Burrows Zeta: Varianten und Evaluation

Schöch, Christof
schoech@uni-trier.de
Universität Trier, Deutschland

Zehe, Albin
zehe@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Calvo Tello, José
jose.calvo@uni-wuerzburg.de
Universität Würzburg, Deutschland

Hotho, Andreas
hotho@informatik.uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Der vorliegende Beitrag enthält methodische Überlegungen und Experimente zu "Zeta", einem von John Burrows (2007) vorgeschlagenen Maß für die Distinktivität oder "keyness" von textuellen Merkmalen (Wortformen, Lemmata, etc.). Mit solchen Maßen werden Merkmale ermittelt, die für eine bestimmte Gruppe von Texten gegenüber einer Vergleichsgruppe charakteristisch sind.

Das Exposé gibt einen Überblick zu solchen Maßen, bevor die Funktionsweise von Zeta erläutert wird. Aufbauend auf einer Neu-Implementierung in Python ("pyzeta", https://github.com/cligs/pyzeta) und Vorarbeiten (Schöch im Druck) liegt der spezifische Forschungsbeitrag dann in den folgenden Schritten: erstens werden mehrere Varianten von Zeta vorgeschlagen und implementiert; zweitens werden Verfahren zum Vergleich und der Evaluation der Ergebnisse erprobt. Ziel ist es, Zeta in seiner Funktionsweise und in seiner Beziehung zu vergleichbaren Maßen besser zu verstehen und vorhandene Nachteile des Maßes durch gezielte Modifikationen zu beheben.

Überblick und Stand der Forschung

Die vergleichende, kontrastierende Analyse zweier Gruppen von Texten ist ein in den Sprach- und Literaturwissenschaften weit verbreitetes Verfahren. Entsprechend wurden zahlreiche Maße der Distinktivität oder "keyness" von Merkmalen entwickelt und für vielfältige Fragestellungen eingesetzt. Die grundlegende Annahme solcher Maße ist, dass ein Merkmal nicht schon durch seine reine Häufigkeit in einer Textgruppe für diese charakteristisch ist, sondern dass dies auch davon abhängt, wie häufig das Merkmal in einer Vergleichsgruppe ist. Diejenigen Merkmale bekommen einen besonders hohen Wert zugewiesen, die in der einen Gruppe sehr häufig sind und zugleich in der Vergleichsgruppe sehr selten sind (Scott 1997, 236). Man kann vier Arten von Verfahren unterscheiden:

1. Verfahren, welche erwartete und beobachtete Werte vergleichen (wie "log-likelihood-ratio"; siehe Rayson und Garside 2000);
2. Verfahren, die eine Gewichtung der Häufigkeiten vornehmen (wie "tf-idf", "term frequency / inverse document frequency"; siehe Robertson 2004);
3. Statistische Hypothesentests, die Verteilungseigenschaften vergleichen (wie "Welch's t-Test"; siehe Bortz und Schuster 2010);
4. Dispersionsmaße, die nicht die Häufigkeit, sondern den Grad der konsistenten Verwendung von Merkmalen in Beziehung setzen (wie "deviation of proportions"; Gries 2008).

Die praktische Bedeutung von Distinktivitätsmaßen ist daran erkennbar, dass Korpusanalyse-Software meist eine entsprechende Funktion anbietet, so "keyness" in WordCruncher (Scott 1997) oder "spécificity" in TXM (Heiden et al. 2012). Kilgariff 2004 und Lijfijt et al. 2014 sind wichtige Arbeiten zur Evaluation von Distinktivitätsmaßen.

Was ist Zeta?

Das von John Burrows (2007) vorgeschlagene "Zeta" beruht auf einem Dispersionsmaß. Vor der Berechnung werden die Texte in kleinere Segmente gesplittet, wobei die Segmentlänge ein wichtiger Parameter ist. Dann wird für jedes Merkmal der Anteile der Segmente erhoben, in denen das Merkmal mindestens einmal vorkommt (die "document proportion"). Von diesem Anteil in der untersuchten Gruppe wird der entsprechende Anteil in der Vergleichsgruppe subtrahiert, woraus sich ein Zeta-Wert zwischen -1 und 1 ergibt.

