Autorschaftsattribution bei nicht-normalisiertem Mittelhochdeutsch. Bessere Erkennungsquoten durch ein Normalisierungswörterbuch

Dimpel, Friedrich Michael
mail@dimpel.de
FAU Erlangen-Nürnberg, Deutschland

Einleitung: Delta im Mittelalter und in der Forschung

Im Bereich der Autorschaftsattribution sind in den letzten Jahren große Fortschritte erzielt worden; insbesondere der Delta-Test nach Burrows' (2002) und Varianten zu Burrows' Verfahren haben sich in vielen Validierungsstudien als sehr erfolgreich erwiesen (Hoover 2004, Eder / Rybicki 2011, Eder 2013a, Eder 2013b, Jannidis / Lauer 2014, Evert / Proisl / Jannidis / Pielström / Schöch / Vitt 2015, Evert / Proisl / Jannidis / Pielström / Reger/ Schöch / Vitt 2016). In mittelalterlichen Texten stellen sich jedoch besondere Probleme: Hier ist die Schreibung weitgehend nicht normiert, das Wort ‚und' wird teilweise mit ‚u' oder ‚v', mit weichem ‚d' oder hartem ‚t' geschrieben; die Genitiv-Form zum nhd. Wort ‚Gott' lautet ‚gotes' oder ‚gotis' (Viehhauser 2015).

Im Rahmen eines Vortrags auf der DHd-Tagung 2016 in Leipzig konnte ich zeigen (Dimpel 2016), dass Delta bei normalisierten mittelhochdeutschen Texten sehr gut funktioniert, insbesondere dann, wenn Texte verwendet werden, die aus mindestens 5.000 Wörtern bestehen, und wenn die Bag-of-Words-Technik (vgl. Eder 2013b) zum Einsatz kommt. Um die Erkennungsquote zu ermitteln, habe ich ein „Ratekorpus" und ein „Validierungskorpus" gebildet. In beiden Sammlungen sind Texte mit bekannter Autorschaft enthalten. Zu jedem Autor, der im Ratekorpus enthalten ist, ist jeweils ein Text des gleichen Autors im Validierungskorpus enthalten. Ermittelt wurde der Prozentsatz der richtig erkannten Autoren. Bei einem Test mit 16 Texten im Validierungskorpus und 15 Texten im Ratekorpus wurde eine Erkennungsquote von 97,1% ermittelt.

Erster Validierungstest nicht-normalisierte Texte

Zu nicht-normalisierten Texten habe ich 2016 erste Zahlen ebenfalls mit positivem Ergebnis vorlegen können, die allerdings nicht valide sind, weil mir zu diesem Zeitpunkt nur sehr wenige nicht-normalisierte Texte digital verfügbar waren: bei einer Textlänge von 5.000 Wörtern konnte ich gegen ein Validierungskorpus mit 14 Texten nur 6 Texte von 5 Autoren prüfen. Nunmehr liegen weitere Texte vor, so dass für einen Validierungstest nun ein Ratekorpus mit 15 Texten von 10 Autoren zur Verfügung steht. Im Validierungskorpus ist je ein Text dieser 10 Autoren enthalten, dazu kommen weitere 10 Texte, die Fehlattributionen auslösen könnten.

Dass Delta bei nicht-normalisierten Texten schlechtere Erkennungsquoten liefert, ist deshalb zu erwarten, weil Delta auf der Verteilung von hochfrequenten Wörtern beruht. Wenn im Werk X überwiegend ‚unt' steht, wenn sich im Werk Y des gleichen Autors jedoch der Abschreiber für die Graphie ‚vnnd' entschieden hat, wird die Zuordnung des richtigen Autors dadurch erschwert. Erwartungsgemäß liegt die Erkennungsquote mit ca. 80% (bei Bag-of-Words mit 5.000 Wörtern; 50 Iterationen zum Ausgleich von Zufallsschwankungen bei der Bag-of-Words-Bildung; davon der Mittelwert für die Vektoren 200, 400, 600 und 800) deutlich unter der Quote für normalisierte Texte. Um eine Verbesserung der Erkennungsquote zu ermöglich, wurden nun Ansätze zur automatischen Teilnormalisierung erprobt.

