



GOEDOC – Dokumenten- und Publikationsserver der Georg-August-Universität Göttingen

2017

Ein Delta-Rätsel

–
Nicht-normalisierte mittelhochdeutsche Texte, Z-Wert-Begrenzung und ein
Normalisierungswörterbuch. Oder: Auf welche Wörter kommt es bei Delta an?

Friedrich Michael Dimpel

TU Darmstadt und FAU Erlangen

DARIAH-DE Working Papers

Nr.25

Dimpel, Friedrich Michael: Ein Delta-Rätsel : Nicht-normalisierte mittelhochdeutsche Texte, Z-Wert-Begrenzung und ein Normalisierungswörterbuch. Oder: Auf welche Wörter kommt es bei Delta an? Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2017 (DARIAH-DE working papers 25)

Verfügbar:

PURL: <http://resolver.sub.uni-goettingen.de/purl/?dariah-2017-5>

URN: <http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2017-5-1>



Bibliographische Information der Deutschen Nationalbibliothek

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

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

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

Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch

Abstract: Bei nicht-normalisierten mittelhochdeutschen Texten liegt die Erkennungsquote bei Autorschaftsanalysen mit Burrows' Delta deutlich unter der Quote für normalisierte Texte. Mit zwei Optimierungsverfahren – der Z-Wert-Begrenzung und dem Einsatz eines Normalisierungswörterbuchs – gelingt es, die Erkennungsquote um 11% zu verbessern. Die Z-Wert-Begrenzung dämpft den Einfluss von Ausreißerwerten, das Normalisierungswörterbuch erlaubt ein Mapping von nicht-normalisierten auf normalisierte Wortformen, um eine bessere Vergleichbarkeit verschiedener Texte herzustellen. Erstaunlich ist jedoch, dass beide Verfahren nicht jeweils anteilig zu einem Verbesserungseffekt beitragen, der sich auf 11% summieren lassen würde: Wendet man nur die Z-Wert-Begrenzung an, kommt es sogar zu einer geringen Verschlechterung. Analysen fördern zwar keine einfache Erklärung dieses Rätsels zutage, sie bieten jedoch einige Erkenntnisse darüber, bei welchen Z-Wert-Konstellationen eine Autorerkennung erschwert wird. Einzeltestanalysen zeigen, welche Konsequenzen jeweils Normalisierungswörterbuch und Z-Wert-Begrenzung für den Vergleich von zwei Texten haben. Beispielsweise führt der Einsatz des Normalisierungswörterbuchs dazu, dass nicht etwa weniger, sondern mehr nicht-normalisierte Wortformen in den Bereich der untersuchten Most Frequent Words (MFWs) gelangen.

Während man annehmen könnte, dass Wortformen, bei denen hohe Z-Wert auftreten, spezifisch für einen Autor sein könnten, zeigt sich, dass bei diesen Wortformen häufig auch hohe Z-Wert-Differenzen zu einem Vergleichstext des gleichen Autors vorliegen. Und während man annehmen könnte, dass hohe Z-Wert-Differenzen zwischen zwei Texten des gleichen Autors einer korrekten Autorerkennung meist entgegenstehen würden, zeigt sich in einer Testreihe, dass bei diesen Wortformen die Differenz der Z-Werte zu einem Vergleichstext eines anderen Autors häufig noch größer sind.

Bildet man erstens die Z-Wert-Differenz zwischen einem Untersuchungstext und einem Vergleichstext eines anderen Autors und zweitens die Z-Wert-Differenz zwischen dem ersten Untersuchungstext und einem Vergleichstext des gleichen Autors und sodann die Differenz dieser beiden Differenzen ('Level-2-Differenz'), so lässt sich sagen, dass bei einer positiven Level-2-Differenz die jeweilige Wortform gut geeignet ist, um den ersten Text vom Vergleichstext des anderen Autors zu unterscheiden; ist die Level-2-Differenz negativ, so steht diese Wortform einer korrekten Autorerkennung entgegen. Das Auftreten negativer Level-2-Differenzen wird durch das Normalisierungswörterbuch deutlich reduziert; die Z-Wert-Begrenzung verstärkt in manchen Bereichen mitunter vorhandene Differenzen. Bei diesen Analysen wurde (als Serendipitätseffekt) nebenbei eine Möglichkeit entdeckt, störende Features zu eliminieren, die bei Use-Case-Szenarios dann interessant sein könnte, wenn kürzere Texte vorliegen, bei denen es zweckmäßig ist, alle Mittel zur Verbesserung der Erkennungsquote auszuschöpfen.

Keywords: Stilometrie, Autorschaftsattribution, Mittelhochdeutsche Texte, Normalisierung

Stylometry, Authorship-Attribution Studies, Middle High German Texts, Normalization

Ein Delta-Rätsel

Nicht-normalisierte mittelhochdeutsche Texte,
Z-Wert-Begrenzung und ein Normalisierungswörterbuch. Oder:
Auf welche Wörter kommt es bei Delta an?

Friedrich Michael Dimpel

TU Darmstadt und FAU Erlangen



Friedrich Michael Dimpel: „Ein Delta-Rätsel“. [DARIAH-DE Working Papers](#) Nr. 25. Göttingen: DARIAH-DE, 2017. URN: <urn:nbn:de:gbv:7-dariah-2017-5-1>.

Dieser Beitrag erscheint unter der
Lizenz [Creative-Commons Attribution 4.0 \(CC-BY\)](#).

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



Zusammenfassung

Bei nicht-normalisierten mittelhochdeutschen Texten liegt die Erkennungsquote bei Autorschaftsanalysen mit Burrows' Delta deutlich unter der Quote für normalisierte Texte. Mit zwei Optimierungsverfahren – der Z-Wert-Begrenzung und dem Einsatz eines Normalisierungswörterbuchs – gelingt es, die Erkennungsquote um 11% zu verbessern. Die Z-Wert-Begrenzung dämpft den Einfluss von Ausreißerwerten, das Normalisierungswörterbuch erlaubt ein Mapping von nicht-normalisierten auf normalisierte Wortformen, um eine bessere Vergleichbarkeit verschiedener Texte herzustellen. Erstaunlich ist jedoch, dass beide Verfahren nicht jeweils anteilig zu einem Verbesserungseffekt beitragen, der sich auf 11% summieren lassen würde: Wendet man nur die Z-Wert-Begrenzung an, kommt es sogar zu einer geringen Verschlechterung. Analysen fördern zwar keine einfache Erklärung dieses Rätsels zutage, sie bieten jedoch einige Erkenntnisse darüber, bei welchen Z-Wert-Konstellationen eine Autorerkennung erschwert wird. Einzeltestanalysen zeigen, welche Konsequenzen jeweils Normalisierungswörterbuch und Z-Wert-Begrenzung für den Vergleich von zwei Texten haben. Beispielsweise führt der Einsatz des Normalisierungswörterbuchs dazu, dass nicht etwa weniger, sondern mehr nicht-normalisierte Wortformen in den Bereich der untersuchten Most Frequent Words (MFWs) gelangen.

Während man annehmen könnte, dass Wortformen, bei denen hohe Z-Wert auftreten, spezifisch für einen Autor sein könnten, zeigt sich, dass bei diesen Wortformen häufig auch hohe Z-Wert-Differenzen zu einem Vergleichstext des gleichen Autors vorliegen. Und während man annehmen könnte, dass hohe Z-Wert-Differenzen zwischen zwei Texten des gleichen Autors einer korrekten Autorerkennung meist entgegenstehen würden, zeigt sich in einer Testreihe, dass bei diesen Wortformen die Differenz der Z-Werte zu einem Vergleichstext eines anderen Autors häufig noch größer sind.

Bildet man erstens die Z-Wert-Differenz zwischen einem Untersuchungstext und einem Vergleichstext eines anderen Autors und zweitens die Z-Wert-Differenz zwischen dem ersten Untersuchungstext und einem Vergleichstext des gleichen Autors und sodann die Differenz dieser beiden Differenzen (**Level-2-Differenz**), so lässt sich sagen, dass bei einer positiven Level-2-Differenz die jeweilige Wortform gut geeignet ist, um den ersten Text vom Vergleichstext des anderen Autors zu unterscheiden; ist die Level-2-Differenz negativ, so steht diese Wortform einer korrekten Autorerkennung entgegen. Das Auftreten negativer Level-2-Differenzen wird durch das Normalisierungswörterbuch deutlich reduziert; die Z-Wert-Begrenzung verstärkt in manchen Bereichen mitunter vorhandene Differenzen. Bei diesen Analysen wurde (als Serendipitätseffekt) nebenbei eine Möglichkeit entdeckt, störende Features zu eliminieren, die bei Use-Case-Szenarios dann interessant sein könnte, wenn kürzere Texte vorliegen, bei denen es zweckmäßig ist, alle Mittel zur Verbesserung der Erkennungsquote auszuschöpfen.

Schlagwörter

Stilometrie, Autorschaftsattribution, Mittelhochdeutsche Texte, Normalisierung

Keywords

Stylometry, Authorship-Attribution Studies, Middle High German Texts, Normalization

Inhaltsverzeichnis

1 Das Rätsel	5
1.1 Nullwertquote bei 200 und 800 MFWs	10
1.2 Verteilung der Z-Wert-Ausreißer in einer Bag-of-Words	11
2 Test Wolfram-Partonopier	14
2.1 Ratedatei-Z-Werte und Z-Wert-Differenzen zum Autor-Vergleichstext	14
2.2 Z-Wert-Differenzen Ratetext–Autor-Vergleichstext und Z-Wert-Differenzen Ratetext–Distraktortext	16
2.3 Z-Wert-Differenzen in 25er-Blöcken zu Ratetext–Autor-Vergleichstext versus Ratetext–Distraktortext	17
3 Test Heinrich-hinvart	19
4 Test Heinrich-Staffeln	20
5 Verteilung normalisierter Wortformen in der Z-Wert-Rangliste	22
5.1 Höherer Anteil nicht-normalisierter Wortformen nach Anwendung des Normalisierungswörterbuchs	22
5.2 Schlüsselprofilthese oder Ausreißerthese: Graduelle Befunde	25
6 Das macht den Unterschied: Differenzen der Z-Wert-Differenzen	26
7 Fazit	29
8 Anhang	31
8.1 Diagramme zum „Test Heinrich-hinvart“	31
8.2 Diagramme zum „Test Heinrich-Staffeln“	32
9 Literaturverzeichnis	33

1 Das Rätsel

Burrows' Delta¹ und Delta-Varianten sind mittlerweile als Testverfahren bei Autorschaftsfragen etabliert, zahlreiche Validierungsstudien haben gezeigt, dass Delta mit sehr großer Zuverlässigkeit das Autorsignal aufspüren kann.² Auch bei normalisierten mittelhochdeutschen Texten liegt die Erkennungsquote in einem Validierungstest, der ausschließlich mit Texten durchgeführt wurde, bei denen die Autorschaft unstrittig ist, bei 91%.³ In diesem Validierungstest wird – wie auch bei den Tests in der vorliegenden Studie – das Korpus in zwei Teilkorpora aufgeteilt. Einige Texte mit bekannter Autorschaft werden in ein Ratekorpus gegeben. In ein Vergleichskorpus werden weitere Texte von eben diesen Autoren, von denen Texte im Ratekorpus enthalten sind, gegeben (zusätzlich können noch weitere Distraktor-Texte im Vergleichskorpus enthalten sein). Für jeden Text im Ratekorpus wird nun geprüft, ob der Algorithmus dazu in der Lage ist, den Text des gleichen Autors im Vergleichskorpus als den am wenigsten unähnlichen Text zu „erraten“. Es befindet sich nur ein Text des Autors, den es zu „erraten“ gilt, im Vergleichskorpus: Würde man mehrere „richtige“ Texte ins Vergleichskorpus einspeisen, wäre die Aufgabe für den Algorithmus etwas leichter. Verwendet wird zudem das Bag-of-Words-Verfahren, bei dem via Zufallsgenerator bspw. 5.000 einzelne Wortformen aus jedem Text in ein Sample gegeben werden – es handelt sich also um eine Zufallsauswahl, bei der Informationen über die Reihenfolge der Wörter im Text nicht mehr verfügbar sind. Um Zufallsschwankungen bei der Bildung der Bag-of-Words auszugleichen, wird jeder Test in mehreren Iterationen durchgeführt und der Mittelwert dieser wiederholten Tests verwendet.⁴

Als Distanzmaß wird Burrows' Delta verwendet. Bei der Delta-Berechnung werden zu einer bestimmten Anzahl der Most Frequent Words jeweils die Häufigkeiten im Untersuchungskorpus ermittelt. Um Häufigkeits-Unterschiede gemäß der Zipf'schen Verteilung auszugleichen, werden im weiteren Verfahren nicht Häufigkeiten, sondern Z-Werte zu den Häufigkeiten verwendet. Für jede Wortform wird die Z-Wert-Differenz zwischen Ratetext und einem Vergleichstext ermittelt. Zu dem Betrag der Z-Wert-Differenzen aller Wortformen wird der Mittelwert gebildet – das Distanzmaß Delta. Die Autorschaftserkennung ist dann erfolgreich, wenn der Ratetext zu dem Text des gleichen Autors im Vergleichskorpus die geringste Distanz aufweist. Um zu prüfen, ob Delta eher auf Ausreißern beruht – also auf Wortformen mit überdurchschnittlichen oder unterdurchschnittlichen Häufigkeiten, die zu besonders hohen oder niedrigen Z-Werten führen (Ausreißerhypothese), oder ob Delta eher auf der breiten Verteilung der Z-Werte beruht (Schlüsselprofilthese), haben Evert et al. hohe Z-Werte auf einen Maximalwert begrenzt.⁵ Ausreißer haben so nur eine begrenzte Auswirkung auf den Delta-Wert. Würde nun Delta vorwiegend auf Ausreißerwerten beruhen, müsste nach dem Begrenzen von Ausreißer-Werten die Erkennungsquote schlechter werden. Allerdings wird die Erkennungsquote hierdurch besser statt schlechter – Ausreißerwerte scheinen also weniger Gewicht zu haben als die allgemeine Schlüsselprofil-Verteilung. Mit der Z-Wert-Begrenzung haben Evert et al. eine Methode entdeckt, mit der die Erkennungsquote weiter

¹Burrows 2002, Burrows 2003. Für wertvolle Hinweise danke ich Stefan Evert, Steffen Pielström und Christof Schöch. Die Mittelhochdeutsche Begriffsdatenbank (Salzburg) hat mir einige digitale Texte überlassen – auch hierfür darf ich mich bedanken.

²Hoover 2004, S. 477–495, Argamon 2008, S. 131–147, Eder / Rybicki 2011, S. 315–321, Eder 2013, S. 603–614, Eder 2015, S. 9f., Jannidis / Lauer 2014, vgl. auch Evert et al. 2015, Büttner et al. 2017.

³Vgl. hierzu Dimpel 2017b; vgl. auch die Kurzfassung Dimpel 2017a. Ein etwas höherer Wert wurde mit einem kleineren Korpus ermittelt; vgl. Büttner et al. 2017. Erstmals zu Delta bei mittelhochdeutschen Texten Viehhauser 2015.

⁴Mit dem Bag-of-Words-Verfahren werden bessere Erkennungsquoten erreicht; vgl. Eder 2015, S. 170, Büttner et al. 2017.

⁵Evert et al. 2016, sowie Büttner et al. 2017.

verbessert werden kann; eine Methode, die gerade für mittelhochdeutsche Texte dann interessant ist, wenn graphische Varianten der gleichen Wortform zu besonders hohen Z-Werten führen, die womöglich eher den Launen des Zufalls oder des Schreibers als dem Autorsignal zu verdanken sind.

