 der häufigsten Wörter erreicht, auch bei über 1000 Wörtern fällt das Qualitätsmaß kontinuierlich ab. Ein Clustering nach Disziplinen hingegen erreicht bei ca. 1200–2500 Wörtern die besten Ergebnisse. Es fällt auf, dass zum einen die besten Ergebnisse mit viel kleineren Wortmengen erreicht werden als bei Studien zur Autorschaftszuschreibung und dass zum anderen die Ergebnisse deutlich schlechter sind.

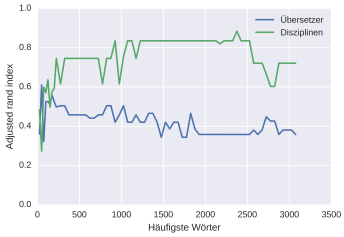

Abb. 26: Clusteringqualität in Abhängigkeit von der Anzahl der häufigsten Wörter [Andreas Büttner, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Da das Hauptziel eine korrekte Zuordnung der Übersetzer ist, soll die Menge der 500 häufigsten Wörter (im Folgenden *MFW500*), mit der ein $ARI_Ü$ von 0,458 erreicht wird, als Vergleichsmaßstab für die weiteren Versuche dienen. Für ein Clustering nach Disziplinen wird mit diesen Wörtern ein ARI_D von 0,746 erreicht.

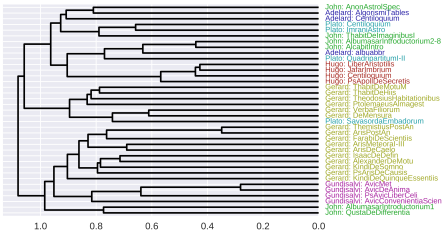

Abb. 27: Dendrogramm für das Clustering mit MFW500, Einfärbung nach Übersetzern [Andreas Büttner, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Durch RFE wählen wir aus der Gesamtmenge weniger als 500 Wörter aus. Mit 495 Wörtern ist eine perfekte, d.h. fehlerfreie, Klassifikation nach Übersetzern möglich. Wenig überraschend erzielen wir mit diesen Wörtern auch ein perfektes Clustering der Texte nach Übersetzern ($ARI_Ü$ = 1,0). Auch für die Disziplinen lässt sich eine Menge von 485 Wörtern finden, bei der die Texte sich vollständig korrekt aufteilen lassen (ARI_D = 1,0). Da die durch RFE bestimmten Wörter teilweise sehr spezifisch sind und dadurch zu befürchten ist, dass Merkmale selektiert werden, die jeweils nur zwei Texte aneinander binden oder voneinander trennen, wählen wir aus den für die Übersetzer RFE-selektierten Merkmalen diejenigen aus, die auch in MFW500 enthalten sind. Mit diesen 74 Wörtern ist immer noch eine recht gute, wenn auch nicht perfekte Unterscheidung der Übersetzer möglich ($ARI_Ü$ = 0,824). Disziplinen lassen sich mit diesen Merkmalen nur sehr schlecht unterscheiden (ARI_D = 0,228).

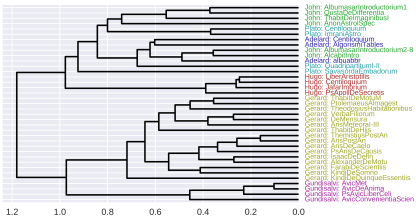

Abb. 28: Dendrogramm für das Clustering mit der Schnittmenge aus RFE und MFW500, Einfärbung nach Übersetzern [Andreas Büttner, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Die Analyse der *z*-Werte dieser Wörter zeigt, dass diese überwiegend bei nur einem einzigen Übersetzer besonders häufig sind. Sie lassen sich deshalb zu dem Übersetzer gruppieren, in dessen Texten der Mittelwert dieser z-Werte am höchsten ist, wodurch sich für jeden Übersetzer eine Liste von spezifischen bevorzugten Wörtern ergibt.

