# „The Vectorian“ – Eine parametrisierbare Suchmaschine für intertextuelle Referenzen

**Burghardt, Manuel**
burghardt@informatik.uni-leipzig.de
Computational Humanities Group, Universität Leipzig

**Liebl, Bernhard**
Bernhard.Liebl@gmx.org
Computational Humanities Group, Universität Leipzig

## Einleitung: Shakespeare, Intertextualität und computergestützte Erkennung von Zitaten

Shakespeare ist überall. Über alle zeitlichen und medialen Grenzen hinweg finden sich intertextuelle Bezüge auf die Werke von Shakespeare (vgl. Garber, 2005; Maxwell & Rumbold, 2018), der damit nicht nur der meistzitierte und meistgespielte Autor aller Zeiten, sondern auch der meistuntersuchte Autor der Welt ist (Taylor, 2016). Doch wenngleich in zahllosen Studien diverse Einzelaspekte von Shakespeares Werk aus Perspektive der Intertextualitätsforschung gründlich mittels *close reading* untersucht wurden, so gibt es bis heute keinen Überblick, kein Gesamtbild, keine systematische Karte intertextueller Shakespeare-Referenzen für größere Textkorpora. Auffällig ist zudem, dass bislang kaum Verfahren der computergestützten Erfassung intertextueller Shakespeare-Referenzen im Sinne des *distant reading* zum Einsatz kommen. Dies verwundert umso mehr, als dass sich im Bereich der Informatik und des *natural language processing* vielfältige Methoden zur Ermittlung der Ähnlichkeit zwischen Texten finden (Bär et al., 2012; Bär et al. 2015) – und nichts anderes ist Intertextualität letzten Endes. Natürlich ist hier anzumerken, dass die volle Bandbreite intertextueller Phänomene mit bloßen Mitteln der Textähnlichkeitsbestimmung nicht abgedeckt werden kann. Für unser Verständnis von Intertextualität berufen wir uns daher auf die Definition von Genette (1993) – “la présence effective d’un text dans un autre” – wobei wir unter der “effektiven Präsenz” eines Texts in einem anderen tatsächlich eine mehr oder weniger objektiv erkennbare, explizite Referenz an der Textoberfläche verstehen. Die textuelle Umschreibung einer Balkonszene mit einem Mann und einer Frau würden wir demnach nicht automatisch “Romeo and Juliet” zuordnen, was vermutlich auch nicht in allen Fällen korrekt wäre. Die folgende Variante eines bekannten Zitats aus Macbeth (Shakespeares Ursprungsvariante steht jeweils in eckigen Klammern) wäre nach unserem Verständnis hingegen objektiv aus dem Text zu erkennen und eindeutig als intertextuelle Referenz einzuordnen:

By the *stinking* [pricking] of my *nose* [thumbs], something *evil* [wicked] this way *goes* [comes]. *(Terry Pratchett: „I Shall Wear Midnight“).*

Eine weitere methodische Einschränkung machen wir, indem wir Phänomene wie strukturelle Ähnlichkeit (Versmaß, Figurenkonstellation) und stilistische Ähnlichkeit[1], wie sie bspw. in der *Parodie* oder im *Pastiche* üblich sind, zunächst außer Acht lassen. In Erweiterung einer ersten Pilotstudie zur Identifizierung von Shakespearezitaten in der Fernsehserie „Dr. Who“ (Burghardt et al., 2019) erproben wir in einem aktuellen Experiment das Potenzial von *word embeddings* (Mikolov et al., 2013), um so zusätzlich semantisch ähnliche oder zumindest “funktional äquivalente” (Bubenhofer, 2019) Wörter und Phrasen zu identifizieren. Durch die Auswahl unterschiedlicher *embeddings*-Modelle und weiterer, damit einhergehender Parameter (bspw. der Gewichtung anhand von Wortarten, dem Festlegen von Ähnlichkeitsschwellwerten, etc.) kann es mitunter zu sehr unterschiedlichen Ergebnissen kommen. Um hier systematisch Parameterkombinationen zu untersuchen, die möglichst optimierte Werte bzgl. *precision* und *recall* liefern, wurde im Sinne von Molnars (2019) Desiderat eines „interpretable machine learning“ eine parametrisierbare Suchmaschine zur Identifizierung von Shakespeare-Referenzen als Vorstufe für einen *embeddings*-basierten Ansatz umgesetzt.