Bei normalisierten mittelhochdeutschen Texten führt eine Begrenzung der Z-Werte ($Z\text{-Werte} > |1,64|$ werden auf $|1,7|$ gesetzt)⁶ zu einer geringfügigen Verbesserung der Erkennungsquote: In einem Test mit 24 Ratetexten, die gegen ein Vergleichskorpus mit 20 Texten getestet wurden, wurden bei einer Bag-of-Words-Bildung mit 5.000 Wörtern und 100 Iterationen zum Ausgleich von Zufallsschwankungen bei der Bag-of-Words-Bildung für 200–800 MFWs (in 100er-Schritten) eine Steigerung der durchschnittlichen Erkennungsquote von 91,5% auf 92,2% erzielt. Bei kleineren Bag-of-Words ist die Steigerung, die durch die Z-Wert-Begrenzung erreicht wird, teilweise etwas größer.

Tabelle 1: Erkennungsquote normalisierte ohne Z-Wert-Begrenzung. BOW = Bag-of-Words

Anzahl MFWs	100	200	300	400	500	600	700	800	MW ohne 100 MFWs
BOW 2000	63,1	73,2	77	77,6	77,6	75,9	76	75	76,0
BOW 3000	75	80,7	84,6	85,5	85,1	84,6	84,7	85,2	84,3
BOW 4000	79,3	84,7	88,1	90	89,7	88,8	88,7	88,4	88,3
BOW 5000	83,1	87,2	91,4	92,3	92,4	92,3	92,6	92,2	91,5

Tabelle 2: Erkennungsquote normalisierte mit Z-Wert-Begrenzung ($Z\text{-Werte} > |1,64|$ werden auf $|1,7|$ gesetzt)

Anzahl MFWs	100	200	300	400	500	600	700	800	MW ohne 100 MFWs
BOW 2000	64,7	73,1	76,7	77,8	78,6	77,5	75,8	75,7	76,5
BOW 3000	74,9	81,5	85,4	85,8	85,7	86,2	86,5	85,5	85,2
BOW 4000	79,5	85,4	89,8	90,2	90,4	90,9	91,4	91,6	90,0
BOW 5000	83,3	88,5	92,3	92,4	92,3	92,6	93,6	94	92,2

Für einen Test von nicht-normalisierten mittelhochdeutschen Texten steht mir gegenwärtig nur ein kleineres Korpus zur Verfügung, das zusätzlich den Makel aufweist, dass nicht für alle nicht-normalisierten Texte auch ein ebenfalls nicht-normalisierter Text vom gleichen Autor vorliegt, so dass im Vergleichskorpus teilweise auf normalisierte Texte des gleichen Autors zurückgegriffen werden muss. Im Ratekorpus sind 15 nicht-normalisierte Texte von 10 Autoren enthalten.⁷ Im Vergleichskorpus sind insgesamt 20

⁶Vgl. Büttner et al. 2017.

⁷Zwischen stark normalisierten Texten und einer diplomatischen Wiedergabe der handschriftlichen Überlieferung gibt es viele Übergangsscheinungen; mittelhochdeutsche Texte liegen auf vielfältige Art mehr oder weniger normalisiert vor.

Texte enthalten: Je ein Text zu den 10 Autoren aus dem Ratekorpus, dazu kommen weitere 10 Texte, die Fehlattributionen auslösen könnten (Distraktortexte). Getestet wurde in diesem Fall mit Bag-of-Words mit 5.000 Wörtern und 50 Iterationen. Der Mittelwert der Erkennungsquote für 200, 400, 600 und 800 MFWs liegt bei **79,5%** und damit deutlich unter der Quote für normalisierte Texte.

Dass bei nicht-normalisierten Texten die Erkennungsquote deutlich schlechter ist, ist erwartbar – überraschend ist allenfalls, dass die Quote nicht noch schlechter wird. Wenn in einem Werk meist *unt* steht, in einem anderen Werk des gleichen Autors jedoch *vnt* und gelegentlich *vnnd*, steht diese Varianz einer korrekten Autorattribution entgegen. Um zu besseren Erkennungsquoten zu gelangen, habe ich erstens eine Z-Wert-Begrenzung vorgenommen (Z-Werte $> |1,64|$ werden auf $|1,7|$ gesetzt). Zweitens habe ich ein Normalisierungswörterbuch erstellt⁸ und nicht-normalisierte Wortformen wörterbuchbasiert durch normalisierte Wortformen ersetzt.⁹ Beim gleichzeitigen Einsatz beider Techniken konnte die Erkennungsquote auf **90,7%** gesteigert werden, also um beachtliche 11%. Im Anschluss an diesen Erfolg habe ich beide Optimierungstechniken jeweils separat getestet, um zu ermitteln, welchen Anteil an der Verbesserung der Quote einerseits die Z-Wert-Begrenzung und andererseits der Abgleich mit dem Normalisierungswörterbuch hat. Als intuitiv plausibel erachtet hätte ich beispielsweise eine Verbesserung durch das Normalisierungswörterbuch um ca. 5% von ca. 80% auf ca. 85% und eine Verbesserung durch die Z-Wert-Begrenzung um ca. 6% von ca. 80% auf ca. 86%. In diesem Fall hätte ich angenommen, beide Optimierungstechniken würden in der Summe die vorgefundene Verbesserung um ca. 11 Prozentpunkte ergeben. Der Befund ist jedoch zunächst recht kontraintuitiv:

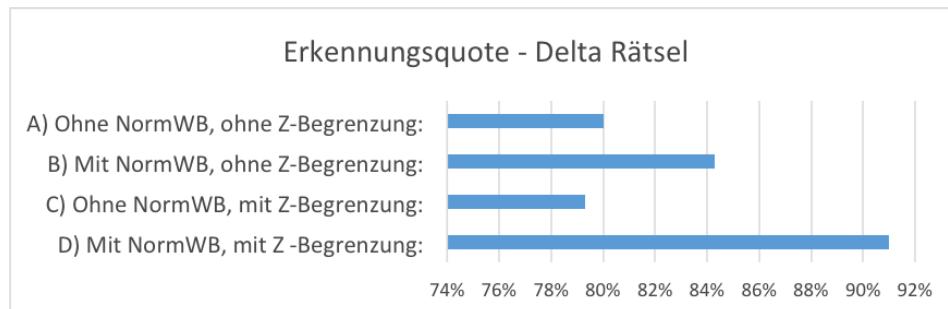


Diagramm 1: Erkennungsquote – Delta-Rätsel

Mit Normalisierungswörterbuch steigt die Quote von **79,5%** auf **85,1%** an. Eine Z-Wert-Begrenzung ohne Normalisierungswörterbuch ergibt allerdings nicht etwa eine Steigerung auf beispielsweise 86%, vielmehr fällt die Erkennungsquote auf **78,9%**; die Z-Wert-Begrenzung ohne Normalisierungswörterbuch führt also bei nicht-normalisierten mittelhochdeutschen Texten zu einer minimalen Verschlechterung anstatt zu einer deutlichen Verbesserung. Hier die Einzelwerte – die Werte für 100 MFWs gehen nicht in den Mittelwert ein; auch bei normalisierten mittelhochdeutschen Texten hat sich gezeigt, dass 100 MFWs für eine gute Erkennungsquote zu wenig sind.

Bei Delta führt bereits eine Abweichung bei einem Buchstaben zu einer Abweichung bei der Worthäufigkeit und somit zu einer größeren Formenvielfalt. Daher betrachte ich bei der Korpuserstellung nur Texte, die sich an Lachmanns Ausgaben orientieren, als normalisiert. Solche Texte, die davon abweichen, verwende ich in diesen Tests als nicht-normalisierte Texte – auch dann, wenn sie einer gewissen Normalisierung unterzogen wurden.

⁸Vgl. hierzu Dimpel 2017b und 2017a.

⁹Aus Laufzeitgründen werden nur die ersten 40.000 Einträge des Normalisierungswörterbuchs verwendet.

Tabelle 3: Einzelwerte zum Delta-Rätsel

A) Ohne Normalisierungswörterbuch

MFWs	100	200	400	600	800	MW
BOW 2000	56,8	62,4	62,8	61,5	60,7	
BOW 3000	69,6	72	72	71,7	71,7	
BOW 4000	73,1	77,7	80,3	76,1	73,6	
BOW 5000	79,3	80,4	81,1	78,8	77,7	79,5

B) Mit Normalisierungswörterbuch, ohne Z-Wert-Begrenzung

MFWs	100	200	400	600	800	MW
BOW 2000	56,8	62,4	62,8	61,5	60,7	
BOW 3000	69,6	72	72	71,7	71,7	
BOW 4000	73,1	77,7	80,3	76,1	73,6	
BOW 5000	79,3	80,4	81,1	78,8	77,7	79,5

C) Ohne Normalisierungswörterbuch, mit Z-Wert-Begrenzung

MFWs	100	200	400	600	800	MW
BOW 2000	61,5	64,4	58,4	56,8	52,1	
BOW 3000	73,5	75,9	71,6	67,5	64,4	
BOW 4000	78,5	78,5	75,9	72,7	72,1	
BOW 5000	82,7	84,8	79,9	75,7	73,6	78,9

D) Mit Normalisierungswörterbuch, mit Z-Wert-Begrenzung

MFWs	100	200	400	600	800	MW
BOW 2000	55,6	72,1	74,7	75,2	74,3	
BOW 3000	68,8	80,7	84,8	86,7	86,3	
BOW 4000	75,7	85,6	88,9	90,3	91,5	
BOW 5000	77,2	88,9	90,8	91,6	91,6	90,7

Mit Normalisierungswörterbuch werden die Quoten für längere Vektoren (mit einer größeren Anzahl an MFWs)¹⁰ zunehmend besser. Dass die Erkennungsquote mit zunehmender Vektorgröße bis zu einem hohen Maximalwert ansteigt, korrespondiert mit Ergebnissen zu anderen Sprachen bzw. Sprachstufen.¹¹ Allerdings ist die Verbesserung der Quoten mit zunehmender Vektorlänge hier deutlich stärker ausgeprägt.

Ohne Normalisierungswörterbuch werden dagegen ab 200 MFWs die Quoten für längere Vektoren schlechter – mit Z-Wert-Begrenzung ist dieser Effekt deutlich stärker ausgeprägt als ohne.

Da es von Interesse ist, wie stabil die Durchschnittswerte bei 50 Iterationen sind, habe ich die Durchschnittswerte für 50 Iterationen wiederum in 50 Iterationen berechnet und dazu die Standardabweichung und das Konfidenzintervall kalkuliert. Ziel war es, zu ermitteln, wie sich etwa die Abweichung der Erkennungsquote ohne Normalisierungswörterbuch und mit Z-Wert-Begrenzung zur Erkennungsquote ohne Normalisierungswörterbuch und ohne Z-Wert-Begrenzung verhält. In jedem Einzeltest (hier ohne Z-Wert-Begrenzung, ohne Normalisierungswörterbuch) wird in 50 Iterationen der Durchschnittswert für die Vektoren 200, 400, 600, 800 berechnet; diese Werte, die auf 50 Iterationen beruhen, werden wiederum 50 Mal erhoben. Für diese 50 x 50 Tests ergibt sich ein Mittelwert der durchschnittlichen Erkennungsquote von 79,5%, die Standardabweichung liegt bei 0,65, das Konfidenzintervall erstreckt sich bei alpha=0,05 von 79,3 bis 79,8.¹² In Dimpel 2017b und 2017a habe ich bei nur einem Durchlauf von 50 Iterationen ein Durchschnitt von 80% statt von 79,5% angegeben – ein Wert, der knapp außerhalb des Konfidenzintervalls liegt; hier werden nun Werte aus einem Test angegeben, der in der Mitte des Konfidenzintervalls liegt. Dieses Verfahren wurde auch für die übrigen Testreihen wiederholt; auch hier werden nun die aktuellen Durchschnittswerte verwendet:

Tabelle 4: Standardabweichung und Konfidenzintervalle zum Test von 50 Iterationen des 50-Iterationentests. (MW=Mittelwert, StDev=Standardabweichung)

	MW 50 Iterationen zu 50 Iterationen	StDev	Konfidenzintervall bei alpha=0,05
A) Ohne Normalisierungswörterbuch, ohne Z-Wert-Begrenzung	79,5	0,65	79,3 - 79,8
B) Mit Normalisierungswörterbuch, ohne Z-Wert-Begrenzung	85,1	0,52	84,9 – 85,3
C) Ohne Normalisierungswörterbuch, mit Z-Wert-Begrenzung	78,85	0,62	78,6 - 79,1
D) Mit Normalisierungswörterbuch, mit Z-Wert-Begrenzung	90,7	0,49	90,5 - 90,9

¹⁰ Die Häufigkeiten von vielen Wortformen in verschiedenen Texten können als vieldimensionaler Vektorraum aufgefasst werden; vgl. Schöch 2017, S. 292–294.

¹¹ Vgl. Schöch 2014, Büttner et al. 2017.

¹² Die Streuung ist also bei 50 Iterationen hier eher gering.

1.1 Nullwertquote bei 200 und 800 MFWs

Gerade bei nicht-normalisierten Texten ist zu erwarten, dass aufgrund der hohen Varianz in der Graphie bei einigen Texten zu einer Wortform aus der Häufigkeitsliste kein Treffer gefunden wird. Die Häufigkeit in diesen Texten wäre dann Null.¹³

Als **These 1** sei geprüft: A) **Eine große Menge an Nullwerten kann zu einer schlechteren Erkennungsquote führen und B) die Anwendung des Normalisierungswörterbuchs führt bei langen Vektoren dazu, dass der Anteil der Nullwerte gegenüber einem Versuch ohne Normalisierungswörterbuch stärker sinkt als bei kurzen Vektoren.**

Die Annahme, dass bei kurzen Vektoren eher normalisierte Wortformen zu erwarten sind, beruht auf der Überlegung, dass Wortformen in Bezug auf den Ratetext und auf das gesamte Vergleichskorpus so hohe Häufigkeiten aufweisen müssen, dass sie bspw. unter die Top-200-Wortformen gelangen. Das ist bei normalisierten Wortformen wahrscheinlicher als bei nicht-normalisierten Wortformen; zudem zeigt eine Sichtung der Worthäufigkeiten,¹⁴ dass in den ersten 200 MFWs überwiegend normalisierte Wortformen vorkommen; allerdings befinden sich unter den ersten 80 MFWs auch zahlreiche nicht-normalisierte Wortformen. Untersucht wurde hier ein einzelner Delta-Test, in dem die ‚Parzival‘-Handschrift-M¹⁵ gegen das oben beschriebene Vergleichskorpus geprüft wurde.

