Handwritten Text Recognition und Word Mover's Distance als Grundlagen der digitalen Edition "Die Kindheit Jesu Konrads von Fußesbrunnen"

Tomasek, Stefan
stefan.tomasek@germanistik.uni-wuerzburg.de
Universität Würzburg, Germany

Reul, Christian
christian.reul@uni-wuerzburg.de
Universität Würzburg, Germany

Wehner, Maximilian
maximilian.wehner@uni-wuerzburg.de
Universität Würzburg, Germany

Gegenstand und Fragestellung des Vortrags

OCR4all und HTR

Zur Zeit entsteht an der JMU Würzburg in einem Kooperationsprojekt zwischen dem Lehrstuhl für ältere deutsche Philologie und dem Zentrum für Philologie und Digitalität ein HTR-Projekt (Handwritten Text Recognition) zur Erfassung mittelhochdeutscher (mhd.) und frühneuhochdeutscher (frnhd.) Handschriften (Hss.) des 11.-15. Jh.s. Ausgangspunkt waren die zunächst in Transkribus[1] erstellten Transkriptionsdaten des Würzburger Editionsprojektes zur 'Kindheit Jesu' Konrads von Fußesbrunnen. Dieses Projekt baut neben den eigentlichen Editionstexten auf einer umfangreichen Datenmenge nicht normalisierten Mhd.s auf (s. u.). Im Open Source Bereich, sowohl mit Blick auf die automatische Texterkennung im Allgemeinen[2] als auch bei der Erkennung vormoderner volkssprachiger Hss., gab es in letzter Zeit erhebliche Fortschritte. Daher kommt mittlerweile für die Erstellung der Datengrundlage für das Editionsprojekt das an Frühdrucken[3] erarbeitete, frei verfügbare Open Source Tool OCR4all[4] zum Einsatz.

Die Grundidee von OCR4all ist es, insbesondere technisch weniger versierten NutzerInnen die Möglichkeit zu geben, anspruchsvolle historische Drucke und Handschriften selbstständig und in höchster Qualität zu erfassen. Dies wird v. a. dadurch ermöglicht, dass einzelne, auf unterschiedliche Schritte des OCR Workflows (Optical Character Recognition) spezialisierte (Kommandozeilen-)Werkzeuge in einem leicht zu installierenden Tool gekapselt und über eine einheitliche Benutzeroberfläche zugänglich gemacht werden. Die Konzeption als Client/Server-Anwendung und die Auslieferung mittels einer Containerlösung erlaubt dabei einen flexiblen Einsatz sowohl lokal beim Einzelnutzer als auch das kollaborative Arbeiten über eine zentralisierte Serverinstanz. Die Bearbeitung eines Werkes kann ebenfalls sehr flexibel erfolgen und an das vorliegende Material und die eigenen Ansprüche angepasst werden. Generell ist ein vollautomatischer Durchlauf möglich, dieser kann allerdings nach jedem Teilschritt unterbrochen und die Ergebnisse kontrolliert und bei Bedarf manuell nachkorrigiert werden, um Folgefehler zu vermeiden.

Im Vergleich zu herkömmlichen OCR-Verfahren ist bereits das Erfassen von Frühdrucken besonders anspruchsvoll, da hier z.T. komplexe Layoutstrukturen vorliegen und der Druck- bzw. Erhaltungszustand erheblich variiert. Zudem unterscheiden sich die verwendeten Drucktypen und Schriftarten innerhalb eines Werkes und zwischen unterschiedlichen Werken. Beide Kriterien gelten für die Erfassung historischer Handschriften in verstärktem Maße, da sich das Layout z.T. innerhalb der gleichen Hs. deutlich ausdifferenziert, die mittelalterlichen Schreiber vielfältige Schreibvarianten verwenden und sich das Schriftbild zwischen den einzelnen Schreibern erheblich unterscheidet bzw. im historischen Längsschnitt weiterentwickelt wurde (siehe hierzu unten).

