Settles, Burr / Zhu, Xiaojin (2012): "Behavioral factors in interactive training of text classifiers", in: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 563-567.

Wiedemann, Gregor / Lemke, Matthias / Niekler, Andreas (2013): "Postdemokratie und Neoliberalismus. Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949-2011", in: *Zeitschrift für Politische Theorie* 4, 1: 99-115.

Wiedemann, Gregor / Niekler, Andreas (2014): "Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries", in: *Proceedings of the Conference on Terminology and Knowledge Engineering 2014*, Berlin.

Wiedemann, Gregor / Niekler, Andreas (2015): "Analyse qualitativer Daten mit dem 'Leipzig Corpus Miner'", in: Lemke, Matthias / Wiedemann, Gregor (eds.):*Text Mining in den Sozialwissenschaften*. Grundlagen und Anwendungen zwischen qualitativer und quantitativer Diskursanalyse. Wiesbaden: Springer VS 63-88.

Zimmermann, Malte (2011): "Discourse particles", in: von Heusinger, Klaus / Maienborn, Claudia / Portner, Paul (eds.): *Semantics 2*(= Handbücher zur Sprach- und Kommunikationswissenschaft 33, 2). Berlin: Mouton de Gruyter 2011-2038.

"Delta" in der stilometrischen Autorschaftsattribution

Evert, Stefan
stefan.evert@fau.de
Universität Erlangen-Nürnberg, Deutschland

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

Dimpel, Friedrich Michael
friedrich.m.dimpel@fau.de
Universität Erlangen-Nürnberg, Deutschland

Schöch, Christof
christof.schoech@uni-wuerzburg.de
Universität Würzburg, Deutschland

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
Universität Würzburg, Deutschland

Vitt, Thorsten
thorsten.vitt@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reger, Isabella
isabella.reger@uni-wuerzburg.de
Universität Würzburg, Deutschland

Büttner, Andreas
andreas.buettner@uni-wuerzburg.de
Universität Würzburg, Deutschland

Proisl, Thomas
thomas.proisl@fau.de
Universität Erlangen-Nürnberg, Deutschland

Die Sektion

Stilometrische Verfahren der Autorschaftsattribution haben eine lange Tradition in den digitalen Geisteswissenschaften: Mit der Analyse der *Federalist Papers* durch Mosteller und Wallace (1963) konnten schon Anfang der 1960er Jahre Erfolge verzeichnet werden. Überblicksbeiträge von Patrick Juola (2006) und Efstathios Stamatatos (2009) belegen die Vielfältigkeit der Bestrebungen, stilometrische Verfahren für die Autorschaftsattribution einzusetzen und weiterzuentwickeln.

Ein jüngerer Meilenstein der stilometrischen Autorschaftsattribution ist ohne Zweifel das von John Burrows (2002) vorgeschlagene “Delta”-Maß zur Bestimmung der stilistischen Ähnlichkeit zwischen Texten. Die beeindruckend gute Performance von Delta in verschiedenen Sprachen und Gattungen sollte allerdings nicht darüber hinwegtäuschen, dass die theoretischen Hintergründe weitgehend unverstanden geblieben sind (Argamon 2008). Anders ausgedrückt: Wir wissen, dass Delta funktioniert, aber nicht, warum es funktioniert. In diesem Kontext möchte die hier vorgeschlagene Sektion den aktuellen Stand der Forschung in der stilometrischen Autorschaftsattribution mit Delta vorstellen und neueste Entwicklungen anhand konkreter, eigener Untersuchungen demonstrieren. Jeder der drei Vorträge der Sektion leistet hierzu einen Beitrag:

- Der Beitrag von Stefan Evert, Thomas Proisl, Fotis Jannidis, Steffen Pielström, Isabella Reger, Christof Schöch und Thorsten Vitt “Burrows Delta verstehen” (vgl. 2.), gibt einen Überlick über den Forschungsstand rund um Delta und analysiert, warum die Veränderung von Delta durch Verwendung des Kosinus-Abstands zwischen den Vektoren (Smith / Aldridge 2012) eine so deutliche Verbesserung der Ergebnisse erbracht hat (Jannidis

et al. 2015). Am Beispiel einer Sammlung deutscher Romane aus dem 19. und 20. Jahrhundert zeigt der Beitrag, wie sich verschiedene Strategien der Normalisierung oder anderweitigen Behandlung des Merkmalsvektors (hier: Wortformen und ihre Häufigkeiten) auf die Attributionsqualität auswirken und inwiefern dies Einblick darin erlaubt, wie sich Information über Autorschaft im Merkmalsvektor manifestiert - was auch einen Aspekt der Leistungsfähigkeit des klassischen Delta erklärt.

- Der Vortrag von Friedrich Michael Dimpel, "Burrows Delta im Mittelalter: Wilde Graphien und metrische Analysedaten" (vgl. 3.), beleuchtet den Einsatz unterschiedlicher Merkmalstypen für die Ähnlichkeitsbestimmung von Texten mit Delta. Er zeigt am Beispiel einer Sammlung mittelhochdeutscher Texte, dass nicht nur die äußerst häufigen Funktionswörter, sondern auch metrische Eigenschaften für die Autorschaftsattribution eingesetzt werden können. Zugleich thematisiert er ein Problem, das immer dann auftritt, wenn Texte älterer Sprachstufen stilometrisch analysiert werden: das der nicht normierten, d. h. variablen Schreibweisen von Wörtern.
- Der Beitrag von Andreas Büttner und Thomas Proisl, "Stilometrie interdisziplinär: Merkmalsselektion zur Differenzierung zwischen Übersetzer- und Fachvokabular" (vgl. 4.), 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 darin, welche Einzelmerkmale für das Autorschaftssignal statistisch gesehen entscheidend sind.

Die drei Beiträge demonstrieren auf diese Weise verschiedene aktuelle Entwicklungen in der stilometrischen Autorschaftsattribution mit Delta und seinen Varianten. Sie zeigen, wie bei der Anwendung stilometrischer Distanzmaße auf ganz unterschiedliche Gegenstandsbereiche ähnliche methodische Fragen zu berücksichtigen sind. Und sie partizipieren direkt an aktuellsten, internationalen Entwicklungen bei der Verwendung von Distanzmaßen wie Delta für die stilometrische Autorschaftsattribution.

Burrows Delta verstehen

Überblick zum Forschungsstand

Burrows Delta ist einer der erfolgreichsten Algorithmen der Computational Stylistics (Burrows 2002). In einer ganzen Reihe von Studien wurde seine Brauchbarkeit nachgewiesen (z. B. Hoover 2004, Rybicki / Eder 2011). Im ersten Schritt bei der Berechnung von Delta werden in einer nach Häufigkeit sortierten Token-Dokument-Matrix alle Werte normalisiert, indem ihre relative Häufigkeit im Dokument berechnet wird, um Textlängenunterschiede auszugleichen. Im zweiten Schritt werden alle Werte durch eine z-Transformation standardisiert:

z i D = f i D - μ i σ i

wobei *f* i(*D*) die relative Häufigkeit des Wortes *i* in einem Dokument, μ i der Mittelwert über die relativen Häufigkeiten des Wortes *i* in allen Dokumenten ist und σ i die Standardabweichung. Durch diese Standardisierung tragen alle Worte in gleichem Maße zum Differenzprofil, das im dritten Schritt berechnet wird, bei. In einem dritten Schritt werden die Abstände aller Texte voneinander berechnet: Für jedes Wort wird die Differenz zwischen dem z-Score für das Wort in dem einen Text und dem anderen Text ermittelt. Die Absolutbeträge der Differenzen werden für alle ausgewählten Wörter aufaddiert:

Δ B = ∑ i = 1 m z i D 1 - z i (D 2)

m steht für die Anzahl der häufigsten Wörter (MFW - *most frequent words*), die für die Untersuchung herangezogen werden. Diese Summe ergibt den Abstand zwischen zwei Texten; je kleiner der Wert ist, 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 (2008) zeigt, dass der dritte Schritt in Burrows Delta sich als Berechnung des *Manhattan*-Abstands zwischen zwei Punkten in einem mehrdimensionalen Raum verstehen lässt, wobei in jeder Dimension die Häufigkeit eines bestimmten Wortes eingetragen ist. Er schlägt vor, stattdessen den Euklidischen Abstand, also die Länge der direkten Linie zwischen den Punkten, zu nehmen, weil dieser „possibly more natural" (Argamon 2008: 134) sei und zudem eine wahrscheinlichkeitstheoretische Interpretation der standardisierten *z*-Werte erlaubt. Bei einer empirischen Prüfung zeigte sich, dass keiner der Vorschläge eine Verbesserung bringt (Jannidis et al. 2015).