Teilnormalisierung: Normalisierungswörterbuch und Vollformenwörterbuch

Ein erster Schritt dabei ist die Eliminierung der Sonderzeichen und der deutschen Umlaute – auch in dem soeben erwähnten Test war diese Bereinigung bereits implementiert. Eine weitere automatische Teilnormalisierung ist deshalb ökonomisch realisierbar, weil für den Delta-Test keine vollständige Normalisierung nötig ist. Weil Delta auf den hochfrequenten Wörtern beruht, sollte bereits eine Normalisierung der häufigen Wörter zu einer Verbesserung führen.

In einem nächsten Schritt wurde ein Skript entwickelt, das aus einigen kürzeren nicht-normalisierten Texten die hochfrequenten Wortformen heraussucht und den User bittet, die normalisierte Form zuzuordnen. Eine Vorschlagsliste aus einem normalisierten Lachmann-Korpus, die mittels Levenshtein-Distanz generiert wurde, hat meiner Hilfskraft das Leben leichter gemacht. Zudem habe ich zwei Datengeschenke bekommen: Sonja Glauch hat mir Daten aus dem Projekt „Lyrik des Mittelalters" gegeben, das eine Zuordnung von normalisierten zu nicht-normalisierten Wortformen herstellt. Mit den Skript-Daten und den Lyrik-Projekt-Daten lag ein Normalisierungswörterbuch mit gut 1.100 Zuordnungen vor, als mir Thomas Klein Daten aus dem Referenzkorpus Mittelhochdeutsch überlassen hat. Die vorbildliche Struktur des Referenzkorpus hat es möglich

gemacht, weitere Zuordnungen zu extrahieren und sie in das Normalisierungswörterbuch einspeisen, das nunmehr gut 120.000 Zuordnungen enthält.

Eine Sichtung des Normalisierungswörterbuchs hat jedoch gezeigt, dass teilweise auch solche diplomatische Wortformen wie ‚sluc' zu ‚sluoc' normalisiert werden, die eigentlich auch selbst als normalisierte Form eines anderen Lemmas stehen könnten: ‚sluc' kann als starkes Femininum etwa nhd. „ein Schluck" heißen und müsste dann nicht durch eine andere normalisierte Form ersetzt werden. Mitunter wurde im Lyrikprojekt und im ReM unerwartet normalisiert: So findet sich bspw. eine Normalisierung der Wortform ‚chunich' zu ‚küninc', während im BMZ und im Lachmann-Parzival meist ‚künec' steht.

Um das Normalisierungswörterbuch zu überprüfen und vereinheitlichen zu können, wurde ein mittelhochdeutsches Vollformenwörterbuch benötigt, das die Wortformen enthält, die zu normalisierten Lemmata durch Flexion gebildet werden können.

Aus der CD „Mittelhochdeutsche Wörterbücher im Verbund" (Trier, Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften 2002) wurden Daten extrahiert und mögliche Flexionsformen zu den Lemmata generiert. Wenn es auf Vollständigkeit und Korrektheit ankommen würde, wäre die Erstellung eines derartigen Vollformenwörterbuchs eine große Herausforderung. Doch geht es sowohl bei dem Normalisierungswörterbuch als auch bei dem Vollformenwörterbuch hier nur darum, eine prozentuale Verbesserung der Delta-Erkennungsquote zu erreichen. Fehler bei seltenen Wortformen sind bei Delta meist zu vernachlässigen, wichtig ist eine Vereinheitlichung der häufigen Wörter. Manche Probleme haben sich überwiegend erfreulich lösen lassen: Zu Ablaut und grammatischem Wechsel sind Informationen BMZ hinterlegt. Schwache Verben mit sogenanntem Rückumlaut generiert das Skript dann, wenn im Singular Präsens ein umgelauteter Vokal sowie Positionslänge oder Naturlänge vorliegen. In der verfügbaren Zeit wurde Vieles nicht vollständig gelöst – für Nomina mit Umlaut müssten im Artikel noch die Belegstellen examiniert werden, hier wird bislang nur der Artikelkopf des Lexer ausgewertet. Funktionswörter sind überwiegend listenbasiert ergänzt.

