Clemens Neudecker, Karolina Zaczynska, Konstantin Baierer,
Georg Rehm, Mike Gerber, Julián Moreno Schneider

Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten

1 Einleitung

Durch die systematische Digitalisierung der Bestände in Bibliotheken und Archiven hat die Verfügbarkeit von Bilddigitalisaten historischer Dokumente rasant zugenommen. Das hat zunächst konservatorische Gründe: Digitalisierte Dokumente lassen sich praktisch nach Belieben in hoher Qualität vervielfältigen und sichern. Darüber hinaus lässt sich mit einer digitalisierten Sammlung eine wesentlich höhere Reichweite erzielen, als das mit dem Präsenzbestand allein jemals möglich wäre. Mit der zunehmenden Verfügbarkeit digitaler Bibliotheks- und Archivbestände steigen jedoch auch die Ansprüche an deren Präsentation und Nachnutzbarkeit. Neben der Suche auf Basis bibliothekarischer Metadaten erwarten Nutzer:innen auch, dass sie die Inhalte von Dokumenten durchsuchen können.

Im wissenschaftlichen Bereich werden mit maschinellen, quantitativen Analysen von Textmaterial große Erwartungen an neue Möglichkeiten für die Forschung verbunden. Neben der Bilddigitalisierung wird daher immer häufiger auch eine Erfassung des Volltextes gefordert. Diese kann entweder manuell durch Transkription oder automatisiert mit Methoden der *Optical Character Recognition* (OCR) geschehen (Engl et al. 2020). Der manuellen Erfassung wird im Allgemeinen eine höhere Qualität der Zeichengenauigkeit zugeschrieben. Im Bereich der Massendigitalisierung fällt die Wahl aus Kostengründen jedoch meist auf automatische OCR-Verfahren.

Die Einrichtung eines massentauglichen und im Ergebnis qualitativ hochwertigen OCR-Workflows stellt Bibliotheken und Archive vor hohe technische Herausforderungen, weshalb dieser Arbeitsschritt häufig an dienstleistende Unternehmen ausgelagert wird. Bedingt durch die Richtlinien für die Vergabepraxis und fehlende oder mangelhafte Richtlinien der digitalisierenden Einrichtungen bzw. entsprechender Förderinstrumente führt dies jedoch zu einem hohen Grad an Heterogenität der Digitalisierungs- bzw. Textqualität sowie des Umfangs der strukturellen und semantischen Auszeichnungen. Diese Heterogenität erschwert die Nachnutzung durch die Forschung, die neben einheitlichen

https://doi.org/10.1515/9783110691597-009

Mindeststandards für die Textqualität vor allem verlässliche Angaben bzgl. Qualität und Umfang der Struktur- und Texterfassung voraussetzt. Eine systematische Qualitätskontrolle findet bei der Massendigitalisierung auf Grund der großen Textmengen allenfalls stichprobenartig statt. Strukturelle Auszeichnungen werden manuell vorgenommen und stehen noch nicht in Verbindung mit von der OCR identifizierten Strukturen. Auf Grund mangelnder ausgereifter Verfahren zur automatischen Qualitätssicherung bleibt Bibliotheken und Archiven nur die Möglichkeit, ausgewählte Dokumente manuell als *Ground Truth* (GT) zu erfassen und mit den OCR-Ergebnissen zu vergleichen, was für die Massendigitalisierung nicht leistbar ist.

Im Folgenden möchte dieser Beitrag zunächst einen Überblick über die für die Bestimmung der OCR-Qualität vorliegenden gängigsten Methoden und Metriken (Abschnitt 2) bieten. Zwei Beispiele illustrieren die Heterogenität der Aussagekraft diverser Metriken und leiten die Diskussion der Vor- und Nachteile der Verfahren ein. Zudem werden alternative Ansätze für eine Qualitätsbestimmung betrachtet, bevor Abschnitt 3 die Relevanz der OCR-Qualität und Metriken aus der Perspektive dreier typischer Anwendungsfälle diskutiert und bewertet. Abschließend werden die gewonnenen Erkenntnisse kurz zusammengefasst und es wird ein Ausblick auf die Möglichkeiten einer Dokumentenanalyse gegeben, die sich durch eine zunehmend stärkere Verflechtung von Verfahren für die OCR, Layoutanalyse und sprachwissenschaftliche Methoden andeutet.

2 OCR-Evaluierung: Methoden und Metriken

Die effiziente und aussagekräftige Bewertung der Qualität von OCR-Ergebnissen ist in mehrerlei Hinsicht problematisch. Zum einen erfordern etablierte Verfahren das Vorliegen geeigneter GT-Daten, die als Referenz für die gewünschte Ergebnisqualität dienen. Vor dem Hintergrund der Massendigitalisierung ist dies jedoch weder sinnvoll noch leistbar. Die Erstellung von GT für historische Dokumente ist zum einen äußerst zeitintensiv, zum anderen würde gerade durch die Erstellung von GT in der Form von Transkriptionen die eigentliche OCR überflüssig gemacht. Es soll aber gerade darum gehen, eine hochqualitative manuelle Transkription durch eine vollautomatisierte OCR-Erkennung zu ersetzen. Wie lässt sich also auf kosten- und zeiteffiziente Weise die Qualität der OCR-Ergebnisse für Millionen von Seiten diverser historischer Dokumente ermitteln, ohne diese in vollem Umfang bereits vorab transkribieren zu müssen?

Hier schließt sich eine weitere Schwierigkeit bei der Bewertung von OCR-Resultaten an – Standards und etablierte Richtlinien, die für die GT-Erstellung

klare und einheitliche Vorgaben machen, sind bisher nur in Teilen vorhanden. Insbesondere bei historischen Dokumenten ergibt sich noch ein weites Feld bislang nicht hinreichend spezifizierter Fälle, die bei der Bewertung der OCR-Qualität auftreten können. Als Beispiele seien hier exemplarisch Ligaturen[1] genannt, die entweder als einzelne Zeichen oder als Zeichenkombination erkannt werden können, oder die Kodierung von historischen Sonderzeichen[2] wie z. B. Umlauten oder Abkürzungen, die noch nicht im Unicode-Standard enthalten sind und bei denen auf Erweiterungen wie die Medieval Unicode Font Initiative[3] oder sogar die *Private Use Area*[4] zurückgegriffen werden muss. Ein erster Versuch, hier zwischen der OCR-Community und den Anforderungen der Wissenschaft an OCR-Ergebnisse zu vermitteln und entsprechende Grundlagen zu fixieren, stellen die *OCR-D Ground Truth Guidelines*[5] dar (Boenig et al. 2018, Boenig et al. 2019). Weitere Fragen ergeben sich bei der praktischen Implementierung: So existieren bislang keinerlei standardisierte Vorgaben, wie mit technischen Details beispielsweise der Zählung von Interpunktion und Leerzeichen bzw. Zeichen, die sich nicht mit einem einzelnen Codepoint[6] darstellen lassen, im Zuge der Qualitätsmessung zu verfahren ist.

Da es sich bei der zu verarbeitenden Menge an digitalisierten und noch zu digitalisierenden historischen Dokumenten um Millionen von Titeln handelt, ist es naheliegend, zunächst auf Verfahren und Methoden zurückzugreifen, die z. B. anhand von Stichproben oder über statistische Verfahren versuchen, Einblicke in die Qualität der OCR zu gewinnen. So liefert die OCR-Software zumeist eine Selbsteinschätzung des Algorithmus in Form eines Konfidenzwertes, der angibt, wie „sicher" der OCR-Algorithmus ist, ein Zeichen richtig erkannt zu haben. Doch wie verlässlich ist diese Angabe im Vergleich zur GT? Hierzu fehlen noch entsprechend aussagekräftige Studien und Auswertungen. Bei der Auswahl von Stichproben ist auf die Verwendung geeigneter Verfahren für Repräsentativität zu achten (siehe unten zum Bernoulli-Experiment und Wernersson 2015), jedoch kann so zumindest die für eine aussagekräftige Bewertung benötigte Menge an GT reduziert werden.

Eine noch vielschichtigere Perspektive auf die OCR-Qualität ergibt sich, wenn man auch die Qualität der Layoutanalyse (bzw. Segmentierung) im angemessenen Maße mitberücksichtigt, da sie ihrerseits einen wichtigen Teil des

1 Vgl. https://ocr-d.de/en/gt-guidelines/trans/trLigaturen2.html (1.12.2020).
2 Vgl. https://ocr-d.de/en/gt-guidelines/trans/trBeispiele.html (1.12.2020).
3 Vgl. https://folk.uib.no/hnooh/mufi/ (1.12.2020).
4 Vgl. The Unicode Standard, Chapter 23: Special Areas and Format Characters. https://www.unicode.org/versions/Unicode13.0.0/ch23.pdf (1.12.2020).
5 Vgl. https://ocr-d.de/en/gt-guidelines/trans/ (1.12.2020).
6 Vgl. https://de.wikipedia.org/wiki/Codepoint (1.12.2020).

OCR-Erkennungsprozesses darstellt und insbesondere bei Dokumenten mit komplexem Layout eine eigene Betrachtung erforderlich macht. Beispielhaft sei hier die Digitalisierung und OCR von Zeitungen genannt, bei denen die Einhaltung der korrekten Reihenfolge von Abschnitten innerhalb von Artikeln im Zuge der OCR aufgrund des komplexen, zumeist mehrspaltigen Layouts von Zeitungen nur durch eine akkurate Layoutanalyse gewährleistet werden kann. Zudem erfordern Methoden für die Texterkennung, die auf tiefen neuronalen Netzen basieren und momentan die beste OCR-Qualität liefern, bereits segmentierte Textzeilen (Neudecker et al. 2019). Dafür ist es erforderlich, dass im OCR-Workflow vor der Texterkennung mittels Layoutanalyse Textbereiche und einzelne Zeilen in der richtigen Reihenfolge erkannt werden.

Zusammenfassend lässt sich festhalten, dass bislang zwar grundlegende Verfahren und Metriken für die GT-basierte Qualitätsbestimmung von OCR-Ergebnissen existieren, diese aber auch für viele Detailfragen noch keine zufriedenstellenden Antworten geben. Zudem ist eine GT-basierte Evaluierung über große, im Kontext von Massendigitalisierung entstehende Bestände nicht effizient durchführbar. Inwieweit Konfidenzwerte und auf Stichproben beruhende statistische Auswertungen imstande sind, belastbare Aussagen zu liefern, muss zudem noch systematisch untersucht werden.

2.1 Stand der Technik

Unterschiedlichste Metriken liegen inzwischen in der Form wissenschaftlicher Beiträge sowie teilweise auch in Implementierungen vor, liefern aber jeweils nur eine Teilperspektive auf die Qualität der OCR. Im folgenden Abschnitt werden häufig genutzte Metriken diskutiert. Anschließend wird mit zwei Beispielen illustriert, inwieweit die sich aus den jeweiligen Metriken ergebenden Aussagen zur Qualität bei Anwendung auf unterschiedliche Dokumentarten und Anforderungen voneinander abweichen bzw. welche Aspekte von Qualität sie jeweils besonders gut oder weniger gut abbilden.