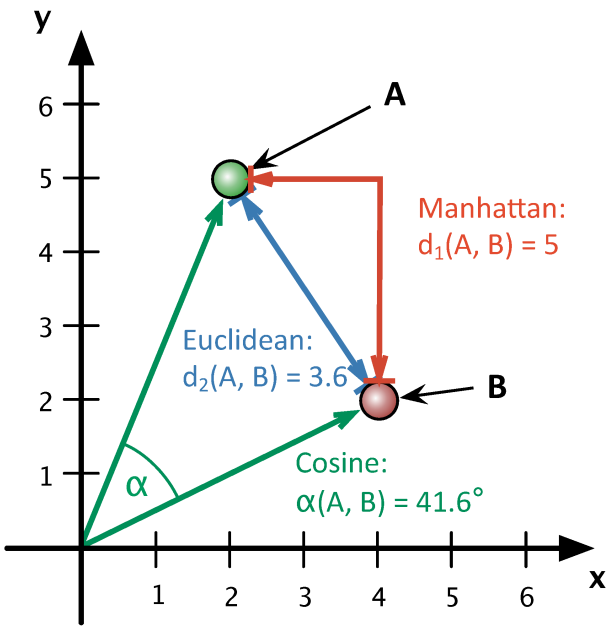

Abb. 1: Darstellung des Abstands zwischen zwei Texten, die nur aus zwei Worten bestehen. Burrows Delta verwendet die Manhattan-Distanz. Argamons Vorschlag, die Euklidische Distanz zu verwenden, sein *Quadratic-Delta*, brachte eine Verschlechterung der Clustering Ergebnisse, während der Vorschlag von Smith und Aldrige, den Cosinus-Abstand bzw. Winkel zwischen den Vektoren zu verwenden, eine deutliche Verbesserung erbrachte.

Smith und Aldrige (2011) schlagen vor, wie im Information Retrieval üblich (Baeza-Yates / Ribeiro-Neto 1999: 27), den Cosinus des Winkels zwischen den Dokumentenverktoren zu verwenden. Die Cosinus-Variante von Delta übertrifft Burrows Delta fast immer an Leistungsfähigkeit und weist, im Gegensatz zu den anderen Varianten, auch bei der Verwendung sehr vieler MFWs keine Verschlechterung auf (Jannidis et. al. 2015). Es stellt sich die Frage, warum Delta $_{Cos}$ besser ist als Delta $_{Bur}$ und ob auf diese Weise erklärt werden kann, warum Delta $_{Bur}$ so überraschend leistungsfähig ist.

Entscheidend für unsere weitere Analyse war die Erkenntnis, dass man die Verwendung des Cosinus-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. Abb. 1). Experimente haben gezeigt, dass eine explizite Vektor-Normalisierung auch die Ergebnisse der anderen Deltamaße erheblich verbessert und Leistungsunterschiede zwischen den Delta-Varianten weitgehend neutralisiert (Evert et al. 2015).

Daraus wurden zwei Hypothesen abgeleitet:

- (H1) Verantwortlich für die Leistungsunterschiede sind vor allem einzelne Extremwerte („Ausreißer“), d. h. besonders große (positive oder negative) *z*-Werte, 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 nahe liegende Erklärung für das schlechte Abschneiden von Argamons „Quadratic Delta“ Delta $_{Q}$. Der positive Effekt der Vektor-Normalisierung wäre dann so zu deuten, dass durch die Vereinheitlichung der Vektorlängen der Betrag der *z*-Werte von textspezifischen Ausreißern deutlich reduziert wird (Ausreißer-Hypothese).
- (H2) Das charakteristische stilistische Profil eines Autors findet sich eher in der qualitativen Kombination bestimmter Wortpräferenzen, 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 (Schlüsselprofil-Hypothese). 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.

Neue Erkenntnisse

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 . Die englischen Texte aus den Jahren 1838 bis 1921 kommen von Project Gutenberg und die französischen Romane von Ebooks libres et gratuits 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.

Experimente

Um die Rolle von Ausreißern und damit die Plausibilität von H1 näher zu untersuchen, ergänzen wir Delta $_{Bur}$ und Delta $_{Q}$ um weitere Delta-Varianten, die auf dem allgemeinen Minkowski-Abstand basieren:

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

Wir bezeichnen diese Abstandsmaße allgemein als L p-Delta. Der Spezialfall $p = 1$ entspricht dem Manhattan-

Abstand (also L_1-Delta = Delta$_{Bur}$), der Spezialfall $p = 2$ dem Euklidischen Abstand (also L_2-Delta = Delta$_Q$). Je größer p gewählt wird, desto stärker wird L p-Delta von einzelnen Ausreißerwerten beeinflusst.

Abbildung 2 vergleicht vier unterschiedliche L p-Abstandsmaße (für p = 1 , 2 , 2 , 4) mit Delta$_{Cos}$. Wir übernehmen dabei den methodologischen Ansatz von Evert et al. (2015): die 75 Texte werden auf Basis der jeweiligen Delta-Abstände automatisch in 25 Cluster gruppiert; anschließend wird die Güte der Autorenschaftszuschreibung 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 Leistung von L p-Delta mit zunehmendem p ab; zudem lässt die Robustheit der Maße gegenüber der Anzahl von MFW erheblich nach.

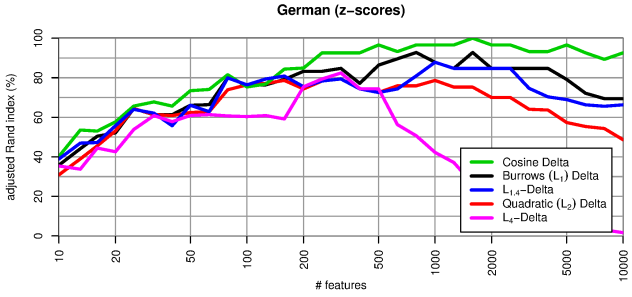

Abb. 2: Clustering-Qualität verschiedener Delta-Maße in Abhängigkeit von der Anzahl von MFW, die als Merkmale verwendet werden. Wie bereits von Janndis et al. (2015) und Evert et al. (2015) festgestellt wurde, liefert Delta$_{Bur}$ (L_1) durchgänig bessere Ergebnisse als Argamons Delta$_Q$ (L_2). Delta$_Q$ erweist sich als besonders anfällig gegenüber einer zu großen Anzahl von MFW. Delta$_{Cos}$ ist in dieser Hinsicht robuster als alle anderen Delta-Varianten und erreicht über einen weiten Wertebereich eine nahezu perfekte Autorenschaftszuschreibung (ARI > 90%).

Eine Vektor-Normalisierung verbessert die Qualität aller Delta-Maße erheblich (vgl. Abb. 3). Argamons Delta$_Q$ ist in diesem Fall identisch zu Delta$_{Cos}$: die rote Kurve wird von der grünen vollständig überdeckt. Aber auch andere Delta-Maße (Delta$_{Bur}$, $L_{1.4}$-Delta) erzielen praktisch dieselbe Qualität wie Delta$_{Cos}$. Einzig das für Ausreißer besonders anfällige L_4-Delta fällt noch deutlich gegenüber den anderen Maßen ab. Diese Ergebnisse scheinen zunächst H1 zu bestätigen.