Eine Evaluierung hat gezeigt, dass eine vollständige Bereinigung des Normalisierungswörterbuchs um Formen wie ‚sluc' zu einer minimalen Verschlechterung der Erkennungsquote führt, so dass die Eliminierung von diplomatischen Wortformen, die auch normalisierte Wortformen sein könnten, bei hoch- und mittelfrequenten Wortformen nicht angewendet wurde.

Delta-Verbesserung: Z-Wert-Begrenzung

Bei Delta berechnet man bspw. für 200 Most-Frequent-Words für jedes dieser Worte einzeln Z-Werte, in die die Abweichung der Häufigkeit eines Wortes in einem Text zur Häufigkeit dieses Wortes im Gesamtkorpus unter Berücksichtigung der Standardabweichung eingeht. Delta ist das arithmetische Mittel der positiven Z-Wert-Differenzen (Burrows 2002). Evert / Proisl / Jannidis / Pielström / Reger/ Schöch / Vitt 2016 haben meinen Verdacht evaluiert, dass Delta weniger aufgrund einzelner Extremwerte funktioniert (Ausreißerhypothese), sondern eher aufgrund einer breiten autorspezifischen Verteilung der Z-Werte (Schlüsselprofilhypothese).

Fehlt eine Wortform in einem Text, kann dies mitunter mit erhöhten negativen Z-Werten einhergehen. Evert et alia haben besonders hohe Z-Werte auf einen Maximalwert begrenzt, so dass „Ausreißer" nur abgemildert eingehen. Wenn der Erfolg von Delta den „Ausreißern" zu verdanken wäre, hätte sich die Erkennungsquote bei einer Begrenzung der Z-Werte verschlechtern müssen. Ein Begrenzen der Z-Werte führt jedoch zu einer Verbesserung der Erkennungsquote. Evert et alia haben so nicht nur die Schlüsselprofilhypothese bestätigen können, sondern zugleich eine Möglichkeit entdeckt, die Erkennungsquote zu verbessern, die gerade bei mittelalterlichen Texten nützlich sein kann: Wenn in einem Text ein Schreiber eine bestimmte Schreibvariante ganz vermeidet, können Nullwerte zu hohen Z-Werten führen; diese Schreibvariantenproblematik kann durch das Begrenzen der Z-Werte gemildert werden.

Zweiter Validierungstest nicht-normalisierte Texte

Bei dem zweiten Validierungstest habe ich einerseits nicht-normalisierte Wortformen mit Hilfe des Normalisierungswörterbuchs bei der Erstellung der jeweiligen Bag-of-Words in eine normalisierte Wortform konvertiert. Andererseits habe ich eine Z-Wert-Begrenzung durchgeführt und Z-Werte ab |1,64| auf den Wert 1,70 gesetzt (dieser Wert hat sich in einer Versuchsreihe mit verschieden gestalteten Validierungs- und Ratekorpora als vorteilhaft erwiesen). Die Erkennungsquote für Bag-of-Words mit 5.000 Wortformen steigt damit von 80% auf 91% an.

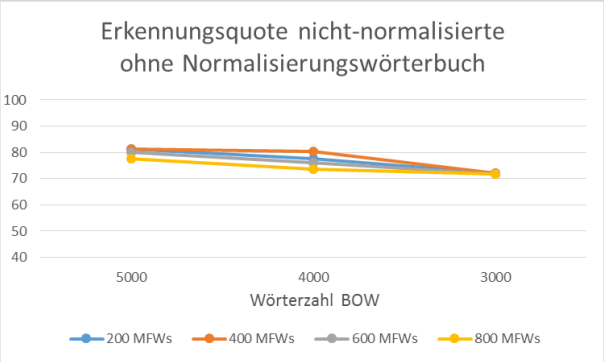

Abb. 1: Erkennungsquoten ohne Normalisierungswörterbuch / Z-Wertbegrenzung für nicht-normalisierte Texte (15 Texte Ratekorpus / 20 Texte Validierungskorpus)

