Digital Humanities

Elisabeth Klein*, Matthias Däumer und Andreas Hütig

Gute Ergebnisse aus „schlechten“ Textvorlagen

Einflüsse der OCR-Qualität auf kulturwissenschaftsgeschichtliche Forschungsergebnisse am Beispiel von Drucken des 18. Jahrhunderts

DOI 10.1515/iwp-2016-0054

Zusammenfassung: Die OCR-Erkennung alter Frakturtexte ist noch immer Gegenstand verschiedener Forschungsarbeiten. Die bisher möglichen Ergebnisse erreichen die Ansprüche deutscher Drittmittelgeber in der Regel nicht. Der vorliegende Beitrag gewährt einen Einblick in ein Projekt, in dem Mainzer Wissenschaftler gemeinsam mit ihrer Universitätsbibliothek die Qualitätssicherung von OCR-Texten ausloten. Neue Forschungsansätze der Digital Humanities, wie das *scalable reading*, eröffnen Möglichkeiten, mit schlechten Scanergebnissen dennoch zu guten Forschungsergebnissen zu kommen.

Deskriptoren: Optische Zeichenerkennung, Scannen, Qualität, Geisteswissenschaften, Forschung

Good results from “bad” text documents – Influences of OCR-quality on cultural research analysing 18th century gothic prints

Abstract: Blackletter prints still challenge computers and scholars alike when it comes to accuracy of text recognition. In this paper we show our approach to evaluate OCR accuracy of old text documents. Thus we challenge funding agencies’ demands on accuracy by arguing that mixed methods like *scalable reading* help to achieve sufficient results from text documents of insufficient accuracy.

Descriptors: Optical character recognition, Scanning, Quality, Humanities, Research

*Kontaktperson: Elisabeth Klein, M.A., Universitätsbibliothek Mainz, Johannes Gutenberg-Universität Mainz, Jakob-Welder-Weg 6, 55128 Mainz, E-Mail: e.klein@ub.uni-mainz.de, <http://orcid.org/0000-0002-1619-3471>

Dr. Matthias Däumer, Universität Tübingen, Deutsches Seminar, Wilhelmstraße 50, 72074 Tübingen, E-Mail: m.daeumer@uni-tuebingen.de

Dr. Andreas Hütig, Johannes Gutenberg-Universität Mainz, Studium generale, Colonel-Kleinmann-Weg 2, 55128 Mainz, E-Mail: ahuetig@uni-mainz.de

De bons résultats à partir de «mauvais» documents de textes – l’influence de la qualité de l’OCR sur des résultats de la recherche scientifique dans le domaine de l’histoire culturelle en prenant pour exemple des imprimés du 18^e siècle

Résumé: La reconnaissance optique de caractères (OCR) de vieux fragments de textes fait toujours l’objet de divers travaux de recherche. Les résultats obtenus jusqu’ à présent parviennent rarement à satisfaire les exigences des bailleurs de fonds tiers allemands. Cet article donne un aperçu d’un projet, dans lequel des chercheurs de Mayence explorent en collaboration avec leur bibliothèque universitaire la qualité de textes «océrisés». Les nouvelles approches de recherche des humanités numériques, comme par exemple le «scalable reading», ouvrent des possibilités pour obtenir de bons résultats de recherche même en partant d’une mauvaise numérisation.

Describeurs: Reconnaissance optique de caractères, Numérisation, Qualité, Sciences humaines Recherche scientifique

1 Alte Textvorlagen als Herausforderung für Mensch und Maschine

Die kulturgeschichtliche Forschung ist vor allem eine Textforschung. Dazu müssen Texte in der Regel intensiv intellektuell durchgearbeitet und aufs Jota analysiert werden. Zwar können viele Texte für die Entwicklung verallgemeinerbarer Ergebnisse zusammengetragen, aber nicht alle in einem Menschenleben gelesen werden, so dass Technologien wie OCR nötig werden, um die intellektuelle Auseinandersetzung mit den Inhalten zu unterstützen.¹ Ins-

1 Vgl. dazu auch Crane (2006).

besondere für alte Drucke ergeben sich dabei zwei Problemstellungen:

Erstens wird der maschinelle Zugang zu Texten umso schwerer, je älter diese sind, da sie vor 1850 auf vorindustriell hergestellten Frakturtypen basieren. Diese stellen eine besondere Herausforderung für die maschinelle Zeichenerkennung dar. Je älter die Texte werden, desto interessanter sind sie allerdings für die Kulturgeschichtsforschung.