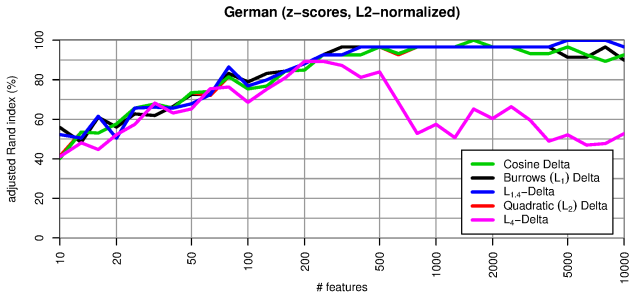

Abb. 3: Clustering-Qualität verschiedener Delta-Maße mit Längen-Normalisierung der Vektoren. In diesem Experiment wurde die euklidische Länge der Vektoren vor Anwendung der Abstandsmaße auf den Standardwert 1 vereinheitlicht.

Ein anderer Ansatz zur Abmilderung von Ausreißern besteht darin, besonders extreme z-Werte „abzuschneiden“. Wir setzen dazu alle $|z| > 2$ (ein übliches Ausreißerkriterium) je nach Vorzeichen auf den Wert +1 oder –1. Abbildung 4 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. Abschneiden großer z-Werte wirkt sich nur auf überdurchschnittlich häufige 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 der Vektor-Normalisierung.

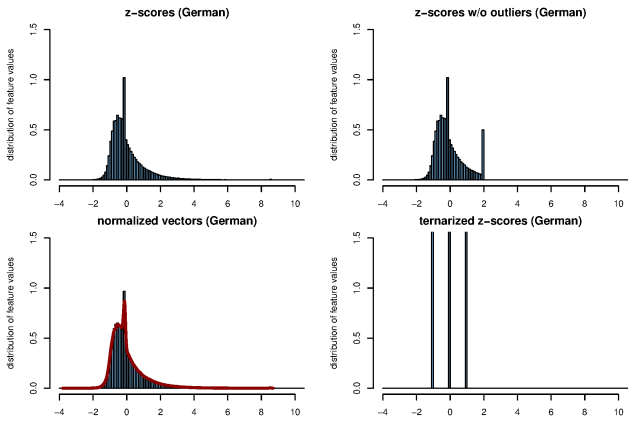

Abb. 4: Verteilung von Merkmalswerten über alle 75 Texte bei Vektoren mit 5000 MFW. Gezeigt wird die Verteilung der ursprünglichen z-Werte (links oben), die Verteilung nach einer Längen-Normalisierung (links unten), die Verteilung beim Abschneiden von Ausreißern mit $|z| > 2$ (rechts oben) sowie eine 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.

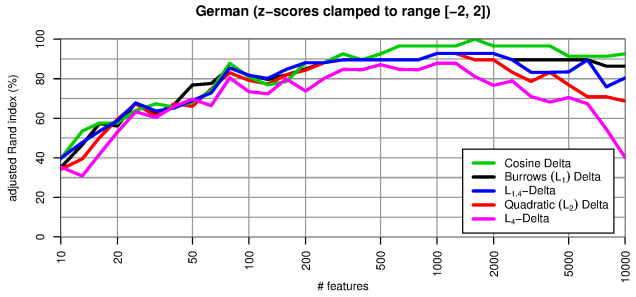

Abb. 5: Clustering-Qualität nach „Abschneiden“ von Ausreißern, bei dem Merkmalswerte $|z| > 2$ je nach Vorzeichen durch die festen Werte –2 bzw. +2 ersetzt wurden.

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. Abb. 4, rechts unten).

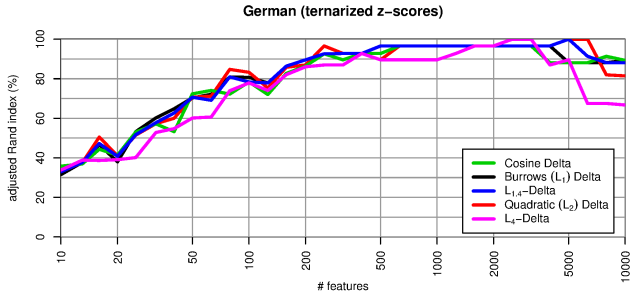

Abb. 6: Clustering-Qualität bei ternärer Quantisierung der Vektoren in überdurchschnittliche (+1, bei $z > 0.43$), unauffällige (0, bei $-0.43 < z < 0.43$) und unterdurchschnittliche (–1, bei $z < -0.43$) Häufigkeit der Wörter.

Abbildung 6 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.

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 der Abweichung über die MFW hinweg. Auffällig ist das unterschiedliche Verhalten der Maße, wenn mehr als 2000 MFW 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, und eher für Themen und Inhalte ist. Die Klärung dieser Fragen wird zusätzliche Experimente erfordern.

Burrows’ Delta im Mittelalter: Wilde Graphien und metrische Analysedaten

Einleitung

Burrows’ Delta (Burrows 2002) hat sich in Autorschaftsfragen etabliert; viele Studien zeigen, dass Delta für germanische Sprachen ausgezeichnet funktioniert (Hoover 2004b; Eder / Rybicki 2011; Eder 2013a; Eder 2013b; für das Neuhochdeutsche zuletzt Jannidis / Lauer 2014; Evert et al. 2015). Beim Mittelhochdeutschen ist jedoch die Schreibung nicht normiert: 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 (2015) hat in einer ersten Delta-Studie zum Mittelhochdeutschen diese Probleme diskutiert: Wolfram von Eschenbach benutzt zum Wort „kommen“ die Präteritalform „kom“, Hartmann von Aue verwendet „kam“, eine Form, die eher in den südwestdeutschen Raum gehört. Die Bedingungen für den Einsatz von Delta auf der Basis der *most frequent words* erscheinen auf den ersten Blick also als denkbar ungünstig; Viehhauser war 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 Eder (2013a), die den Einfluss von Noise (wie z. B. Schreibervarianten) analysiert – mit dem Ergebnis (u. a.) für das Neuhochdeutsche, dass ein zufälliger Buchstabentausch von 12% bei 100-400 MFWs die Ergebnisse kaum beeinträchtigt; bei einer mäßig randomisierten Manipulation der MFWs-Frequenzen verschlechtert sich die Quote der korrekten 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 nachdem 40% des Originalvokabulars ausgetauscht wurden.

Während die 17 Texte, die Viehhauser analysiert hat, in normalisierten Ausgaben vorliegen, habe ich zunächst 37 heterogene Texte von sieben Autoren mit Stylo-R (Eder / Kestemont et al. 2015) getestet sowie drei Texte mit fraglicher Autorzuschreibung zu Konrad von Würzburg. 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.

Per Skript wurden Längenzeichen eliminiert, damit nicht Texte mit und ohne Längenzeichen auseinander sortiert werden. Tustep-Kodierungen etwa für Superskripte habe ich 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 (diesen Parameter verwenden auch Eder 2013b und Viehhauser 2015) gelingt stylo-R ohne Pronomina und bei Culling=50% eine fehlerfreie Sortierung nach Autorschaft; Delta ordnet Rudolf zu Rudolf – ob normalisiert oder nicht.

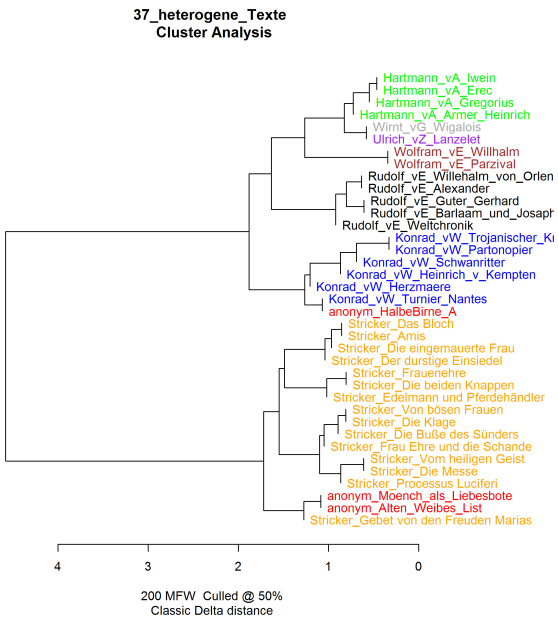

Abb. 7: Clusteranalyse

Validierungstests

Dieser Befund ist Anlass für eine Serie an automatisierten Tests in Anlehnung an Eder (2013b): Bei welchem Vektor und ab welcher Textlänge liefert Delta zuverlässige Ergebnisse? Wie wirkt sich das Einbringen von Noise aus?

