Christof Schöch

Zeta für die kontrastive Analyse literarischer Texte

Theorie, Implementierung, Fallstudie

Abstract: The comparative, contrasting analysis of two or more texts or groups of texts is a method widely used in linguistics and literary studies. In Corpus Linguistics, Computational Linguistics and Digital Literary Studies, relevant measures of distinctiveness or keyness have been developed for the contrastive analysis of texts and have been applied to a wide variety of research questions. One such measure, one that has been developed in the area called Stylometry, is Zeta. It has originally been proposed by John Burrows in 2007. A variant of this measure is being described in some detail in this contribution. Based on a new implementation of the measure that also offers several different ways of visualizing the results, the procedure is illustrated using the example of a comparison of the three dramatic genres of comedy, tragedy and tragicomedy. The aim of this contribution is to foster a more precise understanding of the properties of the Zeta measure and to support its wider use in Digital Literary Studies. It is a measure that is easy to understand on the mathematical level while producing highly interpretable results.

Einleitung

Die vergleichende, kontrastierende Analyse zweier oder mehrerer Texte oder Gruppen von Texten ist ein in den Sprach- und Literaturwissenschaften weit verbreitetes Verfahren. Die Korpuslinguistik, Computerlinguistik und digitale Philologie haben entsprechende Maße der Distinktivität oder *keyness* von Merkmalen für die kontrastive Analyse von Texten entwickelt und für zahlreiche Fragestellungen eingesetzt. Ein solches, im Bereich der Stilometrie entwickeltes Maß ist *Zeta*, das ursprünglich von John Burrows vorgeschlagen wurde.[1] Eine Variante dieses Maßes wird im vorliegenden Beitrag zunächst genauer beschrieben. Auf der Grundlage einer neuen Implementierung des Maßes, die auch eine Reihe von

1 Vgl. John Burrows: »All the Way Through. Testing for Authorship in Different Frequency Strata«, in: *Literary and Linguistic Computing* 22.1 (2007), S. 27–47.

https://doi.org/10.1515/9783110523300-004

Visualisierungsmöglichkeiten anbietet, wird das Verfahren anschließend am Beispiel des Vergleiches der drei dramatischen Gattungen Komödie, Tragödie und Tragikomödie illustriert. Ziel des Beitrags ist es, das genauere Verständnis und den breiteren Einsatz dieses Maßes zu unterstützen, denn es ist mathematisch gut nachvollziehbar und fördert inhaltlich gut interpretierbare Ergebnisse zu Tage.

1 Methode: Kontrastive Textanalyse

Das Grundprinzip der kontrastiven Analyse ist, zunächst die gesamte Datensammlung nach einem bestimmten Klassifikationskriterium – im vorliegenden Kontext ist es die Gattungszugehörigkeit, in anderen Kontexten könnten es Epochen, Autoren, Textsorten, Register oder Ähnliches mehr sein – in zwei »Partitionen« genannte Teilmengen aufzuteilen. Man geht häufig explorativ vor und ermittelt, welche Merkmale (Wortformen, Lemmata oder andere Merkmale) für eine Partition (der Zielpartition) gegenüber der anderen Partition (der Vergleichspartition) besonders deutlich überrepräsentiert sind und damit als typisch, charakteristisch oder distinktiv gelten können. Dabei können die beiden Partitionen entweder gleichrangig sein: beispielsweise könnte man aus einer Sammlung von Dramentexten die Tragödien und die Komödien gegenüberstellen. Alternativ können die beiden Partitionen in einem hierarchischen Verhältnis zueinander stehen: beispielsweise könnte man aus der Sammlung von Dramentexten die Tragödien allen anderen dramatischen Texten der Sammlung gegenüberstellen.

Diese Methode der kontrastiven Gattungsanalyse hat zur Vorbedingung, dass die Zuordnung der Einzeltexte zu den einzelnen Gattungskategorien unproblematisch ist. Diese gewichtige Einschränkung bedeutet, dass sich die Methode insbesondere für die Untersuchung klar umrissener Textklassen eignet. Ist diese Vorbedingung erfüllt, erlaubt die Methode die Extraktion verschiedener distinktiver Merkmale und damit einen vielfältigen Blick auf die Unterschiede zwischen den untersuchten Textklassen. Es ist aber auch ein alternativer Einsatz der Methode möglich, bei dem zunächst an einer Teilmenge von Texten, die klar zugeordnet werden können – beispielsweise weil sie unstrittig prototypische Beispiele für eine Textklasse darstellen –, distinktive Merkmale herausgearbeitet werden. Diese Merkmale können dann in einem zweiten Schritt für die Einordnung von Texten verwendet werden, für die keine klare Zuordnung bekannt ist.

Damit kann dann die Positionierung strittiger Texte auf einem Gradienten zwischen den prototypisch verstandenen Untergattungen erfolgen,[2] wobei sich auch neue Untergruppen von Texten bilden können, deren Übereinstimmung mit etablierten Gattungsvorstellungen dann abgeglichen werden kann.

1.1 Maße für Distinktivität von Merkmalen

Den zahlreichen Maßen für die Extraktion distinktiver Merkmale aus zwei Textgruppen ist gemeinsam, dass sie von der reinen relativen Häufigkeit der Merkmale abstrahieren und auf je unterschiedliche Weise anstreben, die Distinktivität der Merkmale herauszuarbeiten. Die grundlegende Annahme ist, dass ein Merkmal nicht nur durch seine reine Häufigkeit in der Zielpartition für diese charakteristisch ist, sondern dass dies auch davon abhängt, wie häufig das Merkmal in der Vergleichspartition ist. Diejenigen Merkmale bekommen einen besonders hohen Wert zugewiesen, die in der Zielpartition sehr häufig und zugleich in der Vergleichspartition sehr selten sind. Das Phänomen wird in den Sprachwissenschaften auch unter dem Begriff der *marker words* oder *key words* beziehungsweise, abstrakter gefasst, der *keyness*, verhandelt. Mike Scott definiert *key word* als »a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency, by comparison with a reference corpus of some kind«.[3]

Eines der einfachsten dieser Maße ist das Verhältnis der relativen Häufigkeiten der Merkmale in zwei Partitionen (*ratio of relative frequencies*).[4] Für jedes Merkmal *i* ermittelt man die relative Häufigkeit *rf* in der Zielpartition *Z* und der Vergleichspartition *V* und dividiert den Wert in der Zielpartition durch den Wert in der Vergleichspartition. Dies fasst die folgende Formel zusammen:

---

2 Vgl. Klaus W. Hempfer: »Some Aspects of a Theory of Genre«, in: *Linguistics and Literary Studies / Linguistik und Literaturwissenschaft*, hg. v. Monika Fludernik und Daniel Jacobs. Berlin 2014, S. 405–422.

