

Visuelle Textanalyse vom Distant zum Close Reading mit THeMSE

Lehmann, Marina

marina.lehmann@uni-mainz.de
Johannes Gutenberg-Universität Mainz, Deutschland
ORCID: 0000-0002-6818-6169

John, Markus

markus.john@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Kuczera, Andreas

andreas.kuczera@mni.thm.de
Technische Hochschule Mittelhessen, Deutschland
ORCID: 0000-0003-1020-507X

Einleitung

Textvergleiche begegnen uns überall: Von Rechtschreibprüfung und Autokorrekturvorschlägen über Plagiatserkennungssoftware bis hin zu Versionierungssystemen wie Git – alle vergleichen in der ein oder anderen Form Texte. Auch im Bereich Digital Humanities beschäftigen sich Forschende mit der Frage, wie ähnlich sich Texte sind. Ein Beispiel dafür ist das DFG-Projekt *Das Buch der Briefe der Hildegard von Bingen. Genese – Struktur – Komposition*¹, welches die in verschiedenen Handschriften überlieferten Briefe Hildegard von Bingens digital ediert.

Zwei Forschungsfragen sind im Kontext des Hildegard-Projekts von besonderem Interesse:

1. Welche handschriftenübergreifenden Ähnlichkeiten bestehen zwischen den überlieferten Manuskripten? Ziel ist es, Briefe oder Textbausteine zu identifizieren, die in mehreren Fassungen in gleichem oder ähnlichem Wortlaut auftreten.
2. Welche Ähnlichkeiten bestehen zwischen den Briefen innerhalb einer Handschrift? Diese Frage geht einher mit der Hypothese, dass die Briefsammlungen nicht willkürlich zusammengestellt wurden, sondern kompositorischen Prinzipien folgen. Eine Aufgabe der Forschenden besteht darin, Themen und Leitgedanken zu identifizieren, die innerhalb einer Handschrift wiederkehren, um dadurch Erkenntnisse über die Komposition als Ganzes zu gewinnen.

Um diese Fragen auf möglichst systematische Weise zu beantworten, wurde ein lauffähiger Prototyp für ein

visuelles Analysewerkzeug entwickelt (Demo-Version: <https://mlehma03.pages.gitlab.rlp.net/themse/>). Das Werkzeug *THeMSE* (*Text-Hermeneutic Multilevel Similarity Exploration*) orientiert sich dabei an dem Grundprinzip, visuelle Abstraktionen für ein Distant Reading (Moretti 2000, 54–68) der Texte mit der Möglichkeit zum Close Reading (Burdorf et al. 2007, 126) zu verbinden. Solche visuellen Abstraktionen, wie z.B. Wordclouds oder Netzwerke, können komplexe Zusammenhänge sichtbar machen, die verborgen bleiben, wenn die Analyse direkt im Detail ansetzt. Um die Zusammenhänge sowie die im Distant Reading entstandenen Ideen und Hypothesen detaillierter analysieren zu können, ist es jedoch notwendig, auch die zugehörigen Textabschnitte zu studieren. Daher werden die visuellen Abstraktionen mit dem zugrunde liegenden Textgegenstand verknüpft.

Forschungsstand

Im Bereich der visuellen Textanalysewerkzeuge gibt es verschiedene visuelle Abstraktionen, die unterschiedliche Abstufungen von Nähe bzw. Distanz zum Text umsetzen. Visualisierungen wie *CollateX* (Haentjens Dekker et al. 2015, 452–470) oder *TRAViz* (*Text Re-use Alignment Visualization*) (Jänicke et al. 2015a, i83–i99) ermöglichen ein Close Reading, indem sie Dokumente als *Text Variant Graph* (Schmidt und Colomb 2009, 497–514) visualisieren, um Abweichungen zwischen den Textvarianten verschiedener Editionen erkennbar zu machen. Aber auch diff-ähnliche Werkzeuge wie *eComparatio* (Bräckel et al. 2019, 221–238), das speziell auf lateinische und griechische Texte zugeschnitten ist, oder *Tesserae* (Coffee et al. 2013, 221–228), das intertextuelle Bezüge in lateinischen Texten untersucht, fallen in diese Kategorie.