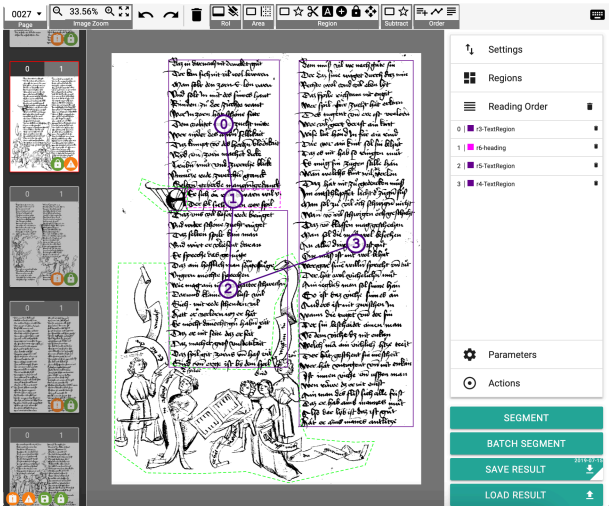

Abb. 1: OCR4all ermöglicht die Segmentierung und Typisierung der Layoutelemente einer Textseite ebenso wie die Festlegung deren Lesereihenfolge.

Abb. 2.: Im Kontext der 'Post Correction' kann automatisch generierter Text mithilfe einer virtuellen Tastatur (rechts) zeichengetreu nachkorrigiert werden.

Digitales Editionsprojekt "Die Kindheit Jesu Konrads von Fußesbrunnen" und Levenshtein-Distanzen bzw. Word Mover's Distance

Das Würzburger HTR-Projekt ist verzahnt mit dem digitalen Editionsprojekt der „Kindheit Jesu Konrads von Fußesbrunnen“ (KJ). Dieses in mhd. Reimpaarversen verfasste apokryphe Kindheitsevangelium (entstanden um 1200) weist erhebliche Varianten zwischen allen Textzeugen auf. Im Editionsprojekt wird daher die vollständige Überlieferungssituation des Textes (vier Haupthss., sieben Fragmente, vier Sekundärzeugnisse) synoptisch abgebildet. Um diese Synopse dennoch lesbar zu halten, soll mit Hilfe von Levenshtein-Distanzen (LevD) und einem Word-Embedding-Verfahren[5] ein Filtersystem ermöglicht werden, mit dem die Genauigkeit des Anzeigemodus’ von den BenutzerInnen der Edition selbst festgelegt werden kann. Zum einen kann über die LevD zeichengenau gefiltert werden. Durch Teilnormalisierungen des Textes sollen zusätzlich verschieden Parameter der Textvarianz (z.B. orthographische Varianz, dialektale Varianz, Graphemvarianten etc.) herauszufiltern sein.[6] Mit diesen LevD-Werten kombiniert werden semantische Distanzwerte: Mithilfe von fastText[7] werden die mhd. Begriffe in ein n-dimensionales Vektorensystem übertragen, in dem durch Word Mover’s Distance[8] (WMD) die semantischen Distanzen zwischen den mhd. Begriffen bestimmt werden. Damit sind jedem Verspaar des mhd. Textes in jeder Hss.-Kombination mehrere Distanzwerte zugewiesen, die wiederum miteinander kombinierbar sind. So ist es für die BenutzerInnen der Edition möglich, die Anzeigegenauigkeit jeweils dem eigenen Leseinteresse entsprechend zu modellieren: Von kleinteiligen Zeichenvarianten über phonetische Unterschiede bis zu abgestuften Bedeutungsvarianten mit niedriger oder hoher semantischer Differenz bzw. Zusatz- oder Fehlversen kann voreingestellt werden, wie genau die Parallelüberlieferung zu der gewählten Lesehandschrift des Textes angezeigt werden soll. Dieser Genauigkeitsfaktor kann jederzeit dynamisch angepasst werden. Gleichzeitig sind die Distanzwerte in die Suchfunktion der Edition integriert, wodurch die historische Überlieferungssituation der KJ nachvollziehbar und abbildbar wird (siehe Abb. 3).

Abb. 3: Die Überlieferungssituation der vier Haupthss. der KJ: Abgebildet sind die Nähe-Distanz-Verhältnisse aller Verspaare in allen Hss.-Kombinationen nach LevD und WMD.