Zweitens stellen Drittmittelgeber Anforderungen an die Genauigkeit digitaler Volltexte, die derzeit für Frakturtexte nicht ohne Weiteres zu erreichen ist. Im Spannungsfeld zwischen wissenschaftlichem Forschungsinteresse, menschlicher wie maschineller Limitationen und Ansprüchen von Drittmittelgebern sind deshalb Herausforderungen zu meistern, die am Projekt „Vor der Kulturgeschichte“ deutlich gemacht und zumindest in Ansätzen entlang der Wirklichkeit des Forschungsalltags gelöst werden können. In dieser Fallstudie soll aufgezeigt werden, wie ein praxisbezogener Ansatz zur Lösung beider Problemstellungen aussehen kann.

1.1 Auf der Suche nach der verlorenen ›Cultur‹ – projektspezifische Grundlagen und Problemstellungen

Die aktuelle Kulturgeschichtsschreibung hebt nach ihrem Selbstverständnis erst in den 1980er Jahren an. Allenfalls gesteht sie zu, dass es um 1900 eine frühe, mittlerweile sehr gut erforschte Phase der Kulturgeschichtsschreibung gegeben hat. Das Projekt „Vor der Kulturgeschichte“ (VdK) setzt mit einer Kritik an diesem Narrativ ein. Eine stichprobenartige Sichtung von kulturgeschichtlichen Druckschriften aus dem Zeitraum von 1760 bis 1810 ergab, dass schon im 18. Jahrhundert eine ebenso intensiv wie extensiv betriebene, heute weitgehend vergessene Kulturgeschichtsschreibung existierte. Die Bedeutung(en) des Begriffs ›Cultur‹ ist im untersuchten Zeitraum jedoch weit aus disparater als die heutige Fassung des Kulturbegriffs, ebenso – wenn auch vergleichsweise von der Forschung stärker aufgearbeitet – das zeitgenössische Verständnis von Geschichtsschreibung.

Um diese Bedeutungen systematisch zu rekonstruieren, wurde zu Beginn des VdK-Projekts ein Korpus aus etwa 200 zumeist deutschsprachigen Texten zusammengestellt, die mindestens eine der folgenden Eigenschaften aufweisen mussten:

- 1) ›Kultur‹ oder ›Cultur‹ kommt als Titelwort vor.
- 2) Ein kulturgeschichtlicher Zuschnitt des Textes lässt sich vermuten.

- 3) Der Text besitzt einen programmatischen Charakter für das Studium der Geschichte – einschließlich der Kulturgeschichte.

Darunter befanden sich bekannte Autoren wie Herder, Schiller und Schlegel, historiographisch wirkmächtige Gelehrte wie Adelung, Gatterer oder Schlözer, intellektuell und politisch aufschlussreiche Figuren wie Weishaupt oder Forster, aber auch eine Menge weniger bekannter bis gänzlich unbekannter Autoren. Die Lektüre der Werke hat über diskursive Verweise zunehmend zu weiteren Texten geführt, so dass sich das Korpus ständig erweiterte – und noch immer erweitert.

Um eine systematische Verlinkung, thematische Verschlagwortung², kollaborative Kommentierung und Suche in den Texten zu ermöglichen, wurde das Textkorpus zunächst als Scansammlung zusammengetragen, in der mittlerweile rund 90 Prozent der Texte als PDF-Dateien zugänglich sind. Von diesen stammen rund 60 Prozent aus dem Bestand der VD18-Plattform, einem DFG-Projekt zur Sammlung von Drucken des 18. Jahrhunderts.³ Die restlichen 40 Prozent wurden vom Projekt ferngeliehen, nach DFG-Vorgaben eingescannt und mit Metadaten versehen. Am Ende schließlich steht das Ziel einer Publikation der prägnantesten Texte in Form einer Anthologie, um die Entwicklung des Begriffs nachzuzeichnen.

1.2 Methoden der Digital Humanities im Umgang großen Textmengen

Aufgrund der nicht enden wollenden Korpusaufschwellung gilt es nach Methoden zu suchen, die bei der Bewältigung solcher Textmengen und der Bearbeitung nach den wissenschaftlichen Bedürfnissen des Projekts helfen können.

Distant reading ist eine Methode, die es erlaubt, große Textmengen zu erfassen und auf projektspezifische Aspekte hin zu untersuchen. Sie soll durch quantitative und statistische Erhebung und Aufbereitung von Daten zu und aus Literatur Erkenntnisse generieren und mehr als eine exemplarische Analyse möglich machen. Franco Morettis Programm eines solchen *distant reading*⁴ war ursprünglich als Provokation gegen das als unzeitgemäß

² Strange et al. 2014 zeigen, wie themen- und genrespezifische Verschlagwortung den analytischen Zugang zu Textsammlungen mit schlechter OCR-Qualität erleichtert.

³ <http://www.vd18.de>.