## The Vectorian

Abb. 1 zeigt die Systemarchitektur der besagten Suchmaschine, die fortan als “The Vectorian”[2] bezeichnet wird. Im *Vectorian* fungieren kurze Shakespeare-Passagen (bspw. „If you prick us, do we not bleed?“) als Queries; Texte, die diese Textteile (wortwörtlich oder als Variante) aufgreifen, stellen im Sinne des Information Retrieval dann die entsprechenden Ergebnisdokumente dar (für einen vergleichbaren Ansatz siehe Manjavacas et al., 2019).

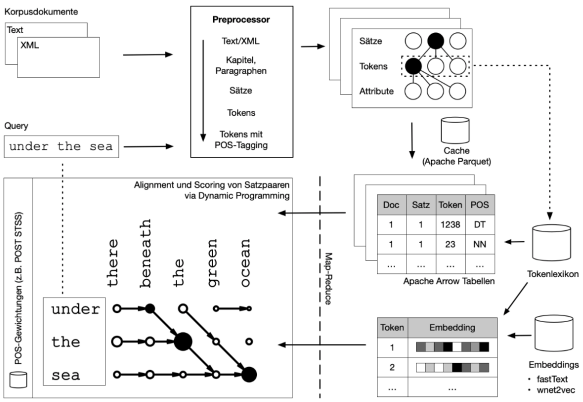


*Abbildung 1: Systemarchitektur der Zitat-Suchmaschine "The Vectorian".*

Kern des *Vectorian* ist die Suche von optimalen semi-globalen *alignments* zwischen Satzpaaren (wobei wir einen Satz als Sequenz von Worten verstehen) über eine Variante des Needleman-Wunsch-Algorithmus (Sellers, 1974) mit sog. *free shift alignment*. Als Bewertungsfunktion nutzen wir eine über *word embeddings* errechnete Distanz zwischen Worten. Diesen Ansatz kombinieren wir mit einer Reihe experimenteller Parameter (siehe die fünf Punkte im nachfolgenden Abschnitt).

Abb. 2 zeigt das Frontend des *Vectorian*. Zu sehen ist ein Eingabefeld für beliebige Suchanfragen, d.h. die Textstellen, die man als intertextuelle Referenzen in anderen Texten finden will. Die Parameter der Suche, die nachfolgend noch näher erläutert werden, können über entsprechende Auswahlmenüs konfiguriert werden. Schließlich gibt der *Vectorian* eine Ergebnisliste zurück, deren Ranking dem jeweils höchsten Ähnlichkeitswert zwischen der Suchanfrage und einer entsprechenden Textstelle entspricht. Wortwörtliche Zitate haben demnach einen höheren Wert als stark abgeänderte Referenzen mit diversen Auslassungen und Substitutionen.

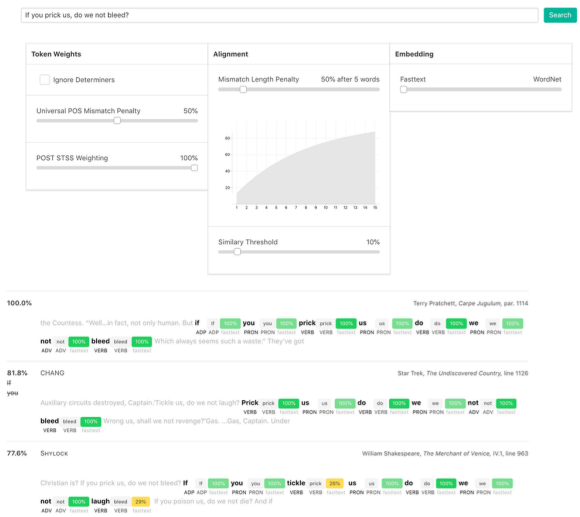


*Abbildung 2: Frontend des Vectorian mit allen möglichen Suchparametern und einer beispielhaften Ergebnisliste für die Suchanfrage "If you prick us, do we not bleed?".*