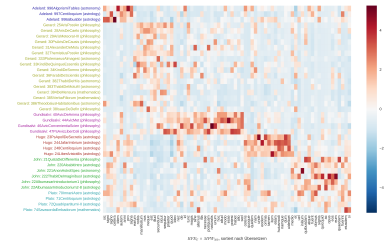

Abb. 29: Heatmap der *z*-Werte aus der Schnittmenge von RFE$_Ü$ und MFW500 [Andreas Büttner, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Die 426 Wörter aus MFW500, die in der Menge der RFE-selektierten Wörter nicht enthalten sind, unterscheiden, wie erwartet, deutlich schlechter zwischen Übersetzern (ARI$_Ü$ = 0,284), dafür aber sehr gut zwischen Disziplinen (ARI$_D$ = 0,795) – überraschenderweise sogar deutlich besser als alle 500 Wörter aus MFW500 (ARI$_D$ = 0,746).

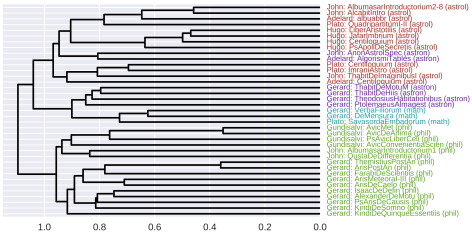

Abb. 30: Dendrogramm für das Clustering mit der Differenzmenge aus MFW500 und RFE, Einfärbung nach Disziplinen [Andreas Büttner, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Bei den Disziplinen erzielt die Schnittmenge der dafür mit RFE ausgewählten Wörter mit MFW500 ebenfalls sehr gute Ergebnisse (Anzahl der Merkmale: 123, ARI$_D$ = 0,836). Auch die Differenzmenge zeigt hier den oben beschriebenen Effekt: Die Clusteringqualität nach Disziplinen sinkt durch den Ausschluss der Wörter ab (ARI$_D$ = 0,593), die nach Übersetzern fällt hingegen besser aus (ARI$_Ü$ = 0,526) als der Wert für alle MFW500 (ARI$_Ü$ = 0,458).

Um die Robustheit der Ergebnisse zu prüfen und insbesondere gegen ein Overfitting durch das RFE-Verfahren abzusichern, kann das bisher Beschriebene mit einem ohne Überschneidungen in ein Trainingsset und ein Testset aufgeteilten Korpus wiederholt werden. Dabei lassen sich die mit dem Gesamtkorpus beschriebenen Effekte reproduzieren, wenn auch – aufgrund der dann sehr kleinen Textanzahl – in schwächerer Ausprägung.

4.5 Ergebnisse

Durch die Experimente wurde gezeigt, dass sich die Menge der *n* häufigsten Wörter, die üblicherweise zur Autorschaftszuschreibung verwendet wird, so in zwei Teilmengen partitionieren lässt, dass die eine die Identifikation der Übersetzer der Texte besser ermöglicht als die Gesamtmenge, während die Wörter aus der anderen Teilmenge zur Identifizierung von Disziplinen verwendet werden können. Die rekursive Merkmalseliminierung erwies sich dabei als wirksames Mittel zur Differenzierung zwischen den zur Bestimmung des Verfassers relevanten und den durch die unterschiedlichen Inhalte der Texte bedingten Merkmalen. Darüber hinaus bietet eine solche Kondensierung der Wortliste die Chance, von einer aus philologischer Sicht undurchschaubaren statistischen Maschinerie zu tatsächlich durch den Leser der Texte intuitiv nachvollziehbaren Kriterien zu gelangen.

Weitere Experimente in diesem Kontext werden dem Versuch dienen, die unterscheidenden Wörter besser zu charakterisieren, sodass idealerweise auch ohne maschinelles Lernen eine Auswahl der Merkmale möglich wird. Zudem steht eine Anwendung der Methode auf andere wissenschaftliche und literarische Textkorpora aus.