⁴ Vgl. Moretti (2013/2016).

empfundene *close reading*, die intensive und sorgfältige, in einzelne Passagen sich vertiefende Interpretationshaltung der herkömmlichen Philologie gedacht – und als solches hat es (hinsichtlich der Diskursrelevanz und des Einflusses auf die US-amerikanische und mittlerweile auch die deutsche Forschungslandschaft und Förderkultur) wunderbar funktioniert. Morettis Provokation, man könnte mit der Unzahl von Texten nur umgehen, indem man sie *nicht* lese,⁵ hat sich, wie der gegenwärtige Stand der Digital Humanities zeigt, aber mittlerweile überlebt und es wurde ein Bewusstsein dafür entwickelt, wie viel mehr sich aus den Zwischentönen der Debatten und Methoden ergeben kann.⁶

Abgesehen von dieser methodologischen Notwendigkeit kamen bei der Auseinandersetzung mit den Möglichkeiten der Digital Humanities aber auch erkenntniskritische Bedenken zum Tragen: Für die Zielsetzung von VdK ist selbst ein graduelles *distant reading* auf mehreren Ebenen problematisch. Zum einen bietet das Verfahren keine Möglichkeit, die Relevanz bestimmter Texte und Autoren für die Bildung eines Kanons zu erfassen. Für die Beschreibung einer Begriffsbildung von ›Culturgeschichte‹ müssen aber für das 18. Jahrhundert Texten von Adelung, Herder, Kant oder Schiller eine größere kulturelle Relevanz zugeschrieben werden als denen von auch damals eher unbekannten Autoren. Große Datenmengen müssen also unter Einbeziehen des Kriteriums der kulturellen Relevanz analysiert werden, um eine Verzerrung der Perspektive zu verhindern und dem Umstand gerecht zu werden, dass auch ein heterogenisierender Kanonisierungsvorgang eine kulturelle Sinnstiftung bedeutet.

Zum anderen installiert das *distant reading* hinsichtlich der zeitgenössischen Rezeption generell eine problematische Perspektive, indem gleichsam ein historischer Rezipient entworfen wird, der alle Texte zumindest potenziell gekannt haben könnte. Eine solche Konstruktion ist an sich nichts Ungewöhnliches, idealisiert doch jeder Forschungsansatz der Kulturgeschichte, der das Entstehen von kultureller Relevanz erst im Blick der Rezipienten verortet, den jeweiligen heuristischen Fluchtpunkt. Man muss sich (entweder implizit oder explizit) einen historischen Rezipienten entwickeln, um das Forschungsobjekt durch seine Augen betrachten zu können. Graduell besteht aber ein enormer Unterschied zwischen einer imaginierten

Rezeption, der man etwa unterstellt, dass sie die wichtigsten kulturgeschichtlichen Denker der jeweiligen Zeit kannte, und derjenigen, deren Wissen man mit einer umfangreichen Daten-Sammlung gleichsetzt. Wo sollte dieser Rezipient, der alles, also jeden kleinsten Text, überblickt und sich mit der Souveränität einer Suchmaschine durch dieses Konvolut fräst, als historische Realität bspw. des 18. Jahrhunderts zu finden gewesen sein? Wenn es diesen idealen Rezipienten aber nicht gegeben hat, ist jede im *distant reading* gewonnene Aussage über Begriffsverwendungen, Bedeutungsnuancen und kulturgeschichtliche Relevanz eine problematische Konstruktion, die dort Bezüge sieht, wo kein Mensch sie gesehen oder intentional gesetzt haben kann.

Angesichts des schieren Umfangs der VdK-Korpus, aber auch vieler Einzelpublikationen, deren Relevanz auf den ersten Blick gar nicht zu bestimmen ist, weil sie sich nur am Rande oder (wie bspw. viele der so genannten ›Universalgeschichten‹) nur über Synonyme mit dem Kulturbegriff auseinandersetzen, wurde schnell klar, dass VdK Verfahren des *distant reading* mit solchen des *close reading* kombinieren muss.

Will man eine kulturelle Entwicklung historisch angemessen beschreiben, muss die Basis der Beschreibung gerade mit Blick auf die zeitgenössischen Rezeptionsvorgänge menschenmöglich bleiben und darf den Umfang der Rezeption nicht (implizit) mit der Archivierungsmacht heutiger elektronischer Erfassung gleichsetzen. Um dies zu erreichen, scheint ein *scalable reading*, die Methode, die Martin Mueller für sein *Shakespeare His Contemporaries*-Projekt beschreibt, aussichtsreich:

„Digital tools and methods certainly let you zoom out, but they also let you zoom in, and their most distinctive power resides precisely in the ease with which you can change your perspective from bird’s eye view to close-up analysis. [...] ›Scalable reading‹ of this kind is at the heart of new forms of intertextual analysis opened up by digital tools that let you discover shared verbal patterns across large data sets.“⁷

Diese Methode, die *distant* mit *close reading* verbindet, galt es für die Textsammlung des Projekts zu operationalisieren. Des Weiteren wurde für unseren Umgang mit den Synonymen und Antonymen von ›Cultur‹ und verwandten Suchbegriffen die Definition von ›Begriff‹ tragend, die der Linguist Jochen Bär für die Zentralbegriffe der klassisch-romantischen Kunstperiode (ZBK) entwickelt hat.⁸ Bei Bär werden Synonyme ebenso wie diskursiv verwandte Begriffe

⁵ „What are we mortal beings supposed to do with all these books? Franco Moretti has a solution: don’t read them.” Katryn Schulz, „What Is Distant Reading?“, in: New York Times, 24.06.2011 (online unter: http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html?_r=0).