Auf der anderen Seite des Spektrums stehen Überblicksvisualisierungen, die für ein Distant Reading entwickelt wurden. Ein Beispiel dafür sind die *Fingerprint Matrizen* von Keim und Oelke (2007, 115–122). Sie bieten pro Text eine pixelbasierte Visualisierung bezogen auf ein bestimmtes Merkmal der Texte (z.B. Satzlänge, Lesbarkeit etc.). Ein weiterer Darstellungsansatz ist der *Text Re-use Grid* (Jänicke et al. 2015b, 153–171), eine Matrixvisualisierung mit Farbcodierung, bei der jede Zelle den Text-Reuse zwischen zwei Dokumenten im Korpus repräsentiert. Je intensiver die Farbe, desto mehr Text wurde wiederverwendet.

Darüber hinaus gibt es integrative Ansätze, welche auf verschiedenen Ebenen sowohl Close Reading als auch Distant Reading ermöglichen. *TextComparator*, eine Visualisierung, die im Projekt *ePoetics* zur vergleichenden Analyse verschiedener deutschsprachiger Poetiken entwickelt wurde, bietet einen Überblicksmodus mit zwei Bändern, die auf abstrakte Weise den Textverlauf mit farblich hervorgehobenen Suchwörtern visualisieren, sowie eine Detailansicht des Textes ebenfalls mit farblichen Markierungen (John 2020, 22f.). Innerhalb des gleichen Projekts ist auch der *Varifocal Reader* entstanden, der eine ganze Reihe

von abstrakten Übersichtsansichten parallel zur Textansicht bietet (Koch et al. 2014, 1727).

THEMSE stellt ebenfalls das Zusammenspiel von Close Reading und Distant Reading in den Mittelpunkt. Das Prinzip, die Visualisierung in verschiedene Abstraktionsebenen einzuteilen – sowohl im *TextComparator* als auch im *Varifocal Reader* grundlegend – wurde für *THEMSE* übernommen. *THEMSE* besteht aus drei Ebenen: einer Überblicks-, einer Explorations- und einer Detailebene. Die Überblicksebene (Abb. 1) mit Heatmap ermöglicht es, handschriftenübergreifende Textähnlichkeiten zu identifizieren. Über die Explorationsebene (Abb. 2) können Nutzende flexibel anpassbare Fingerprints für Briefe erstellen lassen, in denen bestimmte Begriffe oder Begriffskombinationen auftreten. Jeder Fingerprint ist über einen Klick direkt mit der Textansicht verknüpft (Detailansicht, Abb. 3). Die drei Ebenen werden im Folgenden anhand konkreter Anwendungsfälle vorgestellt.

Überblick: Handschriftenübergreifende Textähnlichkeit

[...] [T]he seemingly simple question “How similar are two texts?” cannot be answered independently from asking what properties make them similar. (Bär et al. 2011, 515)

Welche Eigenschaften eines Textes bei der Ermittlung von Textähnlichkeit berücksichtigt werden, hängt stark vom jeweiligen Textvergleichsverfahren ab. Daher bietet *THEMSE* mehrere Verfahren zur Auswahl. Je nach Ansatz wird Textähnlichkeit zeichenbasiert als Editierdistanz (Levenshtein), worthäufigkeitsbasiert (als Bag-of-Words-Modell gewichtet per TF-IDF) oder kontextbasiert (doc2vec) definiert.