Tabelle 5: Differenzen der Nullwertquote bei 200 und 800 MFWs

Nullwertquote	200 MFWs	800 MFWs
A) Ohne Normalisierungswörterbuch:	17,4%	44,2%
B) Mit Normalisierungswörterbuch:	14,1%	41%

Der Anteil der Nullwerte sinkt bei längeren Vektoren durch die Anwendung des Normalisierungswörterbuchs nicht stärker als bei kurzen Vektoren (Tabelle 5). Die Daten aus dieser Stichprobe stehen also in Widerspruch zur These 1 zum Einfluss der Nullwerte: Der Abstand zwischen Vektor 200 und Vektor 800 bleibt jeweils bei etwa 3% – ob mit oder ohne Normalisierungswörterbuch. Denkbar ist jedoch, dass zufällig eine atypische Bag-of-Words-Konstellation untersucht wurde.¹⁶

¹³Solche Nullwerte müssen insbesondere dann, wenn sie bei mehreren Texten im Vergleichskorpus auftreten, nicht zwingend zu sehr hohen Z-Werten führen: Wenn mehrere Messwerte im Vergleichskorpus gleich Null sind, kann das zu einer erhöhten Standardabweichung führen. Zugleich wird der Mittelwert niedriger als ohne Nullwerte. Dadurch ist der Betrag sowohl über als auch unter dem Bruchstrich erhöht – eine gewisse ausgleichende Tendenz:

$$Z_{Wert} = \frac{\text{niedrigerer Messwert(hier:0)} - \text{niedrigerer Mittelwert}}{\text{erhöhte Standardabweichung}}$$

¹⁴Gesichtet wurden Wortformen aus einigen Samples, die mittels Bag-of-Words-Verfahren erstellt wurden.

¹⁵Der ‚Parzival‘ von Wolfram von Eschenbach ist in zahlreichen Handschriften und Fragmenten überliefert. Die Handschrift M weist eine enorme Zahl an Abweichungen zu anderen Handschriften und zur normalisierten Lachmann-Ausgabe auf, so dass die Zuordnung dieser Handschrift zum normalisierten Text von Wolframs ‚Willehalm‘, der sich als Autor-Vergleichstext im Vergleichskorpus befindet, eine große Herausforderung für jeden Algorithmus darstellt. Für die Überlassung der Handschrift M in digitaler Form danke ich Michael Stolz.

¹⁶Diesen Test habe ich noch zwei Mal wiederholt mit schwankenden Einzelwerten, die Tendenzen blieben jedoch gleich.

1.2 Verteilung der Z-Wert-Ausreißer in einer Bag-of-Words

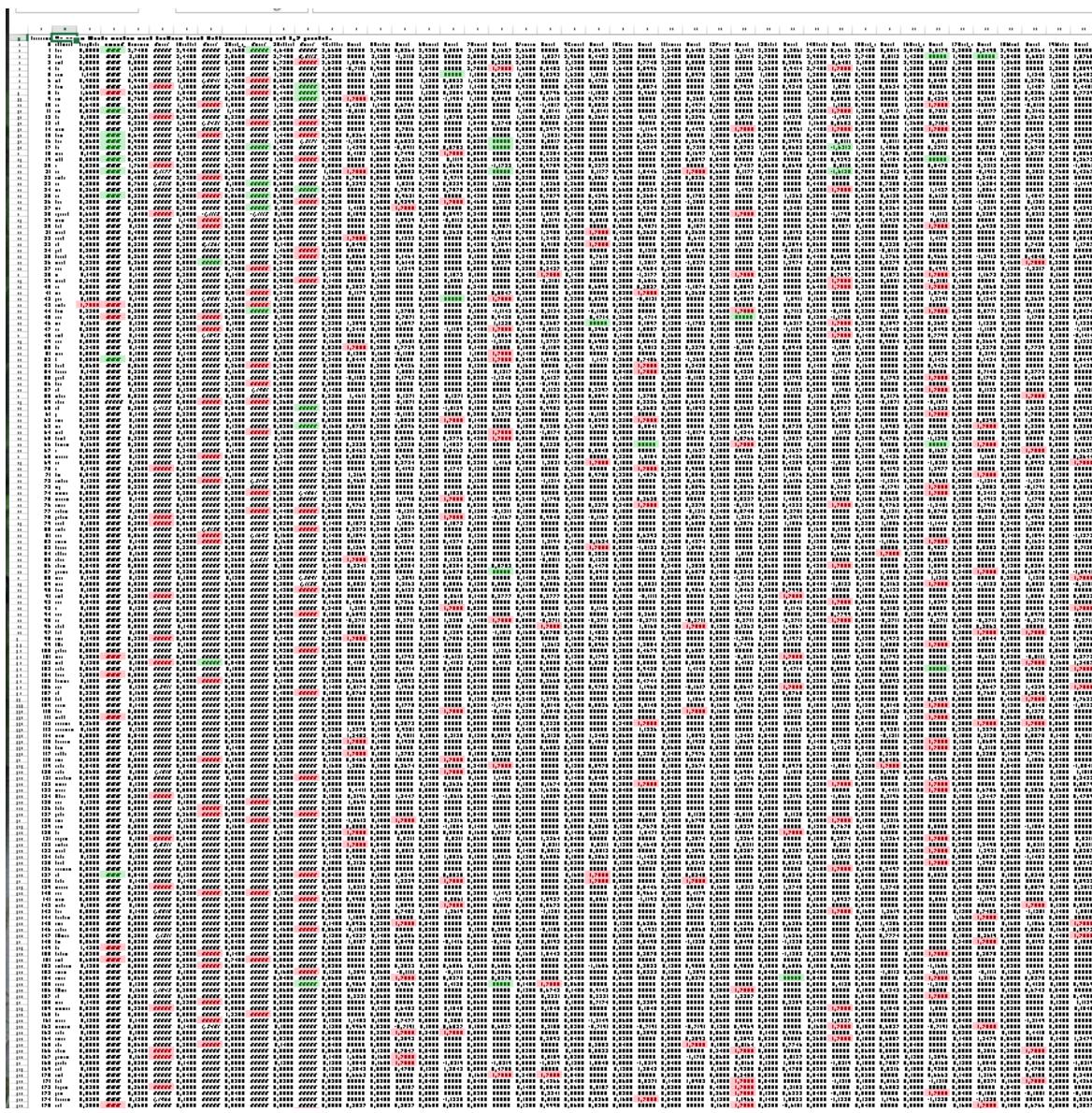


Diagramm 2: Verteilung von Z-Wert-Ausreißern in einer einzelnen Bag-of-Words (sortiert nach der Häufigkeit der Wortformen).
Hohe Z-Werte >1,64 sind rot, negative Z-Werte <-1,64 grün eingefärbt.

Bei einer Sichtprüfung der Z-Werte zu den Wortformen, die sich in einer konkreten Bag-of-Words befinden, die im Abschnitt 1a verwendet wurde (hier ohne Normalisierungswörterbuch), wurden hohe Z-Werte >1,64 rot und negative Z-Werte <-1,64 grün eingefärbt (Diagramm 2). Eine besonders auffällige Häufung von Z-Werten >1,64 und <-1,64 etwa am Anfang oder am Ende der Häufigkeitsliste wird nicht signifikant. Besonders hohe Z-Werte und besonders niedrige Z-Werte verteilen sich wie ein Flickenteppich relativ unregelmäßig über das ganze Feld und über alle MFWs. Dieser Befund ist eher mit der Schlüsselprofilthese als mit der Ausreißerhypothese (sensu Evert et al. 2017) vereinbar: Würde

die Ausreißerhypothese gelten, wäre eine Häufung hoher Z-Werte an bestimmten Positionen der Häufigkeitsliste – etwa an ihrer Spitze – nicht unplausibel.

Die Bag-of-Words, die in Abschnitt 1a verwendet wurde, wurden nach der Höhe der Z-Werte des Ratetextes sortiert und untersucht:

Tabelle 6: Verteilung der Z-Werte in einer einzelnen Bag-of-Words

Sortiert nach Z-Werten (800 MFWs)	Types ohne NormWB	Types mit NormWB
Z > 1,64 (5% Fläche)	41	36
Z > 1,34 (10%)	51	51
Z > 1 (16%)	72	82
Z > 0,52 (30%)	117	140
Z > 0 (positive bis Mitte)	196	263
Z < 0 (negative ab Mitte)	604	537
Z < -0,52	394	326
Z < -1	75	49
Z < -1,34	24	12
Z < -1,64	10	8

Auffällig sind zwei Befunde:

- Mit Normalisierungswörterbuch reduziert sich die Zahl der Types mit besonders hohen oder besonders niedrigen Z-Werten.
- Mit Normalisierungswörterbuch ergeben sich 263 statt 196 positive Z-Werte, die auf einer Type-Häufigkeit im Ratetext beruhen, die höher ist als die Häufigkeit im Korpusdurchschnitt. Die Z-Wert-Liste ist also in beiden Fällen grundlegend anders zusammengesetzt.

Interessant ist auch ein Blick auf die Types an der Spitze der Z-Wert-Liste ($Z > 1,64$): Ohne Normalisierungswörterbuch weisen 23 von 41 Types Besonderheiten auf, die auf fehlender Normalisierung beruhen wie *syn*, *eyn*, *zcu*, *vnd* (statt *sin*, *ein*, *ze*, *unde*). Mit Normalisierungswörterbuch ist die Spitze der Z-Wert-Liste deutlich anders zusammengesetzt. Beim Vergleichstext zur ‚Parzival‘-HS-M handelt es sich um den normalisierten ‚Willehalm‘. Das Normalisierungswörterbuch bringt nicht Graphie-Varianten, sondern diejenigen Wörter in die Spitze der positiven Z-Werte, die auch im normalisierten Vergleichstext vorkommen.

Bei dieser einen Bag-of-Words kann die Z-Wert-Begrenzung ohne Normalisierungswörterbuch 51 von 800 Z-Werten verändern, mit Normalisierungswörterbuch 46 von 800 Z-Werten.

Als These 2 sei geprüft: Das Normalisierungswörterbuch bringt Wortformen an die Spitze der Z-Wert-Rangliste, die eher autorspezifisch sind, als dass sie auf einer nicht-autorspezifischen grafischen Varianz beruhen.

Warum jedoch eine Z-Wert-Begrenzung ohne Normalisierungswörterbuch zu einer leichten Verschlechterung der Erkennungsquote führt, während sie mit Normalisierungswörterbuch zu einer deutlichen Verbesserung führt, kann auch aus diesen Daten nicht erklärt werden. Immerhin ist der folgenden Abbildung zu entnehmen, dass ohne Normalisierungswörterbuch ein relativ breiter Bereich mit sehr hohen positiven Z-Werten (Werte >4) einen erheblichen Einfluss auf den Durchschnitt der Z-Werte hat,¹⁷ während ohne Normalisierungswörterbuch der Bereich der Z-Werte >4 deutlich schmäler ist.

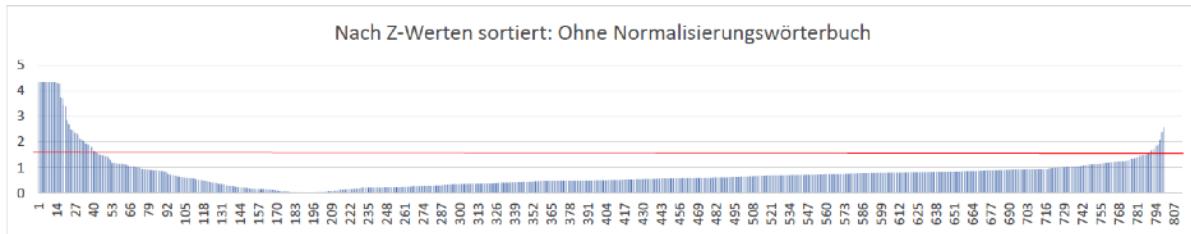


Diagramm 3: Betrag der Z-Werte aus einer zufällig erstellten Bag-of-Words – sortiert nach Z-Werten (Parzival'-Handschrift-M versus 20er Vergleichskorpus). Das linke Segment bis Nr. 196 (bzw. 263 in Diagramm 4) zeigt positive Z-Werte, das rechte Segment zeigt den Betrag der negativen Z-Werte.

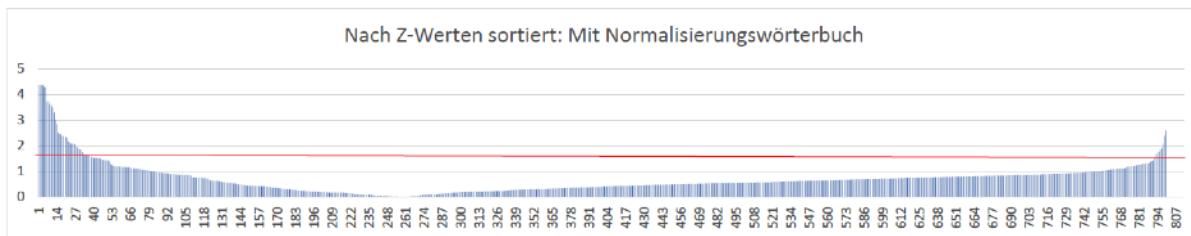


Diagramm 4: Wie Diagramm 3, hier nach Anwendung des Normalisierungswörterbuchs.

Summiert man für alle Types, deren Z-Wert-Betrag 1,64 übersteigt, den Anteil der Z-Werte, der über $|1,64|$ liegt, so summieren sich die Kappungswerte nach der Z-Wert-Begrenzung nach der Anwendung des Normalisierungswörterbuchs sich auf 40,1. Der Durchschnitt aller 800 Z-Werte nach der Z-Wert-Begrenzung fällt damit um 0,05. Würde eine solche Veränderung der durchschnittlichen Z-Werte auf den Mittelwert der positiven Z-Wert-Differenzen (=Delta) durchschlagen, so könnte eine Veränderung um 0,05 mitunter ausschlaggebend sein bei Delta-Werten, die bei korrekter Autorattribution des Öfteren zwischen 0,5 und 0,9 liegen. Es ist allerdings erst noch zu prüfen, inwieweit eine Veränderung der durchschnittlichen Ratetext-Z-Werte auch den Delta-Wert beeinflusst.

¹⁷ In Delta geht allerdings nicht der Durchschnitt der Z-Werte ein, sondern der Durchschnitt von jeweils einzelnen Z-Wert-Differenzen von zwei Wortformen aus zwei Texten. Vgl. dazu auch die weiteren Ausführungen.

2 Test Wolfram-Partonopier

2.1 Ratedatei-Z-Werte und Z-Wert-Differenzen zum Autor-Vergleichstext

Im Abschnitt 1 wurde eine Bag-of-Words zu einem Ratetext untersucht, die möglicherweise auf eine atypischen Verteilung beruhen kann, da die Zusammenstellung der Bag-of-Words randomisiert erfolgt. Die weitere Analyse erfolgt daher nicht auf der Grundlage einer einzelnen Bag-of-Words, sondern auf der Basis eines Tests, in dem ein einzelner Text (hier die ungenutzte „Parzival“-Handschrift-M von Wolfram von Eschenbach) gegen das oben beschriebene Vergleichskorpus (20 ebenfalls ungenutzte Texte) getestet wird. Dieses Vergleichskorpus wird bei allen Tests in dieser Studie beibehalten. Im Vergleichskorpus sollte diese Datei im Erfolgsfall zum „Willehalm“ Wolframs die niedrigsten Distanzwerte aufweisen. Verglichen werden diese Distanzwerte mit den Distanzwerten zu einem Distraktortext von einem anderen Autor – hier mit den Distanzwerten von Wolframs „Parzival“-Handschrift-M zum Partonopier-Roman von Konrad von Würzburg. Da hier ganze Texte und nicht die Bag-of-Words-Technik verwendet werden, ist es nicht nötig, diesen Test wiederholt durchzuführen.

