Kontrastive Textanalyse mit pydistinto

Ein Python-Paket zur Nutzung unterschiedlicher Distinktivitätsmaße

Du, Keli
duk@uni-trier.de
Universität Trier, Germany

Dudar, Julia
dudar@uni-trier.de
Universität Trier, Germany

Rok, Cora
rok@uni-trier.de
Universität Trier, Germany

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

Viele Wissenschaftsbereiche, die sich mit der quantitativen Textanalyse beschäftigen, wie die Korpuslinguistik oder die Computational Literary Studies (CLS) setzen verschiedene statistische Distinktivitätsmaße ein, um Elemente (z.B. Wortformen oder Wortarten) zu bestimmen, die charakteristisch für eine Textgruppe im Vergleich mit einer anderen Textgruppe sind. Tools wie z.B. *WordSmith* (Scott 2020) oder *AntConc* (Anthony 2005), die solche Analysen ermöglichen, sind weit verbreitet, haben jedoch einige Nachteile: Die meisten bieten nur häufigkeitsbasierte Maße (z.B. Log-Likelihood-Ratio Test oder Chi-Squared Test) an, die in vielen Fällen Ergebnisse produzieren, die für die kontrastive (explorative) Textanalyse nicht hilfreich sind (siehe u.a. Baker 2004 und Johnson and Ensslin 2006). Dispersionsmaße wie z.B. DP (Gries 2008) oder dispersionsbasierte Distinktivitätsmaße wie z.B. Zeta (Burrows 2007), die besser interpretierbaren Ergebnisse liefern (siehe Gries 2021; Schöch 2018), werden dagegen nicht implementiert. Eine Ausnahme bildet *stylo* , das Zeta implementiert (Eder et al. 2016). Ein weiterer Nachteil ist, dass bei den meisten Tools nur ein oder zwei Maße für die Analyse ausgewählt werden können, was einen Vergleich der unterschiedlichen Maße erschwert. Gerade wenn Nutzende ihre Analysen anpassen und eigene Parametereinstellungen vornehmen oder verschiedene Datenformate nutzen wollen, erweisen sich die meisten Tools als ungeeignet.

Um den Einsatz relevanter Maße für die kontrastive Textanalyse zu erleichtern und das Bewusstsein für die Vielfalt der Maße zu schärfen, entwickeln wir im Rahmen des Projekts „Zeta and Company“ ein Python-Paket mit dem Namen *pydistinto* .[1] Ziel unseres Projekts ist es, zu einem tieferen Verständnis der verschiedenen Distinktivitätsmaße zu gelangen und Verbesserungen für deren Implementierung und Anwendung vorzuschlagen. Mithilfe von *pydistinto* können zwei Textkorpora mit unterschiedlichen Maßen verglichen werden.

Hierfür haben wir zunächst ein konzeptionelles Framework erstellt, auf dessen Basis die Maße in *pydistinto* implementiert werden (Du et al. 2021a). Das Framework definiert die Bereiche Preprocessing, Berechnung von Häufigkeiten, Korpusaufteilung sowie der eigentlichen Berechnung der Distinktivitätswerte, Visualisierung sowie quantitative und qualitative Evaluation der Ergebnisse.

In der Implementierung umfasst das Preprocessing die Tokenisierung, Lemmatisierung und das POS-Tagging der Texte. Danach werden die Texte je nach Parameter entweder segmentiert (dies wird bei der Berechnung von dispersionsbasierten Maßen empfohlen) oder als ganze Dokumente belassen. Die (absoluten, binären, relativen usw.) Worthäufigkeiten in den Segmenten bzw. Dokumenten werden in einer Matrix zusammengefasst. Als Nächstes werden die Segmente bzw. Dokumente in zwei Gruppen, ein Ziel- und ein Vergleichskorpus, aufgeteilt. Anschließend werden die Distinktivitätswerte auf Basis der Worthäufigkeits-Matrizen berechnet und die distinktiven Wörter für das Zielkorpus visualisiert. Die Implementierung des Moduls zur quantitativen Evaluation steht noch aus. Geplant ist hier, die statistischen Eigenschaften der Wortlisten zu analysieren und die Korrelation verschiedener Maße zu untersuchen (siehe, für Zwischenergebnisse, Du et al. 2021c). Bei der qualitativen Evaluation werden die ausgegebenen Wörter manuell interpretiert und ihre Relevanz für das Zielkorpus wird beurteilt.