Die LevD und WMD-Werte haben sich für mhd. Texte als prinzipiell anwendbar gezeigt. Da die überwiegende Mehrheit der bisherigen Texteditionen, auf deren Grundlage das WMD-Modell bisher trainiert wurde,[9] einen normalisierten mhd. Sprachstand aufweisen, zeigen sich bei zeichengenauen (nicht normalisierten) Texten dagegen noch deutliche Schwächen: Bestimmte Schreibvarianten, die im konstruierten “Normalmittelhochdeutsch” der traditionellen Editionen nicht vorkommen,[10] werden durch die WMDs nicht als Synonyme erkannt und dementsprechend mit hohen semantischen Distanzen ausgezeichnet.

Da die WMDs diesen Anwendungsfall aber prinzipiell abdecken und gleichzeitig auf die (unnormalisierten) Transkriptionsdateien des genannten HTR-Projektes anwendbar sein sollen (s.u.), muss das Trainingscorpus erheblich ausdifferenziert werden. Daher sind beide Teilprojekte doppelt miteinander verzahnt: Die vollständig vorliegenden Transkriptionsdaten aller Hss. der KJ bilden das Grundmodell für OCR4all, auf dem das HTR-Modell trainiert wird. Alle Ground Truth Daten (GT), die aus den werkspezifischen Modellen generiert werden, bilden umgekehrt das Trainingscorpus für die WMDs der KJ. Die Filterstruktur der KJ wird damit auf einer wesentlich umfänglicheren “Datenbank nicht normalisiertes Mittelhochdeutsch” aufbauen, die ihrerseits aus HTR-Daten besteht.

Trotz insgesamt noch relativ geringer Mengen GT wurden durch das HTR-Projekt innerhalb kurzer Zeit bereits gute Texterkennungsergebnisse erzielt: So konnte in ersten Testdurchläufen auf einer anspruchsvollen Bastarden-Handschrift des 15. Jh.s[11] mit etwa 1.500 Zeilen GT eine Zeichenfehlerrate von etwas mehr als 3% erreicht werden. Auf einer Gotischen Buchschrift des 13. Jh.s[12] wurde mit nur knapp 600 Zeilen sogar eine Fehlerrate von ca. 2% erreicht. Durch die im nächsten Arbeitsschritt angestrebte Erweiterung der Trainingsmenge und weitere Ausdifferenzierung des Trainingsmodells sind werkspezifische Transkriptionsmodelle in Reichweite gelangt, mit denen relativ schnell auch umfängliche Handschriften bzw. Handschriftencorpora erschließbar sind. Perspektivisch sollen auch die bisher erstellten Trainingsdaten[13] frei zur Verfügung gestellt werden (sofern dies die jeweiligen Bildrechte zulassen). Veröffentlicht werden darüber hinaus die derzeit entstehenden gemischten HTR Modelle, die sowohl externen NutzerInnen für eine out-of-the-box Anwendung zur Verfügung stehen als auch die Grundlage weiterer Trainingsprozesse darstellen können. Aus dem Würzburger HTR-Projekt gehen damit drei mögliche Anwendungsfälle hervor, die im Folgenden skizziert werden.

Anwendungsfälle

HTR-generierte Texte als Vorstufe für Editionsprojekte

Den gewissermaßen klassischen Anwendungsfall von (HTR-)Transkriptionen in der älteren deutschen Literaturwissenschaft stellt die zeichengenaue Umschrift als Vorstufe für (digitale) Editionsprojekte dar. Diplomatische Volltranskriptionen werden hier z.B. als Grundlage für stemmatologische Analysen verwendet,[14] auf denen die Auswahl des Editionstextes fußt. Sie bilden den Ausgangspunkt für normalisierte Textstufen und stehen v.a. in den neueren digitalen Editionen gleichberechtigt neben dem normalisierten Lesetext.[15] Die diplomatischen Texte erweitern in Hybrideditionen den printbasierten Editionstext[16] oder ergänzen durch die Darstellung von Text-Bild-Beziehungen die traditionellen Editionen[17] etc. Damit kommt den Transkriptionen der mhd. Hss. in der Editionsphilologie ein erheblicher Stellenwert zu.