3 Vgl. Mike Scott: »PC Analysis of Key Words and Key Key Words«, in: *System* 25.2 (1997), S. 233–245.

4 Siehe Stefan T. Gries: »Useful Statistics for Corpus Linguistics«, in: *A Mosaic of Corpus Linguistics. Selected Approaches*, hg. v. Aquilino Sánchez und Moisés Almela. Frankfurt a. M. 2010, S. 269–291. Schon 1944 hatte George Yule einen *difference coefficient* vorgeschlagen, der die gleiche Intuition umsetzt, wenn auch in mathematisch leicht unterschiedlicher Weise (vgl. George Yule: *The Statistical Study of Literary Vocabulary*. Cambridge 1944; Alistair Baron, Paul Rayson und Dawn Archer: »Word frequency and key word statistics in historical corpus linguistics«, in: *Anglistik. International Journal of English Studies* 20.1 (2009), S. 41–67.

$$rrf_i = \frac{rf_i(Z)}{rf_i(V)}$$

Sortiert man anschließend die resultierenden Werte absteigend, enthält der Anfang der Liste die für die Zielpartition am deutlichsten überrepräsentierten oder präferierten Merkmale, das Ende der Liste dagegen die am deutlichsten unterrepräsentierten oder vermiedenen Merkmale. In der Mitte der Liste finden sich diejenigen Merkmale, die in beiden Partitionen etwa gleich häufig sind, also weder vermieden noch präferiert werden. Einige weitere Maße folgen einem ähnlichen Prinzip, auch wenn sie mathematisch ein Wenig komplexer sind: so der *tf-idf*-Score (für *term frequency-inverse document frequency*), ein im *Information Retrieval* weit verbreitetes Maß[5] sowie zwei inbesondere in den Sprachwissenschaften häufig eingesetzte Maße, das X^2-Maß (*chi-square*) und der *llr* (*log-likelihood ratio*[6]).

Verfahren aus dieser Gruppe können wertvolle Hinweise geben, jedoch betrachten sie die Partitionen jeweils als Ganzes und vergleichen lediglich zwei Mittelwerte bzw. einen erwarteten und einen beobachteten Wert. Dadurch wird die spezifische Verteilung der Werte innerhalb der Texte, die eine Partition bilden, nicht berücksichtigt. Solche Maße sind daher zudem nicht in der Lage, die statistische Signifikanz des gefundenen Unterschiedes zu quantifizieren. Je größer die Streuung der Werte aber ist und je weniger symmetrisch sich die Werte um den Mittelwert verteilen, desto weniger gut repräsentieren die Mittelwerte (bzw. die erwarteten Häufigkeiten) die Verteilung und desto weniger überzeugend ist ihr direkter Vergleich.

Zahlreiche Testverfahren berücksichtigen diese Überlegungen, wenn auch in teils recht unterschiedlicher Weise.[7] Beim *Welchs t-Test* fließt die Standardabweichung und damit ein Maß für die Streuung der Häufigkeiten mit in den Vergleich

5 Siehe Stephen Robertson: »Understanding Inverse Document Frequency. On Theoretical Arguments for IDF«, in: *Journal of Documentation* 60.5 (2004), S. 503–520.

6 Siehe Paul Rayson und Roger Garside: »Comparing Corpora Using Frequency Profiling«, in: *Proceedings of the Workshop on Comparing Corpora* (Hong Kong, 2000). Shroudsburg 2000, S. 1–6, DOI:10.3115/1117729.1117730.

7 Einen nützlichen Überblick bieten Jürgen Bortz und Christof Schuster: *Statistik für Human- und Sozialwissenschaftler*. Berlin 2010, Kap. 5. Für eine Untersuchung, in der unterschiedliche Distinktivitätsmaße präzise auf ihre statistischen Annahmen und Eigenschaften hin überprüft wurden, siehe Jefrey Lijffijt, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki und Heikki Mannila: »Significance Testing of Word Frequencies in Corpora«, in: *Digital Scholarship in the Humanities* 31.2 (2014), S. 374–397.

zweier Partitionen ein, allerdings unter der Annahme der Normalverteilung der Häufigkeiten, die häufig nicht gegeben ist.[8] Bei der Berechnung der »spécificités«, wie sie im Korpuswerkzeug TXM implementiert ist, wird ein erwarteter Median berechnet und statt der t-Verteilung eine hypergeometrische Verteilung angenommen.[9] Bei der Verwendung des *Wilcoxon Rangsummen-Tests* oder des eng verwandten *Mann-Whitney-U-Test* wird ebenfalls die Verteilung der Merkmalshäufigkeiten berücksichtigt, und zwar ohne, dass die Annahme einer Normalverteilung der Merkmale nötig ist.[10] Ein Vorteil dieser Gruppe von Maßen ist, dass die Unterschiedlichkeit der beiden Verteilungen eines *Types* in der Zielpartition und der Vergleichspartition nicht nur durch die Feststellung der sogenannten Effektgröße präzisiert, sondern darüber hinaus auch auf ihre statistische Signifikanz hin getestet wird.

1.2 Zeta: Mathematische Beschreibung

Das ursprünglich von John Burrows vorgeschlagene Zeta-Maß berücksichtigt auf andere Weise die Streuung der Werte bzw. die mehr oder weniger konsistente Verwendung der *Types* in zwei Partitionen.[11] Dieses Maß ist in unserem Kontext besonders attraktiv, weil es mathematisch sehr einfach und als einziges in den digitalen Geisteswissenschaften entstanden ist. Das Maß selbst sowie eine Implementierung in Python werden im Folgenden genauer beschrieben und im Anschluss für die Analyse eines Fallbeispiels genutzt.[12]

8 Vgl. Michael P. Oakes: *Statistics for corpus linguistics*. Edinburgh 1998, Kapitel 1.3.

9 Pierre Lafon: »Sur la variabilité de la fréquence des formes dans un corpus«, in: *Mots* 1.1 (1980), S. 127–165, DOI:10.3406/mots.1980.1008.

10 Ursprünglich Frank Wilcoxon: »Individual comparisons by ranking methods«, in: *Biometrics Bulletin* 1.6 (1945), S. 80–83; Henry B. Mann und Donald R. Whitney: »On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other«, in: *The Annals of Mathematical Statistics* 18.1 (1947), S. 50–60.