Vektorlänge

Per Perlskript wurde ein Delta-Test implementiert, der in einer großen Zahl an Iterationen (13.425 Delta-Berechnungen) verschiedene „Ratetexte" mit bekannter Autorschaft gegen ein Validierungskorpus mit bekannter Autorschaft jeweils daraufhin prüft, ob für jeden Text im Ratekorpus tatsächlich der niedrigste Delta-Wert bei einem Text des gleichen Autors im Validierungskorpus herauskommt. Gegen ein heterogenes Validierungskorpus mit 18 Texten wurden 19 normalisierte Ratetexte getestet; gegen ein heterogenes Validierungskorpus mit 15 Texten wurden 13 nicht-normalisierte Ratetexte getestet. Ermittelt wurde der Prozentsatz der richtig erkannten Autoren für jeweils eine Vektorlänge; die Vektorlänge wurde in 100er Schritten bis auf 2.500 MFWs erhöht. Pronomina wurden beseitigt. Bei den normalisierten Ratetexten ist die Erkennungsquote sehr gut bis 200–900 MFWs, bei den nicht-normalisierten sehr gut für 100–600 MFWs.

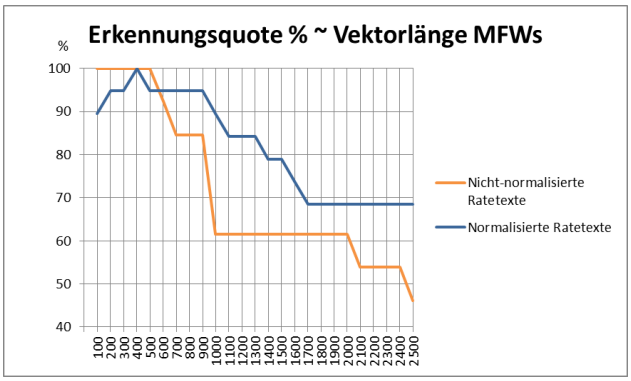

Abb. 8: Vektorlänge

Interessante Fehlattributionen – etwa Strickers ‚Pfaffe Amis' und Konrads ‚Herzmäre' – machen weitere Validierungsläufe nötig: Der normalisierte ‚Pfaffe Amis' wurde gegen einen nicht-normalisierten Stricker-Text getestet; das ‚Herzmäre' ist kurz (2991 Wörter). Während Burrows (2002) davon ausgeht, dass Delta ab einer Textlänge von 1.500 Wörtern anwendbar ist, zeigt Eder (2013b), dass Delta im Englischen ab 5.000 Wörtern sehr gute und unter 3.000 Wörtern teils desaströse Ergebnisse liefert; nur im Lateinischen werden ab 2.500 Wörtern gute Ergebnisse erreicht.

Korrelation Vektorlänge und Textlänge in konventionellen Segmentierungen

Hier wurde die Textlänge linear begrenzt, die Texte wurden nach 1000, 2000 Wörtern usw. abgeschnitten. Das Korpus ist kleiner als zuvor, da zu kurze Texte herausgenommen wurden (normalisierte: 16 Texte Validierungskorpus, 15 Ratekorpus; nicht-normalisierte 14 Validierungskorpus, 6-7 Ratekorpus; 10.056 Delta-Berechnungen).

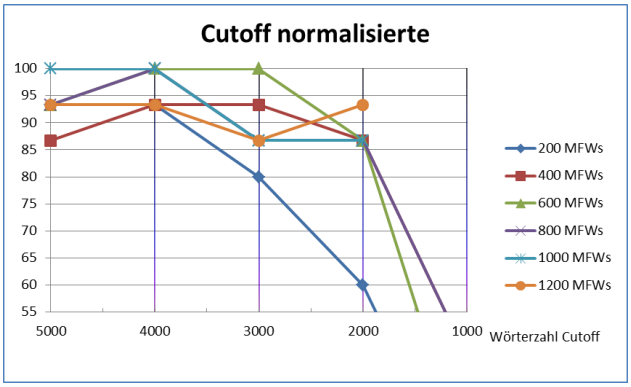

Abb. 9: Cutoff normalisierte

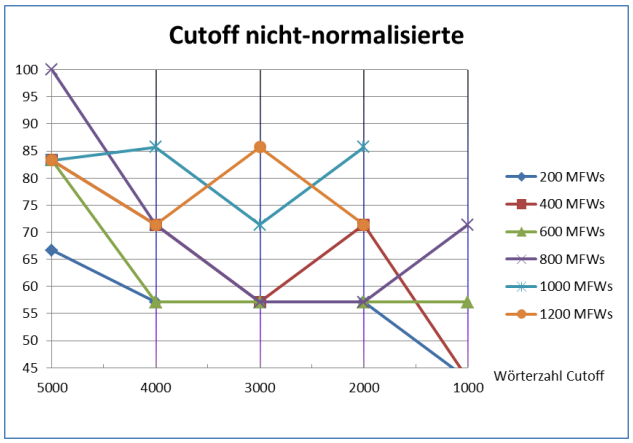

Abb. 10: Cutoff nicht-normalisierte

Korrelation Vektorlänge und Textlänge bei randomisierter Wortauswahl (‚bag-of-words'; vgl. Eder 2013b)

Gleiches Korpus wie zuvor; 167.600 Delta-Berechnungen. Da die bag-of-words randomisiert zusammengestellt wird, schwankt die Erkennungsquote etwas, daher wurde jeder Test pro Textlänge und Wortlistenlänge 25x durchgeführt und der Mittelwert dieser 25 Erfolgsquoten verwendet.