Die grundlegenden und in der wissenschaftlichen Community am weitesten verbreiteten Methoden für die Qualitätsbestimmung von Texterkennungssystemen gehen auf die Doktorarbeit von Stephen V. Rice aus dem Jahre 1996 zurück (Rice 1996). Die Bestimmung der OCR-Qualität wird hier als eine Manipulation von Zeichenketten anhand eines Editieralgorithmus aufgefasst. Rice unterscheidet dabei *character accuracy* (Zeichengenauigkeit) und *word accuracy* (Wortgenauigkeit) wobei Sonderfälle wie *non-stopword accuracy* (Wortgenauigkeit ohne Berücksichtigung von Stoppwörtern) oder *phrase accuracy* (Genauigkeit über eine Sequenz von k Wörtern) bereits berücksichtigt sind. Aus Effizienzgründen

empfiehlt Rice für die Berechnung der jeweiligen Metriken Ukkonens Algorithmus (Ukkonen 1995), eine für lange Zeichenketten optimierte Version der Levenshtein-Distanz. Die Levenshtein-Distanz (Levenshtein 1966) ist eine oft verwendete Metrik, die die Distanz zwischen zwei Zeichenketten, meist Wörtern, bemisst. Sie wird bei der Rechtschreibprüfung und als Suchalgorithmus zur Bildung von Kandidaten bei der Rechtschreibkorrektur angewandt. Die Levenshtein-Distanz gibt die geringste Anzahl an Editieroperationen für eine Zeichenkette an, die notwendig ist, um diese in eine Zielzeichenkette umzuwandeln. Für die vom *Information Science Research Institute* (ISRI) der University of Nevada, Las Vegas 1992–1996 jährlich durchgeführten OCR-Evaluierungen wurden die von Rice vorgestellten Methoden in Form der *ISRI Evaluation Tools*[7] (Rice und Nartker 1996) implementiert, die seitdem das am häufigsten verwendete Werkzeug für die Qualitätsbestimmung von OCR im Rahmen von wissenschaftlichen Artikeln und Wettbewerben darstellen. Zwischen 2015 und 2016 wurden die *ISRI Evaluation Tools* aktualisiert, u. a. durch die Unterstützung des Unicode-Zeichensatzes und die Veröffentlichung des Quellcodes[8] (Santos 2019).

Erste systematische Studien der OCR-Qualität im Kontext von Massendigitalisierung stellen die Arbeiten von Tanner et al. (2009) und Holley (2009) dar. Tanner et al. (2009) untersuchen die OCR-Qualität des digitalisierten Zeitungsarchivs der British Library. Dabei bedienen sie sich der Metriken *character error rate* (CER) und *word error rate* (WER), die den Anteil inkorrekter Buchstaben bzw. Wörter im OCR-Ergebnis im Verhältnis zur GT angeben. Zusätzlich schlagen sie eine *significant word error rate* vor, in der ausschließlich die Anzahl signifikanter Wörter unter den nicht korrekt erkannten Beachtung findet, also der Wörter, die relevant für die Erfassung des Dokumentinhalts sind. Dabei darf die Qualität der Originaltexte nicht außer Acht gelassen werden, denn laut Klijn (2008) sagt die Qualität der mit OCR verarbeiteten Texte oft mehr über die Qualität der Digitalisierung aus als über das verwendete OCR-Verfahren (dies gilt insbesondere für historische Dokumente). Holley (2009) präsentiert eine Untersuchung der OCR-Qualität für das *Australian Newspaper Digitisation Program*. Neben einer Analyse der wichtigsten Einflussfaktoren für die OCR-Qualität werden auch Hinweise gegeben, wie sich diese potenziell verbessern lässt, z. B. durch die Integration von Lexika für unterrepräsentierte Sprachvariationen wie Dialekte. Ein Sprachmodell könnte ebenfalls zur Verbesserung der Ergebnisse führen, indem Wörter bevorzugt werden, die bezüglich ihres Kontextes wahrscheinlicher an dieser Stelle im Satz auftauchen als andere. Der Einsatz frequenzbasierter Sprachmodelle in der OCR wurde jedoch durch den Einzug

7 Vgl. https://code.google.com/archive/p/isri-ocr-evaluation-tools/ (1.12.2020).
8 Vgl. https://github.com/eddieantonio/ocreval (1.12.2020).

sprachunabhängiger tiefer neuronaler Netze im Bereich der OCR weitestgehend verdrängt, auch weil neuere Klassifikationsmodelle durch frequenzbasierte Sprachmodelle sogar an Qualität einbüßen können (Smith 2011). Letzteres ist insbesondere bei historischer Sprache, wo entsprechend robuste und zugleich spezifische Sprachmodelle noch nicht im benötigten Ausmaß zur Verfügung stehen, ein limitierender Faktor.

Das von der Europäischen Kommission geförderte Projekt IMPACT[9] (Improving Access to Text, 2008–2012) stellt den bislang ambitioniertesten Versuch dar, bessere, schnellere und effizientere OCR für historische Dokumente zu entwickeln. Dazu wurden zahlreiche technologische Innovationen erarbeitet, mit dem Ziel, den Zugang zu historischen Dokumenten zu verbessern und die Volltextdigitalisierung deutlich voranzutreiben. Im Laufe des Projekts entstanden auch mehrere Verfahren für die OCR-Evaluierung, darunter das *NCSR Evaluation Tool*,[10] welches auf den *ISRI Evaluation Tools* beruht und diese um Unterstützung von UTF-8, UTF-16 und die Metrik *figure of merit* erweitert. Die *figure of merit* ist eine für das IMPACT-Projekt definierte Metrik, die versucht, den Aufwand für eine manuelle Nachkorrektur der OCR zu beschreiben (Kluzner et al. 2009). Dafür werden Ersetzungen um einen Faktor 5 höher gewichtet als Löschungen, da sie entsprechend aufwendiger zu erkennen und korrigieren sind. Darüber hinaus entstanden die Werkzeuge *INLWordAccuracyTool*[11] und *ocrevalUAtion*[12] sowie eine Anleitung[13] zur Messung von OCR-Qualität. Das Werkzeug *ocrevalUAtion* erlaubt den Vergleich zwischen Referenztext und OCR-Ergebnissen sowie zwischen verschiedenen OCR-Ergebnissen für einen Referenztext und wertet die OCR-Fehler statistisch aus. Dabei werden neben PAGE-XML[14] auch andere OCR-Formate wie ABBYY-XML[15], das in Bibliotheken gebräuchliche ALTO,[16] das in den Digital Humanities favorisierte TEI[17] oder unformatierter Text unterstützt.

Wesentliche Beiträge für die OCR-Evaluierung wurden von der Forschungsgruppe PRImA (Pattern Recognition & Image Analysis Research Lab) der Universität Salford, Greater Manchester erarbeitet. Schon früh wurden dort mehrere

9 Vgl. http://www.impact-project.eu/ (1.12.2020), grant agreement ID 215064.
10 Vgl. https://users.iit.demokritos.gr/~bgat/OCREval/ (1.12.2020).
11 Vgl. https://github.com/JessedeDoes/INLWordAccuracyTool (1.12.2020).
12 Vgl. https://github.com/impactcentre/ocrevalUAtion (1.12.2020).
13 Vgl. https://sites.google.com/site/textdigitisation/home (1.12.2020).
14 Vgl. https://ocr-d.de/en/gt-guidelines/trans/trPage (1.12.2020).
15 Vgl. https://web.archive.org/web/20200924054833/https://abbyy.technology/en:features:ocr:xml. (6.7.2021).
16 Vgl. https://www.loc.gov/standards/alto/ (1.12.2020).
17 Vgl. https://tei-c.org/(1.12.2020).

Standards für die Auszeichnung und die Evaluierung von OCR-Daten entwickelt. Das Format PAGE (Page Analysis and Ground-Truth Elements) ist ein XML-basierter Standard für die Auszeichnungen von GT (Pletschacher und Antonacopoulos 2010). Mit ihm können granulare Informationen für die Bildmerkmale (Bildränder, Verzerrungen und entsprechende Korrekturen, Binarisierung etc.) sowie zur Struktur des Layouts und des Inhalts festgehalten werden. Softwarewerkzeuge[18] für eine einheitliche Evaluierung von OCR-Dokumenten wurden dort u. a. für die Layoutanalyse (Clausner et al. 2011) und die Evaluierung der Lesereihenfolge (*reading order*, Clausner et al. 2013) entwickelt sowie eine Methode zur Evaluation der CER, wenn die Lesereihenfolge nicht korrekt erkannt wurde (Clausner et al. 2020).

Der Standardisierungsprozess und die Etablierung von vergleichbaren Metriken für die Layoutanalyse von historischen Dokumenten wurde durch verschiedene Wettbewerbe und *shared tasks* im Rahmen des IAPR-TC11 vorangetrieben, z. B. für die Erkennung von komplexem Layout, wie in der *Competition on Recognition of Documents with Complex Layouts* (Antonacopoulos et al. 2015, Clausner et al. 2017, Clausner et al. 2019), oder für historische Dokumente, wie in den *shared tasks* zu *Historical Newspaper Layout Analysis* (Antonacopoulos 2013), *Historical Book Recognition* (Antonacopoulos 2013) und der *Historical Document Layout Analysis Competition* (Antonacopoulos et al. 2011).

Neuere Arbeiten evaluieren die Leistung der einzelnen Arbeitsschritte eines kompletten OCR-Workflows (Pletschacher et al. 2015, Clausner et al. 2016), wie er für das Projekt Europeana Newspapers (2012–2015, Neudecker und Antonacopoulos 2016) angewandt wurde, wobei in dem Projekt mehr als 8 Millionen historische Zeitungsseiten mit OCR und zusätzliche 2 Millionen Seiten mit Artikelsegmentierung verarbeitet wurden. Vor dem Hintergrund derartiger Massendigitalisierungsprojekte entwickelten Clausner et al. (2016) Methoden für die Vorhersage der zu erwartenden OCR-Qualität auf der Grundlage geringer Mengen von GT und der Ermittlung von Merkmalen in Dokumenten, die in einer Abhängigkeit zur erwartbaren Güte der Texterkennung stehen.

Im Rahmen des QURATOR-Projekts[19] (Rehm et al. 2020) entstand an der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB) das Werkzeug *dinglehopper*.[20] Es dient zur transparenten und standardisierten GT-basierten Evaluierung von OCR-Qualität mittels CER/WER und bietet eine Visualisierung von fehlerhaft erkannten Zeichen anhand eines *side-by-side*-Vergleichs von GT und OCR-Ergebnis. Die Software interpretiert Texte als Aneinanderreihungen

18 Vgl. https://www.primaresearch.org/tools/PerformanceEvaluation (1.12.2020).
19 Vgl. https://qurator.ai/ (1.12.2020).
20 Vgl. https://github.com/qurator-spk/dinglehopper/ (1.12.2020).

von Graphem-Clustern[21] – Zeichen im Sinne des Unicode-Standards – die zunächst anhand ihrer Gemeinsamkeiten aligniert und anschließend verglichen werden. Die Visualisierung erlaubt die manuelle Inspektion von OCR-Fehlern, so dass Probleme in Kodierung, Normalisierung, Layoutanalyse oder gar GT einfach erkennbar sind.