Ein Effekt dieser Berechnungsweise ist, dass Zeta Inhaltswörter als distinktive Wörter favorisiert, Funktionswörter sowie Eigennamen hingegen penalisiert. Daraus ergibt sich eine hohe Interpretierbarkeit der Ergebnisse, die Zeta im Vergleich zu anderen Maßen für die (digitalen) Literaturwissenschaften besonders attraktiv macht. Ein Nachteil ist, dass Merkmale durch die Subtraktion niemals einen Zeta-Wert bekommen können, der höher ist als ihre "document proportion" in der untersuchten Textgruppe, selbst wenn sie gegenüber der Vergleichsgruppe deutlich überrepräsentiert sind (Abbildung 1, Wörter in den roten Rahmen; Schöch im Druck).

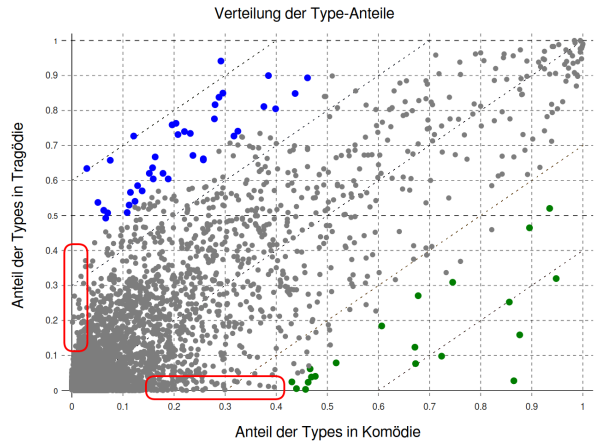

Abbildung 1: Scatterplot der Wörter in zwei Textgruppen (französische Komödien und Tragödien): "document proportions" der Wörter in zwei Textgruppen (x- und y-Achse) und resultierende Zeta-Werte (Distanz von der Diagonale).

Eine bekannte Implementierung von Zeta existiert im stylo-Paket für Rin der Funktion "oppose()" (Eder et al. 2016). Abbildung 2 zeigt für ein Beispiel die Ergebnisdarstellung in der hier verwendeten "pyzeta"-Implementierung.

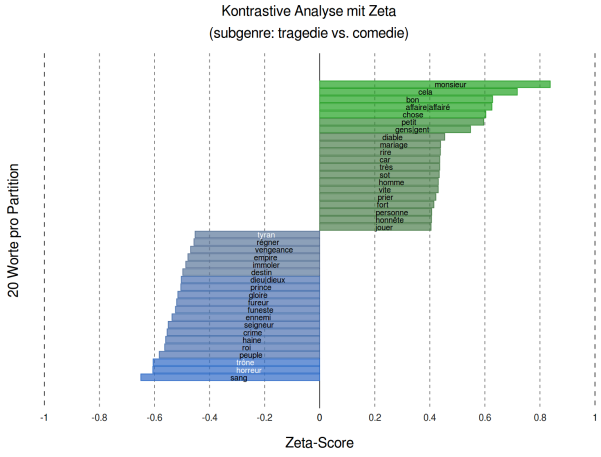

Abbildung 2: Positive und negative Keywords für französische Komödien (rechts) im Vergleich mit Tragödien (links). Zeta-Werte auf der horizontalen Achse.

Anwendungsbeispiele von Zeta gibt es in der Shakespeare-Forschung (Craig und Kinney 2009), der modernen englischsprachigen Literatur (Hoover 2010; Weidman und O'Sullivan 2017) und der Romanistik (Schöch im Druck). In der zuletzt genannten Arbeit zum französischen Theater der Klassik und Aufklärung konnte nicht nur die erwartbare, klare Differenzierung von Komödien und Tragödien gezeigt werden. Vielmehr wurde auch die spezifische Verortung der Tragikomödien deutlich, die nicht als Mischform zwischen Komödien und Tragödien zu verstehen sind, sondern eine besondere Affinität zur Tragödie aufweisen (Abbildung 3).

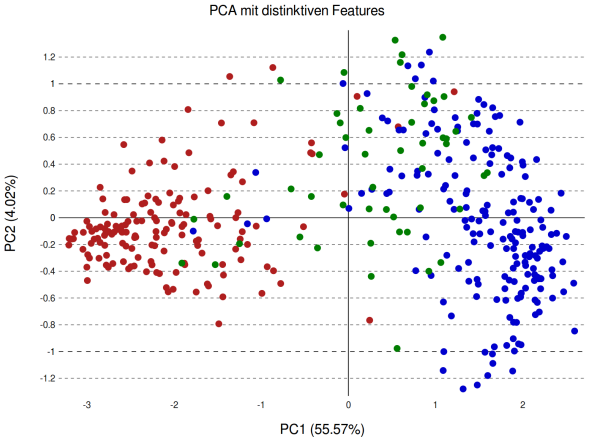

Abbildung 3: Hauptkomponentenanalyse auf Grundlage der 50 Wörter, die für Komödien und Tragödien die höchsten Zeta-Werte erhalten. Komödien in rot, Tragödien in blau, Tragikomödien in grün. Quelle: Schöch im Druck.