11 Vgl. Burrows: »All the Way Through. Testing for Authorship in Different Frequency Strata«.

12 Das ursprünglich von John Burrows vorgeschlagene Maß wurde von Hugh Craig weiterentwickelt (vgl. Burrows: »All the Way Through. Testing for Authorship in Different Frequency Strata«; vgl. Hugh Craig und Arthur F. Kinney: *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge 2009; siehe auch David L. Hoover: »Textual Analysis«, in: *Literary Studies in a Digital Age*, hg. v. Kenneth M. Price und Ray Siemens. New York 2013, https://dls anthology.mla.hcommons.org/textual-analysis/ (20. Januar 2017). Das Maß wurde in mehreren Varianten von Maciej Eder in *stylo* für R implementiert (vgl. Maciej Eder, Mike Kestemont und Jan Rybicki: »Stylometry with R. A Package for Computational Text Analysis«, in: *The R Journal* 16.1 (2016), S. 1–15, https://journal.r-project.org/archive/accepted/ eder-rybicki-kestemont.pdf

Zeta bringt zwei Faktoren ins Spiel: Erstens werden nicht ganze Partitionen verglichen, sondern einzelne Texte in den beiden Partitionen. Mehr noch, auch die einzelnen Texte werden noch einmal in gleich lange kleinere Segmente (z.B. von 2000 oder 5000 Tokens) aufgeteilt. Dies wirkt sich insbesondere dann aus, wenn sehr lange Texte wie beispielsweise Romane verglichen werden, ist aber auch für die deutlich kürzeren Dramentexte sinnvoll. Zweitens werden nicht die Häufigkeiten der Merkmale in den Segmenten erhoben, sondern es wird stattdessen ermittelt, in wie vielen der Segmente ein Merkmal mindestens einmal vorkommt. In den resultierenden Wert fließt damit vor allem ein, wie konsistent ein Merkmal in den beiden Partitionen verwendet wird.

Für jedes Merkmal *i* wird demnach erhoben, in wie vielen Segmenten es der Zielpartition *Z* einerseits, der Vergleichspartition *V* andererseits, vorkommt. Dies ist die *document frequency* (*df*). Dann werden diese Werte zu Anteilen umgerechnet, indem sie durch die Anzahl der Segmente der jeweiligen Partition, *n(Z)* bzw. *n(V)*, geteilt werden. So erhält man die *document proportions* (*dp*):

$$dp_i(Z) = \frac{df_i(Z)}{n(Z)} \text{ bzw. } dp_i(V) = \frac{df_i(V)}{n(V)}$$

Man erhält Zeta, indem man vom Anteil des Merkmals in der Zielpartition den Anteil des Merkmals in der Vergleichspartition subtrahiert:

$$Zeta_i = dp_i(Z) - dp_i(V)$$

Die beiden Anteile, die in die Berechnung von Zeta einfließen, liegen naturgemäß zwischen 0 und 1. Weil sie durch eine Subtraktion miteinander in Bezug gesetzt werden, liegt der resultierende Wert von Zeta zwischen -1 und 1. Der maximale Wert von 1 für Merkmale, die für die Zielpartition besonders charakteristisch sind, entsteht, wenn das Merkmal in allen Segmenten der Zielpartition vorkommt, also sehr konsistent verwendet wird, und zugleich in keinem Segment der Vergleichspartition vorkommt (Zeta = 1-0=1). Der minimale Wert von -1 für Merkmale, die äußerst uncharakteristisch für die Zielpartition sind, entsteht, wenn das Merkmal in keinem Segment der Zielpartition, aber in allen Segmenten der Vergleichspartition vorkommt (Zeta = 0 -1 = -1). Merkmale, die in jedem Segment beider Partitionen mindestens einmal vorkommen, wie viele äußerst verbreitete Funktionswörter oder Merkmale, die in beiden Partitionen im gleichen Anteil von Segmenten vorkommen, bekommen einen neutralen Wert von 0 (Zeta = 1 - 1 = 0

[20. Januar 2017]). Die hier beschriebene Implementierung entspricht mathematisch der Variante, die in *stylo* als »Craig's Zeta« bezeichnet wird.

bzw. allgemein Zeta = x - x = 0). Berechnet man Zeta für alle Merkmale und sortiert sie absteigend nach dem Wert von Zeta, erhält man zu Beginn der Liste die für die Zielpartition charakteristischen Merkmale und am Ende der Liste die für die Zielpartition besonders uncharakterisischen bzw. unterrepräsentierten Merkmale.

Das Zeta-Maß ist symmetrisch in dem Sinne, als die für die Zielpartition besonders uncharakteristischen Merkmale zugleich diejenigen sind, die für die Vergleichspartition besonders charakteristisch sind. Dies ist insbesondere dann nützlich, wenn Ziel- und Vergleichspartition auf der gleichen Abstraktionsebene angesiedelt sind, beispielsweise bei einem Vergleich der Romane zweier Autoren oder der Gegenüberstellung von Tragödien und Komödien. Wenn die Zielpartition eine sehr präzise definierte Gruppe von Texten enthält, die Vergleichspartition hingegen eine breite Auswahl verschiedener Textsorten versammelt und als Referenzkorpus fungiert, sind die uncharakteristischen Merkmale deutlich weniger konturiert und interpretierbar als die charakteristischen Merkmale der Zielpartition.

Das Verfahren hat dabei einige Effekte auf die resultierenden Wortlisten, die insofern als erwünscht gelten können, als sie unserer Intuition von »Distinktivität« entgegenkommen. Weil nur das Vorkommen eines Merkmals (nicht aber seine Anzahl) in einem Textsegment erhoben wird, können Merkmale, die nur in wenigen Texten oder Textsegmenten vorkommen, dort aber extrem häufig sind, keinen hohen Zeta-Wert erreichen. Das hat die Folge, dass beispielsweise die Namen von Figuren oder Orten, die jeweils in nur einem Text sehr häufig vorkommen, nicht als distinktiv für eine ganze Partition erscheinen. Aus demselben Grund können insgesamt sehr häufige Merkmale, die zwar unterschiedliche relative Häufigkeiten haben, aber letztlich doch in fast jedem Textsegment zumindest einmal vorkommen, ebenfalls keine extremen Zeta-Werte bekommen. Dies hat den Effekt, dass die weit verbreiteten Funktionswörter ebenfalls nicht als distinktive Merkmale erscheinen. Damit stellt Zeta die Inhaltswörter von insgesamt mittlerer Häufigkeit in den Vordergrund. Diese spezifischen Eigenschaften von Zeta müssen allerdings erkannt werden, wenn Zeta sinnvoll eingesetzt werden soll.