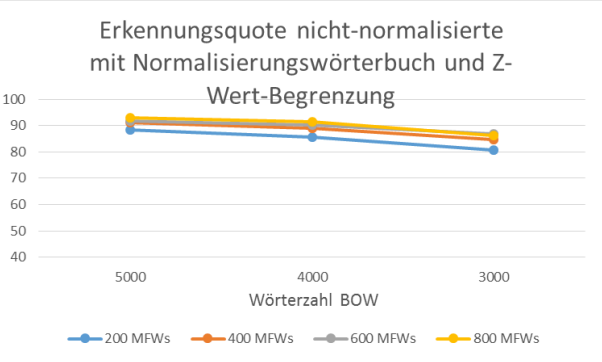

Abb. 2: Erkennungsquoten mit Normalisierungswörterbuch / Z-Wertbegrenzung für nicht-normalisierte Texte (15 Texte Ratekorpus / 20 Texte Validierungskorpus)

Wenn man die mühevoll bereinigten Texte nun wieder mit Fehlern kontaminiert, indem man den Inhalt der Bag-of-Words bspw. durch korpusfremdes Vokabular (teilweise durch altfranzösische Wörter statt mhd. Wörter) austauscht, so sinkt die Erkennungsquote erstaunlich langsam. Wenn 12% der Wörter durch Fremdmaterial getauscht wurden, ist nur ein geringes Absinken erkennbar. Tauscht man 20% aller Wörter durch Noise aus, dann gibt die Erkennungsquote etwas mehr nach als bei normalisierten Texten (vgl. Dimpel 2016) – das ist plausibel, weil hier trotz aller Anstrengungen mit Normalisierungswörterbuch und Z-Wert-Begrenzung noch immer mehr Varianz in den Texten enthalten ist als in Texten, die ein Editor manuell normalisiert hat. Dennoch bleiben die Quoten erstaunlich stabil.

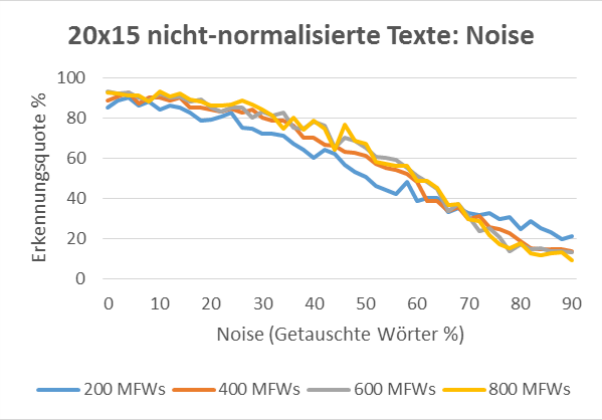

Abb. 3: Noise – Absinken der Erkennungsquote beim Tausch des Wortmaterials der BOW der Ratedatei

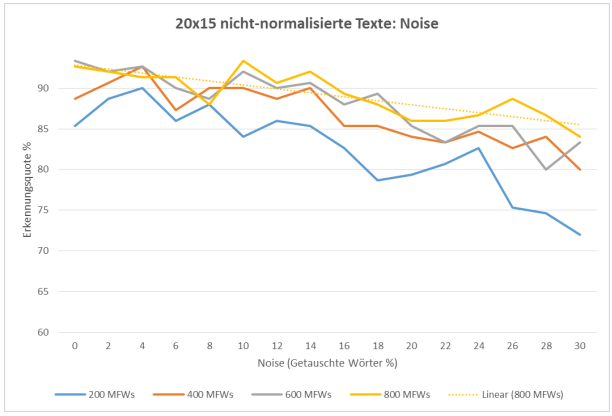

Abb. 4: Noise – Ausschnittvergrößerung 0-30 % Noise

Bibliographie

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 (2016): „Burrows' Delta im Mittelalter: Wilde Graphien und metrische Analysedaten", in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 65–70.

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 / 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, CO: Association for Computational Linguistics, 79–88 10.5281/zenodo.18177 http://www.aclweb.org/anthology/W/W15/W15-0709.pdf [letzter Zugriff 20. August 2015].

Evert, Stefan / Proisl, Thomas / Jannidis, Fotis / Pielström, Steffen / Reger, Isabella/ Schöch, Christof / Vitt, Thorsten (2016): „Burrows Delta verstehen", in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 61–65.

Hoover, David L. (2004): „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*. New York: 29–54.

Jannidis, Fotis / Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2015): „Improving Burrows' Delta - An Empirical Evaluation of Text Distance Measures", in: *DH2015: Global Digital Humanities*

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 ZfdG 1.