Bei der Verwendung der Levenshtein-Distanz (Levenshtein 1966, 707–710) wird Textähnlichkeit anhand der übereinstimmenden Zeichen zwischen zwei Texten gemessen. Zwei Texte sind ähnlich, wenn ihre Editierdistanz möglichst gering ist, d.h. wenn möglichst wenig Zeichen ergänzt, gelöscht oder ausgetauscht werden müssen.

Alternativ kann Textähnlichkeit anhand der für die Texte charakteristischen Wörter gemessen werden. Mit dem Bag-of-Words-Modell werden Texte in Vektoren verwandelt, die auf Worthäufigkeiten basieren. Per TF-IDF (Term Frequency – Inverse Document Frequency) (Salton und Buckley 1988, 513–523) werden die Häufigkeiten gewichtet, sodass Wörter, die in einem Dokument häufig vorkommen, im übrigen Korpus jedoch weniger häufig, stärker gewichtet werden. Mithilfe der Kosinusdistanz wird der Abstand zwischen den Dokumentenvektoren im Raum und damit indirekt die Ähnlichkeit zwischen den Dokumenten gemessen. Zwei Texte sind somit ähnlich, wenn sie ähnliche charakteristische Wörter enthalten.

Auch bei doc2vec (Le und Mikolov 2014, 1–9) wird Ähnlichkeit als Kosinusdistanz zwischen Dokumentenvektoren modelliert. Der Unterschied liegt darin, dass die Vektoren

mithilfe von Machine-Learning auf Basis der Wortkontexte berechnet werden. Die zugrunde liegende Annahme besteht darin, dass Wörter, die in ähnlichen Kontexten auftreten, ähnliche Bedeutungen haben (Distributional Hypothesis). Bei doc2vec sind zwei Texte somit ähnlich, wenn die Wörter in ähnlichen Kontexten auftreten.

Über diese Verfahren kann pro Textpaar ein Ähnlichkeitswert ermittelt werden, dessen Skala von 0 (keine Ähnlichkeit) bis 1 (maximale Ähnlichkeit) reicht. Die Ähnlichkeitswerte bilden den Ausgangspunkt für die Überblicksebene. Sie besteht aus einer Heatmap, die von den Text Re-use Grids (Jänicke et al. 2015b, 4) inspiriert ist, diese jedoch minimalistischer umsetzt. Während bei Jänicke et al. zwischen drei verschiedenen Arten von Text Re-use differenziert wird und die Zellen des Grids in ihrer Größe angepasst werden, beschränkt sich *THEMSE* darauf, die errechneten Ähnlichkeitswerte farblich zu codieren. So lässt sich schnell anhand der Farbintensität erkennen, welche Briefe sich ähneln.

Beim ersten Aufruf erstellt *THEMSE* eine Heatmap mit den Standardeinstellungen (Manuskriptausswahl: R-Wr, Textform: normalisiert, Textvergleichsmethode: Levenshtein). Über das Menü können die drei Parameter angepasst werden. Jede Zelle der Heatmap steht für ein Briefpaar mit jeweils einem Brief aus der Handschrift R sowie Wr. Die Farbintensität der Zellen spiegelt die Höhe des errechneten Ähnlichkeitswerts wider. Über dieses Farbmapping können die Nutzenden schnell erkennen, welche Briefe besonders ähnlich sind. Für R und Wr zeigt sich bei Nutzung des Levenshtein-Verfahrens beispielsweise anhand der dunkelroten Diagonalen, dass es für fast jeden Brief in R einen Brief in Wr gibt, der ihm besonders ähnelt (Abb. 1).