Schließlich sei hier noch auf das Bernoulli-Experiment[22] eingegangen, welches das von den DFG-Praxisregeln *Digitalisierung*[23] (Stand 1.12.2020) empfohlene und für alle Drucke ab 1850 verpflichtende Verfahren zur Qualitätsmessung für Digitalisierungsvorhaben mit OCR darstellt. Da im Allgemeinen keine GT zur Verfügung steht, um die OCR-Qualität umfassend zu messen, werden in einem Experiment eine gewisse Zahl von Stichproben manuell untersucht und so der Fehler statistisch ermittelt. Ein Vorteil dieser Methode ist, dass die Berechnung anhand von randomisierten Stichproben eine statistisch belastbare Aussage über die Qualität erlaubt und es somit eine gewisse Sicherheit gibt, die durch eine manuelle und somit durch Selektionsbias verzerrte Auswahl von exemplarisch geprüften Seiten nicht gegeben wäre. Ein Nachteil dieser Methode liegt jedoch darin, dass die manuelle Kontrolle einzelner Zeichen und damit der manuellen Suche von Textkorrespondenzen im Original nahezu keine Aussage über die Qualität einer Layoutanalyse zulässt, da Fehler in der Lesereihenfolge nicht auffallen. So wird am Ende lediglich eine CER ohne Beachtung einer Lesereihenfolge ermittelt. Zudem kann die randomisierte Stichprobe auch nachteilig sein, da nicht jeder Aspekt eines digitalisierten Werkes gleich bedeutsam ist: Ein fehlerhafter Titel eines Zeitungsartikels ist womöglich schwerwiegender zu bewerten als ein Fehler innerhalb einer Zeitungsannonce.

2.2 Beispiele

An dieser Stelle sollen zwei Beispiele zur Illustration dienen inwieweit die gebräuchlichsten der hier aufgeführten Methoden und Metriken für die OCR-Evaluierung in ihrer Bewertung der Ergebnisse übereinstimmen und inwieweit sie voneinander abweichen. Dabei soll insbesondere der Blick für die Auswirkungen der (Nicht-)Berücksichtigung von Lesereihenfolge im Zuge der Layoutanalyse geschärft werden.

---

21 Vgl. Unicode Standard, Annex #29: Unicode Text Segmentation. https://www.unicode.org/reports/tr29/tr29-37.html (1.12.2020).

22 Vgl. https://en.wikipedia.org/wiki/Bernoulli_trial (1.12.2020).

23 Vgl. DFG-Praxisregeln *Digitalisierung* [12/16], S. 34 ff.

Für das Beispiel wurden für zwei Seiten aus dem digitalisierten Bestand der SBB GT erstellt und die Digitalisate mit der OCR-Software *Tesseract*[24] verarbeitet. Dabei wurde einmal eine Seite aus einer Monografie aus dem VD16 gewählt, die als Besonderheit Marginalien enthält. Kontrastiert wird diese mit einer Seite der Beilage einer Berliner Tageszeitung vom 1. Mai 1930. So wird die Bedeutung der bei Zeitungen und anderen mehrspaltigen Dokumenten bedeutenden Lesereihenfolge besser ersichtlich. Die OCR-Ergebnisse wurden anschließend mit mehreren der oben dargestellten Metriken (sowie je nach Methode unter Hinzuziehung von GT) ausgewertet (siehe Tab. 1).

Auch wenn es sich hier nur um einzelne und zudem ausgewählte Beispiele handelt, so ist die Diversität der Aussagen im Hinblick auf die erzielte Qualität doch enorm. Welche Fehler sind hier im OCR-Prozess aufgetreten, die von den jeweiligen Metriken besser oder weniger gut erfasst wurden?

Betrachten wir zunächst Beispiel (a), ein monografisches Druckwerk aus dem 16. Jahrhundert, wie sie im Rahmen der umfangreichsten von der Deutschen Forschungsgemeinschaft (DFG) geförderten Digitalisierungskampagne, den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke des 16./17./18. Jahrhunderts (VD16,[25] VD17,[26] VD18[27]), in erheblichem Umfang (ca. 106 000 Titel in VD16, 303 000 in VD17 und mindestens 600 000 in VD18) im Entstehen sind. Neben der hier verwendeten Schwabacher[28] treten als Herausforderung für die OCR in erster Linie Marginalien auf, wie sie ebenfalls in den Drucken des 16.–18. Jahrhundert prominent vertreten sind. Laut den *OCR-D GT Guidelines* bilden Marginalien[29] einen Bestandteil der Lesereihenfolge[30] und sind demnach ihrem semantischen Bezug zu den jeweils zugehörigen Absätzen entsprechend zu erfassen. Die in Abb. 1a dargestellte GT erfordert hiernach die Wiedergabe der Sequenz der Textbereiche in der durch den Pfeil visualisierten Lesereihenfolge.

24 Vgl. https://github.com/tesseract-ocr/tesseract (1.12.2020).
25 Vgl. http://www.vd16.de/ (1.12.2020).
26 Vgl. http://www.vd17.de/ (1.12.2020).
27 Vgl. http://www.vd18.de/ (1.12.2020).
28 Vgl. https://de.wikipedia.org/wiki/Schwabacher (1.12.2020).
29 Vgl. https://ocr-d.de/en/gt-guidelines/trans/lyMarginalie.html (1.12.2020).
30 Vgl. https://ocr-d.de/en/gt-guidelines/trans/lyLeserichtung.html (1.12.2020).

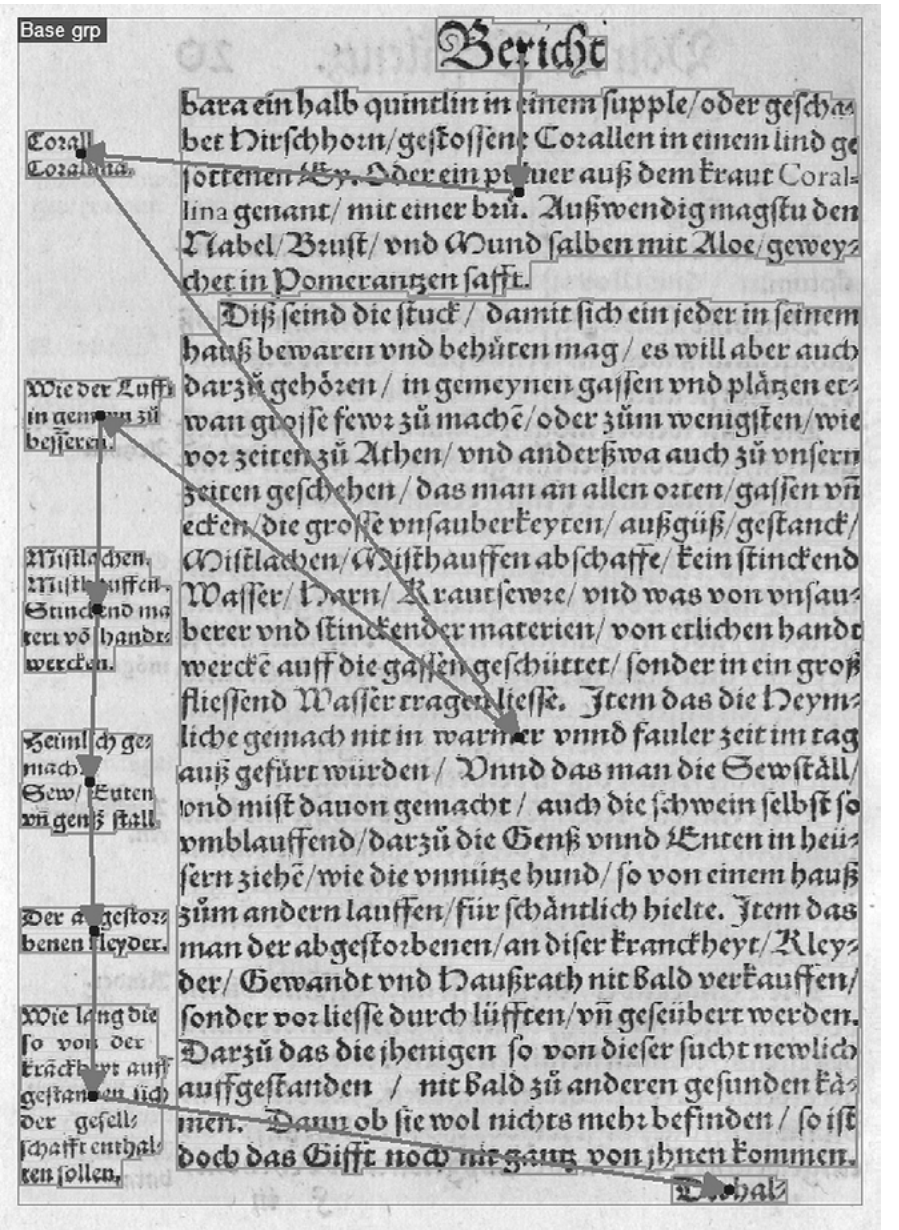

Abb. 1a: Visualisierung der Ground Truth für zwei ausgewählte Beispiele
Winther, Johannes. Bericht und Ordnung in diesen sterbenden Läufen der Pestilenz. 1564. VD16 ZV 1311

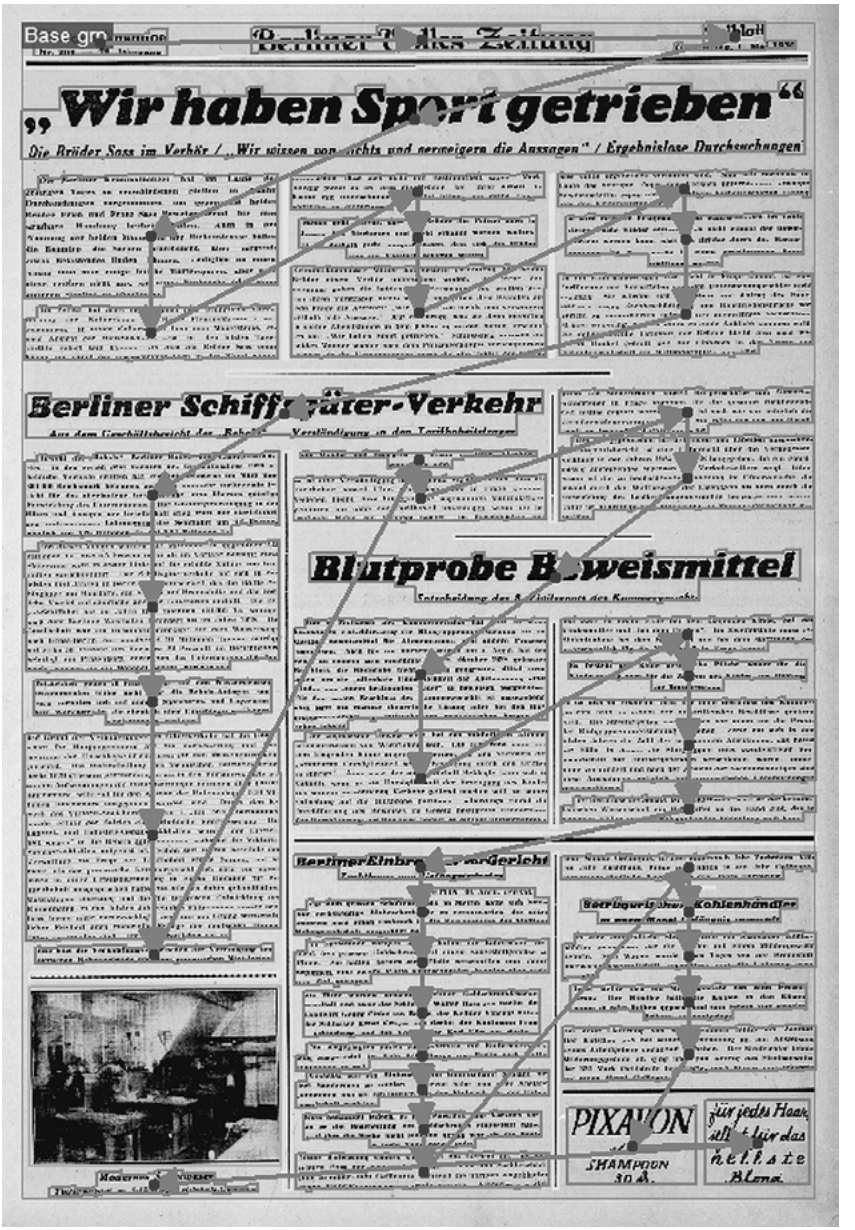

Abb. 1b: Visualisierung der Ground Truth für zwei ausgewählte Beispiele Berliner Volkszeitung. 1. Mai 1930. Morgen-Ausgabe. Beiblatt. ZDB 27971740