In den folgenden Diagrammen werden jeweils vier Varianten dargestellt:

- A) Ohne Normalisierungswörterbuch, ohne Z-Wert-Begrenzung
- B) Mit Normalisierungswörterbuch, ohne Z-Wert-Begrenzung
- C) Ohne Normalisierungswörterbuch, mit Z-Wert-Begrenzung
- D) Mit Normalisierungswörterbuch, mit Z-Wert-Begrenzung

Wie in den Diagrammen 3/4 zeigt die obere Diagrammhälfte die Höhe der Z-Werte für 800 MFWs des Ratetextes (hier die „Parzival“-Handschrift-M). Die Sortierung erfolgt nach der Höhe der Z-Werte und nicht nach der Höhe der Häufigkeiten der Wortformen. Bis zum Abfallen der blauen Balken auf Null (von links aus betrachtet) werden positive Z-Werte dargestellt. Nach dem Nullpunkt wird der Betrag der negativen Z-Werte im rechten Diagrammbereich dargestellt, damit in der unteren Diagrammhälfte weitere Werte Platz haben: Hier stehen – mit negativem Vorzeichen versehen – in orangen Balken die absoluten Z-Wert-Differenzen zwischen dem Ratetext und dem Autotext im Vergleichskorpus, zu dem der Ratetext im Erfolgsfall die niedrigsten Distanzwerte aufweisen sollte (im Folgenden „Autor-Vergleichstext“; hier Wolframs „Willehalm“). Die Reihenfolge der Wortformen ist in beiden Tabellenhälften identisch. In der oberen Hälfte wird der Z-Wert und unten die Z-Wert-Differenz zum Autor-Vergleichstext zur jeweils gleichen Wortform dargestellt.

In dem Test, dessen Daten im Diagramm 5C (ohne Normalisierungswörterbuch, jedoch mit Z-Wert-Begrenzung) dargestellt sind, ist der Delta-Wert zwischen Ratetext und „Gottes Zukunft“ von Heinrich von Neustadt am niedrigsten – eine Fehltribution. In den übrigen Tests wird der Ratetext korrekt zu Wolframs „Willehalm“ sortiert.

Zunächst zur Z-Wert-Verteilung (blaue Balken): Der linke Diagrammbereich ist in allen Fällen kleiner als der rechte; es gibt also überall mehr negative Z-Werte als positive. Positive Z-Werte des Ratetextes beruhen auf einer erhöhten Häufigkeit der Wortform im Ratetext gegenüber dem Korpusdurchschnitt, negative Z-Werte des Ratetextes beruhen auf einer unterdurchschnittlichen Häufigkeit. Durch die

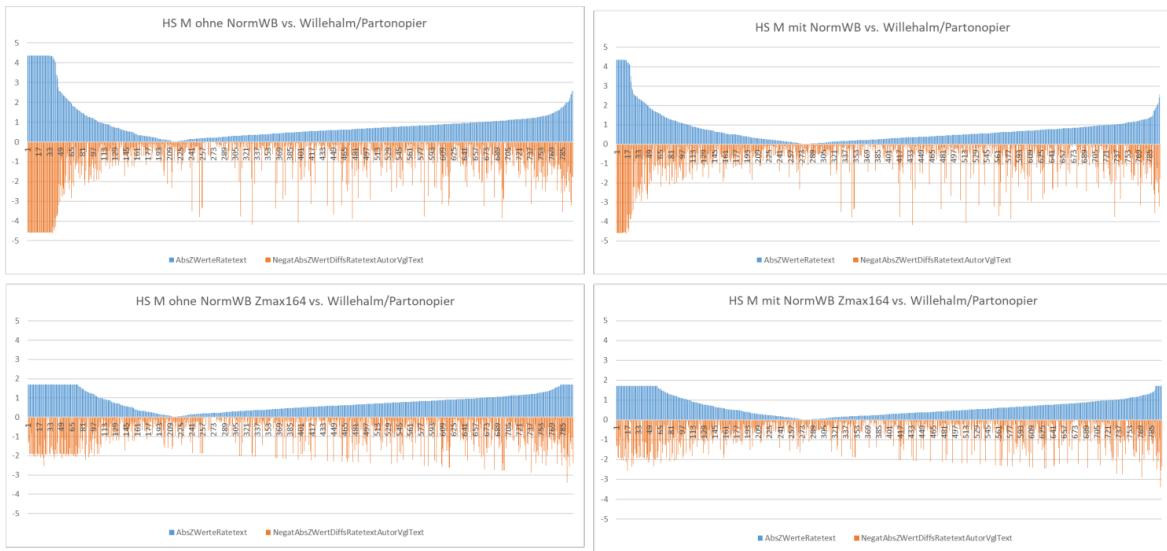


Diagramm 5: (A-D) Ratedatei-Z-Werte und Z-Wert-Differenzen zum Autor-Vergleichstext. HS M = „Parzival“-Handschrift M.

Anwendung des Normalisierungswörterbuchs vergrößert sich der Bereich der positiven Z-Werte – ein plausibler Befund: Durch das Verringern der Graphie-Varianz kann es im Ratetext nun nicht nur vermehrt Okkurrenzen von Wortformen geben, die im Gesamttest zu den TOP-800 MFWs gehören, sondern unter den Okkurrenzen können auch vermehrt überdurchschnittliche Okkurrenzen sein. Bereits im Abschnitt 1.1 wurde nach der Anwendung des Normalisierungswörterbuchs eine Verringerung der Nullwerte beobachtet.

Ähnlich wie in den Diagrammen 3 und 4 ist hier zu sehen, dass der Bereich der sehr hohen positiven Z-Werte (>4) ohne Normalisierungswörterbuch deutlich breiter ist als mit Normalisierungswörterbuch; die Z-Wert-Begrenzung hat also ohne Normalisierungswörterbuch einen höheren Einfluss auf den Durchschnitt der Z-Werte. Auch hier ist noch zu prüfen, inwieweit die Z-Werte des Ratetextes auch auf den Delta-Wert durchschlagen.

Denkbar wäre im Kontext der Ausreißer-Theorie folgendes Szenario: Wenn bei hohen Z-Werten des Ratetextes (blaue Balken) beim Text des gleichen Autors aus dem Vergleichskorpus ebenfalls hohe Z-Werte auftreten, dann wären die Differenzen zwischen den Z-Werten von Ratetext und Autor-Vergleichstext bei diesen Wortformen gering. In diesem Fall wäre bei diesen Wortformen zwar eine signifikante Abweichung beim Ratetext vom Gesamtkorpus gegeben, nicht aber vom Autor-Vergleichstext. In diesem Fall wären die auf hohen Z-Werte basierenden Wortformen potentiell gut dazu geeignet, Ratetext und Autor-Vergleichstext vom übrigen Korpus zu unterscheiden.

Dies trifft aber bei hohen Z-Werten überwiegend nicht zu: Die orangen Balken zeigen die Z-Wert-Differenzen zwischen Ratetext und Autor-Vergleichstext (und nicht die Z-Werte des Autor-Vergleichstextes). Mit einigen Abweichungen zeigen sich im Bereich der hohen blauen Balken auch zahlreiche hohe orange Balken; bei hohen orangen Balken sind die Differenzen zwischen Ratetext und Autor-Vergleichstext ebenfalls groß. Hohe orange Balken stehen also einer korrekten Autorschaftserkennung potentiell entgegen; bei hohen orangen Balken handelt es sich also um einen potentiellen Störfaktor¹⁸ für eine Autorschaftserkennung.

¹⁸ Ein Störfaktor vom Typ-1 (hier) wird im Weiteren von einem Störfaktor vom Typ-2 abgegrenzt.

Dieser Störfaktor-Typ-1 ist ohne Normalisierungswörterbuch und ohne Z-Wert-Begrenzung am häufigsten. Mit Normalisierungswörterbuch und mit Z-Wert-Begrenzung verringert sich dieser Störfaktor. In den Diagrammen mit Z-Wert-Begrenzung sind einzelne orange Balken im Bereich von blauen Z-Werten $>1,64$ niedriger als $-1,64$. In diesen Fällen wurde von einem Ratetext-Z-Wert, der auf $+1,64$ begrenzt wurde, ein erhöhter Z-Wert beim Autor-Vergleichstext abgezogen. Weiterhin sind einzelne orange Balken länger als $-1,64$. In diesen Fällen wurde von einem Ratetext-Z-Wert, der auf $+1,64$ begrenzt wurde, ein erhöhter negativer Z-Wert beim Autor-Vergleichstext abgezogen.¹⁹

Bei den positiven hohen Z-Werten in den linken Diagramm-Hälften liegt der Störfaktor vom Typ-1 deutlich häufiger und stärker vor als bei den negativen Z-Werten.²⁰ Unterdurchschnittliche Häufigkeiten oder gar Nullwerte, die ebenfalls zu Störfaktor-Typ-1 führen, scheinen weniger Gewicht zu haben als überdurchschnittliche Häufigkeiten. Dies ist dann der Fall, wenn bei niedrigeren orangen Balken zugleich der Ratetext deutlich von einem Distraktortext unterschieden werden kann; die absoluten Z-Wert-Differenzen zu einem Distraktortext müssten bei der jeweiligen Wortform höher sein als der zugehörige orange Balken. Denn nur in diesem Fall übersteigt bei dieser Wortform die Differenz der Z-Werte zwischen Ratetext und Autor-Vergleichstext die Differenz zwischen Ratetext und Distraktortext. Ob dies der Fall ist, wird in einem weiteren Schritt untersucht.

Wenn man nach der Anwendung des Normalisierungswörterbuchs eine Z-Wert-Begrenzung ausschließlich der negativen Z-Werte vornimmt, erreicht man lediglich eine Verbesserung der Erkennungsquote auf 86,4%, während eine Z-Wert-Begrenzung ausschließlich der positiven Z-Werte zu einer Verbesserung auf 90,4% führt (zum Vergleich: Nur Normalisierungswörterbuch ohne Z-Wert-Begrenzung: 85,1%; Normalisierungswörterbuch und Begrenzung positiver und negativer Z-Werte: 90,7%). Zu diesem Befund bietet das nach Z-Werten sortierte Diagramm eine Erklärung: Nach der Anwendung des Normalisierungswörterbuch sind noch einige Z-Werte $> 1,64$ vorhanden, jedoch nur noch sehr wenige $<-1,64$.

2.2 Z-Wert-Differenzen Ratetext–Autor-Vergleichstext und Z-Wert-Differenzen Ratetext–Distraktortext

In den folgenden vier Diagrammen wird die untere Hälfte aus dem letzten Abschnitt unverändert beibehalten – und mit ihr auch die Sortierung der Werte nach der Höhe der Z-Werte des Ratetextes. In der oberen Diagrammhälfte werden nun anstelle der Z-Werte des Ratetextes andere Werte dargestellt: Grau erscheinen die absoluten Z-Wert-Differenzen zwischen dem Ratetext und einem Distraktortext aus dem Vergleichskorpus (hier zum ‚Partonopier‘ von Konrad von Würzburg,²¹ bei den Delta-Werten ist der

¹⁹ Die vorstehenden Formulierungen („Z-Wert-Begrenzung auf 1,64“) sind etwas vereinfachend: Z-Werte $>|1,64|$ werden auf $+1,7$ oder $-1,7$ gesetzt. Die maximale Z-Wert-Differenz nach Z-Wert-Begrenzung ist damit 3,4.

²⁰ Im Bereich der hohen positiven Z-Werte finden sich Wörter wie *dasz*, die nicht in den ersten 40.000 Einträgen des Normalisierungswörterbuchs enthalten sind. Orange Peaks in der Mitte nach Anwendung des Normalisierungswörterbuchs betreffen etwa Wörter *iwer* oder *manec* (ebenfalls nicht in den ersten 40.000 Einträgen des Normalisierungswörterbuchs enthalten). Würde man das Normalisierungswörterbuch auf der ‚Parzival‘-Handschrift-M trainieren, würden diese Wortformen ebenfalls nachnormalisiert.

²¹ Der ‚Partonopier‘ wurde deshalb gewählt, weil er einerseits wie der Autor-Vergleichstext (Willehalm) ein normalisierter Text ist und weil er andererseits sowohl mit als auch ohne Normalisierungswörterbuch in der Mitte der sortierten Liste der Delta-Werte steht – die Distanzwerte zum Ratetext liegen also im Mittelfeld des Vergleichskorpus.

Abstand zum ‚Willehalm‘ in allen vier Tests deutlich niedriger als der zum Distraktortext ‚Partonopier‘).

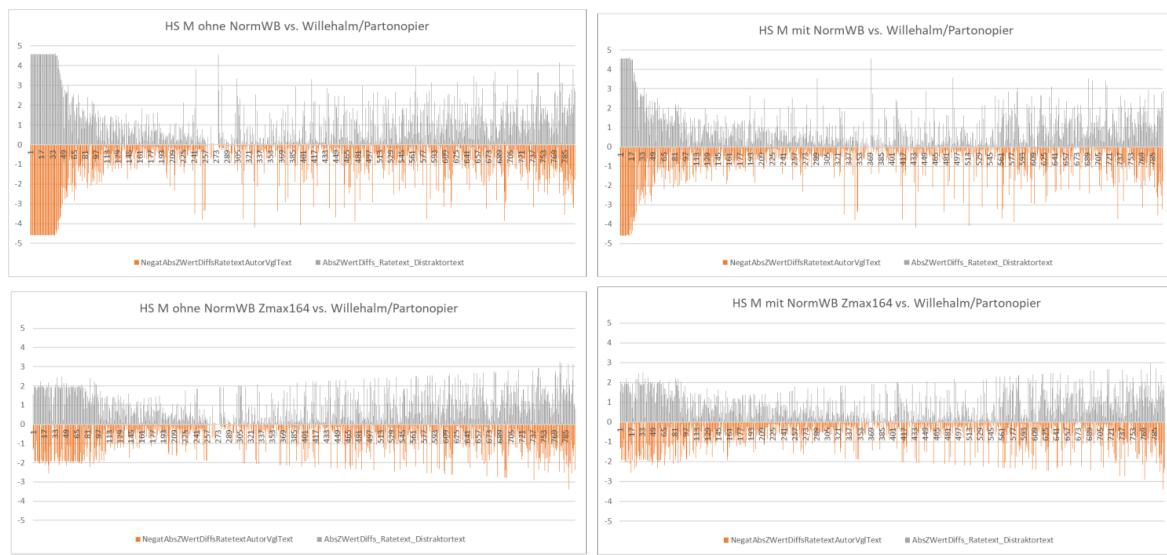


Diagramm 6: (A-D) Z-Wert-Differenzen Ratetext–Autor–Vergleichstext (orange) und Ratetext–Distraktortext (grau)

Bei den hohen positiven und hohen negativen Z-Wert-Differenzen von Ratetext und Autor-Vergleichstext scheinen überwiegend, aber nicht durchgehend auch bei Ratetext und Distraktortext hohe Z-Wert-Differenzen vorzuliegen. Auch bei den Z-Wert-Differenzen Ratetext–Distraktortext führt das Normalisierungswörterbuch dazu, dass der Bereich der sehr hohen Z-Wert-Differenzen (>4) schmäler wird.

Bei den niedrigen und mittleren Differenzen zeigt eine Sichtprüfung sowohl Übereinstimmungen als auch Abweichungen. Bei hohen negativen Z-Wert-Differenzen scheinen die grauen Balken höher zu sein als die orangen; damit wären in diesem Bereich erwartungsgemäß die Unterschiede vom Ratetext zum Distraktortext größer als die zum Autor-Vergleichstext.