Das Würzburger Projekt OCR4all soll hierfür die Datengrundlage für weitere Korrektur- und Bearbeitungsschritte liefern. Um hierbei die Effizienz zu maximieren, wird bei der GT Erstellung iterativ vorgegangen: Nach der initialen Erkennung einiger weniger Seiten mit einem gemischten Grundmodell erfolgt die manuelle Nachkorrektur, die zunächst noch vergleichsweise zeitaufwendig ist (je höher die Fehlerrate, desto größer der Korrekturaufwand). Die so gewonnene werkspezifische GT wird im Anschluss für das Training eines ersten werksspezifischen Modells eingesetzt. Dieses wiederum wird anschließend auf weitere Seiten angewendet und das, im Normalfall bereits deutlich bessere, Ergebnis erneut korrigiert. Dieser Vorgang wird iterativ wiederholt, bis entweder die gesamte Handschrift manuell nachkorrigiert wurde oder ein ausreichend gutes Modell vorliegt, mit dem die übrigens Seiten erkannt werden können. Die Anzahl der beschriebenen Trainings- und Korrekturiterationen sowie der mit ihnen verbundene zeitliche Aufwand hängen stark vom zugrundeliegenden Material und den eigenen/projektspezifischen Qualitätsansprüchen ab. Der oben erwähnte Anwendungsfall einer Gotischen Buchhandschrift deutet einen für OCR/HTR-Modelle typischen Verlauf an: Die initiale Zeichenfehlerrate des gemischten Grundmodells (11%), in dem die zu erfassende Hs. nicht enthalten war, ließ sich in einem werkspezifischen Modell durch die Korrektur und das Training von lediglich drei Seiten (72 Zeilen) bereits auf 3,6% zu reduzieren. In weiteren Iterationen folgten unter Verwendung von sechs, zwölf und 24 Seiten Verbesserungen auf 3,1%, 2,6% und schließlich 2,1%.

Mit diesem Verfahren lässt sich der Zeitaufwand für die Herstellung der Volltranskription einer Hs. erheblich reduzieren. Hierdurch sind nun Editionsprojekte möglich, die auch bei umfänglicher Überlieferungssituation alle Textzeugen transkribieren und in einer Datenbank zur Verfügung stellen können. Daraus resultiert aber das Folgeproblem, dass große Datenmengen entstehen, die ihrerseits von den HerausgeberInnen systematisiert werden müssen. Für diesen Normalfall mhd./frnhd.Texte (mehrere Hss. mit divergenter Überlieferungssituation) soll daher bereits nach der HTR-Transkription mit dem oben beschriebenen kombinierten Filtersystem (LevD und WMD) eine Darstellung der Überlieferungsstruktur geboten werden. Alle Varianten innerhalb der Überlieferung können so systematisch identifiziert und klassifiziert werden. Diese mathematisch nachvollziehbaren Klassifizierungen können wiederum bei der Beschreibung der Editionsrichtlinien als (für die NutzerInnen der Edition überprüfbare) Kriterien angegeben werden. Hierdurch lässt sich, je nach gewünschter Editionsform, beispielsweise die gewählte Leithandschrift begründen und der Anmerkungsapparat erstellen etc. Damit ist das Filtersystem also bereits beim Vorgang der Texterstellung nutzbar. Natürlich kann auch das für die KJ erstellte Filtersystem selbst für eine digitale Edition übernommen werden.