⁶ Vgl. Ross (2014).

⁷ Mueller (2014, Abs. 3.2).

⁸ Vgl. Bär (2014).

fe in die Textsuche miteinbezogen, um sowohl den historischen Charakteristika der Wortsemantik, des Diskurses, als auch spezifischen Formungen einzelner Autoren gerecht zu werden. Für VdK war somit grundlegend abgesteckt, wo man in den Methodenverhandlungen der Digital Humanities anschließen konnte, um den eigenen Ansprüchen gerecht zu werden. Durch die Ansprüche der Forschungsmethoden entstehen jedoch spezifische Anforderungen an die Qualität der Volltexte, auf deren Basis die Methoden erst Anwendung finden können.

2 Probabilistische Aussagen zur prospektiven OCR-Qualität – Studiendesign

2.1 Maschinelle Erkennung von Frakturschriften und deren Rahmenbedingungen

Textarbeit mit sogenannten „gemischten Methoden“ wie dem *scalable reading* erfordert digitale Volltexte, die ein möglichst zuverlässiges Retrieval von Suchtermen ermöglichen. Drucke aus dem für das Projekt relevanten Zeitraum sind jedoch in verschiedener Hinsicht für die maschinelle Texterkennung problematisch. Ein generelles Problem für eine optische Zeichenerkennung (OCR) besteht darin, dass manufakturierte, d. h. handwerklich statt industriell normiert gefertigte Typensätze, die bis ca. 1850 verwendet wurden, sich durch ein hohes Maß an Abweichungen untereinander auszeichnen. Dadurch sind Fraktur-Typen von Druckerei zu Druckerei und damit nicht selten auch von Buch zu Buch unterschiedlich.⁹

Viele Ligaturen sowie ziselierte, sich stark ähnelnde Grapheme oder auch die vor 1900 nicht normierte Orthographie erschweren sowohl die Erkennung selbst als auch den durch die OCR-Software vorgenommenen Wörterbuchabgleich.

Weitere Probleme entstehen aus der Spezifik des VdK-Textkorpus: Beispielsweise werden oftmals fremde Kulturkreise oder -geschichten behandelt, bei denen Namensformen auftauchen, die gar keine Übereinstimmung mit den Abgleich-Wörterbüchern haben können. Das betraf vor allem das Kapitel „Geschichte der Chineser“ aus Johann

Christoph Gatterers *Abriß der Universalhistorie nach ihrem gesamten Umfange von Erschaffung der Welt bis auf unsere Zeiten* (1761), das viele eigenwillige, selbst mit intensivem Rechercheaufwand nicht nachvollziehbare Schreibungen von u. a. Herrscher- und Landschaftsnamen enthält.

Diese Ausgangssituation garantiert „schmutzige“ OCR-Texte, d. h. Volltexte mit hoher Anzahl falsch erkannter Zeichen. Um das Ausmaß der Probleme bei der Texterkennung im Allgemeinen und hinsichtlich der Ansprüche für die Drittmittelförderung zu ermitteln, führten die Projektgruppe und die Universitätsbibliothek Mainz einen Pilotversuch durch, in dem die durchschnittliche Erkennungsrate der digitalen Volltexte prospektiv abgeschätzt werden sollte.

2.2 Stichprobe und Ground Truth

Bei mittleren und größeren Textsammlungen – im Falle des VdK-Konvoluts elf exemplarische Bücher mit einem Umfang von 200 bis 600 Seiten – kommen aufgrund des notwendigen Abgleichs mit den *Ground Truth*-Texten¹⁰ zur Überprüfung der Erkennungsraten nur Stichprobenverfahren infrage. Als methodische Grundlage für die Schätzung diente das Test-Design der Herzog August Bibliothek Wolfenbüttel, das den Anforderungen der vorliegenden Texte und Forschungsfragestellung angepasst wurde.¹¹

Für kleine Projektgruppen oder Bibliotheken ist es aufgrund begrenzter Ressourcen wichtig, Evaluationsverfahren zu entwickeln, die eine Balance zwischen einer möglichst genauen prospektiven Qualitätseinschätzung und den dafür notwendigen Ressourcen schaffen. Eine Berechnung des Verhältnisses von Stichprobengröße und gewünschter Exaktheit der Schätzung der Erkennungsrate ergab, dass eine Stichprobe von 60 Seiten aus dem Gesamtkonvolut sowohl das Kriterium einer möglichst zuverlässigen Schätzung als auch das des tragbaren Ressourcenaufwands erfüllt.