Der *Vectorian* durchsucht aktuell ein Korpus von 230 englischen Einzeltexten, darunter 50 Werke von Shakespeare (Dramen und Sonette) sowie diverse Romane aus unterschiedlichen Epochen und Transkripte von Filmen und Fernsehserien. Das Korpus enthält rund 19,5 Millionen Tokens mit POS-Annotationen (POS = *parts of speech*), die sich auf rund 2,2 Millionen Sätze verteilen. Der *Vectorian* bietet mit *fastText* (Mikolov, 2017) und *wnet2vec* (Saedi, 2018) momentan zwei *embedding*-Varianten zur Auswahl. Wir nutzen für *fastText* bestehende, vortrainierte Modelle ( https://fasttext.cc/ ), für *wnet2vec* wurde ein eigenes *embedding* auf Basis unseres Korpus mit Hilfe einer leicht angepassten Referenzimplementierung von Saedi et al. ( https://github.com/nlx-group/WordNetEmbeddings ) erstellt. Im *Vectorian* kann entweder eines der beiden *embeddings* ausgewählt werden oder eine gewichtete Kombination aus beiden, bspw. 25% *fastText* und 75% *wnet2vec*. Bei der Suche wird auf dem Suchtext zunächst ein POS-Tagging durchgeführt. So können syntaktische Strukturen, die über die reine Wortreihenfolge hinausgehen, in die Suche einfließen.

Neben den beiden *embedding*-Modellen wurden zusätzlich weitere parametrisierbare Optionen umgesetzt, etwa die Berücksichtigung bzw. unterschiedliche Gewichtung von Wortarten, Einschüben sowie generell einer graduellen Anpassung des Ähnlichkeitswerts. Diese Parameter werden nachfolgend kurz erläutert.

1. "**Ignore Determiners**" entfernt alle Worte, die vom POS-Tagging als DT ("the", "this", etc.) erkannt wurden, aus der Suchanfrage.
2. "**Ensure POS Match**" ermöglicht das Ignorieren von Worten in den Korpusdokumenten, deren POS-Tags nicht dem der alignierten Worte im Suchtext entsprechen. Die Auswirkung der Einstellung kann graduell abgeschwächt werden.
3. "**POST STSS Weighting**": Nicht alle Wortarten besitzen gleiches semantisches Gewicht für die Bedeutung eines Satzes. Mittels "POST STSS Weighting" gewichten wir daher Wortähnlichkeiten bei der Suche mit einer an POST STSS („part-of-speech tag-supported short-text semantic similarity", Batanovi#, 2015) angelehnten Gewichtungsmatrix[3]. Die Auswirkung dieser Einstellung kann ebenfalls graduell abgeschwächt werden.
4. "**Mismatch Length Penalty**" konfiguriert, ab welcher Länge eines einzelnen *mismatch* im Ergebnis eine Abschwächung der Bewertung um 50% geschehen soll[4]. Eine Streuung von Matches ohne lokale Nähe führt in einem Ergebnis somit zur mehr oder weniger starken Abwertung. Die gesamte Abwertung für ein

Ergebnis errechnet sich als Summe der Abwertungen für alle *mismatches*.

5. "**Similarity Threshold**" regelt den Schwellwert zur Ähnlichkeitsbewertung zwischen Wörtern. Ein niedriger Schwellwert erlaubt bspw. größere Abweichungen und kann dadurch auch zu einem größeren Rauschen durch mehr *false positives* führen.