HTR-Texte als „neuer Texttyp“

Alle in die „Datenbank nicht normalisiertes Mittelhochdeutsch“ aufgenommenen HTR-Transkriptionen sollen frei zur Verfügung gestellt werden. Die Transkriptionen sind hierbei zeilengenau mit den Digitalfaksimile der jeweiligen Hss. verzahnt. Zudem können über OCR4all ständig neue, von spezifischen Fragestellungen abhängige Corpora generiert werden. Damit gelangt ein neuer Texttyp in den altgermanistischen wissenschaftlichen Diskurs. Dieser weist auf der einen Seite eine höhere Fehlerquote auf als umfänglich (händisch) korrigierte Editionstexte. Auf der anderen Seite bietet er aber (anders als die meisten herkömmlichen Editionen) einen unmittelbaren Zugriff auf die historischen Handschriften und kann so die Grundlage der Erschließung und der Durchsuchbarkeit mhd./frnhd. Hss. darstellen. HTR-Transkriptionen können daher als Schlüssel zur mittelalterlichen Hs. genutzt werden. Das ist in den Fällen besonders relevant, in denen keine vollständig adäquate Editionssituation vorliegt bzw. nicht alle Textzeugen in bereits bestehende Editionen eingegangen sind. Bei Fragestellungen, die durch normalisierte Editionen erschwert werden, können HTR-Transkriptionen zudem als Ergänzung der bestehenden Texteditionen herangezogen werden. Sie nehmen damit generell eine Mittelstellung zwischen der Edition und dem (Digital-)Faksimile ein. Im Vergleich mit der durch Normalisierungen und Konjekturen geprägten Editionspraxis der älteren deutschen Literaturwissenschaft kann beispielsweise auch die Frage aufgeworfen werden, welche Rolle die HerausgeberInnen mhd. Texte eigentlich für unsere moderne Wahrnehmung der Texte und des historischen Sprachstands spielen etc. HTR-Transkriptionen gewähren so einen Blick in die historische Situierung mhd. Texte, der weit über den traditionellen Zugang des kritischen Anmerkungsapparats hinausgeht.

Weitere Anwendungsgebiete von HTR-Transkriptionen und Levenshtein- bzw. WMD-Filtern

Das mit dem Editionsprojekt der KJ verzahnte Würzburger HTR-Projekt soll in drei Anwendungsbereichen Ergebnisse generieren, die für potentielle Folgeprojekte über den Standort Würzburg hinaus frei zur Verfügung gestellt werden: 1. Das gemischte HTR-Grundmodell kann als Grundlage für weitere werkspezifische Erkennungsmodelle verwendet werden, wodurch sich der Transkriptionsaufwand in entsprechenden (externen) Folgeprojekten erheblich reduziert. Hierdurch werden jenseits von Editionsprojekten Fragestellungen ermöglicht, die auf Grundlage der momentan zur Verfügung stehenden Textdaten gar nicht oder nur mit erheblichem Aufwand beantwortet werden könnten (s. u.). Gleichzeitig lässt sich das Grundmodell mit jedem Folgeprojekt (und einer entsprechenden Erweiterung der GT) weiter ausdifferenzieren. 2. Als Folge der GT-Erstellung wächst auch die “Datenbank nicht normalisiertes Mittelhochdeutsch” kontinuierlich an. Die entstehenden Daten können einerseits einschlägigen Datenbanken wie der “Mittelhochdeutsche Begriffsdatenbank” zur Verfügung gestellt werden. Andererseits besteht beispielsweise für corpusanalytische Fragestellungen freier Zugriff auf alle erfassten Texte, die dementsprechend zur Nachnutzung zur Verfügung stehen. 3. Das auf der “Datenbank nicht normalisiertes Mittelhochdeutsch” basierende WMD- und LevD-Filtersystem kann für diverse weitere Fragestellungen angewendet werden (z. B. für diverse Fassungsvergleiche; automatisch erstellbare, auf WMD-Distanzen basierende Textsynopsen o. ä.). Daher werden das HTR-Grundmodell, die “Datenbank nicht normalisiertes Mittelhochdeutsch” und die WMD-/LevD-Daten als Open Source Datenbank zur Verfügung gestellt. Aus diesen drei Anwendungsbereichen folgen weitere mögliche Fragestellungen, die auf dem HTR-Projekt bzw. WMD-/LevD-Projekt fußen. Diese können im Folgenden nur knapp skizziert werden:

1. Die meisten überlieferungsgeschichtlichen Fragestellungen benötigen mehr Datenmaterial, als ein herkömmlicher Lesetext mit Anmerkungsapparat zur Verfügung stellt. Für alle corpusanalytischen Zugänge, die nicht auf die Edition der Corpustexte zielen, ist es zentral, mit möglichst wenig Arbeitsaufwand spezifische Untersuchungscorpora aufbauen zu können. Das ist mit dem HTR-Grundmodell möglich. 2. Durch die Erschließung der mittelalterlichen Hss. sind neue sprachgeschichtliche Erkenntnisse zu erwarten, da mit der “Datenbank nicht normalisiertes Mittel-

hochdeutsch" deutlich mehr nicht normalisiertes Datenmaterial zur Verfügung gestellt werden kann. Die WMDs lassen hierbei beispielsweise neue Perspektiven auf die Semantik historischer Sprachstufen zu. 3. Stilometrische Analysen können durch diese Datenbank quantitativ ausgeweitet und mit den WMDs kombiniert werden.[18] 4. Phraseologische Querschnitte innerhalb eines Untersuchungscorpus' erscheinen durch die WMDs möglich.[19] 5. Überkommene stemmatologische Setzungen sind durch breit angelegte, von HTR-Modellen gestützte Levenshtein- und WMD-Analysen überprüfbar.[20] 6. LevD und WMDs sind für neue Fassungsdefinitionen anwendbar etc.[21]

Bereits diese kursorischen Überlegungen machen deutlich, wie gewinnbringend digitale Methoden, Textkorpora und Editionen besonders für vormoderne Texte nutzbar gemacht werden können. Die interdisziplinäre Zusammenarbeit zwischen den philologischen Disziplinen und den Digital Humanities dürfte hierbei das Potential haben, neue Fragen hervorzubringen und gleichzeitig traditionelle Fragen der Mediävistik neu zu beantworten.

Fußnoten

1. Vgl. Kahle et al. 2017.
2. Vgl. z. B. die DFG-Förderinitiative OCR-D. URL: https://ocr-d.de.
3. Vgl. das digitale Editionsprojekt Narragonien digital. URL: http://www.narragonien-digital.de.
4. Vgl. Reul et al. 2019 und http://ocr4all.de.
5. Vgl. Kusner et al. 2015.
6. Vgl. Dimpel 2017.
7. Vgl. Mikolov et al. 2017; Bojanowski et al. 2017.
8. Vgl. Hung et al. 2016.
9. Verwendet wurde der Datensatz der Mittelhochdeutschen Begriffsdatenbank, vgl. http://www.mhdbdb.sbg.ac.at/.
10. Vgl. Kragl 2015.
11. Vgl. Thomasin von Zerklaere, Der Welsche Gast, München, Bayerische Staatsbibliothek, Cgm 571 (3. Viertel 15. Jh.), vgl. https://digi.ub.uni-heidelberg.de/diglit/bsb_cgm571.
12. Vgl. Priester Wernher, Driu liet von der maget, Krakau, Bibl. Jagiello#ska, Berol. mgo 109 (1. Viertel 13. Jh.), vgl. https://jbc.bj.uj.edu.pl/dlibra/doccontent?id=159362.
13. Entspricht den mit OCR4all erstellten Rohdaten für jede Seite, bestehend aus dem Scan und der zugehörigen XML-Datei, die umfassende Informationen über die Seite enthalten kann, mindestens aber die Koordinaten und die korrekte Transkription einer jeden Zeile. Durch die Verwendung des etablierten Standard Formats PAGE wird eine problemlose und umfangreiche Nachnutzung durch eine Vielzahl von OCR/HTR Programmen sichergestellt.
14. Vgl. Stolz 2006.
15. Vgl. das Editionsprojekt Lyrik des deutschen Mittelalters. Digitale Edition. URL: http://www.ldm-digital.de/.
16. Vgl. das digitale Parzival-Projekt der Universität Bern. URL: http://www.parzival.unibe.ch/.
17. Vgl. das Projekt Welscher Gast digital. URL: https://digi.ub.uni-heidelberg.de/wgd/.
18. Vgl. Krautter 2018.
19. Vgl. grundlegend Friedrich 2006.
20. Vgl. exemplarisch zur KJ Fromm 1971.
21. Vgl. Schiewer 2005.

Bibliographie

Bojanowski, Piotr / Grave, Edouard / Joulin, Armand / Mikolov, Tomas (2017): "Enriching word vectors with subword information", in: *TACL* 5: 135–146.