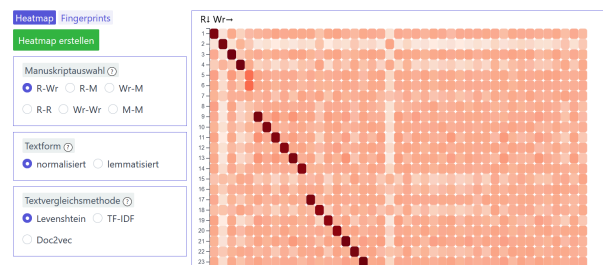


Abb. 1 – Überblicksebene mit Heatmap

Exploration: Handschrifteninterne Ähnlichkeit

Die zweite Ebene (Explorationsebene) implementiert eine Begriffssuche in Kombination mit einer Fingerprint-Visualisierung, durch die das Vorkommen ausgewählter Begriffe und Begriffskombinationen innerhalb einer Handschrift untersucht werden kann. In den Fingerprints von Oelke et al. steht jedes Pixel für eine Texteinheit, z.B. für einen Satz. Über die Farbe des Pixels können Text-

merkmale codiert werden, z.B. Satzlänge oder Lesbarkeit. *TheMSE* sieht pro Wort ein Pixel vor, wobei die farbliche Kodierung das Vorkommen der Suchwörter widerspiegelt (Abb. 2).

Wechselt man zur Fingerprint-Ansicht, erscheint zunächst ein Suchmenü. Neben der Auswahl des Manuskripts für die Suche, können hier Sortier- und Filtereinstellungen vorgenommen werden. Beispielsweise kann festgelegt werden, dass bei mehreren Suchbegriffen nur nach deren Kombination gesucht wird. Zudem kann bestimmt werden, ob die Fingerprints nach Anzahl der Treffer sortiert werden oder nach Anzahl der Wörter zwischen den Treffern. *TheMSE* bietet Suchmöglichkeiten für drei Szenarien:

1) Wissen Forschende bereits, welche Begriffe interessant sind, können sie diese ins Feld der Begriffssuche eingeben.

2) Möchten Forschende stattdessen explorativ vorgehen, können sie sich Begriffsvorschläge basierend auf einem zuvor ausgewählten Brief generieren lassen. Die Vorschläge entsprechen den Top 20 der charakteristischen Begriffe dieses Briefs, d.h. den 20 Begriffen mit den höchsten TF-IDF-Werten.

3) Vielversprechende Begriffe können in einer Merkliste gespeichert werden. Gespeicherte Einträge können ebenfalls für die Suche genutzt werden. Die Suche wird jeweils über den Button “Fingerprints erstellen” gestartet.

Angenommen, die Forschenden möchten in der Handschrift *Wr* suchen, haben jedoch noch keine konkrete Themenvorstellung und betrachten daher die Begriffsvorschläge basierend auf *R26*. Dabei entdecken sie, dass sich aus den Begriffen “pater”, “filius” und “spiritus” das Thema “Trinität” ergeben könnte. Sie wählen die drei Begriffe aus und starten die Suche (Kombinationssuche). Sie vermuten, dass Briefe, in denen die Begriffe in geringer Distanz zueinander auftreten, am wahrscheinlichsten “Trinität” thematisieren. Daher wählen sie die Sortierung nach “Anzahl der Wörter zwischen Treffern” aus. Nun wird für jeden Brief, der diese drei Begriffe enthält, eine Fingerprint Matrix angezeigt (Abb. 2). Die grauen Kästchen der Fingerprint Matrix stehen für ein Wort im Text. Stimmt das Wort mit einem der Suchbegriffe überein, wird es farblich hervorgehoben.

Die Explorationsebene implementiert somit eine weitere Variante des Distant Reading, mit der sich relevante Briefe für das Close Reading identifizieren lassen. Anhand der Profile erkennen die Nutzenden, dass *Wr21* besonders interessant ist, da alle Begriffe häufig und nah beieinander vorkommen, was darauf schließen lässt, dass es in diesem Brief tatsächlich um das gesuchte Thema “Trinität” gehen könnte.