Für die Berechnung des optimalen Verhältnisses zwischen Stichprobengröße und Arbeitsaufwand für die Erstellung der *Ground Truth* wurden t-Tests¹² für unter-

⁹ Die daraus resultierenden Probleme für die OCR werden auch weiterhin durch großangelegte Forschungsprojekte wie etwa Impact (<http://www.impact-project.eu>) und seinen regionalen Nachfolger OCR-D untersucht (<http://www.ocr-d.de/>).

¹⁰ Als *Ground Truth* werden Texte bezeichnet, die manuell abgeschrieben und intellektuell überprüft werden. Sie stimmen dadurch textlich vollständig mit dem Originaldokument überein und dienen als Messlatte für die Genauigkeit des maschinell erkannten Volltexts dieser Dokumente.

¹¹ Stäcker (2013).

¹² Der t-Test wird zur Abschätzung der mittleren Erkennungsrate der Textsammlung verwendet. Er liefert als Ergebnis ein Intervall zurück, in dem die durchschnittliche Erkennungsrate liegen wird.

Abbildung 1: Intervallgrößen nach Stichprobenumfang.

schiedliche Stichprobengrößen mit realitätsnahen durchschnittlichen Erkennungsraten berechnet. Die Berechnung geht von mindestens 30 und maximal 100 Seiten *Ground Truth* aus. Wie Abbildung 1 zeigt, verkleinert sich das Intervall mit zunehmendem Stichprobenumfang, ist aber für Stichprobengrößen zwischen 60 und 80 Seiten relativ konstant (grüne Säulen) und verringert sich danach nur noch langsam. Auf Basis dieser Berechnung ließ sich ein Stichprobenumfang von 60 Seiten als optimales Verhältnis zwischen der Genauigkeit der Schätzung und dem Arbeitsaufwand zum Erstellen der *Ground Truth* errechnen.

Von dieser Festlegung ausgehend wurde das Stichprobendesign weiter angepasst: Statt einer Stichprobe von Seiten, die keinen „hohen Grad von Verschmutzung oder Annotationen bzw. Durchstreichungen“¹³ aufweisen, wurde eine Zufallsstichprobe von 60 Seiten (entsprechend 83.542 Zeichen) aus dem Gesamtkorpus erzeugt, die ein breites Spektrum von Typographien, Verschmutzungen und Scanqualitäten aufwies. Da diese Stichprobe auch problematische Seiten mit einer erhöhten Anzahl von Marginalien, Materialverfärbung oder Verschmutzungen aufwies, wurden realistische Bedingungen für den Einsatz der OCR-Software geschaffen. Zugleich wurden die 60 Seiten der Stichprobe vom Projektteam transkri-

biert. Das Erstellen dieser *Ground Truth* erforderte die manuelle Transkription der Stichprobe nach kooperativ festgelegten Transkriptionskonventionen. Jeder transkribierte Text wurde von einer an dem jeweiligen Arbeitsschritt nicht beteiligten Person gegengelesen. Den Seiten der Stichprobe wurden die Transkriptionen als *Ground Truth* gegenübergestellt zur inferenzstatistischen Schätzung der durchschnittlichen Erkennungsrate für das Gesamtkorpus.

2.3 Software

Die verbreitete OCR-Software *Abbyy Finereader* ist für die Anwender fertiger Plattformen für Digitalisierungsmanagement einfach zugänglich, bietet allerdings nur eingeschränkte Möglichkeiten des Einblicks oder Konfigurierens. Funktionen wie etwa Wörterbuchabgleiche für aktuelle wie ältere Sprachstufen nimmt die Software zwar vor, liefert jedoch nicht immer vorhersehbare Ergebnisse.¹⁴ Freie OCR-Software wie *Tesseract* bzw. *OCROpy*¹⁵ erlaubt zwar umfassendere Einblicke in den Erkennungsprozess, erfordert aber im Gegenzug zeitintensives Einarbeiten in

¹³ Stäcker (2013, S. 132).

¹⁴ Von einem Training der Software wurde abgesehen, da dieses als zeitintensiv, aber wenig effektiv für die Ergebnisverbesserung gilt.

¹⁵ <https://github.com/tmbdev/ocropy>.

Abbildung 2: Erkennungsraten des VdK-Korpus mit Mittelwert und Markierung der Konfidenzintervalle.

die Software. Ausgehend von dieser Abwägung fiel die Wahl auf *ABBYY Finereader*. Die Software war bereits in den Workflow der an der Universitätsbibliothek Mainz genutzten Digitalisierungs- und Präsentationsplattform *Visual Library* eingebunden, um eine nahtlose Verarbeitung im Digitalisierungsprozess zu ermöglichen und wurde für den Test um ein Modul zur Frakturerkennung erweitert.