Varianten von Zeta

Ausgehend von der ursprünglichen Formulierung von Zeta durch Burrows als Subtraktion der "document proportions" lassen sich mehrere Faktoren identifizieren, die zur Formulierung von Varianten von Zeta geeignet erscheinen:

1. Statt "document proportions" werden relative Häufigkeiten verwendet;
2. Statt der Subtraktion erfolgt eine Division;
3. Statt nicht-transformierter Werte wird eine log2-Transformation der Werte vorgenommen.

Die Kombination dieser Faktoren ergibt 8 Varianten von Zeta (Tabelle 1).

	document proportions		relative Häufigkeiten	
	keine Transformation	log2-Transformation	keine Transformation	log2-Transformation
Subtraktion	sd0	sd2	sr0	sr2
Division	dd0	dd2	dr0	dr2

Tabelle 1: Übersicht über die getesteten Varianten von Zeta. Die Variante mit Label „sd0" entspricht Burrows' Zeta.

Einige der Varianten sind mathematisch gut motivierbar und versprechen, den oben genannten Nachteil der begrenzten Werte für bestimmte Wörter auszugleichen

und damit Zeta zu verbessern, es wurden aber alle implementiert und auf zwei Datensätzen evaluiert.

Datensätze

Es wurden zwei unterschiedliche Korpora verwendet. Erstens ein Korpus aus der textbox-Sammlung (Schöch et al. 2017), das Romane enthält, die zwischen 1880 und 1940 veröffentlicht wurden: jeweils 24 Texte aus Spanien und aus Lateinamerika (ca. 2,8 Millionen Tokens). Zweitens, ein Teil der Sammlung *Théâtre classique* (Fièvre 2007-2017) mit französischen Dramen: 134 Tragödien und 158 Komödien aus Klassik und Aufklärung (ca. 4,9 Millionen Tokens).

Evaluation

Die 8 Varianten führen zu unterschiedlichen Wortlisten, geordnet nach absteigenden Zeta-Werten. Vergleicht man den Beginn der Wortlisten für zwei Varianten, fällt auf, dass es wie erwartet zu Verschiebungen im Rang der distinktivsten Wörter kommt.

Ähnlichkeit der Varianten

Um den Grad der Abweichung der Ergebnisse für alle Varianten zueinander auf der Grundlage längerer Wortlisten zu erheben, ist ein quantifizierendes Verfahren unerlässlich. Ein Ansatz ist, ein Clustering der Maße auf Basis der Zeta-Werte ihrer Wörter vorzunehmen (Abbildung 4).

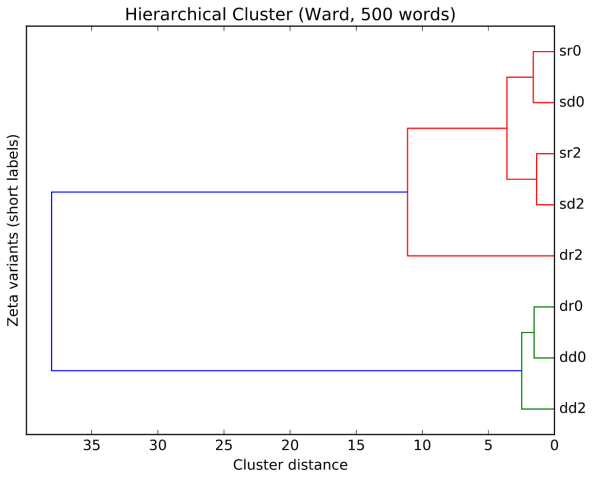

Abbildung 4: Dendrogramm auf Grundlage einer Cluster Analyse der Zeta-Werte für die 8 Zeta-Varianten (Théâtre-classique-Datensatz; 500 distinktive Wörter; Ward-Verfahren).