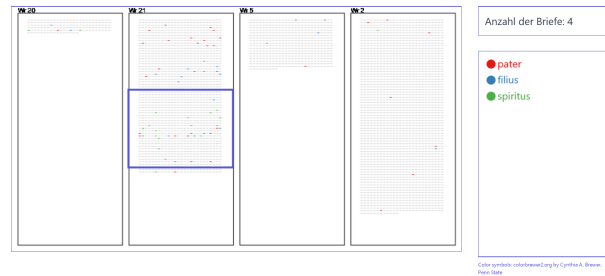


Abb. 2 – Fingerprints in der Explorationsansicht

Detail: Textanalyse im Close Reading

Die Textansicht (Abb. 3) kann per Klick auf die Fingerprints erreicht werden. Längere Fingerprints werden in Abschnitte unterteilt. Durch Hovern über dem Fingerprint lässt sich erkennen, welcher Abschnitt in der Textansicht angezeigt werden wird. Die Textansicht² erscheint als Pop-up, sodass mehrere Briefetexte nebeneinander betrachtet werden können. Die Suchbegriffe sind auch hier farblich hervorgehoben und somit leicht erkennbar. Durch die Detailebene mit Textansicht wird der letzte Schritt vom Distant Reading zum Close Reading vollzogen, indem die abstrakte Textrepräsentation der Fingerprints in eine konkrete Textrepräsentation aufgelöst wird. In unserem Beispiel können die Nutzenden nun lesen, ob es sich wirklich um Texte zum Thema Trinität handelt, und mehrere Briefe zu einem Thema gegenüberstellen und vergleichen.

Brief Wr 21 ✕

Epistola sanctus hildegardus ad idem episcopus Qui sum et qui nichil lateo dico ex pastor non aresco in dulcus fluo odor balsamum qui sum uiredo que prebo sum stultus mens que non habeo uberum mateo misericordius que sugo Que hic non habeo defitio Prebe ergo tuus lampad rex ne in asperitas dispergo et surgo uiuo in lumen Nunc autem o **pater** ego pauperculus forma ad uerus lumen prospeo et secundum quod ibi in uerus uisio uideo et audio qui tu expono peto ita expono non uerbum meus sed uerus lumen qui numquam ullus defectus sum in hic modus transmituo In **pater** maneo eternitas Hoc talis sum Eternitati **pater** nec abscido nec addo sum quia eternitas maneo in similitudo ros que nec incipio nec finis habeo Sed in **pater** sum eternitas ante omnis creatura quia semper et semper eternitas sum Et que sum eternitas Deus sum Eternitas autem non sum eternitas nisi in perfectus uita Ideo deus uiuo in eternitas Uita autem non procedo de mortalitas sed uita sum in uita Arbor enim non flo nisi de uiriditas nec lapis sum sine umor nec ullus creatura sine uis suus Ipsa enim uiuo eternitas non sum sine floriditas Quomodo Uerbum **pater** omnis creatura in officium suus profero Et sic **pater** in fortis uis suus ociosus non sum Unde deus **pater** nomino quoniam omnis ab is nascor Et

Abb. 3 – Pop-up Textansicht

Innovation: Potenziale des visuellen Analysewerkzeugs

THEMSE ist darauf ausgelegt, den hermeneutischen Arbeitsprozess der Nutzenden zu unterstützen und ein produktives Zusammenspiel von qualitativen und quantitativen Herangehensweisen zu ermöglichen (Abb. 4). Seine Modellierung orientiert sich daher am hermeneutischen Zirkel (Schleiermacher 1995; Gadamer 1964) als einem Grundprinzip geisteswissenschaftlichen Arbeitens.