2.3 Z-Wert-Differenzen in 25er-Blöcken zu Ratetext–Autor–Vergleichstext versus Ratetext–Distraktortext

Um den Unzulänglichkeiten der Sichtprüfung zu begegnen, habe ich die Höhe der Z-Wert-Differenzen in 25er-Blöcke zusammengefasst: Jeweils für 25 Z-Wert-Differenzen habe ich den Durchschnittswert gebildet und in Diagrammen dargestellt. Damit ist nun leichter erkennbar, in welchen Segmenten eher die Unterschiede vom Ratetext zum Distraktortext größer sind oder die Unterschiede vom Ratetext zum Autor-Vergleichstext.

In den 25er-Segmenten, in denen die graue Linie (Z-Wert-Differenzen Ratetext–Distraktortext) über der orangefarbenen Linie (Z-Wert-Differenzen Ratetext–Autor–Vergleichstext) liegt, wird eine korrekte Autorerkennung begünstigt: In diesem Bereich sind durchschnittlich die Z-Wert-Differenzen zum ‚Partonopier‘ größer als die zum ‚Willehalm‘. Wenn jedoch die orangefarbene Linie über der grauen Linie liegt, liegt ein Störfaktor-Typ-2 vor: Die Differenzen zum Autor–Vergleichstext sind größer als die zum Distraktortext.

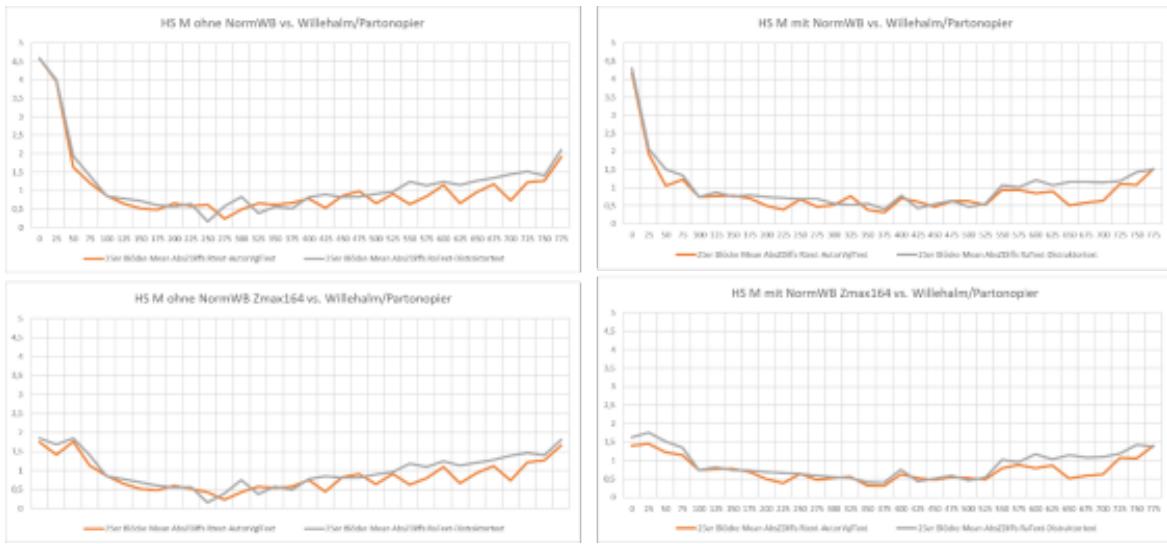


Diagramm 7: (A-D) Z-Wert-Differenzen in 25-Blöcken zu Ratetext–Autor-Vergleichstext (orange) versus Z-Wert-Differenzen Ratetext–Distraktortext (grau)

Z-Wert-Begrenzung und Normalisierungswörterbuch sorgen dafür, dass dieser Störfaktor-Typ-2 kaum noch vorkommt. Das Normalisierungswörterbuch sorgt dafür, dass die graue Linie im Bereich der höheren Z-Werte deutlich über der orangenen liegt.

Aufschlussreich ist ein Blick auf den Bereich der sehr hohen Z-Werte (>4; links): Ohne Z-Wert-Begrenzung liegen die orangenen und grauen Linien praktisch übereinander. Die hohen Z-Wert-Differenzen zwischen Ratetext und Autor-Vergleichstext (Störfaktor-Typ-1) haben offenbar hier keine negativen Auswirkungen auf die Erkennungsquote, da die Z-Wert-Differenzen zwischen Ratetext und Distraktortext mindestens ebenso hoch sind und den Deltawert in vergleichbarer Weise beeinflussen. Die Z-Wert-Begrenzung (noch mehr mit Normalisierungswörterbuch) führt dazu, dass auch im Bereich der hohen Z-Werte die graue Linie deutlich über der orangenen liegt und somit dafür sorgt, dass auch die Wortformen, die im Ratetext mit großen positiven Z-Werten einhergehen, zur korrekten Unterscheidung zum Distraktortext beitragen.

Als Befund sei festgehalten: Ohne Normalisierungswörterbuch scheint es sich bei dem Wortschatz, der zu sehr hohen positiven Z-Werten (>4) beim Ratetext führt, um einen Wortschatz zu handeln, bei dem das textspezifische Signal stärker ist als das autorspezifische Signal (die graue Linie liegt in den beiden ersten 25er-Blöcken nicht über der orangenen); erst mit der Z-Wert-Begrenzung werden hier die Differenzen zwischen Ratetext und Distraktortext größer als die zwischen Ratetext und Autor-Vergleichstext.

Auch im Bereich der hohen Z-Werte (>1,64) lässt die Z-Wert-Begrenzung geringe Unterschiede größer werden. Mit Normalisierungswörterbuch gelangen Wortformen in diesen Bereich, die auch autorspezifische Wortformen sind. Insgesamt ist beim Vergleich dieser drei Texte der Bereich der höheren negativen Z-Werte (etwa das rechte Viertel im Diagramm) stärker autorspezifisch als der Bereich der positiven Z-Werte.

3 Test Heinrich-hinvart

Im Abschnitt 2 wurde die nicht-normalisierte „Parzival“-Handschrift-M mit den normalisierten Texten „Willehalm“ und „Partonopier“ verglichen (kurz: „Test Wolfram-Partonopier“). In einer weiteren Testreihe werden nun bei gleichem Testdesign ausschließlich nicht-normalisierte Texte betrachtet: Als Ratetext dient der „Apollonius von Tyrland“ von Heinrich von Neustadt; Autor-Vergleichstext ist „Gottes Zukunft“ von Heinrich von Neustadt. Als Distraktortext dient „Unser vrouwen hinwart“ von Konrad von Heimesfurt (kurz: „Test Heinrich-hinvart“).²² Die Balkendiagramme, die analog zu den Diagrammen 3 und 4 erstellt wurden, gebe ich in die Anlage. Hier nun gleich die Gegenüberstellung der Z-Wert-Differenzen von Ratetext–Autor-Vergleichstext und Ratetext–Distraktortext:



Diagramm 8: (A-D) Z-Wert-Differenzen in 25er-Blöcken zu Ratetext–Autor-Vergleichstext (orange) versus Z-Wert-Differenzen Ratetext–Distraktortext (grau)

Die graue Linie liegt mit nur einer Ausnahme (hier mit geringem Abstand) stets über der orangefarbenen Linie, und zwar deutlicher als beim „Test Wolfram-Partonopier“. Die Distanzen zwischen Ratetext und Distraktortext sind also fast ausnahmslos größer als die Distanzen Ratetext–Autor-Vergleichstext.

Wie beim „Test Wolfram-Partonopier“ sind die Differenzen zwischen Autor-Vergleichstext und Distraktortext zum Ratetext ohne Normalisierungswörterbuch bei negativen Z-Werten größer als bei positiven. Dabei bleibt es auch mit Normalisierungswörterbuch, auch wenn hier die Differenzen bei den positiven Z-Werten größer werden und die Differenzen bei den negativen kleiner werden in Vergleich zu den Tests ohne Normalisierungswörterbuch. Die Z-Wert-Begrenzung verstärkt diesen Effekt bei den positiven Z-Werten v.a. dadurch, dass auch bei den 25 höchsten Z-Werten ein deutlicher Unterschied zwischen den jeweiligen Z-Wert-Differenzen gegeben ist. **Normalisierungswörterbuch**

²²Ein Distraktor-Text mit relativ großer Distanz; er steht auf Rangplatz 15 der Delta-Liste ohne Normalisierungswörterbuch und auf Rangplatz 20 mit Normalisierungswörterbuch. Im nächsten Abschnitt folgt komplementär ein Test mit einem Distraktortext mit geringerer Distanz: „Die Sieben Staffeln“ von David von Augsburg, der ohne Normalisierungswörterbuch auf Platz 2 der Delta-Rangliste steht bzw. auf Platz 13 mit Normalisierungswörterbuch.

und Z-Wert-Begrenzung sorgen hier also dafür, dass der Bereich der positiven Z-Werte, der eine korrekte Autorschaftsattribution begünstigt, größer wird.

Neben den vorliegenden Befunden wäre auch ein Szenario vorstellbar, in dem links im Diagramm bei den hohen positiven Z-Werten die orange Linie hohe Werte hat, während die graue Linie sehr niedrige Werte hat; ein solcher Befund würde als Störfaktor-Typ-2 einer korrekten Autorzuschreibung deutlich entgegenstehen. Dass ein solches Szenario unwahrscheinlich ist, sei anhand der folgenden Überlegungen dargelegt: Hohe orange Balken sind bei hohen Z-Wert-Differenzen zwischen Ratetext-Autor-Vergleichstext vorhanden; solche Differenzen dürften auf einem starken textspezifischen Signal beruhen. Nur dann, wenn die Differenzen bei diesen Werten zum Distraktortext noch etwas größer sind (vgl. oben), kommt hier zusätzlich noch eine gewisse autorspezifische Komponente hinein, die durch Z-Wert-Begrenzung besser greifbar wird. Wenn hier die graue Distraktor-Linie deutlich niedriger liegen würde, müsste folgender Befund vorliegen: Bei den Wortformen, bei denen es sehr große Differenzen zwischen Ratetext und Autor-Vergleichstext gibt, gibt es zugleich eine große Abweichung der Worthäufigkeit vom Korpusdurchschnitt (sonst würden sich keine sehr hohen Z-Werte im Rate-text ergeben²³). Dennoch müssten ausgerechnet bei diesen Wortformen beim Z-Wert-Vergleich von Ratetext-Distraktortext niedrige Werte herauskommen, die hohe Z-Werte beim Distraktortext voraussetzen. Diese Häufigkeiten vom Distraktortext müssten also vom Korpusdurchschnitt in ähnlicher Weise abweichen wie die Häufigkeiten des Ratetexts (während die Häufigkeiten des Autor-Vergleichstextes weniger abweichen). Wenn ein solcher Befund ausgerechnet bei einem von 19 Distraktortexten nicht nur bei einer Wortform, sondern bei einem ganzen Segment von Wortformen auftreten würde, müsste dies mit einem erstaunlichen Zufall einhergehen, der hier nicht beobachtet wurde. Sehr viel wahrscheinlicher ist es, dass der Distraktortext bei ausgerechnet diesen Wörtern durchschnittsnähere Häufigkeiten und nicht exorbitante Z-Werte aufweist.

4 Test Heinrich-Staffeln

Beim Test „Test Heinrich-hinvart“ im Abschnitt 3 wurde als Distraktortext ein Text mit mittlerem Abstand zum Ratetext gewählt. Nun wird als Distraktortext ‚Die Sieben Staffeln‘ von David von Augsburg (ebenfalls nicht-normalisiert) verwendet, der ohne Normalisierungswörterbuch auf Platz 2 der Delta-Rangliste steht, also ein Text mit geringem Abstand. Als Ratetext dient weiterhin der ‚Apollonius von Tyrland‘ von Heinrich von Neustadt; Autor-Vergleichstext bleibt ‚Gottes Zukunft‘ (kurz: „Test Heinrich-Staffeln“). Die Balkendiagramme, die analog zu den Diagrammen 5 und 6 erstellt wurden, gebe ich wiederum in die Anlage.

²³Diese liegen jedoch vor; vgl. die Diagramme mit der Gegenüberstellung von Ratetext-Z-Werten und ihren Differenzen zu den Autor-Vergleichstext-Z-Werten.



Diagramm 9: (A-D) Z-Wert-Differenzen in 25-Blöcken zu Ratetext–Autor-Vergleichstext (orange) versus Z-Wert-Differenzen Ratetext–Distraktortext (grau)

A) Ohne Normalisierungswörterbuch liegt bei einer Reihe von negativen Z-Werten die orange Linie über der grauen – teils deutlich (Störfaktor-Typ-2). Bei den hohen positiven Z-Werten bleiben nur die ersten 25 Werte im Durchschnitt ohne klare Relevanz für eine Autorschaftsunterscheidung; danach sind die Autor-Vergleichstext-Differenzen größer als die Differenzen zum Distraktortext.

B) Mit Normalisierungswörterbuch und ohne Z-Wert-Begrenzung liegt die orange Linie im Vergleich zu A) seltener und weniger hoch über der grauen – insbesondere bei den hohen negativen Z-Werten. Störfaktor-Typ-2 wird also deutlich reduziert. Zudem stellt sich eine mäßige Verbesserung bei hohen positiven Z-Werten ein.

C) Ohne Normalisierungswörterbuch und mit Z-Wert-Begrenzung kommt es gegenüber A) zu einer Verbesserung bei den positiven Z-Werten. Allerdings bleibt der Befund zum Störfaktor-Typ-2, der in A) beobachtet wurde, erhalten: Abermals liegt bei negativen Z-Werten die orange über der grauen Linie. Bei negativen Z-Werten gibt es nur minimale Abweichungen zu A), da ab Z-Wert Nr. 700 viele negative Z-Werte -1.64 nicht unterschreiten (vgl. hierzu auch das grau-orange Diagramm im Anhang; die 25er-Block-Durchschnittswerte unterschreiten -1,64 nicht). **Auf den Störfaktor-Typ-2 hat die Z-Wert-Begrenzung hier nahezu keine Auswirkungen.**

D) Insbesondere bei den 75 höchsten Z-Werten ergibt sich in der Kombination von Normalisierungswörterbuch und Z-Wert-Begrenzung nun ein recht klares Bild. Störfaktor-Typ-2 tritt nur selten auf – ein ähnlicher Befund wie in B).

Bemerkenswert in Hinblick auf das Delta-Rätsel ist der Befund zum Störfaktor-Typ-2 ohne Normalisierungswörterbuch bei hohen negativen Z-Werten, bei dem die Z-Wert-Begrenzung keine nennenswerte Besserung bringt. In diesem Fall

- sind die einschlägigen Häufigkeiten im Ratetext stark unterdurchschnittlich, damit sich hohe negative Z-Werte ergeben (allerdings zeigt ein Blick in die Diagramme 14A und 14C, dass im

rechten Diagrammabschnitt blaue und orange Balken weniger symmetrisch um die Nullachse verteilt sind als bei den hohen positiven Z-Werten);

- liegen die einschlägigen Häufigkeiten des Autor-Vergleichstextes näher am Durchschnitt als die des Ratetextes (oder sind überdurchschnittlich), sonst käme es nicht zu hohen orangen Balken, die für Differenzen der Z-Werte von Ratetext und Autor-Vergleichstext stehen;
- sind die einschlägigen Häufigkeiten des Distraktortexts deutlich stärker unterdurchschnittlich als die des Autor-Vergleichstextes (vermutlich; oder sie sind stark überdurchschnittlich).