Abbildung 4 zeigt, dass der wichtigste Faktor für die Unterschiedlichkeit der Varianten ist, ob subtrahiert oder dividiert wird (zwei Haupt-Cluster). Die beiden anderen Variablen spielen eine viel kleinere Rolle. (Die Ergebnisse weiterer Analysen, u.a. auf Basis der RBO-Ähnlichkeit ("ranked biased order", Webber et al. 2010), werden aus Platzgründen hier nicht diskutiert.)

Evaluation mit Klassifikationstask

Unabhängig von den Beziehungen der Varianten zueinander stellt sich die Frage, welche der Varianten von Zeta besonders gut distinktive Wörter identifiziert. Dabei kann zur Evaluation nicht auf einen Goldstandard zurückgegriffen werden: eine händische Annotation der Wörter nach dem Grad ihrer Distinktivität ist nicht möglich, weil niemand das zugrunde liegende Korpus überblicken kann. Die Qualität eines Distinktivitätsmaßes kann aber evaluiert werden, indem es als Merkmalsselektor für einen Klassifikationstask verwendet wird.

Wenn die durch Zeta am höchsten bewerteten Wörter als Features für einen Klassifikator verwendet werden, sollte dieser Klassifikator eine höhere Genauigkeit erreichen als bei einfacher Verwendung der häufigsten Wörter. Tatsächlich lässt sich dieser Effekt auf dem Korpus der spanisch-sprachigen Romane nachweisen (Tabelle 2). Zur Ermittlung einer Baseline wurde für die Klassifikation in spanische und lateinamerikanische Romane ein linearer SVM-Classifier auf den häufigsten 80, nach TF-IDF gewichteten Wörtern (ohne Stoppwörter) trainiert. Dieser Classifier erreichte lediglich eine Klassifikationsgüte (F1-Score) von 0.49, ist also nicht vom Zufall zu unterscheiden.

Trainiert man stattdessen auf den 40 distinktivsten Wörtern nach Zeta (oder einer der Varianten), lassen sich Genauigkeiten von deutlich über 90% erzielen. Diese Genauigkeit kann nicht als tatsächliches Klassifikationsergebnis gesehen werden, da die distinktivsten Merkmale auf dem gesamten Korpus extrahiert wurden, ohne Aufteilung in Trainings- und Testdaten. Dennoch zeigt das Ergebnis, dass die von Zeta selektierten Merkmale tatsächlich sehr nützlich für eine Klassifikation sind. Zudem zeigen sich deutliche Unterschiede in der Performanz je nach verwendeter Variante: während mit "sd0" (=Burrows Zeta) 81% Genauigkeit erreicht wird, erhöht sich dieser Wert bei der Variante mit log2-Transformation, "sd2", auf 98%.

baseline	sd0	sd2	sr0	sr2	dd0	dd2	dr0	dr2
0.49	0.81	0.98	0.48	0.83	0.79	0.85	0.75	0.79

Tabelle 2: Klassifikationsergebnisse bei Verwendung einer linearen SVM, trainiert auf den 40 am höchsten gerankten Wörtern verschiedener Maße im Vergleich zur Baseline. Alle Werte sind der Durchschnitt einer dreifachen Kreuzvalidierung.

Fazit

Wichtigste Ergebnisse dieses Beitrags sind ein differenziertes Verständnis davon, wie Zeta im Kontext anderer Distinktivitätsmaße einzuordnen ist und wie bestimmte mathematischen Parameter sich auf die Ergebnislisten auswirken: als ein auf dem Grad der Dispersion der Merkmale beruhendes Maß, dessen entscheidende Eigenschaft die Subtraktion der Werte ist. Ein weiteres wesentliches Ergebnis sind die beiden vorgeschlagenen Strategien zum Vergleich und der Evaluation von Distinktivitätsmaßen, wenn eine direkte Evaluation auf Goldstandard-Daten nicht möglich ist.

Nächste Schritte: Wir möchten als weitere Evaluationsstrategie künstliche Texte generieren, in denen wir kontrolliert einzelne Wörter mit unterschiedlich stark abweichender Verteilung einfügen. So können verschieden Zeta-Varianten direkt dahingehend evaluiert werden, wie gut sie diese Wörter korrekt identifizieren. Zudem möchten wir neben der "document proportion" von Zeta ein weiteres Dispersionsmaß, die von Gries (2008) vorgeschlagene "deviation of proportions" als Grundlage für eine weitere Zeta-Variante verwenden. Schließlich möchten wir untersuchen, ob die hohe Interpretierbarkeit des Original-Zeta bei den Varianten mit noch höherer Klassifikationsgüte erhalten bleibt.