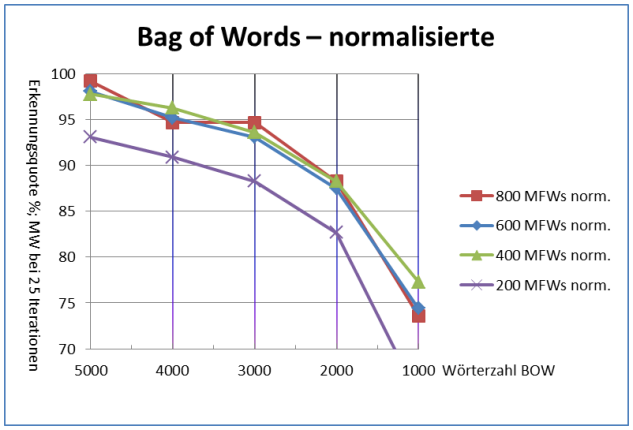

Abb. 11: bag-of-words normalisierte

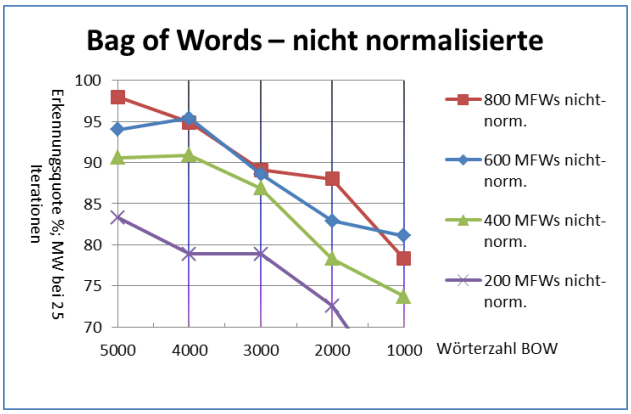

Abb. 12: bag-of-words nicht-normalisierte

Auswirkung bei der Eliminierung von Pronomina

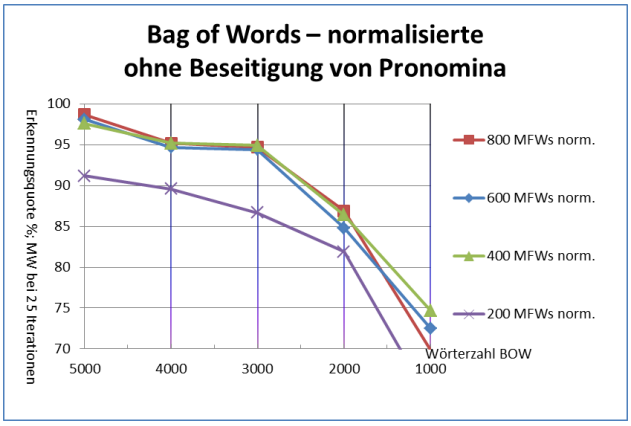

Abb. 13: bag-of-words normalisierte, ohne Beseitigung der Pronomina

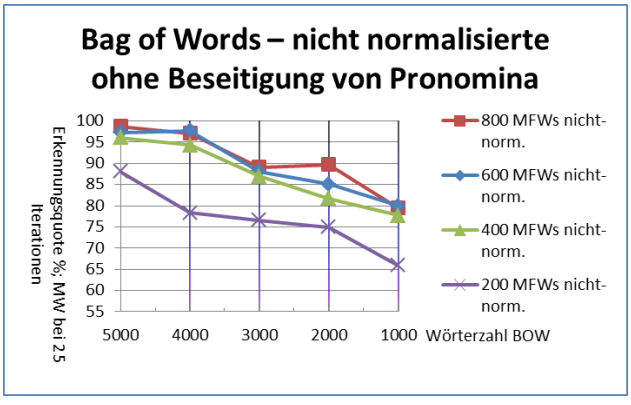

Abb. 14: bag-of-words nicht-normalisierte, ohne Beseitigung der Pronomina

Auswirkungen beim Hinzufügen von Noise

Aus einer Noise-Datei mit >18.000 mittelhochdeutschen und altfranzösischen Wortformen ohne Duplikate werden die Ratetexte prozentual aufsteigend randomisiert: Teile der bag-of-words werden gegen fremdes Sprachmaterial ausgetauscht, um Fehler in der Überlieferungskette zu simulieren. Die Kurve verläuft nicht konstant linear, da für jede bag-of-words-Berechnung erneut Noise randomisiert hinzugefügt wird (hier 10 Iterationen pro Einzelwert; 1.179.360 Delta-Berechnungen).

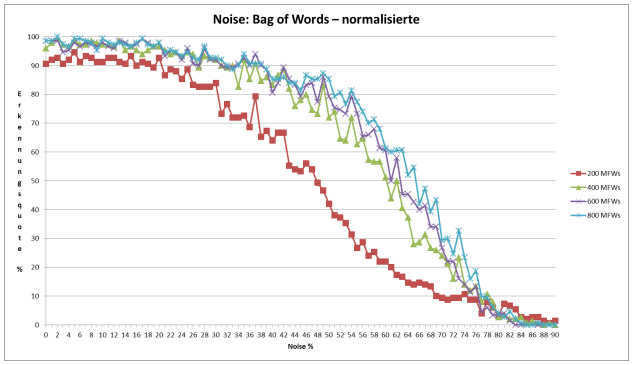

Abb. 15: Noise bei normalisierten

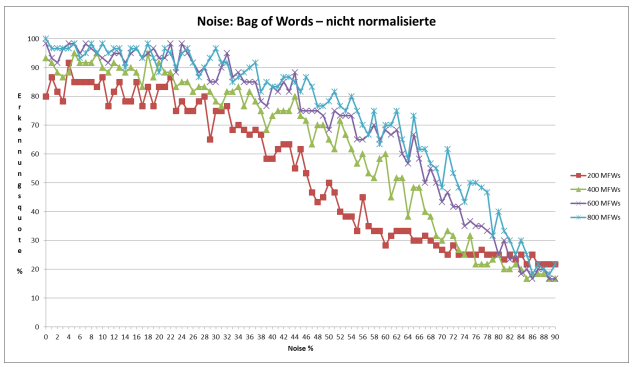

Abb. 16: Noise bei nicht-normalisierten

Beim Test der Vektorlänge (vgl. 3.2.1.) bleiben die Erkennungsquoten bei normalisierten Ratetexten sehr gut bis 200–900 MFWs. Bei den nicht-normalisierten Texten sind die Quoten nur für einen kleineren Bereich sehr gut: für 100–600 MFWs. Bei einer Begrenzung der Textlänge (Cutoff; vgl. 3.2.2.) bleiben die Ergebnisse bei normalisierten Texten nur ab 4000 Wörtern Textlänge weitgehend gut bis sehr gut. Schlecht sieht es bei den nicht-normalisierten Texten aus: Sehr gut ist die Quote nur bei 800 MFWs und 5000 Wörtern, ansonsten weithin desaströs. Bag-of-words (vgl. 3.2.3.) bieten dagegen stabilere Ergebnisse: Bei den normalisierten Texten sind bei einer Textlänge von 5000 die Quoten sehr gut bei

400-800 MFWs. Bei den nicht-normalisierten Texten ist die Quote wiederum nur bei Textlänge 5000 und 800 MFWs sehr gut. Bei kürzeren Texten und anderen Frequenzen verschlechtern sich die Quoten massiv, allerdings bleiben sie noch deutlich besser als beim Cutoff-Test. Bei normalisierten Texten werden durch das Eliminieren von Pronomina geringfügig bessere Quoten erreicht (vgl. 3.2.4.), bei nicht-normalisierten Texten etwas schlechtere Quoten.

Stabil bleiben die Quoten bei normalisierten Texten nach dem Einbringen von Noise (vgl. 3.2.5.): Solange nicht mehr als 17% des Vokabulars ausgetauscht wurden, werden die Erkennungsquoten nur etwas schlechter. 600-800 MFWs liefern sehr gute Erkennungsquoten bis 20%. Auch die Quoten bei nicht-normalisierten Texte sind einigermaßen stabil, solange nicht mehr als 20% Noise eingebracht werden: Der Bereich von 600-800 MFWs liefert bis 9% Noise noch sehr gute und bis 22% noch gute Ergebnisse.

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. Auch wenn die Ergebnisse für nicht-normalisierte Texte etwas zurückfallen, hat mich angesichts der wilden mittelhochdeutschen Graphien doch überrascht, dass die Delta-Performanz derart robust bleibt. Während jedoch etwa Eder Validierungsstudien mit über 60 Texten durchführen konnte, ist es um die digitale Verfügbarkeit von längeren mittelhochdeutschen Texten, von denen mindestens zwei Texte vom gleichen Autor verfasst wurden, derzeit noch deutlich schlechter bestellt. Die Aussagekraft der vorliegenden Studien wird daher durch die Korpusgröße v. a. bei den nicht-normalisierten Texten limitiert.

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 und Feiguina (2007) erzielen Tests auf der Basis von Part-of-Speech–Bigrammen gute Ergebnisse; für das Mittelhochdeutsche ist jedoch noch kein Part-of-Speech-Tagger in Sicht.

2014 habe ich das Metrik-Modul aus meiner Dissertation grundlegend überarbeitet, die Fehlerquote reduziert (nun unter 2%) und es ins Internet zur freien Benutzung eingestellt (Dimpel 2015). Dieses Modul gibt Kadenzen aus (etwa „weiblich klingend“). Die metrische Struktur wird mit „0“ (unbetonte Silben) und „1“ (betonte Silben) ausgegeben; der dritte ‚Parzival‘-Vers hat das Muster „01010011“. Anstatt mit MFWs habe ich Metrikmuster und Kadenzinformationen verwendet und so die Metrikdaten als „Worte“ testen lassen. Da ein weniger variationsreiches Ausgangsmaterial verwendet wird, habe ich wie Hirst und Feiguina (2007) mit Bigrammen gearbeitet.