Das Normalisierungswörterbuch beseitigt dieses Problem. Die orange Linie ist bei den hohen negativen Z-Werten Nr. 725-800 deutlich niedriger nach der Anwendung des Normalisierungswörterbuchs. Entweder sind die Autor-Vergleichstext-Häufigkeiten danach weniger durchschnittsnah (ein eher wenig wahrscheinlicher Effekt des Normalisierungswörterbuchs) oder die Ratetext-Häufigkeiten sind weniger stark unterdurchschnittlich (ein eher plausibler Effekt des Normalisierungswörterbuchs; die Zusammensetzung der TOP 800 MFWs wird hierdurch gründlich geändert; dazu im nächsten Abschnitt).

In Zusammenschau der Abschnitte 2–4 lässt sich zu hohen positiven Z-Werten folgende Beobachtung festhalten: Nach der Anwendung des Normalisierungswörterbuchs ist die Liste der 800 MFWs anders zusammengesetzt als ohne Normalisierungswörterbuch. **Mit Normalisierungswörterbuch liegen deutlich weniger negative Z-Werte unter -1,64 als ohne Normalisierungswörterbuch;**²⁴ die Z-Wert-Begrenzung spielt daher bei den negativen Z-Werten mit Normalisierungswörterbuch eine geringere Rolle.

Hohe positive Z-Wert-Differenzen zwischen Ratetext und Autor-Vergleichstext (Störfaktor-Typ 1) haben offenbar kaum negative Auswirkungen auf Delta, da die Z-Wert-Differenzen zwischen Ratetext und Distraktortext zumindest im 25er-Block-Durchschnitt ebenso hoch oder teils auch deutlich höher sind.

These 2 (vgl. Abschnitt 1.2: Das Normalisierungswörterbuch bringt autorspezifische Wortformen an die Spitze der Z-Wert-Liste) kann vorsichtig bestätigt werden; allerdings profitiert nicht nur die Spitze der Z-Wert-Liste von der Anwendung des Normalisierungswörterbuchs.

5 Verteilung normalisierter Wortformen in der Z-Wert-Rangliste

5.1 Höherer Anteil nicht-normalisierter Wortformen nach Anwendung des Normalisierungswörterbuchs

Bereits oben im Abschnitt 1.2 ist aufgefallen, dass an der Spitze der Z-Wert-Liste oft Wörter stehen, die Besonderheiten aufweisen, die auf fehlender Normalisierung beruhen wie *syn*, *eyn*, *zcu*, *vnd*. Nun soll geprüft werden, in welchen Segmenten der Z-Wert-Liste normalisierte und nicht-normalisierte Wortformen stehen. Dazu wird die Z-Wert-Liste in einem ersten Schritt mit dem Vollformenwörterbuch abgeglichen, das ich automatisch, unvollständig und fehlertolerant aus dem Mittelhochdeutschen

²⁴Z-Werte <-1,64: „Test Wolfram-Partonopier“ ohne Normalisierungswörterbuch: 20; mit: 10. „Test Heinrich-hinvar“ ohne Normalisierungswörterbuch: 21; mit: 11. „Test Heinrich-Staffeln“ ohne Normalisierungswörterbuch: 19; mit: 11.

Wörterbuch von Benecke-Müller-Zarncke (BMZ) generiert habe.²⁵ In einem zweiten Schritt erfolgt neben dem Abgleich mit dem Vollformenwörterbuch noch zusätzlich ein Abgleich mit einem Korpus, das aus dem Wortschatz der normalisierten Ausgaben von ‚Parzival‘, ‚Titorel‘, ‚Willehalm‘ (Wolfram von Eschenbach) und ‚Erec‘, ‚Iwein‘, ‚Gregorius‘, ‚Armer Heinrich‘ (Hartmann von Aue) erzeugt wurde. Diese Ergänzung ist nicht nur wegen der Unzulänglichkeiten des Vollformenwörterbuchs sinnvoll, sondern auch deshalb, weil der Begriff ‚normalisiert‘ in der Realität der Textausgaben gewisse Schwankungen einschließt – in Lachmanns normalisierter ‚Parzival‘-Ausgabe stehen etwa *und*, *unde* und *unt*. Normalisiert ist nicht gleich normalisiert...

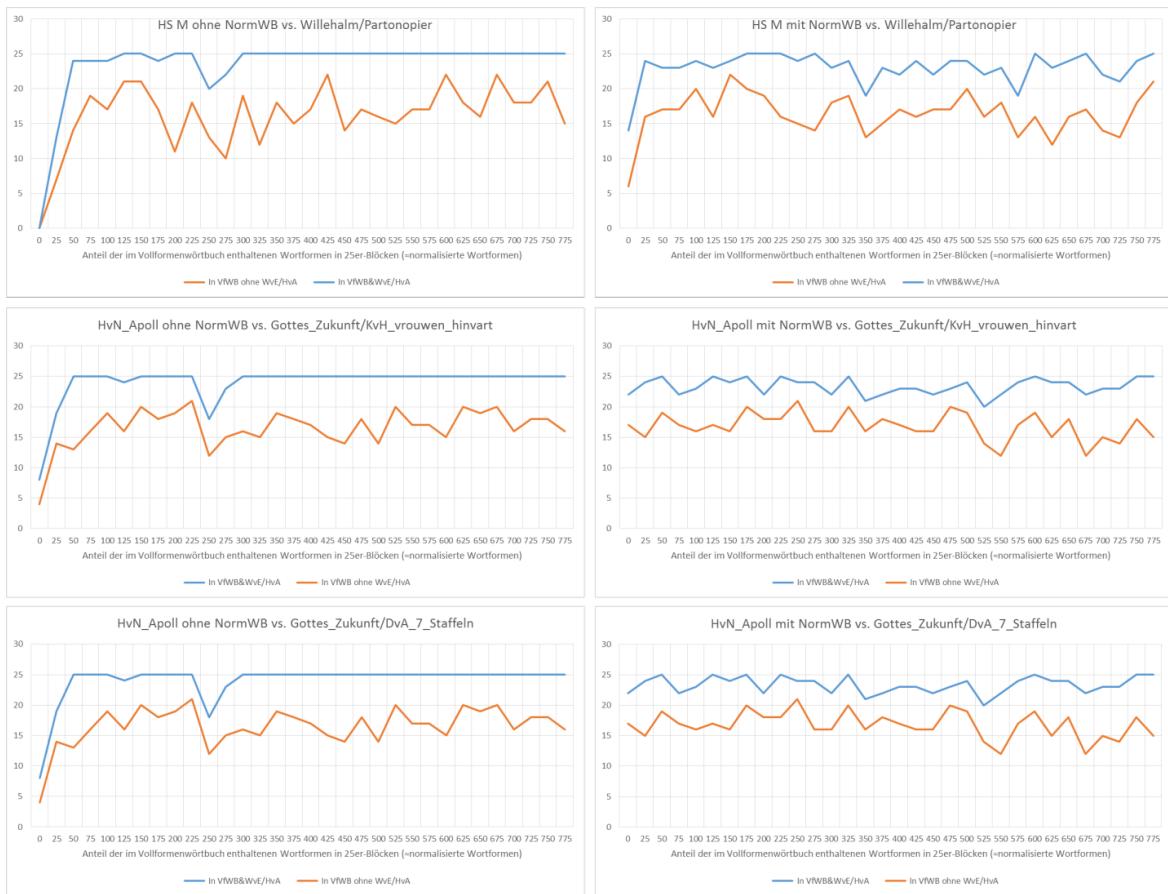


Diagramm 10: (A-F) Anteil normalisierter bzw. nicht-normalisierter Wortformen der 800 MFWs: Okkurrenz der MFWs in einem Vollformenwörterbuch sowie in einem Vollformenwörterbuch, das um den Hartmann-Wolfram-Wortschatz erweitert wurde.

Auch hier erfolgt die Sortierung der Werte nach der Höhe der Z-Werte des Ratetextes. Segmentiert wird wiederum in 25er-Blöcken. Die Höhe der Linien zeigt an, wie viele der Wortformen im jeweiligen Segment im Vollformenwörterbuch (orange) oder im kombinierten Vollformen-Hartmann-Wolfram-Wörterbuch (blau) zu finden sind.

Der bereits referierte Befund, dass ohne Normalisierungswörterbuch vor allem bei den hohen positiven Z-Werten des Ratetextes nicht-normalisierte Wortformen vorkommen, wird hier bestätigt (Diagramme 10,

²⁵Vgl. hierzu Dimpel 2017b und 2017a.

jeweils links: beide Linien zunächst mit niedrigen Werten). Diese Wortformen müssen im Gesamtkorpus so häufig vorkommen, dass sie in die Top 800 MFWs geraten.

Das Normalisierungswörterbuch (Diagramm 10, jeweils rechts) verändert die Zusammensetzung der TOP 800 MFWs dergestalt, dass unter den Wortformen mit hohen positiven Z-Werten beim „Test Wolfram-Partonopier“ ein etwas größerer Anteil an normalisierten Wortformen zu finden ist, während bei Heinrich von Neustadt sogar ein ähnlich hoher Anteil von normalisierten Wortformen unter den Wortformen mit hohen positiven Z-Werten zu finden ist wie in den übrigen 800 MFWs. Während bei Heinrich von Neustadt die Wortformen mit hohen positiven Z-Werten nach der Anwendung des Normalisierungswörterbuchs deutlich zur Autorschaftserkennung beitragen (vgl. oben, Diagramme 8 und 9: graue über orangener Linie), ist dieser Effekt beim „Test Wolfram-Partonopier“ zwar auch vorhanden (Diagramm 7), doch deutlich schwächer ausgeprägt; dass ein Zusammenhang zwischen der Okkurrenz im erweiterten bzw. nicht-erweiterten Vollformenwörterbuch (Diagramm 10) und den Z-Wert-Differenzen von Ratetext zu Autor-Vergleichstext bzw. Distraktortext gegeben ist, scheint mir plausibel zu sein.

Nochmals zu den Wortformen mit hohen positiven Z-Werten nach der Anwendung des Normalisierungswörterbuchs zwischen dem „Test Wolfram-Partonopier“ und Heinrich von Neustadt: Die Unterschiede bei den hohen Werte bei Heinrich (links in den Diagrammen 10D und 10E) zu den mäßig hohen Werten beim „Test Wolfram-Partonopier“ (Diagramm 10B) dürften darauf beruhen, dass bei der Erstellung des Normalisierungswörterbuchs die Heinrich-von-Neustadt-Texte als Trainingstexte verwendet wurden, die ‚Parzival‘-Handschrift-M jedoch nicht: Bei der Erstellung des Normalisierungswörterbuchs wurde bei einigen nicht-normalisierten Texten für die häufigsten 400 Wortformen, die nicht im Hartmann-Wolfram-Korpus enthalten sind, manuell ein Mapping zu normalisierten Wortformen erstellt.

Die blaue Linie, bei der auch ein Abgleich mit dem Wolfram-Hartmann-Korpus erfolgt, liegt deutlich über der orangenen Linie. Das kann einerseits auf Defiziten beim Vollformenwörterbuch (etwa bei Pluralformen der Nomina) beruhen und andererseits darauf, dass im Wolfram-Hartmann-Korpus Wortformen stehen, die nicht im Vollformenwörterbuch stehen wie etwa *das*; im Mittelhochdeutschen Wörterbuch von Benecke-Müller-Zarncke verzeichnete Varianten zum Lemma wurden bei der Erstellung des Vollformenwörterbuchs intentional nicht berücksichtigt.

Dass die blaue Linie ohne Normalisierungswörterbuch bereits ab Wortform Nr. 50 das theoretische Maximum von 25 erreicht,²⁶ während sie mit Normalisierungswörterbuch nur selten den Wert 25 erreicht, überrascht auf den ersten Blick. Wiederum gilt: das Anwendung des Normalisierungswörterbuchs bewirkt, dass andere Wörter in den Delta-Test eingehen als ohne Anwendung des Normalisierungswörterbuchs. Abgesehen von Wortformen mit sehr hohen Z-Werten gelangen in Tests ohne Normalisierungswörterbuch normalisierte Formen in die Top 800 MFWs, wie sie im Wolfram-Hartmann-Korpus vorkommen. Wenn durch das Normalisierungswörterbuch drei Varianten einer Wortform auf eine normalisierte Wortform reduziert werden, kann diese Wortform nunmehr eine Häufigkeit erreichen, mit der die Wortform eher zu den Top 800 MFWs gehören kann, als wenn die Wortform in drei verschiedenen Varianten geschrieben steht. Wenn mehrere Varianten durch das Normalisierungswörterbuch zu einer Wortform zusammengefasst werden, gelangen Wortformen, die ursprünglich auf einem unteren Rangplatz der Häufigkeitsrangliste standen, nun in die Top 800.

²⁶In einem 25er-Block sind 25 Wörter enthalten, von denen maximal 25 Wörter in einem Normalisierungswörterbuch enthalten sein können.

Ohne Normalisierungswörterbuch werden nur solche Wortformen im Test berücksichtigt, die im Korpus hohe Häufigkeiten aufweisen (entweder in der Breite des Korpus recht hohe Häufigkeiten oder in einzelnen Texten extrem hohe Häufigkeiten; letzteres ist bei Wortformen mit hohen positiven Z-Werten des Ratetexts zu beobachten²⁷). Nicht-normalisierte Wortformen haben schlechtere Chancen, überhaupt in den oberen Bereich der Häufigkeitsliste zu gelangen. **Der Anteil der nicht-normalisierten Wortformen, die im Test mit 800 MFWs berücksichtigt werden, ist nach der Anwendung des Normalisierungswörterbuchs höher statt niedriger** (mit Ausnahme der Wortformen mit sehr hohen Z-Werten) – auf den ersten Blick ein kontraintuitiver Befund.²⁸ Da das Normalisierungswörterbuch – bildlich gesprochen – den oberen Bereich der Häufigkeitsliste quasi auf engerem Raum zusammenfasst, haben nun nicht-normalisierte Wortformen, die zuvor hinter dem Rangplatz 800 auf der Häufigkeitsliste lokalisiert waren, ebenfalls Platz in den 800 MFWs. Dies geht mit einer etwas besseren Erkennungsquote einher.

5.2 Schlüsselprofilthese oder Ausreißerthese: Graduelle Befunde

Zu Überlegungen bezüglich Schlüsselprofil- versus Ausreißer-Hypothese lassen sich mit Blick auf die Darstellung der Z-Wert-Differenzen in 25er-Blöcken (ohne Z-Wert-Begrenzung) noch weitere Aussagen treffen: Beim „Test Wolfram-Partonopier“ sind die Ratetext–Distraktor-Differenzen (graue Linie in Diagramm 7A) im Ausreißer-Bereich nur unwesentlich größer als die Ratetext–Autor-Vergleichstext-Differenzen (orange Linie) – mit Normalisierungswörterbuch (Diagramm 7B) sind die Differenzen dieser Differenzen etwas größer als ohne Normalisierungswörterbuch. Sowohl mit als auch ohne Normalisierungswörterbuch sind die Differenzen der Differenzen im Bereich zwischen den Ausreißer-Bereichen größer als in den Ausreißer-Bereichen. Dieser Befund gilt auch für die Heinrich-von-Neustadt-Tests (Diagramme 8AB und 9AB), auch wenn es hier beim „Test Heinrich-Staffeln“ (9AB) zu größeren Differenzen auch im Bereich der hohen positiven Z-Werten kommt.