Tab. 1: Vergleich der Metriken* für zwei ausgewählte Beispiele

	Konfidenz	CER	WER	BOW	RO	Bernoulli
(a) Monografie	82.87 %	44.29 %	79.20 %	50.07 %	24.87 %	94.2 %
(b) Zeitung	60.85 %	62.19 %	87.63 %	55.73 %	69.27 %	89.0 %

* Erläuterung: Konfidenz = Konfidenzwerte des OCR-Algorithmus über die Seite gemittelt; CER: *character error rate*; WER: *word error rate*; BOW: *bag-of-words word index success rate*; RO: *F1 score reading order*; Bernoulli: Genauigkeit Bernoulli-Experiment mit Stichprobe von 500 randomisierten Zeichen. Berechnung der CER/WER/BOW mit PRImA Text Eval v.1.5, Berechnung des *F1 score* mit *reading order* mit PRImA Layout Eval v.1.9 und Evaluationsprofil: *document structure*, vgl. https://www.primaresearch.org/tools/PerformanceEvaluation (1.12.2020).

Bei Abb. 1b handelt es sich um ein Beiblatt einer vom Mikrofilm digitalisierten Tageszeitung aus dem frühen 20. Jahrhundert. Auch im Bereich der Zeitungsdigitalisierung soll mit DFG-geförderten Programmen die Menge im Volltext digital verfügbarer historischer Zeitungen massiv erhöht werden. Eine vorab durchgeführte OCR-Evaluierung ist dabei für die Antragstellung verpflichtend.[31] Das Beispiel veranschaulicht typische Schwierigkeiten für die OCR bei Zeitungen – neben einem mehrspaltigen Layout sind die Erkennung von Abbildungen und Zwischenüberschriften sowie Separatoren für die Ermittlung der korrekten Lesereihenfolge entscheidend (die GT ist erneut durch den Pfeil visualisiert).

Betrachten wir nun die Evaluierungsergebnisse unter Zuhilfenahme der verschiedenen Metriken, so können wir zunächst feststellen, dass die Bewertung anhand des Bernoulli-Experiments gemäß DFG-Praxisrichtlinien *Digitalisierung* zur optimistischsten Einschätzung kommt. Die anhand des Bernoulli-Experiments ermittelten Aussagen liegen dabei sogar noch deutlich über den Konfidenzwerten des verwendeten OCR-Algorithmus. Beide Verfahren kommen zu einer prinzipiell positiven Bewertung der OCR-Qualität in dem Sinne, dass mehr Zeichen richtig als falsch erkannt wurden. Die für die inhaltliche Qualität wesentliche WER liegt deutlich über der CER, da sich einzelne Zeichenfehler auf mehrere der vorkommenden Wörter verteilen. Anhand der BOW-Metrik kommt man in beiden Fällen zu der Einschätzung einer durchschnittlichen Qualität. Betrachtet man die WER genauer, so ergibt sich eine mehr oder weniger drastisch negative Bewertung, der zufolge nur 20 % (a) bzw. 12 % (b) der Wörter von der OCR korrekt erkannt wurden. Zieht man zusätzlich Kriterien wie die Lesereihenfolge für die Layoutanalyse heran, so ergibt sich ein nochmals heterogeneres Bild. Vor allem überrascht die deutlich positivere Bewertung von (b) mit

31 Vgl. https://www.dfg.de/foerderung/info_wissenschaft/2018/info_wissenschaft_18_08/ (1.12.2020).

einem *F1 score* von 69 %, was (b) wiederum eine erhebliche bessere Qualität als die in (a) erreichten 25 % bescheinigt. Dies lässt sich dadurch erklären, dass in Beispiel (a) die Marginalien im Zuge der Layoutanalyse mit den Zeilen des Fließtextes vermischt wurden. Dadurch wird an den jeweiligen Stellen die Lesereihenfolge inhaltlich unterbrochen, was sich gravierend auf diejenigen Metriken auswirkt, welche alignierte Inhalte erfassen.

Im Ergebnis der Layoutanalyse für Beispiel (b) wurden Separatoren überwiegend gut erkannt und der Vorgabe der GT im Hinblick auf die Sequenz der Textbereiche und -zeilen besser entsprochen. Während inhaltliche und typografische Kriterien bei der Festlegung der Lesereihenfolge in Beispiel (a) eindeutig sind, so kann für Beispiel (b) zu Recht hinterfragt werden, inwieweit sich eine GT für die Lesereihenfolge einer Zeitung objektiv definieren lässt. Hierfür bietet das PAGE-XML Format die flexiblen Konzepte *OrdererdGroup* (geordnete Gruppe) und *UnorderedGroup* (ungeordnete Gruppe) an, die auch miteinander kombiniert bzw. verschachtelt werden können. So kann z. B. die Reihenfolge der Artikel einer Zeitungsseite in einer *UnorderedGroup* abgebildet werden, wohingegen die korrekte Abfolge der Absätze innerhalb einzelner Artikel, dazugehörige Illustrationen oder Tabellen und dergleichen in einer strikt festgelegten Lesereihenfolge einer *OrderedGroup* repräsentiert wird. Hierbei entstehende Fehler sind entsprechend komplex und damit nur schwer in einer einzelnen Metrik darzustellen. Nur die Evaluierung nach *F1 score reading order* kann hier ein entsprechend differenziertes Bild liefern. Somit ist die Aussagekraft aller anderen Metriken als gering einzustufen, sobald die korrekte Erfassung der inhaltlichen Zusammenhänge auf Satz- und Abschnittsebene für die Weiterverwendung der OCR-Ergebnisse von Bedeutung ist.

Mit ebenso großer Vorsicht müssen die Konfidenzwerte des OCR-Algorithmus betrachtet werden. Da die OCR-Konfidenzen im Zuge des OCR-Prozesses automatisch entstehen, ist jedoch zumindest eine Bereitstellung dieser Information in den Metadaten empfehlenswert. Liegt z. B. der über alle Seiten eines Dokuments gemittelte Konfidenzwert unter 50 %, über 70 % oder gar über 90 %, so kann dies bereits hilfreich für die Zusammenstellung von Datensätzen für die Forschung sein (Padilla et al. 2019).

2.3 Alternative Ansätze

Nachdem in unserer bisherigen Betrachtung die gängigen Verfahren für die OCR-Evaluierung allesamt noch Defizite bei der differenzierten Ergebnisbewertung aufweisen und zudem GT benötigen, sollen in diesem Kapitel alternative Ansätze diskutiert werden. Lassen sich andere Wege und (z. B. heuristische

oder sprachwissenschaftliche) Verfahren für eine Qualitätsbewertung von (historischen) Volltexten finden, die hochgradig automatisierbar sind und dennoch belastbare Aussagen treffen?

Alex und Burns (2014) entwickelten eine GT-freie Heuristik zur Bestimmung der Textqualität der OCR englischsprachiger wirtschaftlicher Fachliteratur des 19. Jahrhunderts anhand der relativen Häufigkeit von Ziffern und Wörtern in einem Lexikon. Dazu schlagen sie einen *simple quality* (SQ) *score* vor, der sich aus dem Verhältnis der Anzahl von „guten" Wörtern (Wörtern, die in einem Lexikon vorkommen) zu allen Wörtern berechnet. Ein Problem mit diesem Ansatz im Kontext von historischen Dokumenten ist jedoch, dass keine geeigneten historischen Wörterbücher zur Verfügung stehen, durch die im Zuge der OCR korrekt erkannte, valide historische Schreibweisen in die Kategorie „guter" Wörter eingeordnet werden könnten. Im eMOP-Projekt[32] (2012–2014) wurde eine Methode gewählt, die durch die Layoutanalyse generierte *bounding boxes* heranzieht (Gupta et al. 2015). Die zugrundeliegende Annahme ist, dass für Dokumente mit guter OCR-Qualität vor allem solche *bounding boxes* vorliegen, die Text enthalten, während diese im Falle schlechter OCR-Qualität vor allem Rauschen (Pixel) enthalten. Auf dieser Grundlage wurde ein Klassifikator entwickelt, der mit einer Genauigkeit von 93 % *bounding boxes*, die Text enthalten, von denjenigen mit Rauschen unterscheiden kann. Baumann (2014) schlägt vor, die Spracherkennungsbibliothek *langid*[33] (Lui und Baldwin 2012) zu nutzen, um die Qualitätsmessung von OCR-Ergebnissen zu automatisieren. Diese Bibliothek gibt zusätzlich zur erkannten Sprache für jede Zeile eines Dokuments einen Konfidenzwert an, also eine numerische Einschätzung, wie sicher sich das Modell bei der Vorhersage ist. Die Idee ist, dass in Texten mit mehr OCR-Fehlern mehr dem Sprachmodell unbekannte Wörter enthalten sind und damit der Konfidenzwert sinkt. Ein Vorteil an diesem Verfahren ist, dass das Sprachmodell von *langid* einfach neu auf unbekannte Sprachen trainiert werden kann. Ein Ansatz für eine GT-freie Evaluierung der Bildvorverarbeitung ist derjenige von Sing, Vats und Anders (2017), welcher verschiedene Verfahren aus dem Bereich des maschinellen Lernens kombiniert, um aus einer geringen Menge von GT (Bild und Bewertung) Modelle abzuleiten, die dann auf neue Daten übertragbar sind. Diese Methode wäre prinzipiell auch auf die OCR-Qualitätsbewertung übertragbar, indem z. B. prozessrelevante technische Metadaten (Komprimierung, Bildgröße und Auflösung, Farbtiefe etc.), etwaige Bildstörungen, die die OCR-Qualität negativ beeinflussen, und relevante Merkmale (verwendete Schriftart, ein- bzw. mehrspaltiges Layout, Dokumentart etc.) herangezogen

32 Vgl. https://emop.tamu.edu/ (1.12.2020).

33 Vgl. https://github.com/saffsd/langid.py (1.12.2020).

werden, um dafür regelbasierte Heuristiken zu entwickeln oder aus den Daten ein Modell für die Qualitätsbestimmung zu trainieren.