Größere Herausforderungen ergaben sich in Bezug auf die Softwareauswahl für den maschinellen Abgleich von OCR- und zugehörigen *Ground Truth*-Texten. Bislang existieren Prototypen von Programmen, die Abgleiche und Fehleranalysen vornehmen. Diese sind derzeit jedoch noch immer nicht ohne weiteres handhabbar. Entsprechend wurde vor Ort für den OCR-/*Ground-Truth*-Abgleich eine angepasste Implementierung des Levenshtein-Algorithmus in Form eines Perl-Skripts aufgesetzt, mit der die Erkennungsraten bestimmt wurden.

2.4 Ergebnisse und Implikationen

Für die Stichprobe ergab sich eine durchschnittliche Erkennungsrate von 92,76 Prozent mit Konfidenz-

intervallen¹⁶ zwischen 92,71 und 95,35 Prozent ($p = 0.05$), respektive zwischen 92,23 und 95,65 Prozent ($p = 0.01$) als prospektive durchschnittliche Erkennungsrate für das gesamte VdK-Korpus. Abbildung 2 zeigt die Verteilung der Erkennungsraten der Stichprobe mit Konfidenzintervallen (grüne bzw. blaue Linien) und Mittelwert (rote Linie) und markiert sowohl geringe (orange) als auch kritisch geringe Erkennungsraten (rot), die 90 bzw. 80 Prozent Erkennung unterschreiten.

Von diesen Ergebnissen ausgehend wurde im Rahmen des *distant reading* die Suche nach dem Begriff ›Cultur‹ und seinen semantischen Verwandten vorgenommen. Um die Probleme der Zeichenerkennung eingrenzen und im weiteren Verlauf berücksichtigen zu können, wurden die schlecht und kritisch schlecht erkannten Seiten zusätzliche einer qualitativen Fehleranalyse durch das Projektteam unterzogen. Es ließ sich eindeutig feststellen, dass in erster Linie die Kombination aus Sprach- und Typenwechsel innerhalb eines Textes sowie Abkürzungen und

¹⁶ Der Test wurde zu Vergleichszwecken jeweils für einen Konfidenzwert von 0.05 und 0.01 berechnet.

Passagen zu Kulturen, deren Eigennamen sich mit dem lateinischen Schriftsystem nur schwer fassen lassen, die größten Herausforderungen für die Zeichenerkennung darstellten, wohingegen die erwarteten Erkennungsfehler durch Verschmutzung und Gegendrucke nicht nachweisbar waren. Dabei ließen sich ganze Werke ausmachen – hier zwei von den elf digitalisierten Büchern –, die verstärkt Abkürzungen kombiniert mit einem Typenwechsel aufwiesen. Diese wurden zusätzlich mittels *close reading* bearbeitet. Wie sich herausstellte, wurden in den OCR-Texten alle gesuchten Begriffe gefunden, keine Fundstelle blieb unerfasst. Für Forschungsfragestellungen, die eindeutige ein-Wort-Suchterme formulieren können, zeichnet sich folglich ab, dass auch die erreichte Erkennungsrate von 92,26 Prozent bereits ausreichend sein kann.

Weiterhin kann *fuzzy search*¹⁷ bei der Suche im Text einen Teil der Erkennungsfehler abfangen. Für eine einfache Implementierung solch unscharfer Suchalgorithmen reicht es, die Texte öffentlich zugänglich zu machen und durch Suchmaschinen wie Google oder Bing indexieren zu lassen. Auf diese Weise können die Textsammlungen ohne zusätzlichen Aufwand mittels der leistungsfähigen Suchmaschinenalgorithmen durchsucht werden.

Die DFG-Digitalisierungsrichtlinien betrachten bislang nur Textgenauigkeiten von 99,95 Prozent als „wissenschaftlich zuverlässig“¹⁸, da nur unter dieser Voraussetzung in den Texten sogenannte „Negativsuchen“, d. h. das vollständige Retrieval der gesuchten Zeichenkette, garantiert sind. Aufgrund der besonderen Bedingungen, die aus technischen Gründen für Texte mit Produktionsdatum vor 1850 gelten, akzeptiert die DFG deren Digitalisierung und Texterkennung nur „als Überbrückung bis zur Herstellung besserer Texte“¹⁹. Mit der gemischten Methode des *scalable reading*, die in dieser Fallstudie zum Einsatz kamen, kann aus solchen „unzuverlässigen“ Texten jedoch so viel Information extrahiert werden, dass ein nahezu perfekter Text (99,95 %) nicht mehr zwingend nötig ist.