Die Inspektion der Diagramme lässt sich noch durch Zahlen belegen, die nun nicht wie die Diagramme 7, 8 und 9 auf den Durchschnittswerten bei 25er-Blöcken beruhen. Hier habe ich für jede Wortform von der Z-Wert-Differenz zwischen Ratetext und Autor-Vergleichstext einerseits und der Z-Wert-Differenz zwischen Ratetext und Distraktortext andererseits wiederum die Differenz gebildet. Diese Einzelwerte habe ich in drei Gruppen aufsummiert: für die positiven Z-Werte $> 1,64$, für die negativen Z-Werte $< -1,64$ sowie für das Mittelfeld dazwischen. Da sehr viele Werte aus dem Mittelfeld („Test Wolfram-Partonopier“ ohne Normalisierungswörterbuch: 706), einige aus dem Bereich der positiven hohen Z-Werte (ebd. ohne Normalisierungswörterbuch: 74) und wenige aus dem Bereich der negativen hohen Z-Werte (ebd. ohne Normalisierungswörterbuch: 20) vorliegen, werden die Differenzen durch das Aufsummieren jeweils anteilig in einer Form berücksichtigt, wie sie auch die Differenzen zwischen den beiden Delta-Werten von Ratetext–Autor-Vergleichstext und Ratetext–Distraktortext beeinflussen.²⁹

²⁷Etwa bei *dasz* statt *daz* oder *jch* statt *ich* (hier im „Test Wolfram-Partonopier“).

²⁸Das Absinken der blauen Linie (kombiniertes Wörterbuch) nach der Anwendung des Normalisierungswörterbuchs in den Diagrammen rechts gegenüber der blauen Linie in den Diagrammen links ist augenfällig. Beim reinen Vollformenwörterbuch (orange) ergibt sich durchschnittlich für die Wortformen Nr. 101-800 ein Rückgang pro 25er-Block von 17,0 auf 16,7 in Diagramm 10A zu 10B beim „Test Wolfram-Partonopier“ und von 17,2 auf 16,8 in beiden Heinrich-von-Neustadt-Tests (Rückgang von 10C zu 10D, von 10E zu 10F).

²⁹Die Zahl der Wortformen mit Z-Werten $> 1,64$ (etc.) unterscheidet sich von der Zahl unter <Analyse 1.2>: Dort wurde eine

Tabelle 7: Summenwerte der Differenzen der Z-Wert-Differenzen

	Z-Werte > +1,64	Z-Werte <-1,64	Mittelfeld
Wolfram-Partonopier ohne NormWB	8,9	0,4	104,19
Wolfram-Partonopier mit NormWB	15,6	-3,8	104,21
Heinrich-hinvart mit NormWB	30,9	12,4	223,4
Heinrich-hinvart ohne NormWB	34,4	10,2	227,7
Heinrich-Staffeln ohne NormWB	39,0	-6,3	58,2
Heinrich-Staffeln mit NormWB	51,8	-1,4	175,2

Negative Werte in Tabelle 7 zeigen, dass in der Bereichssumme die Z-Wert-Differenzen zum Distraktortext geringer sind als zum Autor-Vergleichstext (Störfaktor-Typ-2). Das größte Gewicht haben positive Z-Wert-Ausreißer beim „Test Heinrich-Staffeln“. Doch selbst hier hat das Mittelfeld noch einen größeren Einfluss auf die Differenzen zwischen beiden Deltawerten. Während negative Z-Wert-Ausreißer teils einen negativen Einfluss auf die Autorerkennung haben, ist der Einfluss der positiven Z-Wert-Ausreißer teils mehr, teils weniger vorhanden; größeren Einfluss haben jedoch die Differenzen im Mittelfeld.

Die Schlüsselprofilthese kann also auch auf diesem Weg zumindest für die untersuchten Texte bestätigt werden; darüber hinaus kann gezeigt werden, dass positive Z-Wert-Ausreißer die Differenzen zwischen den untersuchten Delta-Werten zwischen 8% und 43% beeinflussen, während negative Z-Wert-Ausreißer nur wenig oder gar einen negativen Einfluss auf die Autorschaftserkennung haben.

6 Das macht den Unterschied: Differenzen der Z-Wert-Differenzen

Im vorausgehenden Abschnitt wurde für jede Wortform von der Z-Wert-Differenz zwischen Ratetext und Autor-Vergleichstext einerseits und der Z-Wert-Differenz zwischen Ratetext und Distraktortext andererseits wiederum die Differenz gebildet; diese Werte wurden summiert, um zu ermitteln, welchen Einfluss die jeweiligen Z-Wert-Bereiche auf Delta-Differenzen zwischen Ratetext–Autor-Vergleichstext und Ratetext–Distraktortext haben. Hier werden diese Differenzen erneut verwendet. Allerdings wird diesmal die Liste der Wortformen nach der Größe dieser Differenzen sortiert. An der Spitze (links im Diagramm) stehen damit die Wortformen, die einen besonders wichtigen Beitrag dazu leisten, dass der Deltawert von Ratetext–Autor-Vergleichstext niedriger ist als der Deltawert Ratetext–Distraktortext.³⁰

Bag-of-Words untersucht; hier beruhen die Zahlen auf Tests mit den vollständigen Texten.

³⁰Für den „Test Heinrich-Staffeln“ stehen an der Spitze der Liste der unterscheidenden Wörter (ohne Normalisierungswörterbuch, ohne Z-Wert-Begrenzung): *vnde also se wile wie denne schin deste manig disem wann umb schone gewesen mit oder minnen eime libe*. Mit Normalisierungswörterbuch und mit Z-Wert-Begrenzung: *swie var niht gewesen umb ueber hort also dri luetzel schoene mac stein maniger minnen niwan swes ander we magt*

Am Ende der Liste (rechts im Diagramm) stehen die Wortformen, die einer Autorerkennung entgegenstehen;³¹ würde man nur das Drittel mit negativen Differenzen verwenden, würde man einen Deltawert von Ratetext–Autor–Vergleichstext erhalten, der höher ist als der Deltawert Ratetext–Distraktortext.

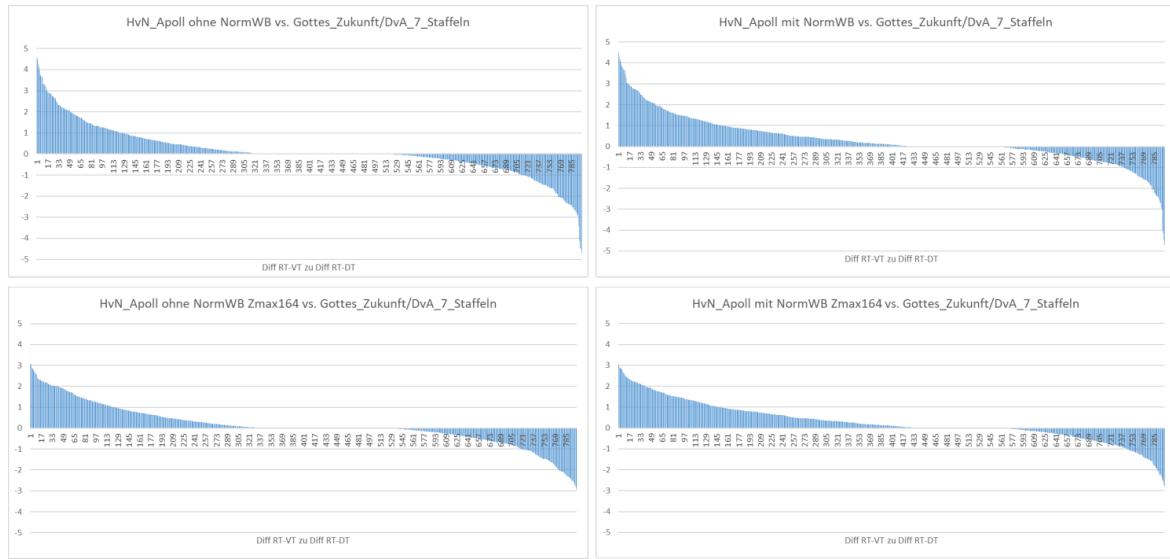


Diagramm 11: (A-D) Differenzen aus der Z-Wert-Differenz zwischen Ratetext und Autor-Vergleichstext einerseits und der Z-Wert-Differenz zwischen Ratetext und Distraktortext andererseits (sortiert nach der Höhe dieser Differenzen); hier zum „Test Heinrich-Staffeln“)

Ohne Normalisierungswörterbuch und ohne Z-Wert-Begrenzung liegen A) 334 positive Differenzen der Z-Wert-Differenzen vor. Ohne Z-Wert-Begrenzung und mit Normalisierungswörterbuch sind es B) 428, mit Z-Wert-Begrenzung sind es C) 338 positive Differenzen. Mit Z-Wert-Begrenzung und mit Normalisierungswörterbuch sind es D) 436. Summiert man die Differenzen, kommt man auf A) 329,6, B) 411,9 C) 299,5, D) 376,2. Mit Blick auf das Delta-Rätsel kann hieran zwar illustriert werden, dass eine Z-Wert-Begrenzung ohne Normalisierungswörterbuch einen negativen Effekt auf die Erkennungsquote hat. Warum jedoch Z-Wert-Begrenzung und Normalisierungswörterbuch zu einer deutlichen Verbesserung führen, kann anhand dieser Zahlen nicht nachvollzogen werden.

In dieser Sortierung (nach der Höhe der Differenzen aus der Z-Wert-Differenz zwischen Ratetext und Autor-Vergleichstext einerseits und der Z-Wert-Differenz zwischen Ratetext und Distraktortext andererseits) nun nochmals die Diagramme aus Abschnitt 2.1 und 2.2 – hier zum „Test Heinrich-Staffeln“:

Da die Sortierung der Diagramme verändert wurde, ist nun nicht mehr erkennbar, wo positive und wo negative Z-Werte dargestellt sind. Gut zu erkennen ist nun jedoch, dass die überwiegende Zahl der Wortformen mit sehr hohen Z-Werten (>4; hohe Balken blau) nicht autorspezifisch sind: Sie liegen in einem Bereich, in dem die Differenzen aus der Z-Wert-Differenz zwischen Ratetext und Autor-

³¹Für den „Test Heinrich-Staffeln“ stehen am Ende der Liste der unterscheidenden Wörter (ohne Normalisierungswörterbuch, ohne Z-Wert-Begrenzung): *reht es daz biz get engel verlorn laz himel geborn gedaht dri bereit lit das wiz wisheit dez rehten waz*. Mit Normalisierungswörterbuch und mit Z-Wert-Begrenzung: *e kueneges daz giht dez abe engel gelegen das wuerde es ez wile krist wisheit erkorn waz verlorn bereit lit*.

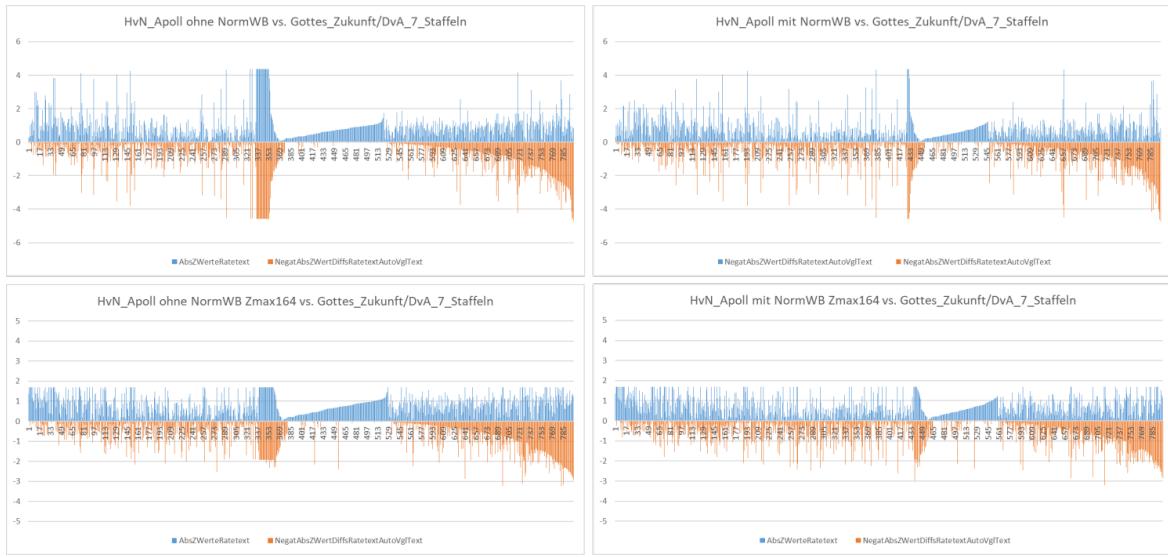


Diagramm 12: (A-D) Z-Werte (blau) und Z-Wert-Differenzen Ratetext–Autor–Vergleichstext (orange))

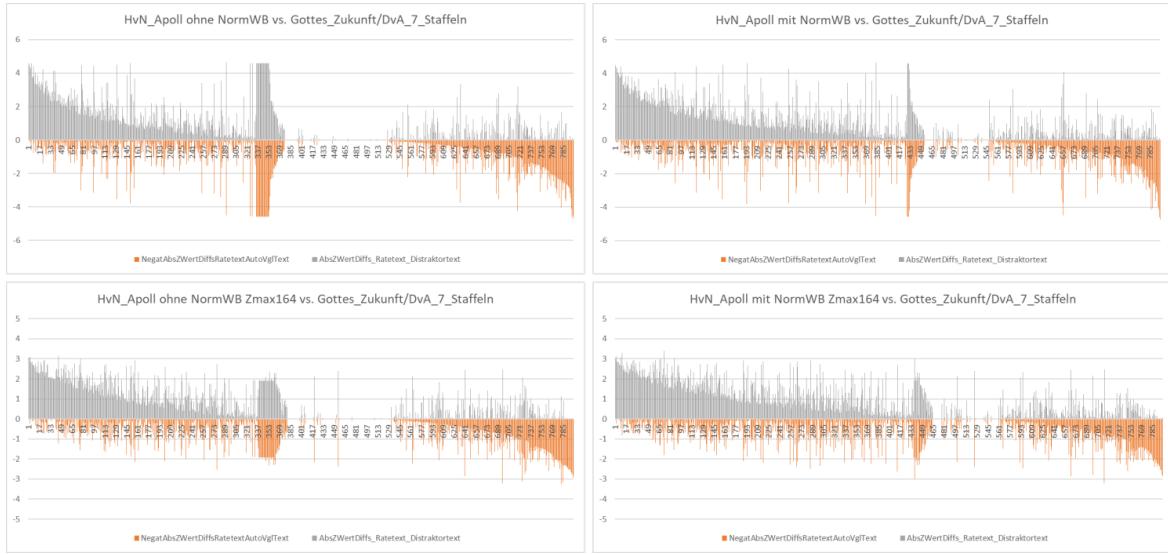


Diagramm 13: (A-D) Z-Wert-Differenzen Ratetext–Autor–Vergleichstext (orange) und Z-Wert-Differenzen Ratetext–Distraktortext (grau); jeweils zum Text „Heinrich-Staffeln“)

Vergleichstext einerseits und der Z-Wert-Differenz zwischen Ratetext und Distraktortext gleich Null sind – auch dieser Befund spricht gegen die Ausreißerhypothese.³²

Prinzipiell könnte man dieses Verfahren auch für alle die weiteren Distraktortexte im Vergleichskorpus wiederholen und auf diesem Weg in der vergleichenden Zusammenschau auch ermitteln, welche Wortformen besonders autorspezifisch für den jeweiligen Ratetext sind. Wenn es bei einer Delta-Studie nicht um Validierungsstudien geht, sondern um ein reelles Use-Case-Szenario, könnte man auf diesem Weg Wörter mit negativer Level-2-Differenz aus der MFWs-Liste eliminieren, die einer Autorerkennung entgegenstehen – insbesondere dann, wenn das gesamte Verfahren mit mehreren Texten des Autors als Ratetexte und Autor-Vergleichstexte durchlaufen werden kann. Möglich wäre es auch, vorwiegend Wörter mit positiver Level-2-Differenz zu verwenden, auf denen in Relation zum konkreten Distraktortext die korrekte Autorerkennung beruht. Dieses Verfahren könnte insbesondere dann interessant sein, wenn es darum geht, kürzere Texte zu analysieren, da hier noch Potential für eine Verbesserung der Erkennungsquote gegeben ist. Eine Studie dazu ist in Arbeit.