# Beispielabfragen

Der *Vectorian* wurde als parametrisierbare und interpretierbare Suchmaschine konzipiert, um einen explorativen Zugang zur Analyse unterschiedlicher Parameterkonfigurationen auf potenzielle Suchergebnisse, also in unserem Falle Shakespeare-Referenzen, zu ermöglichen. Nachfolgend illustrieren wir einige Auswirkungen unterschiedlicher Parametereinstellungen am Beispiel der kurzen Shakespeare-Phrase "under the greenwood tree" (aus Shakespeares „As you like it").

Die am besten bewerteten Ergebnisse sind zunächst viele Varianten nach dem Schema „under the X tree", bspw. „under the *chestnut* tree". Mit dem Parameter *mismatch length penalty* kann man zusätzlich steuern, wie viele Einfügungen in den Treffern erlaubt sind. Werden Einfügungen nur in geringem Umfang erlaubt, dann erhält man vor allem Sätze bei denen die Präposition variiert wird, bspw. „**beneath** the *beech* tree". Erlaubt man hingegen mehr Einfügungen, kommt es entsprechend auch zu Ergebnissen wie "under the **dear old** *plane* tree".

Beim Parameter der *embeddings*-Wahl sieht man sehr gut, wie *FastText* und *WordNet* ganz unterschiedliche Präferenzen bei der Auswahl von alternativen „trees" liefern (*FastText*: „*chestnut*" > „*beech*" vs. *WordNet*: „*beech*" > „*oak*"). Das *mixed embedding* (also eine Aktivierung beider *embeddings* zu gleichen Teilen) scheint Vorteile beider *embeddings* optimal zu kombinieren, indem z.B. „oak tree" höher gewertet wird als „bodhi tree", wobei es sich bei Letzterem um einen spezifischen Baum aus einem religiösen Kontext handelt.

POST-STSS, ein Parameter der unterschiedliche POS unterschiedlich stark gewichtet, ist in Kombination mit dem *WordNet embedding* am aufschlussreichsten: Mit POST STSS werden im Zweifel reine Baumphrasen bevorzugt ("the fir tree", "the yew tree"). Ohne POST-STSS werden auch Substantive hoch bewertet, die mit Bäumen zwar nichts zu tun haben, dafür aber eine hohe semantische Nähe zu anderen Wörtern aufweisen, z.B. „greenwood" und „garden".

# Fazit und Ausblick

Im aktuellen Stadium dient der *Vectorian* wie eingangs geschildert zunächst als Experimentierplattform, mit deren Hilfe man explorativ die Auswirkungen unterschiedlicher Einstellungsparameter erproben kann. Im nächsten Schritt soll eine systematische Evaluierung der Suchmaschine erfolgen, indem gegen eine vorab definierte *ground truth* an Shakespeare-Zitaten in einem Teilkorpus aus Fantasy-Romanen gesucht wird. Dabei werden alle möglichen Parameterkonfigurationen (insgesamt 72 Kombinationsmöglichkeiten) nacheinander durchgerechnet und die jeweiligen Bewertungen der einzelnen Sätze dokumentiert. Weiterhin soll berücksichtigt werden, wie viele *false positives* sich unter die *true positives* aus der *ground truth* mischen. Ziel ist es, diejenige Konfiguration zu identifizieren, die für möglichst viele Sätze der *ground truth* einen hohen *alignment score* aufweist und dabei die Zahl der *false positives* minimiert. Im nächsten Schritt sollen dann mit der bestbewerteten Konfiguration systematisch mehrere hundert Shakespeare-Zitate, die aus bestehenden Zitate-Datenbanken wie *WikiQuote* (https://en.wikiquote.org/) extrahiert werden, in einem großen Korpus von Fantasy-Literatur und Transkripten von Filmen und TV-Serien gesucht werden [5].