Das Python-Paket wird auf Github veröffentlicht und steht somit zur freien Nutzung, eigenen Anpassung und weiteren Entwicklung zur Verfügung (Du et al. 2021b). Im pydistinto sind derzeit folgende Distinktivitätsmaße implementiert: Zeta, Ratio of Relative Frequencies, Gris’ Deviation of Proportions based measure (Eta, siehe Du et al. 2021c), Welch's T-test, Wilcoxon Rank-sum Test, Kullback-Leibler Divergence, Chi-Squared Test, Log-Likelihood-Ratio Test, TF-IDF. Ein besonderer Vorteil des Pakets ist, dass es in einem Beginner-Modus und einem Profi-Modus genutzt werden kann. Im Beginner-Modus können auch weniger erfahrene Nutzende mit geringen Programmier- und Statistikkenntnissen Textkorpora vergleichen. Ziel- und Vergleichskorpus müssen hierfür lediglich als ‘plain text’ vorbereitet und einige Parameter wie z. B. Segmentlänge, Feature-Typen oder Anzahl der Top-Features eingestellt werden. Die Analyse wird dann automatisch durchgeführt und eine Visualisierung angeboten. Wer sich für die statistischen Eigenschaften der unterschiedlichen Maße interessiert und diese vergleichen möchte, kann den Profi-Modus verwenden. Die Nutzenden können dann selbst darüber bestimmen, welche Maße und statistischen Eigenschaften der Features (z.B. absolute Häufigkeit, relative Häufigkeit, Dispersion) für die Berechnung der Distinktivität kombiniert werden sollen. Es gibt in diesem Modus außerdem zusätzliche Möglichkeiten, die Daten zu visualisieren: so kann die Abhängigkeitsstruktur zweier statistischer Merkmale (z.B. Zeta-Wert und absolute Häufigkeiten der Features) auch durch ein Streudiagramm dargestellt werden.

Durch die Entwicklung des Pakets möchten wir auf der einen Seite eine reflektierte Nutzung statistischer Distinktivitätsmaße für die kontrastive Textanalyse erleichtern. Auf der anderen Seite soll das Paket ermöglichen, die Eigenschaften und Leistungsfähigkeit der Maße empirisch zu ermitteln und systematisch zu vergleichen.

Fußnoten

1. Das Projekt gehört zum DFG-geförderten Schwerpunktprogramm “Computational Literary Studies” (SPP 2207) und läuft

von 2020-2023. Weitere Informationen unter https://zeta-project.eu/de/.

Bibliographie

Baker, Paul (2004): "Querying keywords: questions in difference, frequency, and sense in keyword analysis", in: *Journal of English Linguistics* 32 (4), pp. 346–59

Du, Keli / Dudar, Julia / Rok, Cora / Schöch, Christof (2021a): Implementation framework of measures of distinctiveness. Zenodo. http://doi.org/10.5281/zenodo.5092328

Du, Keli / Dudar, Julia / Schöch, Christof (2021b): pydistinto. Version 0.1.0. Verfügbar unter: https://github.com/Zeta-and-Company/pydistinto . DOI: https://doi.org/10.5281/zenodo.5094346 .

Du, Keli / Dudar, Julia / Rok, Cora / Schöch, Christof (2021c): " Zeta & Eta: An Exploration and Evaluation of Two Dispersion-based Measures of Distinctiveness", in: *CHR 2021: Computational Humanities Research Conference* , November 17–19, 2021, Amsterdam, The Netherlands, https://2021.computational-humanities-research.org/conference/

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): " Stylometry with R: a package for computational text analysis", in: *R Journal* , 8(1): 107-21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.htm l

Gries, Stephan (2008): "Dispersions and adjusted frequencies in corpora", in: *International Journal of Corpus Linguistics, Volume* 13(4): 403–437. DOI: https://doi.org/10.1075/ijcl.13.4.02gri

Gries, Stephan (2021): "A New Approach to (Key) Keywords Analysis: Using Frequency, and Now Also Dispersion", in: *Research in Corpus Linguistics*, 9, 1–33. DOI: https://doi.org/10.32714/ricl.09.02.02

Johnson, Sally / Ensslin, Astrid (2006): "Language in the news: some reflections on keyword analysis using WordSmith Tools and the BNC", in: *Leeds Working Papers in Linguistics and Phonetics* 11, pp. 96–109. https://www.latl.leeds.ac.uk/wp-content/uploads/sites/49/2019/05/Johnson-Ensslin_2006.pdf

Laurence, Anthony (2005): " AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit", in: *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning* . 7–13.

Schöch, Christof (2018): " Zeta für die kontrastive Analyse literarischer Texte. Theorie, Implementierung, Fallstudie", in: Bernhart, T., et al. (eds.), Quantitative Ansätze in der Literatur- und Geisteswissenschaften. Berlin: de Gruyter, 77– 94. https://www.degruyter.com/viewbooktoc/product/479792 .

Scott, Mike (2020): WordSmith Tools, Version 8, Stroud: Lexical Analysis Software.