Welche Verbesserungspotenziale liegen in der Texterkennung selbst? Methoden der Mustererkennung, die den graphematischen Kontext miteinbeziehen und deshalb in der Handschrifterkennung (HTR) genutzt werden,²⁰

können neue Möglichkeiten bieten, die Erkennungsqualität zu verbessern. In Bezug auf Texte auf Basis manueller Typen sind solche Verfahren bisher noch nicht in größerem Umfang getestet worden, sollten aber berücksichtigt werden. In diesem Zusammenhang lohnt auch der direkte Vergleich mit OCR-Systemen, die wie *OCRopy* auf der Basis von Schrift- statt Sprachmodellen operieren. Hier sind Projekte wie OCR-D gefragt, die die Weiterentwicklung von Texterkennung koordinieren.

3 Fazit

Das mit der Weiterentwicklung computergestützter Verfahren in den Geistes- und Kulturwissenschaften verbundene Versprechen digitaler Auswertung trifft sich mit dem Bedürfnis nach einer weitergehenden Kontextualisierung und Entkanonisierung in den genannten Wissenschaften. Dieses ist nur zu erfüllen, wenn es gelingt, auch größere Analyseobjekte – Textkorpora, serielle Quellen, ganze Archivbestände – nicht nur zugänglich, sondern auch analysierbar zu machen. Bei historischen Beständen, die in der Regel nicht ohne Weiteres tauglich für die computerisierte Erfassung sind, wird das Problem der defizitären Zugänglichkeit für die Datenverarbeitung verstärkt durch die problematischen Aspekte automatisierter, ›blinder‹ Zugriffe, d. h. Zugriffe, die weder von einem Bewusstsein für die geschichtliche Spezifik noch für die zeitgenössische oder historische Bedeutung der Objekte geprägt sind und so einen Über-Rezipienten implizieren, der jenseits der angewandten Methode keine kulturelle Relevanz für sich in Anspruch nehmen kann.

Aus dem dargelegten Versuch mit den VdK-Texten zeigt sich, dass gemischte Methoden durchaus auch bei problematischen Korpora und unter Vermeidung methodischer Einseitigkeiten gute Ergebnisse produzieren können. Daraus ließe sich ein Appell an Methoden der Digital Humanities nutzende Wissenschaftlerinnen und Wissenschaftler, aber auch an die Vergabepraktiken und Kriterien der Drittmittelförderung ableiten: ein Appell nämlich zur vorsichtigen Einschränkung des Perfektionismus. Zum einen können kulturhistorische und literaturwissenschaftliche Projekte wie VdK ihre Suchanfragen recht eindeutig formulieren (›Cultur‹, ›Culturgeschichte‹ und die sich daraus ergebenden Begriffsfelder) und so durchaus die Möglichkeiten informatikgestützter Analyseverfahren nutzen. Zum anderen können unscharfe Suchen sowie der Methodenmix aus *close* und *distant reading* bereits einen Teil der Probleme sogenannter „schmutziger“ OCR-Texte abfangen und müssen zugleich nicht auf die Qualitäten intensiver Interpretation verzichten. Das kann man nun als traum-

17 Bei der *fuzzy search* oder *unscharfen Suche* werden nicht nur exakt mit dem Suchterm übereinstimmende Zeichenketten als Treffer zurückgeliefert, sondern auch ähnliche. Die Abweichungstoleranz für die Klassifizierung als Treffer ist variabel.

18 Deutsche Forschungsgemeinschaft (2013, S. 32).

19 Ebd.

20 Vgl. die Softwaremodule der Virtuellen Forschungsumgebung Transkribus zur Handschrifterkennung aus dem READ-Projekt (<http://read.transkribus.eu/>).

wandlerisches Glück eines kulturwissenschaftsgeschichtlichen Ansatzes sehen, doch kann man ebenso versuchen zu generalisieren, für welche Forschungsansätze das *scalable reading* eine effektive Arbeit an OCR-erfassten Texten keinesfalls hindert, sondern zu relativ zuverlässigen Ergebnissen führen kann.