Das Verhalten des Zeta-Verfahrens ist von drei Parametern beeinflusst. Erstens die Länge der Segmente: je kürzer die Segmente sind, desto feiner wird die Verteilung der Merkmale betrachtet, desto wahrscheinlicher wird es aber auch, dass zahlreiche Merkmale in vielen Segmenten nicht vorkommen. Zweitens die minimale Häufigkeit, die ein Merkmal haben muss, um überhaupt berücksichtigt zu werden: hier können Merkmale ausgeschlossen werden, die zwar aus dem einen oder anderen Grund nur in einer Partition vorkommen, aber insgesamt dennoch sehr selten sind und damit kaum wirklich charakteristisch sein können.

Dieser Parameter kann auch dazu genutzt werden, um bei größeren Textsammlungen die Berechnung zu beschleunigen, indem beispielsweise alle *hapax legomena* (Merkmale, die nur ein einziges Mal vorkommen) von vorneherein ausgeschlossen werden. Und drittens ist von den Forschenden zu entscheiden, bis zu welchem Wert von Zeta von einer nennenswerten Distinktivität ausgegangen werden soll. Das Maß hat weder einen intrinsischen Schwellenwert, noch ist ein Signifikanztest Teil des Verfahrens. Die beobachteten Extremwerte in einem konkreten Anwendungsfall können zudem deutlich unter 1 bzw. über -1 liegen. Auch ist die Spanne der Werte nur bei anderweitig vergleichbaren Parametern und Eigenschaften der Textsammlungen ein verlässlicher Hinweis auf das Ausmaß der Unterschiedlichkeit der Texte in Zielpartition und Vergleichspartition.

1.3 Zeta: Implementierung in Python

Die Ausgangsbasis für die Berechnungen ist in der für den vorliegenden Beitrag entwickelten Implementierung die Textsammlung, die in Form von einer Textdatei pro Text sowie einer Metadatentabelle mit der Zuordnung jedes Stücks in eine Gruppe (Autoren, Gattungen, Epochen, etc.) vorliegt. Die Implementierung beinhaltet dann die folgenden Arbeitsschritte:

1. Alle Texte werden zunächst vorbereitet, d.h. tokenisiert, lemmatisiert und nach Wortarten ausgezeichnet (hierfür wird der TreeTagger eingesetzt[13]). Je nach gewähltem Zuordnungskriterium wird jeweils eine Liste der Texte erstellt, die der Zielpartition respektive der Vergleichspartition zugeordnet sind.
2. Es findet eine Auswahl der zu berücksichtigenden Merkmale statt: auf Grundlage der Wortlänge, der Worthäufigkeit, der Wortart und/oder durch eine Stoplist. Alle Texte einer Partition werden in Segmente einer festgelegten Länge geteilt.
3. Wie bereits beschrieben wird erhoben, in wie vielen Segmenten der Zielpartition und der Vergleichspartition jedes Merkmal vorkommt; diese Anzahl wird zu Anteilen in Bezug auf die jeweilige Partition umgerechnet. Auf dieser Grundlage wird der Wert von Zeta für jedes Merkmal berechnet.

13 Helmut Schmid: »Probabilistic Part-of-Speech Tagging Using Decision Trees«, in: *Proceedings of International Conference on New Methods in Language Processing*. Manchester 1994, n. p.

4. Als Ergebnis werden eine Tabelle mit den Rohdaten sowie mehrere interaktive Visualisierungen der Merkmale mit den am deutlichsten distinktiven bzw. nicht-distinktiven Werten erstellt.

Gegenüber anderen Implementierungen von Zeta – derjenigen von Maciej Eder in *stylo* für R und derjenigen von David Hoover als Excel-Makro – zeichnet sich die vorliegende Implementierung in Python dadurch aus, dass zusätzlich zu den üblichen Parametern wie der Segmentlänge auch die Möglichkeit gegeben ist, verschiedene Präprozessierungsschritte direkt in den Prozess einzubinden, was eine flexible Definition der zu verwendenden Merkmale ermöglicht. Außerdem berücksichtigt die Implementierung eine Metadaten-Tabelle, in der zu jedem Dokument verschiedene Klassenzugehörigkeiten festgehalten werden können. Die jeweils zu vergleichenden Partitionen innerhalb einer größeren Textsammlung werden auf Grundlage der Metadaten dynamisch generiert, was große Flexibilität bietet. Derzeit in Entwicklung befinden sich Erweiterungen, die einen Konfidenzintervall der zu erwartenden Zeta-Werte ermitteln sowie eine parametrisierte Berücksichtigung der relativen Häufigkeiten bei der Berechnung der Zeta-Werte.[14]

2 Fallstudie zum französischen Theater

In den folgenden Abschnitten wird das Zeta-Maß für die kontrastive Analyse der drei dramatischen Gattungen Komödie, Tragödie und Tragikomödie eingesetzt. Ziel ist es, die Funktionsweise des Zeta-Maßes zu illustrieren und seine Nützlichkeit für kontrastive Analysen literarischer Texte aufzuzeigen. Dabei wird auf Grundlage einer Textsammlung von 391 Dramen der französischen Klassik und Aufklärung auf einer stilistisch-lexikalischen Ebene der Unterscheidung von Tragödie, Komödie und Tragikomödie nachgegangen.[15]

14 Die Implementierung erfolgte in Python 3 und ist in einem GitHub-Repository der CLiGS-Gruppe frei verfügbar: https://github.com/cligs/pyzeta (DOI:10.5281/zenodo.208178).

15 Die hier verwendeten Texte stammen aus der Sammlung *Théâtre classique*, die von Paul Fièvre herausgegeben wird (vgl. Paul Fièvre: *Théâtre classique*. 2007, http://www.theatre-classique.fr [20. Januar 2017]) und derzeit gut 1000 Theaterstücke aus der Zeit von 1600 bis 1810 umfasst. Hier wurden nur Dramentexte aus der Zeit 1630–1780 und mit einer Länge von 3 oder 5 Akten berücksichtigt: 189 Komödien, 150 Tragödien und 52 Tragikomödien, also 391 Texte. Für die hier durchgeführten Analysen wurde ausschließlich der Sprechertext unter Ausschluss von Vorworten, Anmerkungen, Bühnenanweisungen oder Sprechernamen extrahiert.