5. Ausblick

Auch wenn die drei vorstehenden Abschnitte unterschiedliche Fragestellungen im Delta-Kontext beleuchtet haben, zeigen sie doch verschiedene Optionen auf, wie die Erkennungsgenauigkeit von quantitativen Verfahren der Autorschafts- bzw. Übersetzerattribution verbessert werden kann: Durch rekursive Merkmalseliminierung gelingt es, trotz starken Autorsignals auch Spuren des Übersetzersignals zu orten (Abschnitt 4). Durch den Einsatz von Metrik-Daten kann das Manko der nicht-genormten mittelalterlichen Graphie abgefedert werden (Abschnitt 3). Durch eine Vektornormalisierung kann die Erkennungsquote weiter verbessert werden – bei diesem Befund des ersten Abschnitts handelt es sich gewissermaßen um einen Serendipitätseffekt; eigentlich gesucht war eine Antwort auf die Frage, ob der Erfolg von Delta eher auf wenigen Extremwerten oder eher auf einer breiten Schlüsselprofilverteilung beruht.

In weiteren Studien soll versucht werden, eine weitere Verbesserung der Erkennungsquote insbesondere bei kürzeren Texten zu erreichen. Im Bereich der vormodernen Texte wäre zu erproben, inwieweit über eine automatische Teilnormalisierung die Problematik von Schreibereinflüssen und dialektalen Eigenheiten bei nicht-normalisierten Texten reduziert werden kann. Bei der Frage nach dem Übersetzersignal kann in weiteren Experimenten versucht werden, die unterscheidenden Wörter besser zu charakterisieren, sodass idealerweise auch ohne maschinelles Lernen eine Auswahl der Merkmale möglich wird – auch bei einer Anwendung der Methode auf andere Textkorpora. Insgesamt ergibt sich auch die Frage, ob man, ausgehend von der Schlüsselprofil-These, nicht wieder den Anschluss zu Konzepten der Stilistik herstellen kann, nämlich das Verständnis von Stil als individueller Abweichung von einer – wie auch immer konzeptualisierten – sprachlichen Norm.

Bibliographische Angaben

Shlomo Argamon: Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. In: Literary and Linguistic Computing 23 (2008), H. 2, S. 131–147. DOI: 10.1093/llc/fqn003

Ricardo Baeza-Yates / Berthier Ribeiro-Neto: Modern Information Retrieval. Harlow 1999. [Nachweis im GBV]

Roland Barthes: Der Tod des Autors. [Nachweis im GBV] In: Texte zur Theorie der Autorschaft. Hg. von Fotis Jannidis / Gerhard Lauer / Matias Martinez / Simone Winko. Stuttgart 2000 (1968), S. 184–193. [Nachweis im GBV]

John F. Burrows: Computers and the Idea of Authorship. In: Rückkehr des Autors. Erneuerung eines umstrittenen Begriffs. Hg. von Fotis Jannidis / Gerhard Lauer / Matias Martinez / Simone Winko. Tübingen 1999, S. 167–182. [Nachweis im GBV]

John F. Burrows: ›Delta‹: A Measure of Stylistic Difference and a Guide to Likely Authorship. In: Literary and Linguistic Computing 17 (2002), H. 3, S. 267–187. DOI: 10.1093/llc/17.3.267

Hugh Craig: Is the author really dead? An empirical study of authorship in English Renaissance drama. In: Empirical Studies of the Arts 18 (2000), H. 2, S. 119–134. [Nachweis im GBV]

Friedrich Michael Dimpel: Automatische Mittelhochdeutsche Metrik 2.0. In: Philologie im Netz 73 (2015), S. 1–26. [online]

Friedrich Michael Dimpel: Computergestützte textstatistische Untersuchungen an mittelhochdeutschen Texten, Tübingen 2004. [Nachweis im GBV]

Maciej Eder: Mind Your Corpus: systematic errors in authorship attribution. In: Literary and Linguistic Computing 28 (2013), S. 603–614. DOI: 10.1093/llc/fqt039