Literatur

- Bär, Jochen A. (Hrsg.): *Zentralbegriffe der klassisch-romantischen »Kunstperiode« (1760–1840). Wörterbuch zur Literatur- und Kunstreflexion der Goethezeit*. <http://www.zbk-online.de/> [22.08.2016].
- Bär, Jochen A.: Der romantische Kritik-Begriff aus linguistischer Sicht. In: Breuer, Ulrich; Tabarasi-Hofmann, Ana-Stanca (Hrsg.): *Der Begriff der Kritik in der Romantik*. Paderborn: Schöningh, 2014, S. 93–128.
- Crane, Gregory: *What Do You Do with a Million Books?* In: *D-Lib Magazine* 12 (2006), Nr. 3. <http://www.dlib.org/dlib/march06/crane/03crane.html> [22.08.2016].
- Deutsche Forschungsgemeinschaft: *DFG-Praxisregeln „Digitalisierung“* DFG-Vordruck 12.151 – 02/13. http://www.dfg.de/formulare/12_151/12_151_de.pdf [22.08.2016].
- IMPACT: <http://www.impact-project.eu> [22.08.2016].
- Moretti, Franco: *Distant Reading*, London / New York: Verso, 2013. (deutsch: *Distant Reading*, Konstanz: University Press, 2016)
- Mueller, Martin: *Shakespeare His Contemporaries: collaborative curation and exploration of Early Modern drama in a digital environment*. In: *DHQ: Digital Humanities Quarterly* 8 (2014), Nr. 3. <http://www.digitalhumanities.org/dhq/vol/8/3/000183/000183.html> [22.08.2016].
- OCR-D: URL <http://www.ocr-d.de/> [22.08.2016].
- OCRopy: <https://github.com/tmbdev/ocropy> [22.08.2016].
- READ (Recognition and Enrichment of Archival Documents): <http://read.transkribus.eu/> [22.08.2016].
- Ross, Shawana: *In Praise of Overstating the Praise*. In: *DHQ: Digital Humanities Quarterly* 8 (2014), Nr. 1. <http://www.digitalhumanities.org/dhq/vol/8/1/000171/000171.html> [22.08.2016].
- Stäcker, Thomas: *Erfahrungsbericht Helmstedter Drucke Online an der Herzog August Bibliothek Wolfenbüttel*. In: Staatsbibliothek zu Berlin – Preussischer Kulturbesitz (Hrsg.): *Volltext via OCR – Möglichkeiten und Grenzen: Testszenarien zu den Fundamentschriften der Staatsbibliothek zu Berlin – Preussischer Kulturbesitz Mit einem Erfahrungsbericht von Thomas Stäcker aus dem Projekt „Helmstedter Drucke Online“ der Herzog August Bibliothek Wolfenbüttel* (Beiträge aus der Staatsbibliothek zu Berlin – Preussischer Kulturbesitz). Berlin, 2013, S. 123–137.
- Strange, Carolyn; McNamara, Daniel; Wodak, Josh; Wood, Ian: *Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers*. In: *Digital Humanities Quarterly* 8 (2014), Nr. 1, Rn. 1–52. <http://digitalhumanities.org/8081/dhq/vol/8/1/000168/000168.html> [22.08.2016].
- VD 18 (Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 18. Jahrhunderts): <http://www.vd18.de> [22.08.2016].

Elisabeth Klein, M.A.
Universitätsbibliothek Mainz
Johannes Gutenberg-Universität Mainz
Jakob-Welder-Weg 6
55128 Mainz
e.klein@ub.uni-mainz.de
<http://orcid.org/0000-0002-1619-3471>

Elisabeth Klein, M.A. studierte Linguistik, Philosophie, Soziologie und Bibliotheks-/Informationswissenschaften in Mannheim, Frankfurt und Berlin. Nach Stellen in Mannheim, Frankfurt und Bremen arbeitet sie derzeit an einer quantitativ-qualitativen Sprachstudie an der Universität Mannheim und beschäftigt sich an der Universitätsbibliothek Mainz mit den Themen Digital Humanities, OCR und Forschungsdaten. Ihr Schwerpunkt liegt auf quantitativen und qualitativen Methoden.

Dr. Matthias Däumer
Universität Tübingen
Deutsches Seminar
Wilhelmstraße 50
72074 Tübingen
m.daeumer@uni-tuebingen.de

Dr. Matthias Däumer studierte Literatur- und Theaterwissenschaft in München und Mainz. 2011 wurde er in Mainz mit einer interdisziplinären Dissertation zur Vortragspraxis höfischer Romane promoviert. Nach Stellen am *Zentrum für Literatur- und Kulturforschung* in Berlin und am *FSP Historische Kulturwissenschaften* in Mainz lehrt er momentan am Germanistischen Seminar der Universität Tübingen. Sein laufendes, aus voller Überzeugung dem *close reading* verschriebenes Forschungsprojekt widmet sich den Jenseits- und Höllenreisen des Hochmittelalters.

Dr. Andreas Hütig
Johannes Gutenberg-Universität Mainz
Studium generale
Colonel-Kleinmann-Weg 2
55128 Mainz
ahuetig@uni-mainz.de

Dr. Andreas Hütig studierte Philosophie, Politikwissenschaften und Komparatistik in Mainz, wo er 2009 mit einer kulturphilosophischen Arbeit zu Cassirer promoviert wurde. Nach Tätigkeiten am Philosophischen Seminar der JGU und am *FSP Historische Kulturwissenschaften* arbeitet er derzeit am Studium generale der JGU und betreut Vortragsreihen und interdisziplinäre Lehrveranstaltungen zu kulturtheoretischen, wissenschaftsreflexiven und ethischen Themen.