Die normative Poetik der französischen Klassik betont die klare Unterscheidung der dramatischen Untergattungen, insbesondere zwischen Komödie und Tragödie, lässt zugleich aber zumindest zu Beginn ihrer Wirkmächtigkeit die Möglichkeit zu, dass die Tragikomödie eine solche klare Unterscheidung unterläuft.[16] Vor diesem Hintergrund, und da in vorigen Arbeiten zum gleichen Gegenstandsbereich bereits zahlreiche Unterschiede zwischen Tragödie und Komödie auf der Ebene der Funktionswörter, der Topics, der Textstrukturierung und der Figurenanzahl herausgearbeitet werden konnten, ist es das Ziel des vorliegenden Beitrags, dies nun auch auf der lexikalischen Ebene zu zeigen.[17] Darüber hinaus soll aber auch gezeigt werden, wie sich die Tragikomödie zwischen Komödie und Tragödie verorten lässt.[18] Hierfür werden die distinktiven Merkmale dieser drei dramatischen Untergattungen herausgearbeitet und insbesondere die Verortung der Tragikomödie im Gattungssystem datenbasiert genauer in den Blick genommen.

2.1 Gegenüberstellung von Tragödie und Komödie

In einem ersten Schritt werden Ergebnisse aus der Gegenüberstellung von Tragödie und Komödie präsentiert, für die eine sehr klare Differenzierung zu erwarten ist. Zunächst sei aber ein für das Verständnis des Zeta-Maßes aufschlussreicher Scatterplot (Abb. 1) gezeigt, in dem die Merkmale mit ihren Anteilen in den Komödien einerseits, den Tragödien andererseits aufgetragen sind.

16 Für eine detaillierte Beschreibung der formalen Poetik der Gattung, siehe Jacques Scherer: *La dramaturgie classique en France*. Paris 2001; für einen Überblick über die literarische Produktion der Zeit, siehe Charles Mazouer: *Le théâtre français de l'âge classique, II: L' apogée du classicisme*. Paris 2010.

17 Christof Schöch: »Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik«, in: *Literaturwissenschaft im digitalen Medienwandel*, hg. v. Christof Schöch und Lars Schneider. http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf. Beihefte von Philologie im Netz 7 (2014) (20. Januar 2017); Allen Riddell und Christof Schöch: »Progress through Regression«, in: *Digital Humanities 2014: Conference Abstracts*. Lausanne, http://dharchive.org/paper/DH2014/Paper-60.xml (20. Januar 2017); Christof Schöch: »Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama«, in: *Digital Humanities Quarterly* 11.2 (2017), http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (10. Juni 2017).

18 Zu Fragen der Definition und Geschichte dieser Gattung, siehe Frank H. Ristine: *English Tragicomedy. Its Origin and History*. New York 1963.

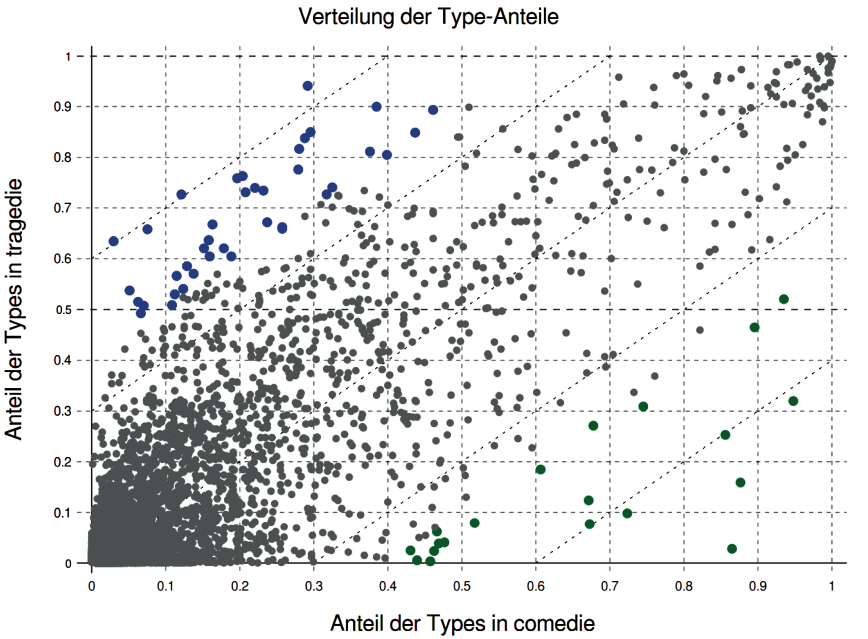

Abb. 1: Scatterplot der Verteilung der Lemmata-Anteile in Komödien und Tragödien

In diesem Scatterplot ist jeder Punkt ein Lemma. Lemmata, die einen besonders hohen oder niedrigen Zeta-Wert haben (hier mit einen arbiträren *cut-off* von +/- 0.40), sind farbig hervorgehoben. Ein Zeta-Wert von 0 entsteht, wenn ein Lemma in beiden Partitionen den gleichen Anteil hat, was für alle Kombinationen von Anteilen zutrifft, die sich auf der mittleren Diagonale befinden. Man sieht, dass ein substantieller Teil der Lemmata wegen ihrer ähnlichen Anteile in beiden Partitionen einen Zeta-Wert um 0 haben. Zudem wird deutlich, dass höhere Zeta-Werte dann zustande kommen, wenn die Anteile des Lemmas in den beiden Partitionen sich deutlich unterscheiden. Je weiter ein Punkt von der 0-Diagonalen entfernt ist, desto extremer ist der resultierende Zeta-Wert. Zwei Lemmata können auch dann einen vergleichbaren Zeta-Wert haben, wenn ihre Anteile sich auf unterschiedlichen Niveaus bewegen: Ein Lemma mit Anteilen von 0.68 und 0.08 in Tragödien respektive Komödien erhält den gleichen Zeta-Wert wie ein insgesamt eher häufiges Lemma mit den Anteilen 0.87 und 0.27, nämlich 0.60. Dies ist kein zweifelsfrei erwünschter Effekt der Berechnungsweise.

Welche inhaltlichen Ergebnisse sind nun aber zu konstatieren? Bei Lemmatisierung und Berücksichtigung aller *Types* ergeben sich für die Gegenüberstellung von Komödien und Tragödien sehr klare und interpretierbare Ergebnisse (Abb. 2).