# Fußnoten

1. Für eine Systematisierung von text reuse Methoden anhand der Kategorien inhaltliche, strukturelle und stilistische Ähnlichkeit vgl. Bär et al. 2012.
2. "The Vectorian" ist als Prototyp auf Anfrage verfügbar.
3. Beispiel: Eine Ähnlichkeit auf einem Adjektiv (Tag JJ) wird mit dem Faktor 0.7 gewichtet, während ein Verb (Tag VB) mit 1.2 gewichtet wird.
4. Die Abwertung über andere Längen erfolgt ausgehend vom gegebenen Basiswert exponentiell in der Länge des *mismatches*, was uns intuitiv und aufgrund der Beobachtungen in (Beeferman, 1997) sinnvoll erscheint. Der genaue Kurvenverlauf für gängige Längen wird im UI als Plot dargestellt.
5. Die Dokumentbasis des *Vectorian* kann flexibel erweitert werden solange die Texte in einem grundlegend bereinigten *plain text*-Format vorliegen.

# Bibliographie

**Bär, D. / Zesch, T. / Gurevych, I.** (2012): Text Reuse Detection using a Composition of Text Similarity Measures. Proceedings of COLING 2012, 167-184.

**Bär, D. / Zesch, T. / Gurevych, I.** (2015): Composing Measures for Computing Text Similarity. Technical Report TUD-CS-2015-0017, TU Darmstadt.

**Batanovi#, V. / Boji#, D.** (2015): "Using Part-of-Speech Tags as Deep Syntax Indicators in Determining Short Text Semantic Similarity". In Computer Science and Information Systems, 12(1), S. 1–31.

**Beeferman, D. / Berger, A. / Lafferty, J.** (1997): A model of lexical attraction and repulsion. In Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, S. 373-380.

**Bubenhofer, N.** (2019): Word Embeddings: Funktionale Äquivalenz statt Synonymie. Publiziert auf Sprechtakel-Blog (2.3.2019), online verfügbar unter https://www.bubenhofer.com/sprechtakel/2019/03/02/.word-embeddings-funktionale-aequivalenz-statt-synonymie/

**Burghardt, M. / Meyer, S. / Schmidtbauer, S. / Molz, J.** (to appear in 2019): "The Bard meets the Doctor" – Computergestützte Identifikation intertextueller Shakespearebezüge in der Science Fiction-Serie Dr. Who. In Book of Abstracts, DHd 2019.

**Garber, M.** (2005): Shakespeare after All. New York: Anchor Books.

**Genette, G.** (1993): Palimpseste. Die Literatur auf zweiter Stufe. Frankfurt am Main: Suhrkamp. Translation of the revised second edition. [Genette, G. (1982). Palimpsestes. La littérature au second degré. Paris: Éditions de Seuil. Revised 2nd edition 1983.]

**Kusner, M. / Sun, Y. / Kolkin, N. / Weinberger, K.** (2015): "From Word Embeddings To Document Distances". In Proceedings of the 32nd International Conference on Machine Learning. Lille, Frankreich.

**Manjavacas, E. / Long, B. / Kestemont, M.** (2019): "On the Feasibility of Automated Detection of Allusive Text Reuse". ArXiv: 1905.02973 [Cs], 8. Mai 2019. http://arxiv.org/abs/1905.02973 .

**Maxwell, J. / Rumbold, K. (eds.)** (2018): Shakespeare and Quotation. Cambridge: Cambridge University Press.

**Mikolov, T. / Chen, K. / Corrado, G. / Dean, J.** (2013): Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

**Mikolov, Tomas, et al.** "Advances in Pre-Training Distributed Word Representations." ArXiv:1712.09405 [Cs], Dec. 2017. arXiv.org, http://arxiv.org/abs/1712.09405.

**Molnar, C.** (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. eBook online verfügbar unter https://christophm.github.io/interpretable-ml-book/

**Saedi, C. / Branco, A. / Rodrigues, J. A. / Silva, J.** (2018, July). Wordnet embeddings. In Proceedings of The Third Workshop on Representation Learning for NLP (pp. 122-131).

**Sellers, Peter H.** (1974): „On the Theory and Computation of Evolutionary Distances". *SIAM Journal on Applied Mathematics* 26, Nr. 4 (Juni 1974): 787–93. https://doi.org/10.1137/0126070 .