7 Fazit

In den verschiedenen Studien konnte keine einfache Erklärung für das Delta-Rätsel gefunden werden, doch immerhin wurden einige Erkenntnisse über die Struktur der untersuchten Wortlisten gewonnen. Inwieweit sich die Aussagen für andere Texte verallgemeinern lassen, wäre erst noch zu untersuchen.

Zentral ist, dass die Anwendung des Normalisierungswörterbuchs zu einer anderen Zusammensetzung der Häufigkeitsrangliste führt. Bei 800 MFWs reduziert sich mit Normalisierungswörterbuch die Zahl der Types mit besonders hohen oder besonders niedrigen Z-Werten. Zudem wächst der Anteil der Wortformen mit positiven Z-Werten; dennoch sind mehr Wortformen mit negativen als positiven Z-Werten im Vektor enthalten.

Da der Bereich der sehr hohen positiven Z-Werte ohne Normalisierungswörterbuch deutlich breiter ist als mit Normalisierungswörterbuch, hat die Z-Wert-Begrenzung ohne Normalisierungswörterbuch einen höheren Einfluss auf den Durchschnitt der Z-Werte. Geprüft wurde, wie sich hohe Z-Werte des Ratetextes auswirken. Überwiegend liegen bei hohen Z-Werten des Ratetextes niedrige Z-Werte des Autor-Vergleichstextes und damit hohe Z-Wert-Differenzen zwischen Ratetext und Autor-Vergleichstext vor; solche Differenzen sind ein Störfaktor-Typ-1 für eine korrekte Autorerkennung. Allerdings sind bei den untersuchten Texten die Z-Wert-Differenzen von Ratetext und Distraktortext im Durchschnitt der 25-Blöcke nicht kleiner, sondern gleich oder größer, so dass sich der Störfaktor-Typ-1 kaum negativ auf die Autorerkennung auswirkt.

Virulent ist dagegen der Störfaktor-Typ-2, der dann vorliegt, wenn bei der jeweiligen Wortform die Z-Wert-Differenzen Ratetext–Distraktortext größer sind als die Z-Wert-Differenzen Ratetext–Autor-Vergleichstext. Dieser Effekt wird durch den Einsatz des Normalisierungswörterbuchs deutlich reduziert; die Z-Wert-Begrenzung verstärkt in manchen Bereichen (v.a. bei hohen positiven Z-Werten) mitunter

³²Dass diese sehr hohen Werte, die sich ungefähr in der Mitte der Diagramme befinden, nebeneinander liegen, erklärt sich durch die Sortierung: Die primäre Sortierung erfolgt absteigend nach der Höhe der Differenzen der Z-Wert-Differenzen. Zuvor war die Liste nach der Höhe der Z-Werte sortiert.

vorhandene Differenzen. Beim „Test Heinrich-Staffeln“ tritt Störfaktor-Typ-2 ohne Normalisierungswörterbuch vorwiegend bei hohen negativen Z-Werten auf. Sie liegen jedoch überwiegend nicht unter -1,64, so dass die Z-Wert-Begrenzung hier nahezu keine Auswirkungen hat. Normalisierungswörterbuch und Z-Wert-Begrenzung sorgen in allen drei Tests („Wolfram-Partonopier“ und in beiden Tests zu Heinrich von Neustadt) dafür, dass der Bereich der positiven Z-Werte, der eine korrekte Autorschaftsattribution begünstigt, größer wird als ohne diese Optimierungstechniken.

8 Anhang

8.1 Diagramme zum „Test Heinrich-hinvart“



Diagramm 14: (A-D) Ratedatei-Z-Werte und Z-Wert-Differenzen zum Autor-Vergleichstext

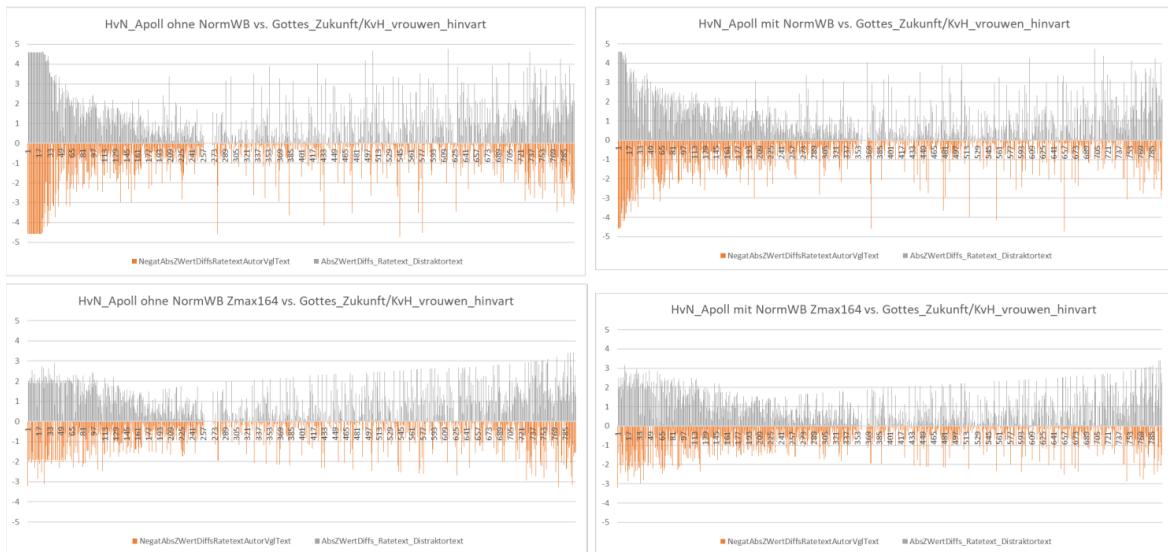


Diagramm 15: (A-D) Z-Wert-Differenzen Ratemtext-Autor-Vergleichstext (orange) und Ratemtext-Distraktortext (grau)

8.2 Diagramme zum „Test Heinrich-Staffeln“

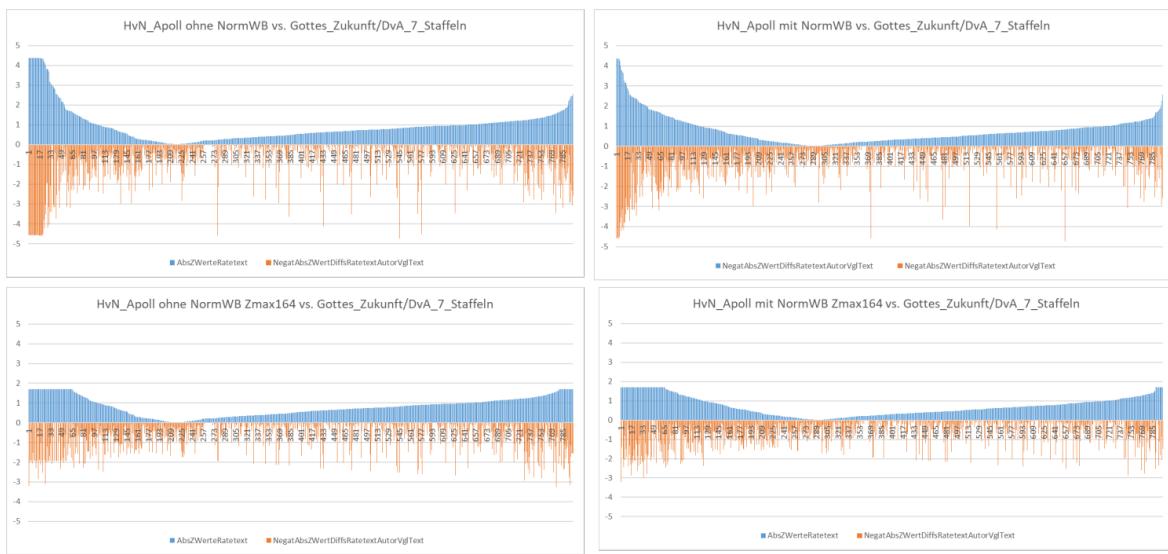


Diagramm 16: (A-D) Ratedatei-Z-Werte und Z-Wert-Differenzen zum Autor-Vergleichstext

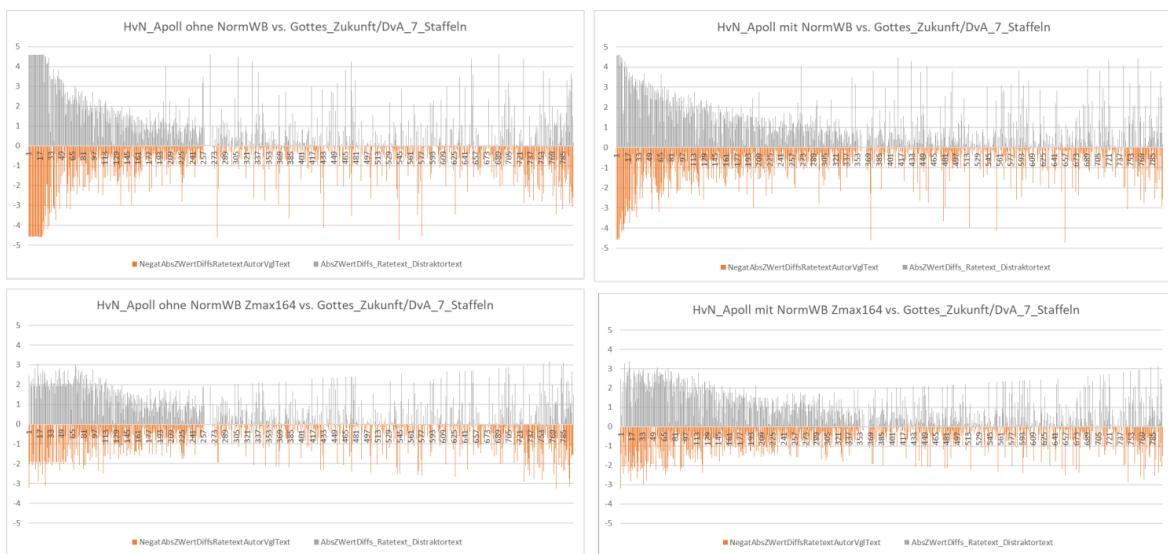


Diagramm 17: (A-D) Z-Wert-Differenzen Ratetext–Autor-Vergleichstext (orange) und Ratetext–Distraktortext (grau)

9 Literaturverzeichnis

- Argamon 2008: Shlomo Argamon: Interpreting Burrows's Delta: Geometric and Probabilistic Foundations. In: Literary and Linguistic Computing 23 (2008), S. 131–147. DOI: 10.1093/linc/fqn003
- Büttner et al. 2017: Andreas Büttner / Friedrich Michael Dimpel / Stefan Evert / Fotis Jannidis / Steffen Pielström / Thomas Proisl / Isabella Reger / Christof Schöch / Thorsten Vitt: „Delta“ in der stilometrischen Autorschaftsattribution. In: ZfdG n.n. (2017), S. n.n.
- Burrows 2003: John F. Burrows: Questions of Authorship: Attribution and Beyond. A Lecture Delivered on the Occasion of the Roberto Busa Award, ACH-ALLC 2001, New York. In: Computers and the Humanities 37 (2003), S. 5–32. <http://www.jstor.org/stable/30204877>
- Burrows 2002: John F. Burrows: ‚Delta‘. A Measure of Stylistic Difference and a Guide to Likely Authorship. In: Literary and Linguistic Computing 17 (2002), S. 267–287. DOI: 10.1093/linc/17.3.267
- Dimpel 2017a: Friedrich Michael Dimpel: Autorschaftsattribution bei nicht-normalisiertem Mittelhochdeutsch. Bessere Erkennungsquoten durch ein Normalisierungswörterbuch. In: Konferenzabstracts DHd 2017 Bern. Digitale Nachhaltigkeit. Hg. von Michael Stolz. Bern 2017, S. 100–103. <http://www.dhd2017.ch/programm>
- Dimpel 2017b: Friedrich Michael Dimpel: Stabile Autorschaft trotz handschriftlicher Varianz? Die Erfolgsquote von Burrows' Delta bei nicht-normalisierten mittelhochdeutschen Texten optimieren [n.n.]
- Eder / Rybicki 2011: 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), S. 315–321. DOI: 10.1093/linc/fqr031
- Eder 2015: Maciej Eder: Does size matter? Authorship attribution, small samples, big problem. In: Digital Scholarship Humanities 30 (2015), S. 167–182. DOI: 10.1093/linc/fqt066
- Eder 2013: Maciej Eder: Mind Your Corpus: systematic errors in authorship attribution. In: Literary and Linguistic Computing 28 (2013), S. 603–614. DOI: 10.1093/linc/fqt039
- Eibl 2013: Karl Eibl: Ist Literaturwissenschaft als Erfahrungswissenschaft möglich? Mit einigen Anmerkungen zur Wissenschaftsphilosophie des Wiener Kreises. In: Empirie in der Literaturwissenschaft. Hg. von Philip Ajouri / Katja Mellmann / Christoph Rauen. Münster 2013, S. 19–45
- Evert et al. 2016: Stefan Evert / Fotis Jannidis / Steffen Pielström / Isabella Reger / Christof Schöch / Thorsten Vitt: Burrows' Delta verstehen. In: Konferenzabstracts DHd 2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Hg. von Elisabeth Burr. Leipzig 2016, S. 62–65. <http://dhd2016.de>
- Evert et al. 2015: Stefan Evert / Thomas Proisl / Fotis Jannidis / Steffen Pielström / Christof Schöch / Thorsten Vitt: Towards a better understanding of Burrows's Delta in literary authorship attribution. In: Proceedings of the Fourth Workshop on Computational Linguistics for Literature. Association for Computational Linguistics. Denver 2015, S. 79–88. DOI: 10.5281/zenodo.18177

Hoover 2004: David L. Hoover: Delta Prime? In: *Literary and Linguistic Computing* 19 (2004), S. 477–495.
DOI: 10.1093/llc/19.4.453

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

Schneider 2002: Martin Schneider: Teflon, Post-it und Viagra. Große Entdeckungen durch kleine Zufälle. Weinheim 2002

Schöch 2014: Christof Schöch: Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik. In: *Literaturwissenschaft im digitalen Medienwandel*. Hg. von Christof Schöch / Lars Schneider. Berlin 2014, S. 130–157

Schöch 2017: Christof Schöch: Quantitative Analyse. In: Fotis Jannidis, Hubertus Kohle und Malte Rehbein (Hrsg.): *Digital Humanities. Eine Einführung*. Stuttgart 2017, S. 279–298

Viehhauser 2015: Gabriel Viehhauser: Historische Stilometrie? Methodische Vorschläge für eine Annäherung textanalytischer Zugänge an die mediävistische Textualitätsdebatte. In: *Grenzen und Möglichkeiten der Digital Humanities*. Hg. von Constanze Baum / Thomas Stäcker. 2015. DOI: 10.17175/sb001_009