Springmann et al. (2016) schlagen zwei Metriken für eine GT-freie OCR-Evaluierung vor: Zum einen betrachten sie Konfidenzwerte des OCR-Algorithmus unter einer Studentschen t-Verteilung[34], zum anderen Lexikalität unter Verwendung von dokumentspezifischen Sprachprofilen wie sie Reffle und Ringlstetter (2013) verwenden und in denen auch historische Schreibvarianten Berücksichtigung finden. Bei der OCR-Verarbeitung von historischen Korpora können die darin enthaltenen Dokumente einen großen Zeitraum und damit verschiedene Schreibvarianten (bezüglich der Schreibweise und verwendeten Lexik) ein und derselben Sprache beinhalten. Reffle und Ringlstetter (2013) stellen eine Evaluierungs- und Korrekturmethode ohne GT vor, bei der ein Profiler[35] für historische Dokumente statistische Sprachprofile errechnet und für die jeweiligen Profile Muster für historische Schreibvarianten, Vokabeln, Worthäufigkeiten sowie für typische OCR-Fehler verwendet werden. In ihren Experimenten zeigen sie eine starke Korrelation zwischen der Verteilung von Schreibvarianten und OCR-Fehlern auf und demonstrieren beispielhaft, wie die Profile OCR-Systeme bei der Nachkorrektur verbessern. Diese Schreibvarianten können bei der Volltextsuche ebenfalls ein Problem darstellen, weshalb die Sprachprofile auch für eine Annäherung von Suchanfrage und in Dokumenten befindliche Lexik genutzt werden. Fink et al. (2017) verbessern die Methode, indem sie die Möglichkeit hinzufügen, adaptiv auf Feedback (z. B. durch manuelle Korrekturen) zu reagieren, sowie durch das Hinzufügen zusätzlicher historischer Schreibvarianten und die Behandlung nicht interpretierbarer Token als vermutete Fehler.

Mehrere Forschungsgebiete in der automatischen Sprachverarbeitung (*Natural Language Processing*, NLP) beschäftigen sich mit der Analyse und Weiterverarbeitung von nicht wohlgeformter Sprache. Dies geschieht insbesondere in Bezug auf die Anwendungsbereiche maschinelle Übersetzung, automatische Erstellung von Textzusammenfassungen und Frage-Antwort-Systeme. Für maschinell erstellte Zusammenfassungen basieren Evaluierungen meist auf zwei eng verwandten Metriken, BLEU (Papineni et al. 2002) und ROUGE (Lin 2004). Beide messen die lexikalische Überschneidung zwischen einer oder mehreren Referenz-Zusammenfassungen und dem maschinell erstellten Text. Eine weitere Evaluierungsmethode ist METEOR (Banerjee und Lavie 2005), welche zusätzlich die Wörter auf ihre Stammform zurückführt (*Stemming*) und unter Verwendung von Wörterbüchern und Wissensbasen Synonyme beim Vergleich zum Referenztext mit in Betracht zieht.

34 Vgl. https://de.wikipedia.org/wiki/Studentsche_t-Verteilung (1.12.2020).
35 Vgl. https://github.com/cisocrgroup/Profiler (1.12.2020).

Auch für die maschinelle Übersetzung oder für Frage-Antwort-Systeme werden diese Metriken benutzt, die jedoch alle vom Vorhandensein von GT-Daten abhängig sind. Diese NLP-Bereiche grenzen sich insofern von einer OCR-Evaluierung ab, als hier jeweils mindestens eine, oft jedoch auch *mehrere* Übersetzungen oder Zusammenfassungen bzw. *mehrere* korrekte Antworten auf eine Frage möglich sind – so enthalten Texte linguistische und lexikalische Variationen sowie verschiedene Wortstellungen. Zusätzlich sollte der generierte Ausgabetext idealerweise auch auf der semantischen Ebene evaluiert werden, zum Beispiel im Hinblick auf Informationsgehalt, Kohärenz und ob der Inhalt des dahinterliegenden Textes adäquat und korrekt wiedergegeben wird. Dies macht eine einheitliche Beurteilung schwierig und ist noch eine offene Frage in der Forschung.

Einen Schritt hin zu einer Evaluierung, die zur Einschätzung der Qualität zahlreiche unterschiedliche Aspekte in Betracht zieht, stellt MQM (Lommel et al. 2013) dar. MQM, *Multidimensional Quality Metrics*, ist ein Framework zur Bewertung der Übersetzungsqualität, mit dem Nutzer:innen ihre eigenen Bewertungsmetriken anpassen können. Hierbei werden, im Gegensatz zu den oben genannten Metriken, verschiedene Klassen von Fehlertypen definiert, die sich an die Ansprüche der Nutzer:innen an den übersetzten Text richten. Ein möglicher Fall wäre eine inkonsistente Terminologie innerhalb einer Übersetzung, in der zum Beispiel dieselbe Entität als *PC* in einem Satz und als *Computer* in einem anderen Satz übersetzt wurde. Dies würde in einem Referenzhandbuch als Fehler gesehen werden, in einem Zeitungsartikel aber kann diese Variation durchaus akzeptabel sein, vielleicht sogar bevorzugt werden. Die Übersetzungsqualität ist somit immer relativ zu dem beabsichtigten kommunikativen Zweck und Kontext des Textes zu sehen (Burchardt et al. 2016). Die formale Spezifikation sowie unterschiedliche Gewichtung von Fehlerklassen, wie sie durch MQM ermöglicht wird, kann dazu dienen, die Qualitätseinschätzung der automatischen Übersetzung zu verbessern. Weitere Forschungsarbeiten werden zeigen müssen, ob Metriken wie z. B. BLEU, ROUGE, METEOR und MQM in sinnvoller und zielführender Weise für die Evaluation von OCR-Qualität eingesetzt werden können.

So kann man feststellen, dass zwar verschiedene alternative, auch GT-freie Ansätze für die OCR-Evaluierung existieren, diese sich jedoch in den meisten Fällen vorhandene Sprachressourcen zunutze machen. Sind die Sprachressourcen oder Methoden aber nicht auf die besonderen Anforderungen historischer Sprache zugeschnitten, so ist deren Anwendbarkeit für die OCR-Evaluierung begrenzt oder noch nicht systematisch untersucht. Inwieweit eine anwendungsbezogene Perspektive auf die OCR-Ergebnisse den Blick für die Qualität weiter zu schärfen vermag, wird im folgenden Abschnitt ausgeführt.

3 Anwendungsbezogene Perspektiven auf OCR-Qualität

Vor dem Hintergrund der bislang dargestellten Methoden, Metriken und Ansätze stellt sich nun erneut die Frage, wie Qualität am besten messbar ist. Kann eine Metrik überhaupt genug Aussagekraft für sämtliche Anforderungen haben? Und unter Beachtung der prohibitiven Kosten für die GT-Erstellung und der eingeschränkten Leistungsfähigkeit alternativer Ansätze: Kann die Erwartung einer replizierbaren und belastbaren Methode für die OCR-Qualitätsmessung ohne GT und für unterschiedliche Dokumenttypen überhaupt aufrechterhalten werden? Oder müssen nicht viel eher konkrete Anwendungsfälle für die OCR unterschieden und für den jeweiligen Fall die bedeutsamsten und praktikabelsten Verfahren angewendet werden? Während für die Indexierung in einer Suchmaschine für eine Keyword-Suche die Reihenfolge von Wörtern in einem OCR-Ergebnis keine bzw. nur eine sehr untergeordnete Rolle spielt, so ist die Einhaltung der korrekten Lese-, Satz- und Wortreihenfolge für die semantische Analyse und Weiterverarbeitung der OCR-Ergebnisse mit NLP oder den in den Digital Humanities verwendeten Methoden von besonderer Bedeutung. Dementsprechend werden hier drei typische Anwendungsfälle für die Nutzung von OCR-Ergebnissen kurz dargestellt und es wird diskutiert, welche der vorgestellten Methoden und Metriken für die jeweiligen Anwendungsfälle besonders geeignet sind bzw. welche Aspekte der OCR-Qualitätsbewertung aus der Perspektive der Anwendungsfälle momentan noch nicht oder nur unzureichend durch die vorgestellten Metriken abgedeckt werden.

3.2 Natural Language Processing

Named Entity Recognition (NER), die automatische Erkennung von Eigennamen in Texten, stellt eine Schlüsseltechnologie für den Zugriff auf die Inhalte digitaler Bibliotheksbestände dar, da die Suchbegriffe der Nutzeranfragen häufig Personen- oder Ortsnamen sowie Zeitangaben beinhalten (Crane und Jones 2006)[36]. Bislang stellt die Anreicherung von OCR-Ergebnissen in Bibliotheken mit NER jedoch noch einen Sonderfall dar. Lediglich im Bereich der Zeitungsdigitalisierung sind auf Projektbasis größere Bestände einer NER unterzogen worden (Neudecker et al. 2014, Mac Kim und Cassidy 2015). Grund dafür war die meist

36 Siehe hierzu auch *Named Entity Linking mit Wikidata und GND – Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten* (in diesem Band).

als zu gering eingeschätzte Erfolgsrate auf Basis bereits unzufriedenstellender OCR-Ergebnisse (Kettunen et al. 2017, Kettunen und Pääkkönen 2016). Demnach werden *Named Entities* auch noch nicht in nennenswertem Umfang in die digitalen Sammlungen, Suchmaschinen und Bibliothekskataloge integriert.

Gleichzeitig ist NER ein klassischer Anwendungsfall innerhalb der automatischen Sprachverarbeitung. Für die Genauigkeit der Eigennamenerkennung spielt die OCR-Qualität der Texte eine entscheidende Rolle. Hamdi et al. (2019) erstellten einen Datensatz zur Eigennamenerkennung und integrierten darin verschiedene Klassen von OCR-Fehlern. Sie beobachteten ein Absinken der Genauigkeit des NER-Modells, das auf Basis eines neuronalen Netzes mit Long Short-Term Memory und einer Conditional Random Field-Schicht (LSTM-CRF) trainiert wurde, von 90 % auf 60 %, wenn die Fehlerrate der Worterkennung von 1 % auf 7 % bzw. die Fehlerrate der Zeichenerkennung von 8 % auf 20 % erhöht wurde. Eine Evaluierung der eingereichten Systeme des *shared task HIPE* (*Identifying Historical People, Places and other Entities*) kam zu ähnlichen Ergebnissen bezüglich der Abhängigkeit zwischen NER-Ergebnissen und der OCR-Qualität für historische Texte (Ehrmann et al. 2020). Der *shared task* wurde 2020 im Rahmen der elften Conference and Labs of the Evaluation Forum (CLEF2020)[37] organisiert und untersucht NER in historischen Zeitungen (auf Deutsch, Französisch und Englisch). Wie schon erwähnt ist der Umgang mit einer größeren Sprachvarietät eine Herausforderung bei diachronen historischen Korpora. Deshalb ist ein Ziel des *shared task*, die Robustheit von NER-Systemen auch für Eingaben außerhalb der Standardsprache zu stärken. Zudem soll eine Vergleichbarkeit der Performance von NER-Verfahren für historische Digitalisate ermöglicht werden. Ein längerfristiges Ziel ist eine inhaltliche Erschließung der in Dokumenten vorkommenden Eigennamen, die die Auffindbarkeit der Dokumente für Suchanfragen verbessern könnte. Die Evaluierung der Systeme ergab u. a., dass Verfahren, die tiefe Netze sowie vortrainierte Modelle nutzten (insbesondere BERT, Devlin et al. 2019), zu besseren Ergebnissen kamen als symbolische oder musterbasierte Verfahren.