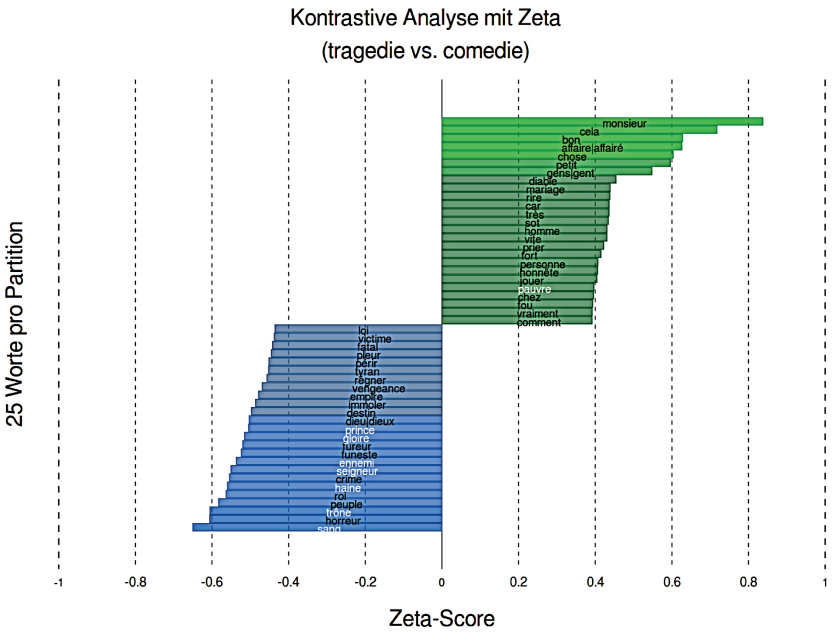

Abb. 2: Lemmata mit den extremsten Zeta-Scores für Tragödie und Komödie

Auf Seiten der für die Tragödie distinktiven Lemmata sind unter den 20 am stärksten distinktiven Begriffen folgende Einordnungen möglich. Die Begriffe *seigneur*, *roi* und *prince* beziehen sich auf das Figureninventar der Tragödie, das sich durch einen hohen Stand auszeichnet. Auch die Begriffe *trône* und *empire* sowie indirekt *peuple* beziehen sich klar auf diesen Kontext königlicher Herrschaft. Die Begriffe *gloire* und *vertu* beziehen sich auf tragische Leitwerte, die zu den wesentlichen Motoren zahlreicher tragischer Konflikte gehören. Die Begriffe *haine*, *fureur*, *horreur*, *malheur* und *pleurs* beziehen sich auf tragische Leit-Emotionen, die sich insbesondere dadurch auszeichnen, dass sie fast alle intensive und negative Emotionen sind. Bleiben hier noch Begriffe wie *sang* (mit dem deutlichsten Wert überhaupt), *crime* und *bras*, die man dem Wortfeld des physischen Konflikts zuordnen kann.

Auf Seiten der Komödie ergibt sich ein völlig anderes Bild: Wiederum beziehen sich mehrere Begriffe auf das Figureninventar, es handelt sich nun aber um den bürgerlichen *monsieur* (direkt dem adeligen *seigneur* der Tragödie gegenübergestellt) sowie um *gens*, *homme* und *personne*. Auffallend sind außerdem die Interjektionen (*oh*, *voilà*). Auch *diable* ist in diesem Sinne verwendbar (»Au diable!«) und dürfte sich auf diese Weise erklären.

Nimmt man eine Selektion der Merkmale in Abhängigkeit der Wortarten vor, kann man den Blick auf die distinktiven Substantive, Verben oder Adjektive jeweils für sich genommen lenken. Ein solches Vorgehen bestätigt und konkretisiert die Ergebnisse aus den Überblicksstudien im Wesentlichen. Blicken wir einmal nur auf die Adjektive (Abb. 3).

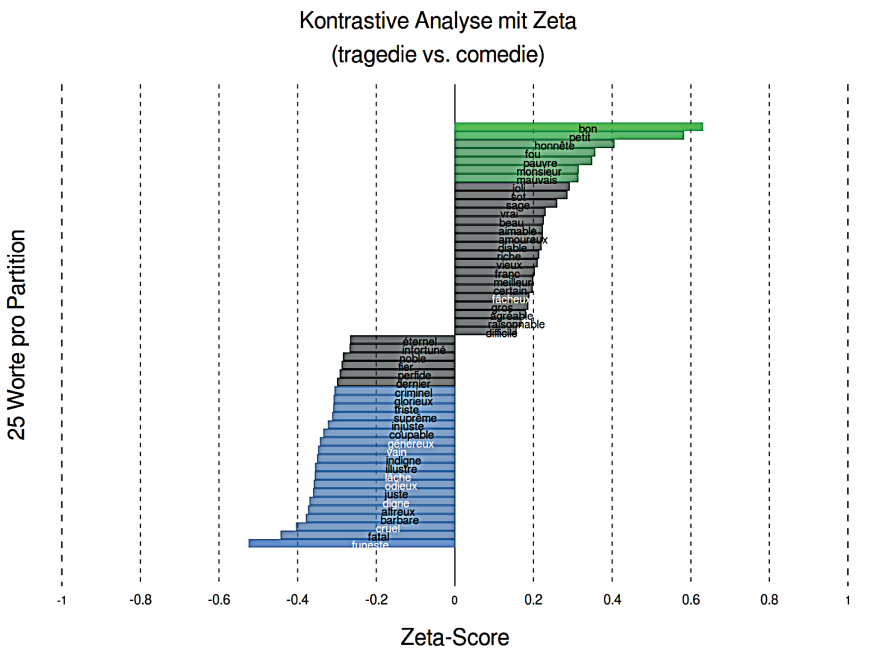

Abb. 3: Adjektive mit den extremsten Z-Scores für Komödien und Tragödien

Die Gegenüberstellung der Adjektive, die jeweils für Tragödien und Komödien distinktiv sind, zeigt, dass hier die großen, edlen, extremen Emotionen der Tragödie einerseits (*implacable*, *auguste*, *magnanime*, *victorieux*, *triomphant*, *intrépide*, *ambitieux*, *inévitable*, *altier*, *immortel*), die einfachen, lustigen Qualitä-

ten der Komödie andererseits (*joli, sot, coquin, vilain, impertinent, gai*) herausgestellt werden, aber auch Aspekte der klassischen *honnêteté* eine Rolle spielen (*galant, honnête, plaisant*). Zugleich ist erkennbar, dass der Abfall der Zeta-Werte naturgemäß viel steiler ist, als wenn alle Wortarten berücksichtigt werden.