Maciej Eder: Does size matter? Authorship attribution, small samples, big problem. In: Digital Scholarship Humanities 30 (2015), H. 2, S. 167–182. DOI: 10.1093/llc/fqt066

Maciej Eder / Mike Kestemont / Jan Rybicki: Stylometry with R: a suite of tools. In: Digital Humanities 2013: Conference Abstracts. Lincoln: University of Nebraska (2013), S. 487–489. [online]

Maciej Eder / Mike Kestemont / Jan Rybicki: Stylometry with R: A Package for Computational Text Analysis. In: The R Journal 16 (2016), H. 1, S. 1–15. [online]

Maciej Eder / Jan Rybicki: Deeper Delta across genres and languages: do we really need the most frequent words? In: Literary and Linguistic Computing 26 (2011), H. 1, S. 315–321. DOI: 10.1093/llc/fqr031

Stefan Evert / Thomas Proisl / Fotis Jannidis / Steffen Pielström / Christof Schöch /Thorsten Vitt: Towards a better understanding of Burrows's Delta in literary authorship attribution. DOI: 10.5281/zenodo.18177 In: Proceedings of the Fourth Workshop on Computational Linguistics for Literature. (NAACL HLT 2015, Denver, CO, 4.6.2015) Denver, CO. 2015, S. 79–88. [online] [Nachweis im GBV]

Stefan Evert / Fotis Jannidis / Thomas Proisl / Steffen Pielström / Thorsten Vitt / Isabella Reger / Christof Schöch: Understanding and Explaining Distance Measures for Authorship Attribution. In: Digital Scholarship in the Humanities, 2017 (Advance Articels). DOI: 10.1093/llc/fqx023

Isabelle Guyon / Jason Weston / Stephen Barnhill / Vladimir Vapnik: Gene Selection for Cancer Classification using Support Vector Machines. DOI: 10.1023/A:1012487302797 In: Machine Learning 46 (2002), H. 1-3, S. 389–422. [online] [Nachweis im GBV]

Ruth Haag: Noch einmal. Der Verfasser der »Nachtwachen von Bonaventura«, 1804. In: Euphorion 81 (1987), S. 289–297. [Nachweis im GBV] und [Nachweis im GBV]

Dag Nikolaus Hasse / Andreas Büttner: Notes on the Identity of the Latin Translator of Avicenna's Physics and on Further Anonymous Translations in Twelfth-Century Spain. Vorabversion 2012. PDF. [online]

Graeme Hirst / Olga Feiguina: Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. In: Literary and Linguistic Computing 22 (2007), H. 4, S. 405-417. DOI: 10.1093/llc/fqm023

David L. Hoover (2004a): Testing Burrows's Delta. In: Literary and Linguistic Computing 19 (2004), H. 4, S. 453–475. DOI: 10.1093/llc/19.4.453

David L. Hoover (2004b): Delta Prime? In: Literary and Linguistic Computing 19 (2004), H. 4, S. 477–495. DOI: 10.1093/llc/19.4.477

Fotis Jannidis: Der Autor ganz nah – Autorstil in Stilistik und Stilometrie. In: Theorien und Praktiken der Autorschaft. Hg. von Matthias Schaffrick / Marcus Willand. Berlin 2014, S. 169–195. [Nachweis im GBV]

Fotis Jannidis / Gerhard Lauer: Burrows's Delta and Its Use in German Literary History. [Nachweis im GBV] In: Distant Readings. Topologies of German Culture in the Long Nineteenth Century. Hg. von Matt Erlin / Lynne Tatlock. New York 2014, S. 29–54. [Nachweis im GBV]

Fotis Jannidis / Steffen Pielström / Christof Schöch / Thorsten Vitt: Improving Burrows' Delta – An Empirical Evaluation of Text Distance Measures. In: Digital Humanities Conference 2015, Sydney, Australien. [online]