Zusätzlich zur Erkennung von Eigennamen kann die Verlinkung dieser in Wissensbasen (*Named Entity Linking*, NEL) zur Disambiguierung von mehrdeutigen Eigennamen genutzt werden, um so Inhaltserschließung und Suchergebnisse zu optimieren. Hier spielt die OCR-Qualität ebenso eine bedeutende Rolle (Pontes et al. 2019). Auch für andere Aufgabenbereiche des NLP ist eine gute OCR-Qualität entscheidend (Mieskes und Schmunk 2019, van Strien et al. 2020). Eine Hoffnung liegt in vortrainierten Transformer-Modellen, die sich aufgrund

37 Vgl. https://impresso.github.io/CLEF-HIPE-2020/ und https://clef2020.clef-initiative.eu/ (1.12.2020).

der größeren Datengrundlage, auf der sie trainiert wurden und der Subword-Tokenisierung, die sie benutzen, als robuster gegenüber OCR-Fehlern erweisen sollten. Es fehlen jedoch noch Studien, die dies bestätigen bzw. den erreichbaren Mehrwert evaluieren.

Zusammenfassend kann festgehalten werden, dass trotz des großen Einflusses, den OCR-Fehler auf die Sprachverarbeitung haben, nur wenige Arbeiten versucht haben, diesen Einfluss systematisch zu evaluieren (van Strien et al. 2020, Smith und Cordell 2018). Denkbar wären z. B. mit Alex und Burns (2014) Empfehlungen für Qualitätsschwellenwerte von OCR, bei denen man noch zufriedenstellende Ergebnisse für weitere Prozessierungsschritte erwarten kann. Die Verbesserung und Einbeziehung sprachtechnologischer Methoden für historische Dokumente ist auch für Bibliotheken und Archive erstrebenswert, da Ergebnisse aus einer Informationsextraktion für die Erstellung von Metadaten und die automatische Inhaltserschließung genutzt werden können.

3.2 Digital Humanities

Die Digital-Humanities-Forschung nutzt computergestützte Tools und Methoden für Anwendungszwecke in den Geistes- und Kulturwissenschaften und erforscht zudem, welche Bedeutung und Konsequenzen digitale Werkzeuge für die Methoden und die Forschung in den Geisteswissenschaften haben.

Entsprechende Forschungsarbeiten können von digitalisierten Archiven insofern profitieren, weil diese leichter verfügbar sind und auf den Texten quantitative Methoden angewandt werden können, um Forschungsfragen zu bearbeiten. Damit OCR-Fehler jedoch nicht zu verfälschten Ergebnissen führen, sind Forscher von qualitativ hochwertigen Digitalisaten abhängig, bei denen gegebenenfalls auftretende Fehler auch transparent nachvollzogen werden können. Eine qualitative Studie, die mit Historikern durchgeführt wurde, zeigt, dass gerade dieser Aspekt in der Praxis ein großes Problem darstellt (Traub et al. 2015). Drei der vier Befragten gaben an, ihre quantitativen Studien auf Grundlage von digitalisierten Dokumenten nicht zu publizieren, weil sie potenziell nicht vertrauenswürdig seien und die Ergebnisse angezweifelt werden könnten. In einer quantitativen Studie von Hill und Hengchen (2019) wirkte sich die Qualität der OCR zwar nicht so signifikant auf das *Topic Modeling* aus, dafür aber auf andere statistische Verfahren, die in den Digital Humanities typischerweise benutzt werden wie Kollokationsanalyse und Stilometrie (insbesondere bei einer OCR-Genauigkeit unter 70 %–75 %).

In den von Traub et al. (2015) evaluierten digitalen Dokumenten waren die Konfidenz-Angaben für die OCR-Resultate in die ALTO-Metadaten integriert,

jedoch sei es nicht möglich gewesen herauszufinden, wie diese berechnet wurden. Hinzu kommt, dass eine Evaluierung der Nachkorrektur der Dokumente durch das OCR-Tool nicht möglich war, da die ursprünglichen Volltexte vor der Nachkorrektur nicht mehr vorhanden waren. Dies macht die Nachvollziehbarkeit der Ergebnisse unmöglich und erschwert die Arbeit derjenigen Zweige der Digital Humanities, die quantitative Methoden auf OCR-prozessierte Dokumente anwenden wollen.

Ein wichtiger Gegenstand der Digital Humanities sind Forschungsdaten, die im Zuge einer wissenschaftlichen Tätigkeit entstehen. Es geht dabei um die Aufbereitung, Verarbeitung und Verwaltung der Daten, auch Forschungsdatenmanagement genannt. Ziel ist es, die Daten langfristig für die Nachnutzung zugänglich zu machen und nachprüfbar zu halten, und das unabhängig von der Quelle der Daten. Die digitalen Objekte sind oft auch digitale Texte, weshalb viele Fragen, die in diesem Kapitel behandelt wurden, auch wichtig für die Digital Humanities sind. Erwähnt sei hier beispielhaft *NFDI4Culture*,[38] das mit verschiedenen universitären Einrichtungen und Kulturinstitutionen eine nationale Forschungsinfrastruktur für Kulturdaten vorantreiben will.

In Zukunft könnte die Frage nach der weiteren Entwicklung von Metadaten bzw. der Relevanz bestimmter Metadaten für die Digitalisate auch davon abhängig sein, welche Anforderungen innerhalb der Digital Humanities an digitale Archive ausgearbeitet werden (Smith und Cordell 2018). Diese Anforderungen betreffen Aspekte der Standardisierung von Metadaten sowie inhaltliche Aspekte, also die Überlegung, welche Metadaten dabei helfen können, Forschungsfragen in den Digital Humanities zu beantworten.

3.3 Information Retrieval

Betrachten wir hingegen Szenarien aus dem Bereich des Information Retrieval wie z. B. eine klassische Schlagwortsuche (Keyword-Suche), so spielt die Lesereihenfolge dafür kaum eine Rolle. Im Zuge der Indexierung werden Wörter typischerweise einzeln verarbeitet, zudem wird häufig noch eine Tokenisierung sowie teilweise auch *Stemming* durchgeführt (für Inhalte des Feldes *TextField* in Lucene ist z. B. per Default Tokenisierung vorgesehen). OCR-Fehler wirken sich daher vor allem auf das Ranking der Treffer aus (van Strien et al. 2020).

Bis zu einem gewissen Grad existieren auch in den gängigen Suchmaschinen bereits Funktionen, die die Auffindbarkeit fehlerhafter OCR-Daten verbessern können, wie z. B. *Fuzzy Search*. Damit können Wörter, die durch OCR-

38 Vgl. https://nfdi4culture.de/ (1.12.2020).

Fehler in einzelnen Zeichen von dem gesuchten Schlagwort abweichen, trotzdem gefunden werden. Hier sind Metriken für die OCR-Qualität auf Wortebene zumeist ausreichend, insbesondere BOW, *significant word error rate* und *flexible character accuracy measure* bieten sich an. Für eine Phrasensuche, die mehrere Wörter im Zusammenhang zu finden versucht, bestehen hingegen bereits höhere Anforderungen an OCR-Ergebnisse. Denkbar wäre hier z. B. die Verwendung von Metriken, die auf *n*-Gramme zurückgreifen.

Perspektivisch kann aber auch das klassische Information Retrieval von der Einbindung durch OCR erzeugter Merkmale und Metadaten profitieren, indem diese z. B. für alternative Browsing-Einstiege oder die Facettierung genutzt werden. Für historische Schreibvarianten kann z. B. *Query Expansion* (Ernst-Gerlach und Fuhr 2007, Traub et al. 2016) angewandt werden, um Nutzer:innen bei der Formulierung von Suchanfragen zusätzlich historische Schreibvarianten anzubieten.

4 Zusammenfassung & Ausblick

Blicken wir auf die vorangegangenen Betrachtungen zurück, so zeigt sich, dass bei der Bewertung von OCR-Qualität viele Dimensionen berücksichtigt werden müssen. Unterschiedliche Metriken haben unterschiedliche Perspektiven auf und Aussagekraft über verschiedene Aspekte der OCR-Qualität. Insbesondere der Einfluss der Layoutanalyse auf die OCR-Ergebnisse wird bislang durch die meisten gängigen Metriken nicht in ausreichendem Maße abgebildet. Dies hat damit zu tun, dass sich geeignete Konzepte und Standards noch nicht im benötigten Umfang etabliert bzw. durchgesetzt haben, weil sie eine komplexe Auseinandersetzung mit den vielschichtigen Qualitätsaspekten erfordern. OCR beinhaltet neben Texterkennung immer auch eine Layoutanalyse, also die Unterteilung des Dokuments in Abschnitte wie bspw. Text, Abbildungen und Tabellen, und definiert die Begrenzungen auf Pixel-Ebene (*Document Layout Analysis*). Ein weiterer Schritt ist das Identifizieren von logischen Bereichen von Dokumenten, bei dem die semantische Funktion von Textabschnitten ausgezeichnet wird, z. B. Titel, Einleitung, Haupttext oder Zitate. Die Erfassung bzw. (Re-)Konstruktion seitenübergreifender Strukturen wie Inhaltsverzeichnissen oder Registern sind weitere Beispiele. Dies wäre ein Schritt hin zu einem *Document Understanding System*, das eine umfassende automatische Informationsextraktion aus Dokumenten ermöglicht, die nicht nur auf Textebene arbeitet, sondern in der Prozessierung auch visuelle Informationen des Dokuments mit einbezieht.

Andererseits sind die Anforderungen an die OCR-Qualität je nach Anwendungsfall sehr unterschiedlich. Während die Qualität der Layoutanalyse für die Schlagwortsuche kaum eine Bedeutung hat, so ist sie für die semantische Verarbeitung der OCR-Resultate entscheidend. Einen Ausweg können für spezifische Anwendungsfälle individuell definierte Profile für die Evaluierung darstellen, die auf standardisierte und transparente Metriken zurückgreifen bzw. diese kombinieren. Um auch die Nachvollziehbarkeit der Ergebnisse zu gewährleisten, werden zusätzliche freie Referenzdatensätze mit GT sowie quelloffene und gut dokumentierte Implementierungen der Evaluierungsmethoden und Metriken benötigt, so dass die verschiedenen Communities sich auf eine gemeinsame Grundlage für optimale Verfahren verständigen können. Erste Datensätze und Methoden entstehen derzeit primär für stark konventionalisierte Textsorten, z. B. wissenschaftliche Artikel, bei denen Layout-Informationen bereits in XML oder im LaTeX-Format neben den gerenderten PDF-Dateien vorliegen und als GT herangezogen werden können (Zhong et al. 2019). Für andere Textsorten sowie insbesondere historische Dokumente besteht allerdings noch eine große Lücke.