Zusammenfassend lässt sich konstatieren, dass die kontrastive Analyse von Tragödien und Komödien im Wesentlichen bekanntes Wissen über das Verhältnis der beiden Gattungen repliziert. Damit ist immerhin der Nachweis geleistet, dass das Verfahren in sinnvoller Weise funktioniert. Mehr noch, aus dem statistisch bedingten Fokus auf die Inhaltswörter (siehe Abschnitt 1.2) ergibt sich die hohe inhaltliche Interpretierbarkeit der Ergebnisse, die das Zeta-Maß attraktiv macht.

2.2 Kontrastive Analyse der Tragikomödie

Der Vergleich der Tragikomödie mit der Tragödie einerseits, der Komödie andererseits, verspricht Ergebnisse, die nicht unmittelbar auf bekanntes Wissen zurückführbar sind. Sie erlauben insbesondere auch Rückschlüsse auf die Frage, welcher der beiden Untergattungen die Tragikomödie auf dieser inhaltlich-lexikalischen Ebene ähnlicher ist.

Hier manifestiert sich nun deutlich die (einleuchtende und gewünschte) Eigenschaft Zetas – wie im Übrigen aller Distinktivitätsmaße –, dass die charakteristischen Merkmale für die Zielpartition (hier: die Tragikomödien) stark davon abhängen, welche Texte in der Vergleichspartition vorhanden sind (hier: Komödien oder Tragödien). Die Gegenüberstellung der Tragikomödie mit der Komödie bringt Begriffe zu Tage, die sich in großen Teilen mit den oben ermittelten distinktiven Begriffen der Tragödie überschneiden: *roi, prince, crime, sang, gloire, dieux, bras, fureur* sind darunter (und auch *couronne* passt dazu). Offenbar spielt aber die reine physische Auseinandersetzung oder zumindest der möglicherweise auch symbolische Machtkampf in der Tragikomödie eine noch größere Rolle als in der Tragödie: *trépas, mort, vainqueur, ennemi, victoire, puissant, puissance* kommen hier hinzu und setzen einen anderen Akzent als in der Tragödie. Die Gegenüberstellung der Tragikomödie mit der Tragödie bringt ebenfalls Überschneidungen mit den oben ermittelten distinktiven Begriffen der Komödie ans Licht, allerdings sind es deutlich weniger: *bon, fort, chose*. Die Figurenbezeichnungen treten hier (anders als bei der Komödie) nicht als distinktiv zu Tage, was auf ein ähnliches Figureninventar von Tragikomödie und Tragödie hinweist. Erneut finden wir hier eine Interjektion (*ha*), allerdings nicht dieselbe wie in der Komödie. Die Tragikomödie ähnelt also in Bezug auf das Inventar der distinktiven lexikalischen Merkmale der Tragödie deutlich mehr als der Komödie.

Eine weitere Versuchsanordnung geht von den jeweils 25 am deutlichsten distinktiven *Types* von Tragödie und Komödie aus. Dann wird für jedes Stück in der Textsammlung erhoben, in welchem Anteil der Textsegmente dieses Stücks die insgesamt 50 *Types* vorkommen. Mit diesen Daten wird eine Hauptkomponentenanalyse durchgeführt, die wie zu erwarten eine klare Trennung zwischen Tragödien und Komödien zeigt, aber auch den spezifischen Ort der Tragikomödien: nahe bei den Tragödien (Abb. 4).

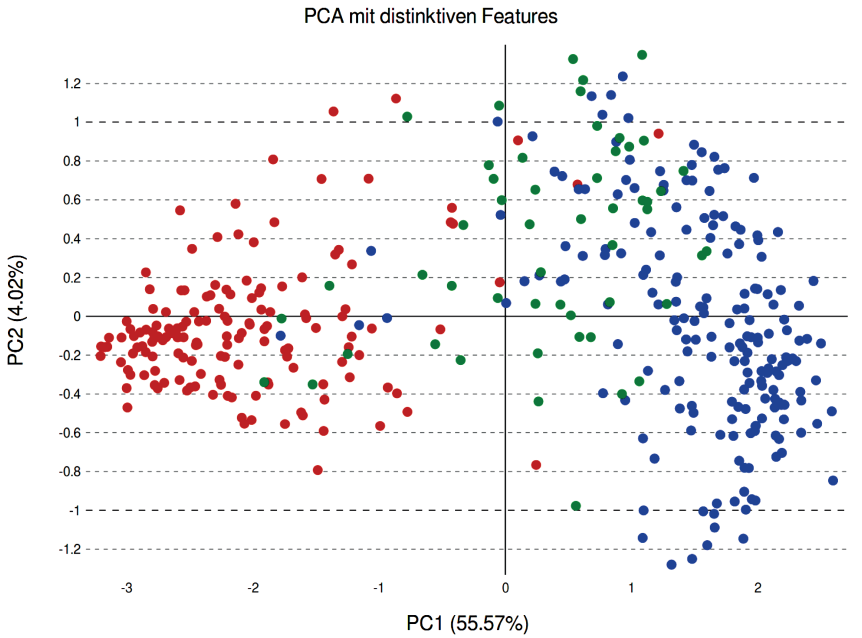

Abb. 4: Vergleich von Komödie (rot), Tragödie (blau) und Tragikomödie (grün)

Die erste Hauptkomponente (PC1) versammelt über 55% der Varianz in den Daten und trennt klar zwischen Tragödien und Komödien. Die Tragikomödien liegen zwar insgesamt zwischen Tragödie und Komödie, ein großer Teil der Tragikomödien aber liegt wesentlich näher bei den Tragödien als bei den Komödien. Die drei Verteilungen der Werte auf der ersten Hauptkomponente unterscheiden sich statistisch hochsignifikant voneinander (*Mann-Whitney-U-Test*). Hier nicht sichtbar, setzt sich die Tragikomödie in der dritten Hauptkomponente leicht von Komödie

und Tragödie gleichermaßen ab. Dennoch kann festgehalten werden, dass ein großer Teil der Tragikomödien den Tragödien sehr viel mehr ähnelt als den Komödien.