Eine separate Untersuchung ist in Vorbereitung zu zwei eng zusammenhängenden Fragen: wie sich unterschiedliche Segmentlängen einerseits auf die Ergebnisse auswirken, und wie sich die Ergebnisse verändern, wenn unterschiedlich lange Texte nicht mit allen Segmenten in die Berechnung eingehen, sondern aus jedem Einzeltext zufällig eine identische Anzahl von Segmenten gesampelt wird.

Übergeordnetes Ziel all dieser Arbeiten zu Zeta ist es letztlich weniger, ein perfektes Distinktivitätsmaß zu identifizieren, als ein justierbares Maß vorzuschlagen, bei dem in Abhängigkeit von Daten und Forschungsfragen dynamisch Parameter verändert und die resultierenden Verschiebungen in den Ergebnissen visualisiert werden können.

Bibliographie

Bortz, Jürgen, and Christof Schuster (2010). *Statistik für Human- und Sozialwissenschaftler*. 7. Auflage. Berlin: Springer.

Burrows, John (2007). "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22, no. 1: 27–47. doi:10.1093/llc/fqi067.

Craig, Hugh, and Arthur F. Kinney, eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. 1st ed. Cambridge University Press.

Eder, Maciej, Mike Kestemont, and Jan Rybicki(2016). "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 16, no. 1: 1–15.

Fièvre, Paul, ed. (2007-2013). "Théâtre classique." Paris: Université Paris-IV Sorbonne. http://www.theatre-classique.fr.

Gries, Stefan Th. (2008). "Dispersions and Adjusted Frequencies in Corpora." *International Journal of Corpus Linguistics* 13, no. 4: 403–37. doi:10.1075/ijcl.13.4.02gri.

Heiden, Serge (2010). "The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme." In *24th Pacific Asia Conference on Language, Information and Computation - PACLIC24*, edited by Ryo Otoguro, Kiyoshi Ishikawa, Hiroshi Umemoto, Kei Yoshimoto, and Yasunari Harada, 389–98. Sendai: Waseda University. https://halshs.archives-ouvertes.fr/halshs-00549764/en.

Hoover, David L. (2010). "Teasing out Authorship and Style with T-Tests and Zeta." In *Digital Humanities Conference*. London: ADHO. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-658.html.

Kilgarriff, Adam (2001). "Comparing Corpora." *International Journal of Corpus Linguistics* 6, no. 1: 97–133. doi:10.1075/ijcl.6.1.05kil.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2014). "Significance Testing of Word Frequencies in Corpora." *Digital Scholarship in the Humanities* 31, no. 2: 374–97. doi:10.1093/llc/fqu064.

Rayson, Paul, and R. Garside (2000). "Comparing Corpora Using Frequency Profiling." In *Proceedings of the Workshop on Comparing Corpora*, 1–6. Hong Kong: ACM.

Robertson, Stephen (2004). "Understanding Inverse Document Frequency: On Theoretical Arguments for IDF." *Journal of Documentation* 60, no. 5 : 503–20.

Schöch, Christof (im Druck). "Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie." In *Quantitative Verfahren in der Literaturwissenschaft. Von einer Scientia Quantitatis zu den Digital Humanities*, edited by Andrea Albrecht, Sandra Richter, Marcel Lepper, Marcus Willand, and Toni Bernhart. Berlin: de Gruyter. https://cligs.hypotheses.org/files/2017/09/Schoech_2017-preprint_Zeta-fuer-die-kontrastive-Analyse.pdf.

Schöch, Christof, José Calvo Tello, Ulrike Henny-Krahmer, and Stefanie Popp (angenommen). "The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in XML-TEI." *Journal of the Text Encoding Initiative* http://cligs.hypotheses.org/files/2017/09/Schoech-et-al_2017_Textbox.pdf.

Scott, Mike (1997). "PC Analysis of Key Words and Key Key Words." *System* 25, no. 2: 233–45.

Webber, William, Alistair Moffat, and Justin Zobel (2010). "A Similarity Measure for Indefinite Rankings." *ACM Trans. Inf. Syst.* 28, no. 4: 20:1–20:38. doi:10.1145/1852102.1852106.