Fotis Jannidis / Gerhard Lauer / Matias Martinez und Simone Winko: Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs. Tübingen 1999. [Nachweis im GBV]

Patrick Juola: Authorship Attribution. Boston, Mass. 2006. (= Foundations and Trends in Information Retrieval, 1,3) [Nachweis im GBV]

Patrick Juola: How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling. Author of the Harry Potter books has a distinct linguistic signature. In: Scientific American (20.8.2013). [online]

Mike Kestemont / Kim Luyckx / Walter Daelemans / Thomas Crombez: Cross-Genre Authorship Verification Using Unmasking. In: English Studies 93 (2012), H. 3, S. 340–356. DOI: 10.1080/0013838X.2012.668793

August Klingemann: Nachtwachen von Bonaventura. Freimüthigkeiten. Hg. und kommentiert von Jost Schillemeit. Göttingen 2012. [Nachweis im GBV]

Frederick Mosteller / David L. Wallace: Inference in an Authorship Problem. In: Journal of the American Statistical Association 58 (1963), H. 302, S. 275–309. [Nachweis im GBV]

Jan Rybicki: The great mystery of the (almost) invisible translator: stylometry in translation. Preprint. [online[In: Quantitative Methods in Corpus-Based Translation Studies. Hg. von Michael Philip Oakley / Meng Ji. Amsterdam u.a. 2012, S. 231–248. (= Studies in corpus linguistics, 51) [Nachweis im GBV]

Christof Schöch: Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In: Literaturwissenschaft im digitalen Medienwandel 7 (2014). PDF. [online]

Peter W. H. Smith / W. Aldridge: Improving Authorship Attribution: Optimizing Burrows' Delta Method*. In: Journal of Quantitative Linguistics 18 (2011), H. 1, S. 63–88. [Nachweis im GBV]

Efstathios Stamatatos: A Survey of Modern Authorship Attribution Methods. In: Journal of the American Society for Information Science and Technology 60 (2009), H. 3, S. 538–556. [Nachweis im GBV]

Efstathios Stamatatos / Nikos Fakotakis / George Kokkinakis: Automatic Text Categorization in Terms of Genre and Author. DOI: 10.1162/089120100750105920 In: Computational Linguistics 26 (2000), H. 4, S. 471–497. [online]

Peter Strohschneider: Situationen des Textens. Okkasionelle Bemerkungen zur New Philology. In: Zeitschrift für deutsche Philologie 116 (1997), S. 62–86. [Nachweis im GBV]

Gabriel Viehhauser: Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte. DOI: 10.17175/sb001_009 In: Grenzen und Möglichkeiten der Digital Humanities. Hg. von Constanze Baum / Thomas Stäcker. 2015 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1) DOI: 10.17175/sb01

Dieter Wickmann: Computergestützte Philologie: Bestimmung der Echtheit und Datierung von Texten. In: Computerlinguistik. Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen. Hg. von István S. Bátori / Winfried Lenders / Wolfgang Putschke. Berlin u.a. 1989, S. 528–534. (= HSK, 4) [Nachweis im GBV]

Dieter Wickmann: Eine mathematisch-statistische Methode zur Untersuchung der Verfasserfrage literarischer Texte. Durchgeführt am Beispiel der »Nachtwachen. Von Bonaventura« mit Hilfe der Wortartenübergänge. Köln u.a. 1969. [Nachweis im GBV]

Dieter Wickmann: Zum Bonaventura-Problem. In: Sprachlehrforschung. Hrsg. von Karl-Richard Bausch / Helmut Kreuzer. Göttingen 1974, S. 8–29. (= Zeitschrift für Literaturwissenschaft und Linguistik, 13) [Nachweis im GBV]