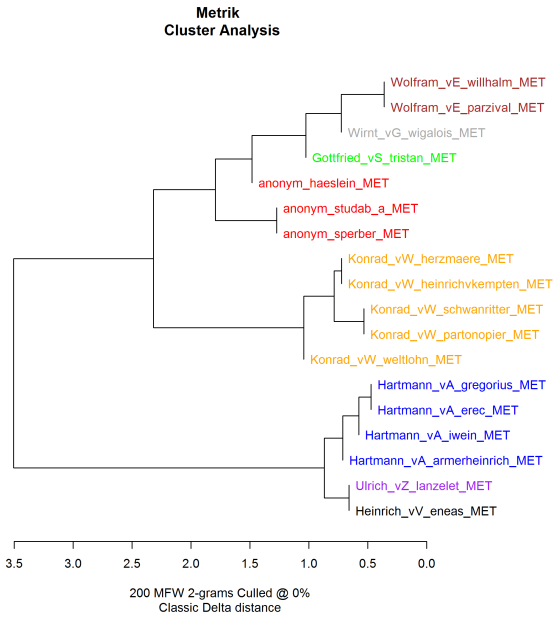

Abb. 17: Clusteranalyse auf Basis von Metrik-Bigrammen

Auch ein Metrik-Delta-Plot mit Stylo-R clustert Autoren hier fehlerlos. Validierungstests sind bislang nur mit einem kleineren Korpus möglich, da das Metrikmodul Längenzeichen benötigt und nur für Texte mit vierhebigen Reimpaarversen konstruiert ist. Bei 13 Ratedateien und 11 Validierungsdateien ergibt sich bei 250–300 „MFWs“ eine Erkennungsquote von 92,3%. Ein erfreuliches Ergebnis: (1) Bei Tests auf Grundlage von Metrik-Daten ist eine etwas geringere Abhängigkeit von Schreibergraphie und von Normalisierungsgewohnheiten gegeben. Zwar hat es mitunter metrische Eingriffe der Herausgeber gegeben, aber längst nicht immer. 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“ oder „unt“. (2) Zudem kann Autorschaft offenbar nicht nur mit dem vergleichsweise einfachen Parameter MFWs dargestellt werden: Nicht nur eine pure Wortstatistik führt zum Ziel, vielmehr erweist sich auch die Kompetenz zum philologischen Programmieren und zur filigranen Textanalyse als fruchtbar. (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 MFWs autorspezifisch sind, gelingt es nun auch über ein wohl oft bewusst gestaltetes Stilmerkmal, Autorschaft zu unterscheiden. Man muss also den Dichtern nicht nur einen unbewussten stilistischen Fingerabdruck zutrauen, vielmehr lässt sich Autorschaft zumindest hier über ein

Merkmal erfassen, das dem bewussten künstlerischen Zugriff unterliegen kann.

Stilometrie interdisziplinär: Merkmalsselektion zur Differenzierung zwischen Übersetzer- und Fachvokabular

Einleitung

Stilometrie ist der Versuch, sprachliche Besonderheiten durch statistische Methoden herauszustellen und zu vergleichen, um damit unter anderem Rückschlüsse auf die Urheberschaft eines Textes ziehen zu können. Als probates Mittel bei der Autorschaftsattribuierung hat sich die Analyse der Verwendung der häufigsten Wörter bewährt. Insbesondere Varianten des von Burrows (2002) vorgeschlagenen Deltamaßes haben sich als sehr erfolgreich erwiesen (Hoover 2004a; Eder / Rybicki 2011). Faktoren der Zusammensetzung des Textkorpus, die sich negativ auf die Qualität der Ergebnisse auswirken können, sind unter anderem zu kurze Texte (Eder 2015), unterschiedliche Genres der Texte (Schöch 2013) und eine Überlagerung von Autor- und Übersetzerstilen (Rybicki 2012). Gerade inhaltliche Unterschiede zwischen Texten stellen ein Hindernis bei der Erkennung der Autoren dar, das nur mit erheblichem technischen Aufwand überwunden werden kann (Stamatatos et al. 2000; Kestemont et al. 2012).

In unserem Beitrag verwenden wir Deltamaße zur Identifikation von Übersetzern. Textgrundlage ist eine Sammlung von im 12. Jahrhundert entstandenen arabisch-lateinischen Übersetzungen wissenschaftlicher Texte aus verschiedenen Disziplinen. Wir zeigen eine Möglichkeit auf, wie die aus den oben genannten Faktoren resultierenden Limitierungen durch den Einsatz maschineller Lernverfahren kompensiert werden können. Gleichzeitig eröffnet sich dadurch eine Möglichkeit, unter den häufigsten Wörtern solche zu identifizieren, die eher Informationen zum Übersetzer oder eher zur Disziplin tragen.

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 (Hasse / Büttner in Vorbereitung)[1] . Es handelt sich dabei um Texte unterschiedlicher Autoren 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 5 Übersetzern, wobei die Texte aus 4 Disziplinen stammen (siehe Abb. 18). 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 200000 Wörtern; insgesamt sind die Texte auch deutlich kürzer als diejenigen der oft verwendeten Romankorpora (vgl. etwa Jannidis et al. 2015).

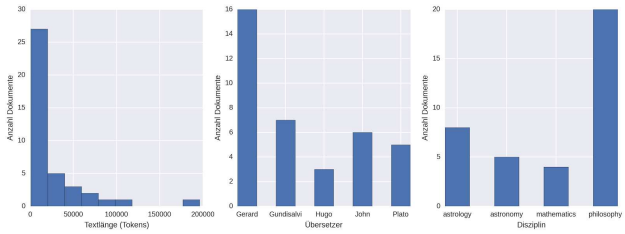

Abb. 18: Verteilung der Textlängen, Übersetzer und Disziplinen im verwendeten Teilkorpus

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

Methoden

Delta-Maße

Ausgehend von Burrows ursprünglichem Deltamaß (Burrows 2002) wurde eine ganze Reihe von Deltamaßen für die Autorschaftszuschreibung vorgeschlagen (bspw. Hoover 2004b, Argamon 2008, Smith / Aldridge 2011, Eder et al. 2013). Alle Maße operieren auf einer Term-Dokument-Matrix der *n* häufigsten Terme im Korpus, die die relativen Häufigkeiten der Terme in den einzelnen Dokumenten enthält. In einem ersten Schritt werden die relativen Häufigkeiten der Terme standardisiert (üblicherweise durch eine *z*-Transformation) um die Größenordnungsunterschiede, die sich durch die Zipfsche Verteilung der Worthäufigkeiten ergeben, zu beseitigen. Im optionalen zweiten Schritt können die Dokumentvektoren normalisiert, d. h. auf Länge 1 gebracht werden. Im dritten Schritt wird die

Ähnlichkeit zwischen zwei Dokumentvektoren durch ein Ähnlichkeits- oder Abstandsmaß bestimmt (bei Burrows Delta wird bspw. die Manhattan-Distanz verwendet, bei Kosinus-Delta der Kosinus des Winkels zwischen den beiden Dokumentvektoren). Auf Basis der so erhaltenen Ähnlichkeitswerte können die Dokumente dann geclustert werden, wobei idealerweise Texte desselben Autors im selben Cluster landen.

Für die folgenden Experimente verwenden wir Kosinus-Delta, das sich unter anderem bei Jannidi, Pielström, Schöch und Vitt (2015) sowie Evert, Proisel und Jannidis et al. (2015) als das robusteste Mitglied der Delta-Familie erwiesen hat.