Dimpel, Friedrich Michael (2017): "Ein Delta-Rätsel. Nichtnormalisierte mittelhochdeutsche Texte, Z-Wert-Begrenzung und ein Normalisierungswörterbuch. Oder: Auf welche Wörter kommt es bei Delta an", in: *DARIAH-DE Working Papers* 25. URL: https://cris.fau.de/converis/portal/publication/120046124

Friedrich, Jesko (2006): *Phraseologisches Wörterbuch des Mittelhochdeutschen. Redensarten, Sprichwörter und andere feste Wortverbindungen in Texten von 1050-1350*. Tübingen: Niemeyer 2006.

Fromm, Hans (1971): "Stemma und Schreibnorm. Bemerkungen anläßlich der "Kindheit Jesu" des Konrad von Fußesbrunnen", in: Hennig, Ursula / Kolb, Herbert (eds.): *Mediaevalia litteraria. FS für Helmut de Boor zum 80. Geburtstag*. München: C.H. Beck 193-210.

Huang, Gao / Guo, Chuan / Kusner, Matt / Sun, Yu / Sha, Fei / Weinberger, Kilian (2016): "Supervised Word Mover's Distance", in: *NIPS* 29. URL: https://proceedings.neurips.cc/paper/2016/hash/10c66082c124f8afe3df4886f5e516e0-Abstract.html

Kahle, Philip / Colutto, Sebastian / Hackl, Günter / Mühlberger, Günter (2017): "Transkribus-a service platform for transcription, recognition and retrieval of historical documents", in: *IAPR* 4: 19-24.

Kragl, Florian (2015): "Normalmittelhochdeutsch. Theorieentwurf einer gelebten Praxis", in: *ZfdA* 144: 1-27.

Krautter, Benjamin (2018): "Über die Attribution hinaus. Forschungsperspektiven der Stilometrie als Anwendungsfeld in der Literaturwissenschaft", in: Bernhart, Toni / Willand, Marcus / Richter, Sandra / Albrecht, Andrea (eds.): *Quantitative Ansätze in den Literatur- und Geisteswissenschaften. Systematische und historische Perspektiven*. Berlin/Boston: De Gruyter 289-314.

Kusner, Matt J. / Sun, Yu / Kolkin, Nicholas I. / Weinberger, Kilian Q. (2015): "From Word Embedding To Document Distances", in: *ICML* 37: 957-966. URL: https://dl.acm.org/doi/10.5555/3045118.3045221

Mikolov, Tomas / Grave, Edouard / Bojanowski, Piotr / Puhrsch, Christian / Joulin, Armand (2018): "Advances in Pre-Training Distributed Word Representations", in: *LREC* 2018. https://aclanthology.org/L18-1008/

Neudecker, Clemens / Baierer, Konstantin / Federbusch, Maria / Boenig, Matthias / Würzner, Kay-Michael / Hartmann, Volker / Herrmann, Elisa (2019): "OCR-D: An end-to-end open source OCR framework for historical printed documents", in: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*: 53-58.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): "OCR4all - An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings", in: *Applied Sciences* 9,22. URL: https://www.mdpi.com/2076-3417/9/22/4853

Schiewer, Jans-Jochen (2005): "Fassung, Bearbeitung, Version und Edition", in: Schubert, Martin J. (ed.): *Deutsche Texte des Mittelalters zwischen Handschriftennähe und Rekonstruktion. Berliner Fachtagung 1.-3. April 2004*. Tübingen: Niemeyer 35-50.

Stolz, Michael (2006): “Vernetzte Varianz. Mittelalterliche Schriftlichkeit im digitalen Medium”, in: Giuriato, Davide / Stingerlin, Martin / Zanetti, Sandro (eds.): *System ohne General. Schreibszenen im digitalen Zeitalter*. München: Wilhelm Fink Verlag: 217-244.

Internetadressen

http://www.narragonien-digital.de

http://ocr4all.de

http://www.mhdbdb.sbg.ac.at/

https://digi.ub.uni-heidelberg.de/diglit/bsb_cgm571

https://jbc.bj.uj.edu.pl/dlibra/doccontent?id=159362

http://www.ldm-digital.de/

http://www.parzival.unibe.ch/

https://digi.ub.uni-heidelberg.de/wgd/