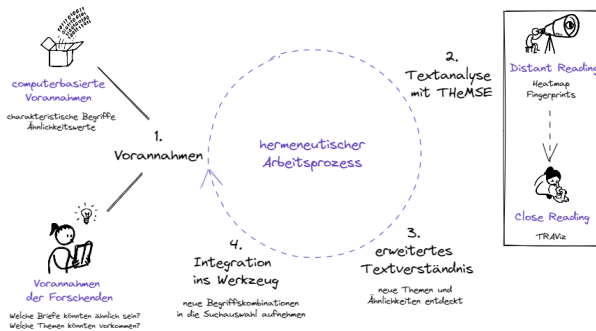


Abb. 4 – Hermeneutischer Arbeitsprozess mit THEMSE

Die Nutzenden bringen ihre Vorannahmen mit, z.B. darüber, welche Briefe ähnlich sind oder welche Themen vorkommen könnten. Zusätzlich enthält das Werkzeug selbst Vorannahmen in Form der Ähnlichkeitswerte sowie der charakteristischen Begriffe. Unter diesen Voraussetzungen analysieren die Nutzenden die Texte im Distant und Close Reading und gewinnen dadurch ein erweitertes Textverständnis. Ihre Erkenntnisse können sie wiederum ins Werkzeug integrieren, indem sie gefundene Themen in Form von Begriffskombinationen abspeichern. Die nächste Analyse baut dann auf dem neuen Wissensstand auf und der geisteswissenschaftliche Erkenntnisprozess wird – unterstützt durch das digitale Werkzeug – weiter fortgesetzt.

Darüber hinaus berücksichtigt *THEMSE*, dass Texte häufig in Sammlungen organisiert sind. Integrierte Ansätze wie der *Varifocal Reader* und *TextComparator* stellen sowohl Textansichten für eine Analyse auf Wortebene als auch abstrakte Darstellungen für eine Analyse auf Textebene bereit. Es fehlt jedoch ein Vergleich auf Sammlungsebene. Die Überblicksansicht in *THEMSE* hilft zu entscheiden, welche Texte aus verschiedenen Handschriften einen Vergleich lohnen – ein Feature, das bei anderen Ansätzen fehlt.

Zuletzt ist *THEMSE* in Bezug auf die Textvergleichsmethoden sehr flexibel: Es wurden drei gängige Methoden ausgewählt, die sehr verschieden arbeiten und somit unterschiedliche Erkenntnisse fördern können. Die Verfahren könnten auch leicht durch andere Verfahren ersetzt oder ergänzt werden, z. B. durch Transformer-Modelle oder Large Language Models.

THEMSE ist somit ein Textanalysewerkzeug, das ausgehend von einem flexibel definierbaren Begriff von Textähnlichkeit Forschenden die Möglichkeit gibt, sowohl handschriftenintern als auch -übergreifend Briefe auf verschiedenen Abstraktionsebenen zu visualisieren und analysieren. Die Verbindung aus hermeneutischem Arbeiten und systematischer datenbasierter Analyse fördert einen erkenntnisorientierten Forschungsprozess, der relevante Merkmale zielgerichtet identifiziert und zugleich Raum zur Exploration lässt.

Fußnoten

1. Beschreibung des Hildegard-Projekts: <https://www.adwmainz.de/projekte/das-buch-der-briefe-der-hildegard-von-bingen-genese-struktur-komposition/projektbeschreibung.html>.
2. Im Prototyp wird in der Textansicht die lemmatisierte Textfassung angezeigt. Perspektivisch wird hier jedoch der Originaltext erscheinen.