Eine vielversprechende Perspektive stellen Verfahren für die Qualitätsvorhersage dar, die auf vergleichsweise kleinen, aber repräsentativ ausgewählten Stichproben, für die GT erstellt wird, mit Dokumentmerkmalen und relevanten Metadaten trainiert werden. So kann zumindest die Menge an benötigten GT-Daten für die Evaluierung deutlich reduziert werden, ohne damit die Qualitätsmessung auf zu unsichere Methoden zu stützen.

Für die Inhaltserschließung können somit durch die OCR-Evaluierung relevante Informationen zur Qualität der durch die OCR erstellten Texte gewonnen werden, um z. B. die automatisierte Verschlagwortung oder Indexierung zu unterstützen. Für eine weitergehende inhaltliche Erschließung, wie etwa die Anreicherung mit semantischen Informationen oder die Verknüpfung mit Wissensbasen müssen im Zuge der OCR-Evaluierung immer auch die Ergebnisse der Layoutanalyse Betrachtung finden, da nur so die Qualität der inhaltlichen Ebene adäquat bewertet werden kann.

Aber auch für die Metadatenanreicherung von Bibliotheksdaten ist eine Layoutanalyse mit der Auszeichnung von semantischen Funktionen von Abschnitten sinnvoll, da Informationen wie Titel, Autor:innen oder Abschnitte die Suche und Arbeit mit Digitalisaten erleichtern. Bereits jetzt können Qualitätsmerkmale und Metadaten aus dem OCR-Prozess für die Kataloganreicherung genutzt werden. Selbst die wenig verlässlichen und zudem schon vorliegenden OCR-Konfidenzen stellen für Nutzer:innen einen Mehrwert dar. Detaillierte Metadaten zur technischen Provenienz, wie der für die OCR verwendeten Software, Version sowie benutzter Modelle und Konfigurationsparameter erlauben

es, den Entstehungsprozess der in den Digital Humanities als Forschungsdaten verwendeten OCR-Daten transparent nachvollziehbar zu machen. Mittelfristig sind dabei auch entsprechende technische Konzepte für die granulare und persistente Zitierbarkeit und Versionierung von OCR-Ergebnissen zu berücksichtigen.

Zuletzt sei hier noch auf aktuelle Forschungsarbeiten zu einer hybriden Dokumenterkennung verwiesen. Während es für Menschen normal ist, Informationen aus Dokumenten auch anhand von Layout-Aspekten zu extrahieren (Größe als Hinweis auf Wichtigkeit eines Satzteiles, Einrückungen und Kursivsetzungen für Zitate etc.), wurde dieser Aspekt lange Zeit in der Forschung außen vorgelassen. Diese Zusatzinformationen können jedoch ein wichtiger Bestandteil für verschiedene Bereiche des NLP sein, wie die Erkennung relevanter Segmente für eine automatische Zusammenfassung von Texten oder für die Übersetzung von Text in *Leichte Sprache*. Inzwischen gibt es mehrere Methoden, die einen hybriden Ansatz für die Dokumenterkennung verfolgen. Einerseits werden dabei mathematische Abbildungen von Textmerkmalen (sogenannte *Text Embeddings*), wie sie in der NLP-Forschung genutzt werden, und andererseits Abbildungen auf Pixel-Ebene, wie sie im Bereich der *Computer Vision* genutzt werden, dazu verwendet, hybride Modelle zu trainieren. Erste vielversprechende Ergebnisse sieht man in Xu et al. 2019 und Garncarek et al. 2020. Umgekehrt können OCR und Layoutanalyse von sprachwissenschaftlichen Methoden und Modellen profitieren. Ein Beispiel dafür stellt die Artikelsegmentierung und Überprüfung sowie ggf. Korrektur der im Zuge der Layoutanalyse ermittelten Lesereihenfolge mit multimodalen Modellen dar (Barman et al. 2020). Auch die großen Technologieunternehmen (z. B. Microsoft *OneOCR*,[39] Google *Cloud Vision OCR*,[40] Baidu *PaddlePaddle*[41]) setzen schon seit einigen Jahren verstärkt auf *End-to-End*-Systeme für die Dokumenterkennung. Diese Entwicklungen gilt es aufmerksam zu beobachten und ggf. erzielte Fortschritte auf den Bereich der Digitalisierung historischer Dokumente und Kulturdaten zu übertragen.

5 Danksagung

Dieser Beitrag wurde im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projektes QURATOR (Unternehmen Region,

39 Vgl. https://icdar2019.org/keynote-speakers/ (1.12.2020).

40 Vgl. Ashok Popat: OCR for Most of the World's Languages. 3. September 2015. https://ewh.ieee.org/r6/scv/sps/20150903AshokPopat.pdf (1.12.2020).

41 Vgl. https://github.com/PaddlePaddle/PaddleOCR (1.12.2020).

Wachstumskern, Projektnr. 03WKDA1A) und des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projektes SoNAR (IDH) (Projektnr. 414792379) erstellt.

6 Literaturverzeichnis

Alex, Beatrice und John Burns: Estimating and rating the quality of optically character recognised text. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (2014), S. 97–102. https://doi.org/10.1145/2595188.2595214.

Baierer, Konstantin und Philipp Zumstein: Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. In: 0.27 Zeitschrift für Bibliothekskultur (2016) Bd.4 Nr. 2. S. 72–83. https://doi.org/10.12685/027.7-4-2-155.

Banerjee, Satanjeev und Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. Hrsg. v. Jade Goldstein, Alon Lavie, Chin-Yew Lin, Clare Voss. Ann Arbor, Michigan: Association for Computational Linguistics 2005. S. 65–72. https://www.aclweb.org/anthology/W05-0909 (1.12.2020).

Barman, Raphaël, Maud Ehrmann, Simon Clematide, Sofia Ares Oliveira und Frédéric Kaplan: Combining visual and textual features for semantic segmentation of historical newspapers. arXiv preprint arXiv:2002.06144. (2020). https://arxiv.org/abs/2002.06144 (1.12.2020).

Boenig, Matthias, Konstantin Baierer, Volker Hartmann, Maria Federbusch und Clemens Neudecker: Labelling OCR Ground Truth for Usage in Repositories. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019). New York, NY: Association for Computing Machinery 2019. S. 3–8. https://doi.org/10.1145/3322905.3322916.

Boenig, Matthias, Maria Federbusch, Elisa Herrmann, Clemens Neudecker und Kay-Michael Würzner: Ground Truth: Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities? In: Konferenzabstracts, Digital Humanities im deutschsprachigen Raum (2018). Hrsg. v. Georg Vogeler. S. 219–223. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf#page=221 (1.12.2020).

Burchardt, Aljoscha, Kim Harris, Georg Rehm und Hans Uszkoreit: Towards a systematic and human-informed paradigm for high-quality machine translation. In: Proceedings of the LREC 2016 Workshop – Translation evaluation: From fragmented tools and data sets to an integrated ecosystem. Hrsg. v. Georg Rehm, Aljoscha Burchardt, Ondrej Bojar, Christian Dugast, Marcello Federico, Josef van Genabith, Barry Haddow, Jan Hajič, Kim Harris, Philipp Köhn, Matteo Negri, Martin Popel, Lucia Specia, Marco Turchi und Hans Uszkoreit. 2016. S. 35–42. http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-MT%20Evaluation_Proceedings.pdf#page=45 (1.12.2020).

Clausner, Christian, Stefan Pletschacher und Apostolos Antonacopoulos: Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods. In: Proceedings of the 11th International Conference on Document Analysis and Recognition. 2011. S. 1404–1408. https://doi.org/10.1109/ICDAR.2011.282.

Clausner, Christian, Stefan Pletschacher und Apostolos Antonacopoulos: The Significance of Reading Order in Document Recognition and its Evaluation. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. 2013. S. 688–692. https://doi.org/10.1109/ICDAR.2013.141.

Clausner, Christian, Stefan Pletschacher und Apostolos Antonacopoulos: Flexible character accuracy measure for reading-order-independent evaluation. In: Pattern Recognition Letters (2020) Bd. 131. S. 390–397. https://doi.org/10.1016/j.patrec.2020.02.003.

Clausner, Christian, Stefan Pletschacher und Apostolos Antonacopoulos: Quality Prediction System for Large-Scale Digitisation Workflows. In: Proceedings of the 12th IAPR International Workshop on Document Analysis Systems. 2016. https://doi.org/10.1109/DAS.2016.82.

Crane, Gregory und Alison Jones: The challenge of Virginia Banks: an evaluation of named entity analysis in a 19th-century newspaper collection. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (JCDL'06). New York, NY: Association for Computing Machinery 2006. S. 31–40. https://doi.org/10.1145/1141753.1141759.

Ehrmann, Maud, Matteo Romanello, Alex Flückiger und Simon Clematide: Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers. In: Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum. Hrsg. v. Linda Cappellato, Carsten Eickhoff, Nicola Ferro und Aurélie Névéol. 2020. CEUR Bd. 2696. http://ceur-ws.org/Vol-2696/paper_255.pdf (1.12.2020).

Engl, Elisabeth, Matthias Boenig, Konstantin Baierer, Clemens Neudecker und Volker Hartmann: Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke. In: Zeitschrift für Historische Forschung (2020) Bd. 47 H. 2. S. 223–250. https://doi.org/10.3790/zhf.47.2.223.

Ernst-Gerlach, Andrea und Norbert Fuhr: Retrieval in text collections with historic spelling using linguistic and spelling variants. In: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL'07). New York, NY: Association for Computing Machinery 2007. S. 333–341. https://doi.org/10.1145/1255175.1255242.

Federbusch, Maria, Christian Polzin und Thomas Stäcker: Volltext via OCR. Möglichkeiten und Grenzen. In: Beiträge aus der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (2013) Bd. 43. https://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/historische_drucke/pdf/SBB_OCR_STUDIE_WEBVERSION_Final.pdf (1.12.2020).

Fink, Florian, Klaus U. Schulz und Uwe Springmann: Profiling of OCR'ed Historical Texts Revisited. In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2017). New York, NY: Association for Computing Machinery 2017. S. 61–66. https://doi.org/10.1145/3078081.3078096.

Garncarek, Łukasz, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama und Filip Graliński: LAMBERT: Layout-Aware language Modeling using BERT for information extraction. arXiv preprint arXiv:2002.08087. (2020). https://arxiv.org/abs/2002.08087 (1.12.2020).

Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas und Frank Wiegand: TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv. In: Jahrbuch für Computerphilologie (2012). http://computerphilologie.digital-humanities.de/jg09/geykenetal.pdf (1.12.2020).

Gupta, Anshul, Ricardo Gutierrez-Osuna, Matthew Christy, Boris Capitanu, Loretta Auvil, Liz Grumbach, Richard Furuta, und Laura Mandell: Automatic assessment of OCR quality in

historical documents. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15). AAAI Press 2015. S. 1735–1741. https://psi.engr.tamu.edu/wp-content/uploads/2018/01/gupta2015aaai.pdf (1.12.2020).