Abbildungslegenden und -nachweise

Abb. 1: Z-Transformation [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 2: Delta. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 3: Veranschaulichung des Abstands zwischen zwei Texten A und B, wenn nur die zwei häufigsten Wörter berücksichtigt werden (also $m = 2$). Burrows Delta verwendet den Manhattan-Abstand. Argamons Vorschlag, den euklidischen Abstand zu verwenden (von ihm als Quadratic Delta bezeichnet), verschlechterte die Clustering-Ergebnisse, während der Vorschlag von Smith / Aldridge, den Kosinus-Abstand bzw. Winkel zwischen den Vektoren zu verwenden, eine deutliche Verbesserung erbrachte. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 4: Delta auf Basis des Minkowski-Abstands [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 5: Clustering-Qualität verschiedener Delta-Maße in Abhängigkeit der Anzahl der MFWs, die als Merkmale verwendet werden. Wie bereits von Jannidis et al. 2015 und Evert et al. 2015 festgestellt wurde, liefert Δ_{Bur} (L_1) durchgängig bessere Ergebnisse als Argamons Δ_{Q} (L_2) (vgl. Jannidis et al. 2015; Evert et al. 2015). Δ_{Q} erweist sich als besonders anfällig gegenüber einer zu großen Anzahl von MFWs. Δ_{Cos} ist in dieser Hinsicht robuster als alle anderen Delta-Varianten und erreicht über einen weiten Wertebereich eine nahezu perfekte Autorschaftszuschreibung (ARI > 90%). [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 6: Clustering-Qualität verschiedener Delta-Maße mit Längen-Normalisierung der Vektoren. In diesem Experiment wurde die euklidische Länge der Vektoren vor Anwendung der Abstandsmaße auf den Standardwert 1 vereinheitlicht. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 7: Verteilung von Merkmalswerten über alle 75 Texte bei Vektoren mit $m = 5000$ MFWs. Gezeigt wird die Verteilung der ursprünglichen z-Werte (links oben), die Verteilung nach einer Längen-Normalisierung (links unten), die Verteilung beim Abschneiden von Ausreißern mit $|z| > 2$ (rechts oben) sowie eine ternäre Quantisierung in Werte –1, 0 und +1 (rechts unten). Im linken unteren Bild gibt die rote Kurve die Verteilung der z-Werte ohne Vektor-Normalisierung wieder; im direkten Vergleich ist deutlich zu erkennen, dass die Normalisierung nur einen minimalen Einfluss hat und Ausreißer kaum reduziert. Grenzwerte für die ternäre Quantisierung sind $z < -0.43$ (–1), $-0.43 \leq z \leq 0.43$ (0) und $z > 0.43$ (+1). Diese Grenzwerte sind so gewählt, dass bei einer idealen Normalverteilung jeweils ein Drittel aller Merkmalswerte in die Klassen –1, 0 und +1 eingeteilt würde. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 8: Clustering-Qualität nach »Abschneiden« von Ausreißern, bei dem Merkmalswerte $|z| > 2$ je nach Vorzeichen durch die festen Werte –2 bzw. +2 ersetzt wurden. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 9: Clustering-Qualität bei ternärer Quantisierung der Vektoren in überdurchschnittliche (+1, bei $z > 0.43$), unauffällige (0, bei $-0.43 < z < 0.43$) und unterdurchschnittliche (–1, bei $z < -0.43$) Häufigkeit der Wörter. [Stefan Evert, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 10: Clusteranalyse [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 11: Vektorlänge [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 12: Cutoff normalisierte Texte [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 13: Cutoff nicht-normalisierte Texte [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 14: *Bag-of-words* normalisierte Texte [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 15: *Bag-of-words* nicht-normalisierte Texte [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 16: *Bag-of-words*, normalisierte Texte, ohne Beseitigung der Pronomina [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 17: *Bag-of-words*, nicht-normalisierte Texte, ohne Beseitigung der Pronomina [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 18: Noise bei normalisierten Texten [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 19: Noise bei nicht-normalisierten Texten [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]

Abb. 20: Noise bei nicht-normalisierten Texten ohne Beseitigung von Pronomina [Friedrich Michael Dimpel, 2017. Lizenziert unter Creative Commons Namensnennung 4.0 International (CC-BY)]