Bibliographie

- Bär, Daniel, Torsten Zesch und Iryna Gurevych.** 2011. „A Reflective View on Text Similarity“. In *Proceedings of Recent Advances in Natural Language Processing*, 515–520. Association for Computational Linguistics. <https://aclanthology.org/R11-1071>.
- Bräckel, Oliver, Hannes Kahl, Friedrich Meins und Charlotte Schubert.** 2019. „eComparatio – a Software Tool for Automatic Text Comparison“. In *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, 221–238. Berlin: De Gruyter.
- Burdorf, Dieter, Christoph Fasbender und Burkhard Moennighoff, Hrsg.** 2007. „Close Reading“. In *Metzler Lexikon Literatur: Begriffe und Definitionen*, 126. Stuttgart: J.B. Metzler. <https://doi.org/10.1007/978-3-476-05000-7>.
- Coffee, Neil, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde und Sarah L. Jacobson.** 2013. „The Tesserae Project: Intertextual Analysis of Latin Poetry“. *Literary and Linguistic Computing* 28 (2): 221–228. <https://doi.org/10.1093/lc/fqs033>.
- Gadamer, Hans-Georg.** 1965. *Wahrheit und Methode – Grundzüge einer philosophischen Hermeneutik*. 2. Aufl. Tübingen: Mohr Siebeck.
- Haentjens Dekker, Ronald, Dirk van Hulle, Gregor Middell, Vincent Neyt und Joris van Zundert.** 2015. „Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project“. *Digital Scholarship in the Humanities* 30 (3): 452–470. <https://doi.org/10.1093/lc/fqu007>.
- Jänicke, Stefan, Annette Geßner, Greta Franzini, Melissa Terras, Simon Mahony und Gerik Scheuermann.** 2015a. „TRAViz: A Visualization for

Variant Graphs“. *Digital Scholarship in the Humanities* 30 (suppl_1): i83–99. <https://doi.org/10.1093/llc/fqv049>.

Jänicke, Stefan, Thomas Efer, Marco Büchler und Gerik Scheuermann. 2015b. „Designing Close and Distant Reading Visualizations for Text Re-Use“. In *Computer Vision, Imaging and Computer Graphics – Theory and Applications*, herausgegeben von Sebastiano Battiato, Sabine Coquillart, Julien Pettré, Robert S. Laramée, Andreas Kerren, und José Braz, 550:153–171. *Communications in Computer and Information Science*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-25117-2_10.

John, Markus. 2020. „Interaktive visuelle Analysetechniken für die Exploration narrativer Texte“. Diss. Stuttgart: Universität Stuttgart.

Keim, Daniel A. und Daniela Oelke. 2007. „Literature Fingerprinting. A New Method for Visual Literary Analysis“. In *2007 IEEE Symposium on Visual Analytics Science and Technology*, 115–122. IEEE. <https://doi.org/10.1109/VAST.2007.4389004>.

Koch, Steffen, Markus John, Michael Worner, Andreas Müller und Thomas Ertl. 2014. „VarifocalReader — In-Depth Visual Analysis of Large Text Documents“. *IEEE Transactions on Visualization and Computer Graphics* 20 (12): 1723–1732. <https://doi.org/10.1109/TVCG.2014.2346677>.

Le, Quoc und Tomas Mikolov. 2014. „Distributed Representations of Sentences and Documents“. In *Proceedings of the 31st International Conference on Machine Learning*, 32: 1–9. *JMLR*. <https://doi.org/10.48550/arXiv.1405.4053>.

Levenshtein, Vladimir. 1966. „Binary Codes Capable of Correcting Deletions, Insertions and Reversals“. In *Soviet Physics Doklady* 10 (8): 707–710. <https://www.bibsonomy.org/bibtex/220546d80ce76f58c6ef6ece9dd5f5056/jimregan>.

Moretti, Franco. 2000. „Conjectures on World Literature“. *New Left Review*, Nr. 1: 54–68.

Salton, Gerard und Christopher Buckley. 1988. „Term-Weighting Approaches in Automatic Text Retrieval“. *Information Processing & Management* 24 (5): 513–523. [https://doi.org/10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0).

Schleiermacher, Friedrich. 1995. *Hermeneutik und Kritik*. Herausgegeben von Manfred Frank. 6. Aufl. Suhrkamp Taschenbuch Wissenschaft 211. Frankfurt am Main: Suhrkamp.

Schmidt, Desmond und Robert Colomb. 2009. „A Data Structure for Representing Multi-Version Texts Online“. *International Journal of Human-Computer Studies* 67 (6): 497–514. <https://doi.org/10.1016/j.ijhcs.2009.02.001>.