Hamdi, Ahmed, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty und Antoine Doucet: An analysis of the performance of named entity recognition over OCRed documents. In: Proceedings. 2019 ACM/IEEE Joint Conference on Digital Libraries. 2019. S. 333–334. https://doi.org/10.1109/JCDL.2019.00057.

Hill, Mark J. und Simon Hengchen: Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. In: Digital Scholarship in the Humanities (2019) Bd. 34 H. 4. S. 825–843. https://doi.org/10.1093/llc/fqz024.

Holley, Rose: How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. In: D-Lib Magazine (2009). Bd. 15 Nr. 3/4. http://www.dlib.org/dlib/march09/holley/03holley.html (1.12.2020).

Jurish, Bryan und Henriette Ast: Using an alignment-based lexicon for canonicalization of historical text. In: Historical Corpora. Challenges and Perspectives. Hrsg. v. Jost Gippert und Ralf Gehrke. (2015), S. 197–208.

Kettunen, Kimmo und Tuula Pääkkönen: Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association 2016. S. 956–961. https://www.aclweb.org/anthology/L16-1152/ (1.12.2020).

Kettunen, Kimmo, Eetu Mäkelä, Teemu Ruokolainen, Juha Kuokkala und Laura Löfberg: Old Content and Modern Tools-Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. In: Digital Humanities Quarterly (2017) Bd. 11 Nr. 3. http://www.digitalhumanities.org/dhq/vol/11/3/000333/000333.html (1.12.2020).

Kluzner, Vladimir, Asaf Tzadok, Yuval Shimony, Eugene Walach und Apostolos Antonacopoulos: Word-based adaptive OCR for historical books. In: 10th International Conference on Document Analysis and Recognition. IEEE 2009. S. 501–505. https://doi.org/10.1109/ICDAR.2009.133.

Levenshtein, Vladimir I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics–doklady (1966) Bd. 10 Nr. 8. S. 707–710. https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf (1.12.2020).

Lin, Chin-Yew: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. Association for Computational Linguistics 2004. S. 74–81. https://www.aclweb.org/anthology/W04-1013/ (1.12.2020).

Lui, Marco und Timothy Baldwin: langid.py: An off-the-shelf language identification tool. In: Proceedings of the ACL 2012 system demonstrations. Association for Computational Linguistics 2012. S. 25–30. https://www.aclweb.org/anthology/P12-3005/ (1.12.2020).

Mac Kim, Sunghwan und Steve Cassidy: Finding names in Trove: named entity recognition for Australian historical newspapers. In: Proceedings of the Australasian Language Technology Association Workshop 2015. 2015. S. 57–65. https://www.aclweb.org/anthology/U15-1007/ (1.12.2020).

Mieskes, Margot und Stefan Schmunk: OCR Quality and NLP Preprocessing. In: Proceedings of the Workshop on Widening NLP 2019. 2019. S. 102–105. https://www.winlp.org/wp-content/uploads/2019/final_papers/176_Paper.pdf (1.12.2020).

Neudecker, Clemens, Lotte Wilms, Willem Jan Faber und Theo van Veen: Large-scale refinement of digital historic newspapers with named entity recognition. In: Proceedings of the IFLA

Newspapers/GENLOC Pre-Conference Satellite Meeting 2014. 2014. https://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-neudecker_faber_wilms-en.pdf (1.12.2020).

Neudecker, Clemens, Konstantin Baierer, Maria Federbusch, Kay-Michael Würzner, Matthias Boenig, Elisa Hermann und Volker Hartmann: OCR-D: An end-to-end open-source OCR framework for historical documents. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019). New York NY: Association for Computing Machinery 2019. S. 53–58. https://doi.org/10.1145/3322905.3322917.

Neudecker, Clemens und Apostolos Antonacopoulos: Making Europe's Historical Newspapers Searchable. In: 2016 12th IAPR Workshop on Document Analysis Systems. IEEE 2016. S. 405–410. https://doi.org/10.1109/DAS.2016.83.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke und Stewart Varner: Final Report – Always Already Computational: Collections as Data. 2019. http://doi.org/10.5281/zenodo.3152935.

Papineni, Kishore, Salim Roukos, Todd Ward und Wei-Jing Zhu: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL'02). Association for Computational Linguistics 2002. S. 311–318. https://doi.org/10.3115/1073083.1073135.

Pletschacher, Stefan und Apostolos Antonacopoulos: The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. In: Proceedings of the 20th International Conference on Pattern Recognition 2010. IEEE 2010. S. 257–260. https://doi.org/10.1109/ICPR.2010.72.

Pletschacher, Stefan, Christian Clausner und Apostolos Antonacopoulos: Europeana Newspapers OCR Workflow Evaluation. In: Proceedings of the 4th Workshop on Historical Document Imaging and Processing (HIP'15). New York, NY: Association for Computing Machinery 2015. S. 39–46. https://doi.org/10.1145/2809544.2809554.

Pontes, Elvys Linhares, Ahmed Hamdi, Nicolas Sidere und Antoine Doucet: Impact of OCR Quality on Named Entity Linking. In: Digital Libraries at the Crossroads of Digital Information for the Future. 21st International Conference on Asia-Pacific Digital Libraries (ICADL 2019). Cham: Springer 2019. S. 102–115. https://doi.org/10.1007/978-3-030-34058-2_11.

Reffle, Ulrich und Christoph Ringlstetter: Unsupervised Profiling of OCRed Historical Documents. In: Pattern Recognition (2013) Bd. 46, H. 5. S. 1346–1357. https://doi.org/10.1016/j.patcog.2012.10.002.

Rehm, Georg, Peter Bourgonje, Stefanie Hegele, Florian Kintzel, Julián Moreno Schneider, Malte Ostendorff, Karolina Zaczynska, Armin Berger, Stefan Grill, Sören Räuchle, Jens Rauenbusch, Lisa Rutenburg, André Schmidt, Mikka Wild, Henry Hoffmann, Julian Fink, Sarah Schulz, Jurica Seva, Joachim Quantz, Joachim Böttger, Josefine Matthey, Rolf Fricke, Jan Thomsen, Adrian Paschke, Jamal Al Qundus, Thomas Hoppe, Naouel Karam, Frauke Weichhardt, Christian Fillies, Clemens Neudecker, Mike Gerber, Kai Labusch, Vahid Rezanezhad, Robin Schaefer, David Zellhöfer, Daniel Siewert, Patrick Bunk, Lydia Pintscher, Elena Aleynikova und Franziska Heine: QURATOR: Innovative Technologies for Content and Data Curation. In: QURATOR 2020 – Conference on Digital Curation Technologies. Proceedings of the Conference on Digital Curation Technologies, Berlin 2020. Hrsg. v. Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus und Lydia Pintscher. CEUR Bd. 2535. http://ceur-ws.org/Vol-2535/paper_17.pdf (1.12.2020).

Rice, Stephen V: Measuring the Accuracy of Page-Reading Systems. UNLV Retrospective Theses & Dissertations, 3014. Las Vegas: University of Nevada 1996. https://doi.org/10.25669/hfa8-0cqv.

Rice, Stephen V. und Thomas A. Nartker: The ISRI analytic tools for OCR evaluation. In: UNLV/Information Science Research Institute (1996), TR-96-02. Version 5.1. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.9427&rep=rep1&type=pdf (4.1.2021).

Santos, Eddie Antonio: OCR evaluation tools for the 21st century. In: Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages. Bd. 1: Papers. Association for Computational Linguistics 2019. S. 23–27. https://www.aclweb.org/anthology/W19-6004/ (1.12.2020).

Schlarb, Sven und Clemens Neudecker: A heuristic measure for detecting influence of lossy JP2 compression on Optical Character Recognition in the absence of ground truth. In: Proceedings of the Archiving Conference 2012. Society for Imaging Science and Technology 2012. S. 250–254. https://www.ingentaconnect.com/contentone/ist/ac/2012/00002012/00000001/art00055 (1.12.2020).

Singh, Prashant, Ekta Vats und Anders Hast: Learning surrogate models of document image quality metrics for automated document image processing. In: 13th IAPR International Workshop on Document Analysis Systems 2018. IEEE 2018. S. 67–72. https://doi.org/10.1109/DAS.2018.14.

Smith, David und Ryan Cordell: A Research Agenda for Historical and Multilingual Optical Character Recognition. Final report and supporting materials for a 2017–2018 project supported by the Andrew W. Mellon Foundation. 2018. http://hdl.handle.net/2047/D20296774 (1.12.2020).

Smith, Ray: Limits on the application of frequency-based language models to OCR. In: Proceedings of the International Conference on Document Analysis and Recognition 2011. IEEE 2011. S. 538–542. https://research.google/pubs/pub36984.pdf (1.12.2020).

Springmann, Uwe, Florian Fink und Klaus U. Schulz: Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings. arXiv preprint arXiv:1606.05157. (2016). https://arxiv.org/abs/1606.05157 (1.12.2020).

Stollwerk, Christoph: Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR Software im Bereich „Alter Drucke“ zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930. In: DARIAH-DE Working papers (2016) Nr. 16. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2016-2-8 (1.12.2020).

Tanner, Simon, Trevor Muñoz und Pich Hemy Ros: Measuring mass text digitization quality and usefulness. In: D-lib Magazine (2009) Bd. 15, Nr. 7/8. http://www.dlib.org/dlib/july09/munoz/07munoz.html (1.12.2020).

Traub, Myriam C., Jacco Van Ossenbruggen und Lynda Hardman: Impact analysis of OCR quality on research tasks in digital archives. In: Research and Advanced Technology for Digital Libraries. 19th International Conference on Theory and Practice of Digital Libraries (TPDL 2015). Cham: Springer 2015. S. 252–263. https://doi.org/10.1007/978-3-319-24592-8_19.

Traub, Myriam C., Thaer Samar, Jacco van Ossenbruggen, Jiyin He, Arjen de Vries und Lynda Hardman: Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL'16). New York, NY: Association for Computing Machinery 2016. S. 7–16. https://doi.org/10.1145/2910896.2910907.

Ukkonen, Esko: On-line construction of suffix trees. In: Algorithmica (2015) Bd. 14 Nr. 3. S. 249–260. https://doi.org/10.1007/BF01206331.

van Strien, Daniel, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray und Giovanni Colavizza: Assessing the Impact of OCR Quality on Downstream NLP Tasks.

In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020). Hrsg. v. Ana Rocha, Luc Steels und Jaap van den Herik. Bd.1. S. 484–496. https://doi.org/10.5220/0009169004840496.

Wernersson, Maria: Evaluation von automatisch erzeugten OCR-Daten am Beispiel der Allgemeinen Zeitung. In: ABI Technik (2015) Bd. 35 Nr. 1. S. 23–35. https://doi.org/10.1515/abitech-2015-0014.

Xu, Yiheng, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei und Ming Zhou: LayoutLM: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'20). New York, NY: Association for Computing Machinery 2020. S. 1192–1200. https://doi.org/10.1145/3394486.3403172.

Zhong, Xu, Jianbin Tang und Antonio Jimeno Yepes: PubLayNet: largest dataset ever for document layout analysis. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR 2019). IEEE 2019. S. 1015–1022. https://doi.org/10.1109/ICDAR.2019.00166.