Rekursive Merkmalseliminierung

Rekursive Merkmalseliminierung (recursive feature elimination, RFE) ist eine von Guyon, Weston, Barnhill und Vapnik (2002) vorgeschlagene Methode zur Selektion einer möglichst kleinen Teilmenge von Merkmalen, mit der trotzdem möglichst optimale Ergebnisse mit einem überwachten maschinellen Lernverfahren erzielt werden können. Evert, Proisel und Jannidis et al. (2015) experimentieren zur Autorschaftszuschreibung mit durch RFE ermittelten Termen als Alternative zu den üblichen *n* häufigsten Termen.

Da RFE auf einem überwachten Lernverfahren (üblicherweise einem *Support Vector Classifier*) 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 (Termen), 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 die 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 Merkmalsmenge zu bestimmen.

Experimente

Zunächst führen wir mit dem Testkorpus einige Versuche zur Anpassung der stilometrischen Methoden durch. Als Maß der Qualität des Clusterings dient dabei 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 19 dargestellt, wird bei Verwendung von Kosinus-Delta der höchste ARI für das Clustering der Übersetzer bereits bei etwa 300–400 der häufigsten Wörter erreicht, bei über 1000 Wörtern fällt das Qualitätsmaß stark ab. Ein Clustering nach Disziplinen hingegen erreicht bei ca. 500–700 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.

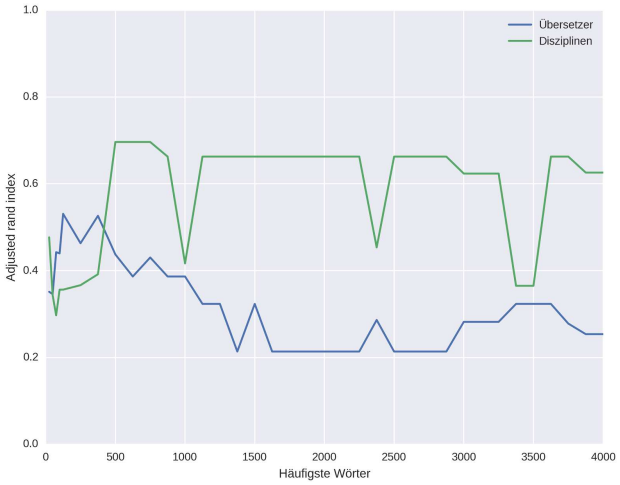

Abb. 19: Clusteringqualität in Abhängigkeit von der Anzahl der häufigsten Wörter

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 0,437 erreicht wird, als Vergleichsmaßstab für die weiteren Versuche dienen. Für ein Clustering nach Disziplinen wird mit diesen Wörtern ein ARI $_{D}$ von 0.696 erreicht.

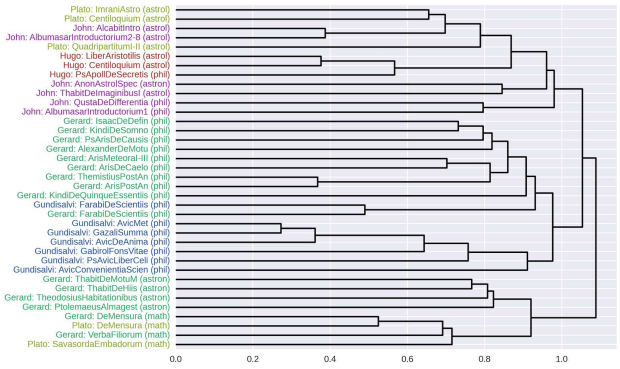

Abb. 20: Dendrogramm für das Clustering mit MFW500, Einfärbung nach Übersetzern

Durch RFE wählen wir aus der Gesamtmenge weniger als 500 Wörter aus. Mit 483 Wörtern ist eine perfekte 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 475 Wörtern finden, bei der die Texte sich perfekt 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 68 Wörtern ist immer noch eine sehr gute, wenn auch nicht perfekte Unterscheidung der Übersetzer möglich (ARI Ü=0,910). Disziplinen lassen sich mit diesen Merkmalen nur sehr schlecht unterscheiden (ARI D=0,162).

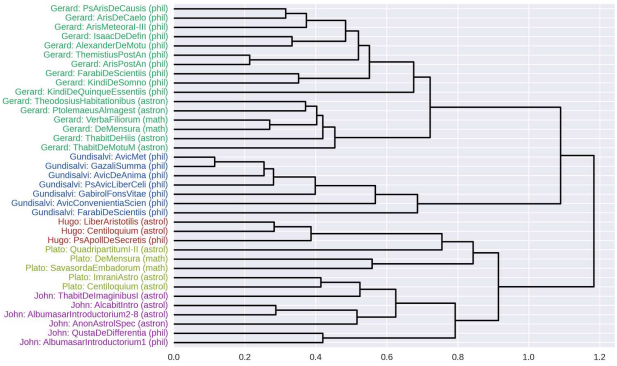

Abb. 21: Dendrogramm für das Clustering mit der Schnittmenge aus RFE und MFW500, Einfärbung nach Übersetzern

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.

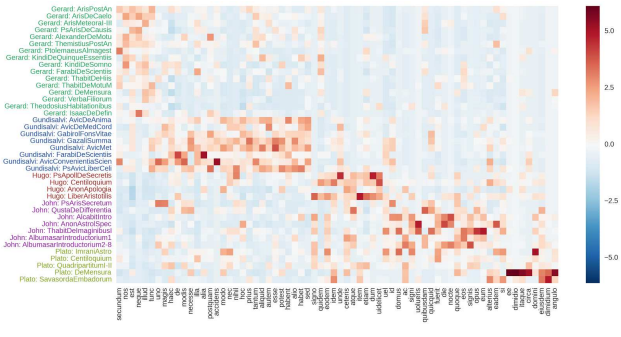

Abb. 22: Heatmap der *z*-Werte aus der Schnittmenge von RFE und MFW500

Die 432 Wörter aus MFW500, die in der Menge der RFE-selektierten Wörter nicht enthalten sind, unterscheiden, wie erwartet, deutlich schlechter zwischen Übersetzern (ARI Ü=0,222), dafür aber sehr gut zwischen Disziplinen (ARI D=0,727) – überraschenderweise sogar deutlich besser als alle 500 Wörter aus MFW500.

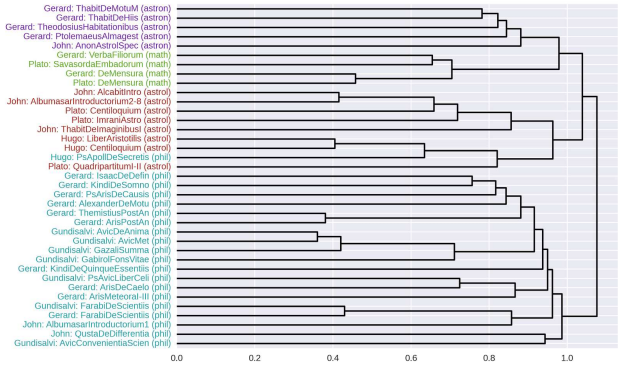

Abb. 23: Dendrogramm für das Clustering mit der Differenzmenge aus MFW500 und RFE, Einfärbung nach Disziplinen

Bei den Disziplinen erzielt die Schnittmenge der dafür mit RFE ausgewählten Wörter mit MFW500 sogar perfekte Ergebnisse (Anzahl der Merkmale: 109, ARI=1,0). Die Differenzmenge zeigt hier allerdings nicht den oben beschriebenen Effekt. Zwar ist die Clusteringqualität nach Disziplinen deutlich schlechter als der mit MFW500 erzielte Wert (ARI D=0,384), die nach Übersetzern aber ebenfalls (ARI Ü=0,198).

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

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 Textkorpora aus.

Notes

1. Siehe hierzu auch die Projekthomepage des Digital Humanities-Zentrums KALLIMACHOS der Universität Würzburg http://kallimachos.de/project/doku.php/kallimachos:identifizierunguebersetzer:start

Bibliographie

Argamon, Shlomo (2008): "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations", in: *Literary and Linguistic Computing* 23, 2: 131–47. 10.1093/llc/fqn003 .