3 Fazit

Auf methodischer Ebene lässt sich konstatieren, dass das Zeta-Maß für distinktive Merkmale trotz seiner mathematischen Einfachheit die Information über die mehr oder weniger konsistente Verwendung von Wörtern in sinnvoller Weise für eine kontrastive Analyse von Texten einsetzt. Allerdings steht eine systematische Überprüfung der statistischen Eigenschaften von Zeta, wie sie von Lijffijt und Kollegen für einige andere Distinktivitätsmaße durchgeführt wurde, bislang noch aus.[19] Dennoch kann dazu ermuntert werden, die nun für R und Python vorhandenen Implementierungen von Zeta für sprach- und literaturwissenschaftliche Untersuchungen zu nutzen und die Ergebnisse mit anderen Maßen, insbesondere den von Lijffijt empfohlenen Welchs t-Test und dem Wilcoxon Rangsummentest, zu vergleichen.

Auf die Beispielanalyse bezogen kann man feststellen, dass sich mit Zeta Unterschiede zwischen Tragödie und Komödie bezüglich ihrer lexikalischen Präferenzen herausarbeiten lassen, die unseren Erwartungen zu diesen beiden Gattungen entsprechen. Außerdem zeigt sich recht deutlich, dass sich Tragikomödie und Tragödie bezüglich ihrer lexikalischen Präferenzen deutlich näher stehen als Tragikomödie und Komödie, was die These von der Tragikomödie als Mischform wiederlegt und vielmehr zeigt, dass die untersuchten Tragikomödien sich besser als eine besondere Form der Tragödie beschreiben lassen. Damit erscheint das Verfahren geeignet, einen Beitrag zur Gattungsgeschichte des französischen Dramas zu leisten.

Hinweise

Die für diese Studie verwendeten Daten und alle Zwischenergebnisse und (interaktiven) Visualisierungen sind auf Github verfügbar, unter https://github.com/cligs/projects (Ordner: 2016/zeta-tc; DOI:10.5281/zenodo.208180). Die Arbeit an diesem Beitrag wurde vom BMBF unter dem FKZ 01UG1508 gefördert.

19 Lijffijt et al.: »Significance Testing of Word Frequencies«, S. 374–397.

Der Autor dankt den Mitgliedern der CLiGS-Gruppe, Ulrike Henny, Katrin Betz, José Calvo und Daniel Schlör für die erhellenden Diskussionen rund um die Distinktivitätsmaße im Allgemeinen und Zeta im Besonderen.

Bibliographie

Baron, Alistair, Paul Rayson und Dawn Archer: »Word frequency and key word statistics in historical corpus linguistics«, in: *Anglistik. International Journal of English Studies* 20.1 (2009), S. 41–67.

Bortz, Jürgen und Christof Schuster: *Statistik für Human- und Sozialwissenschaftler*. Berlin 2010.

Burrows, John: »All the Way Through. Testing for Authorship in Different Frequency Strata«, in: *Literary and Linguistic Computing* 22.1 (2007), S. 27–47.

Craig, Hugh und Arthur F. Kinney: *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge 2009.

Eder, Maciej, Mike Kestemont und Jan Rybicki. »Stylometry with R. A Package for Computational Text Analysis«, *The R Journal* 16.1 (2016), S. 1–15, https://journal.r-project.org/archive/accepted/eder-rybicki-kestemont.pdf (20. Januar 2017).

Fièvre, Paul: *Théâtre classique*. 2007, http://www.theatre-classique.fr (20. Januar 2017).

Gries, Stefan T.: »Useful Statistics for Corpus Linguistics«, in: *A Mosaic of Corpus Linguistics. Selected Approaches*, hg. v. Aquilino Sánchez und Moisés Almela. Frankfurt a. M. 2010, S. 269–291.

Hempfer, Klaus W.: »Some Aspects of a Theory of Genre«, in: *Linguistics and Literary Studies / Linguistik und Literaturwissenschaft*, hg. v. Monika Fludernik und Daniel Jacobs. Berlin 2014, S. 405–422.

Hoover, David L.: »Textual Analysis«, in: *Literary Studies in a Digital Age*, hg. v. Kenneth M. Price und Ray Siemens. New York 2013, https://dlsanthology.mla.hcommons.org/textual-analysis/ (20. Januar 2017).

Lafon, Pierre: »Sur la variabilité de la fréquence des formes dans un corpus«, in: *Mots* 1.1 (1980), S. 127–165, DOI:10.3406/mots.1980.1008.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki und Heikki Mannila: »Significance Testing of Word Frequencies in Corpora«, in: *Digital Scholarship in the Humanities* 31.2 (2014), S. 374–397.

Mann, Henry B. und Donald R. Whitney: »On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other«, in: *The Annals of Mathematical Statistics* 18.1 (1947), S. 50–60.

Mazouer, Charles: *Le théâtre français de l'âge classique, II: L' apogée du classicisme*. Paris 2010.

Oakes, Michael P.: *Statistics for corpus linguistics*. Edinburgh 1998.

Rayson, Paul und Roger Garside: »Comparing Corpora Using Frequency Profiling«, in: *Proceedings of the Workshop on Comparing Corpora* (Hong Kong 2000). Shroudsburg: ACL, 2000, S. 1–6.

Riddell, Allen und Christof Schöch: »Progress through Regression«, in: *Digital Humanities 2014: Conference Abstracts*. Lausanne, http://dharchive.org/paper/DH2014/Paper-60.xml (20. Januar 2017).

Ristine, Frank H.: *English Tragicomedy: Its Origin and History*. New York 1963.

Robertson, Stephen: »Understanding Inverse Document Frequency: On Theoretical Arguments for IDF«, in: *Journal of Documentation* 60.5 (2004), S. 503–520.
Scherer, Jacques: *La dramaturgie classique en France*. Paris 2001.
Schmid, Helmut: »Probabilistic Part-of-Speech Tagging Using Decision Trees«, in: *Proceedings of International Conference on New Methods in Language Processing*. Manchester 1994, n.p.
Schöch, Christof: »Corneille, Molière et les autres. Stilometrische Analysen zu Autorschaft und Gattungszugehörigkeit im französischen Theater der Klassik«, in: *Literaturwissenschaft im digitalen Medienwandel*. hg. v. Christof Schöch und Lars Schneider. Beihefte von Philologie im Netz 7 (2014), http://web.fu-berlin.de/phin/beiheft7/b7t08.pdf (20. Januar 2017).
Schöch, Christof: »Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama«, in: *Digital Humanities Quarterly* 11.2 (2017), http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (10. Juni 2017).
Scott, Mike: »PC Analysis of Key Words and Key Key Words«, in: *System* 25.2 (1997), S. 233–245.
Wilcoxon, Frank: »Individual comparisons by ranking methods«, in: *Biometrics Bulletin* 1.6 (1945), S. 80–83.
Yule, George: *The Statistical Study of Literary Vocabulary*. Cambridge 1944.