Baeza-Yates, Ricardo / Ribeiro-Neto, Berthier (1999): *Modern Information Retrieval*. Harlow: Addison-Wesley.

Burrows, John (2002): "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship", in: *Literary and Linguistic Computing* 17, 3: 267–87. 10.1093/llc/17.3.267 .

Dimpel, Friedrich Michael (2015): "Automatische Mittelhochdeutsche Metrik 2.0", in: *Philologie im Netz* 73: 1–26 http://web.fu-berlin.de/phin/phin73/p73i.htm [letzter Zugriff 26. Januar 2016].

Eder, Maciej (2013a): "Mind Your Corpus: systematic errors in authorship attribution", in: *Literary and Linguistic Computing* 28: 603-614. 10.1093/llc/fqt039 .

Eder, Maciej (2013b): "Does size matter? Authorship attribution, small samples, big problem", in: *Literary and Linguistic Computing Advanced Access* 29: 1-16. 10.1093/llc/fqt066 .

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

Eder, Maciej / Kestemont, Mike / Rybicki, Jan (2013): "Stylometry with R: a suite of tools", in: *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska 487–489 http://dh2013.unl.edu/abstracts/ab-136.html [letzter Zugriff 26. Januar 2016].

Eder, Maciej / Kestemont, Mike / Rybicki, Jan(2015): "stylo R package" https://sites.google.com/site/computational-stylistics/stylo [letzter Zugriff 20. März 2015].

Eder, Maciej / Rybicki, Jan (2011): "Deeper Delta across genres and languages: do we really need the most frequent words?", in: *Literary and Linguistic Computing* 26, 3: 315–321. 10.1093/llc/fqr031 .

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): "Towards a better understanding of Burrows's Delta in literary authorship attribution", in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, Denver 79–88. 10.5281/zenodo.18177 . http://www.aclweb.org/anthology/W/W15/W15-0709.pdf [letzter Zugriff 20. August 2015].

Guyon, Isabelle / Weston, Jason / Barnhill, Stephen / Vapnik, Vladimir (2002): "Gene Selection for Cancer Classification using Support Vector Machines", in: *Machine Learning* 46, 1: 389–422. 10.1023/A:1012487302797 .

Hasse, Dag Nikolaus / Büttner, Andreas (in Vorbereitung): "Notes on the Identity of the Latin Translator of Avicenna's Physics and on Further Anonymous Translations in Twelfth-Century Spain." Vorabversion: https://go.uniwue.de/hassevigoni [letzter Zugriff 17. Februar 2016].

Hirst, Graeme / Feiguina, Ol'ga (2007): "Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts", in: *Literary and Linguistic Computing Advance Access* 22: 1–13. 10.1093/llc/fqm023 .

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

Hoover, David L. (2004b): "Delta Prime?", in: *Literary and Linguistic Computing* 19, 4: 477–495. 10.1093/llc/19.4.477 .

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

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): "Improving Burrows' Delta - An Empirical Evaluation of Text Distance Measures", in: *Digital Humanities Conference 2015*, Sydney http://dh2015.org/abstracts/xml/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empi/JANNIDIS_Fotis_Improving_Burrows__Delta___An_empirical_.html [letzter Zugriff 26. Januar 2016].

Juola, Patrick (2006): "Authorship Attribution", in: *Foundations and Trends in Information Retrieval* 1, 3: 233–334.

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

Mosteller, Frederick / Wallace, David L. (1963): "Inference in an Authorship Problem", in: *Journal of the American Statistical Association* 58, 302: 275–309. 10.2307/2283270 .

Rybicki, Jan (2012): "The great mystery of the (almost) invisible translator: stylometry in translation", in: Oakley, Michael P. / Ji, Meng (eds.):

Quantitative Methods in Corpus-Based Translation Studies. Amsterdam: John Benjamins 231–248 https://sites.google.com/site/computationalstylistics/preprints/Rybicki%20Great%20Mystery.pdf [letzter Zugriff 26. Januar 2016].

Schöch, Christof (2013): "Fine-Tuning our Stylometric Tools: Investigating Authorship and Genre in French Classical Drama", in: *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska 383–386 http://dh2013.unl.edu/abstracts/ab-270.html [letzter Zugriff 26. Januar 2016].

Smith, Peter W. H. / Aldridge, W. (2011): "Improving Authorship Attribution: Optimizing Burrows' Delta Method*", in: *Journal of Quantitative Linguistics* 18, 1: 63–88. 10.1080/09296174.2011.533591 .

Stamatatos, Efstathios (2009): "A Survey of Modern Authorship Attribution Methods", in: *Journal of the Association for Information Science and Technology* 60, 3: 538–56. 10.1002/asi.v60:3 .

Stamatatos, Efstathios / Fakotakis, Nikos / Kokkinakis, George (2000): "Automatic Text Categorization in Terms of Genre and Author", in: *Computational Linguistics* 26, 4: 471–497. 10.1162/089120100750105920 .

TextGrid Konsortium (2006–2015): *TextGrid*. Virtuelle Forschungsumgebung für die Geisteswissenschaften. Göttingen: https://textgrid.de .

Viehhauser, Gabriel (2015): "Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte", in: Baum, Constanze / Stäcker, Thomas (eds.): *Grenzen und Möglichkeiten der Digital Humanities* (Sonderband der Zeitschrift für digitale Geisteswissenschaften 1).

Mobile Anwendungen als multimodale Medien zur Vermittlung vormoderner Artefakte. Die ‚Historisches Paderborn'-App – ein interdisziplinäres Forschungs- und Lehrprojekt

Greulich, Markus
markus.greulich@uni-paderborn.de
Universität Paderborn, Deutschland

Oberthür, Simon
oberthuer@uni-paderborn.de
SICP – Software Innovation Campus Paderborn, Universität Paderborn, Deutschland

Karthaus, Nicola
karthaus@ieman.de
Universität Paderborn, Deutschland

Schmidt, Ariane
arianes@mail.uni-paderborn.de
Universität Paderborn, Deutschland

Wilk, Nicole M.
nicole.m.wilk@upb.de
Universität Paderborn, Deutschland

Stog, Kristina
kristina.stog@uni-paderborn.de
Universität Paderborn, Deutschland

Senft, Björn
bjoern.senft@uni-paderborn.de
SICP – Software Innovation Campus Paderborn, Universität Paderborn, Deutschland

Zusammenfassung der Sektion

Die Vermessung der Welt mittels digitaler Medien hat längst begonnen. Von der Durchdringung der Gesellschaft zeugen nicht nur Street View und weltweit verfügbare Satellitenaufnahmen, sondern auch Twitter und Facebook und nicht zuletzt die Auswirkungen auf die Wissenschaftskulturen, insbesondere auf die der Geisteswissenschaften. Hierfür hat sich der Begriff der Digital Humanities etabliert, der schillernd und komplex zugleich ist. Während zunächst historisches Material und Artefakte digitalisiert wurden, rückte in den letzten Jahren vor allem die Annotation von Digitalisaten und die Anlage und Aufbereitung von Datenbanken ins Zentrum des Interesses. Derzeit fächert sich das Spektrum der Digital Humanities weiter auf.

Anne Burdick, Johanna Drucker und andere (Burdick et al. 2012) weisen in ihrem intensiv rezipierten und vielfach zitierten Buch zum Konzept der Digital Humanities darauf hin, dass die Möglichkeiten und Chancen der Digital Humanities quasi einer Erweiterung der Geisteswissenschaften gleichkommen, die sowohl Werte, interpretative Praxis und Strategien der Bedeutung als auch die Ambiguitäten der menschlichen Existenz betreffen. Innerhalb der Sektion soll der Blick insbesondere auf zwei auch von Burdick und Drucker thematisierte Aspekte gelenkt werden: Zum einen ermöglicht die Erweiterung der Digital Humanities neue Wege der transmedialen Erforschung durch interdisziplinäre Kooperationen. Zum anderen darf nicht