Perspektive Bibliothek 7.1 (2018), S. 33-54 doi: 10.11588/pb.2018.1.48394

AUTOMATISIERTE VOLLTEXTERSCHLIEßUNG VON RETRODIGITALISATEN AM BEISPIEL HISTORISCHER ZEITUNGEN

Anna Kugler
Bayerische Staatsbibliothek[1] / Bibliotheksakademie Bayern
annakugler@gmx.net

1. Einleitung

„Content is King. Context is Queen." So lautet der Leitsatz des Generaldirektors der Bayerischen Staatsbibliothek Klaus Ceynowa, wie er in einem seiner ersten Interviews nach der Amtsübernahme für B.I.T. Online aussagte. Weiter führte er aus, wer über „Content" verfüge, der ein gewisses Alleinstellungsmerkmal aufweist, und in der Lage ist, darauf „intelligente Services aufzusetzen", der habe eine Zukunft.[2] Für eine Kontextualisierung und damit Weiterverarbeitung dieses contents mit Hilfe von Text- und Datamining oder auch semantischer Analysen ist der maschinenlesbare Volltext der digitalisierten Werke notwendig, der in weitestgehend automatisierter Weise gewonnen werden sollte.

Die Deutsche Forschungsgemeinschaft stellt gleichermaßen in ihren Praxisregeln von 2013 fest, dass eine wissenschaftliche Nachnutzung des digitalen Contents auf dem „Angebot des digitalen Volltextes" basiert: „Wo immer möglich und sinnvoll, wird die DFG-Förderung daher im textuellen Bereich auf die Bereitstellung digitaler Volltexte abzielen. [...] Für Druckwerke ab Erscheinungsjahr 1850 gilt ver-

[1] Dieser Text entstand bereits 2016 im Rahmen des Bibliotheksreferendariats an der Bibliotheksakademie München und kann nun erst veröffentlicht werden.
[2] Lübbers (2015, S. 182).

pflichtend, dass Volltext hergestellt werden muss und eine bloße Bilddigitalisierung nicht ausreicht."[3]

Um ein besseres Verständnis dafür zu erlangen, was hinter dem Schlagwort „Volltext" steht, möchte dieser Artikel einen kleinen Einblick in die technischen Verfahren zur automatisierten Volltexterschließung von Retrodigitalisaten liefern. Es soll aufgezeigt werden, welche Fortschritte, aber auch Grenzen aktuell bestehen und wie Qualität in diesem Zusammenhang überhaupt bemessen bzw. verglichen werden kann. Wenn hier von Volltexterschließung die Rede ist, ist die Binnenerschließung der digitalisierten Dokumente gemeint, also die optische Erkennung sowohl des Textes als auch des Layouts, aber nicht ihre bibliothekarische Erschließung. Der gesamte Bereich der Ausgabeformate bzw. Datenformate (zum Beispiel ALTO oder hOCR) sowie die Fragen der Archivierung von Volltexten sind aufgrund des begrenzten Umfangs nicht Gegenstand dieser Arbeit.

Die automatisierten Verfahren zur Volltexterschließung werden aus verschiedenen Gründen am Beispiel historischer Zeitungen vorgestellt. Erstens ist bei diesem Forschungsmedium das Desiderat nach einer besseren Zugänglichkeit sehr hoch. Das Medium der Zeitung hat für viele historische Disziplinen in den letzten Jahrzehnten an Quellenwert gewonnen. Besonders der Presse des 17. und 18. Jahrhunderts kommt ein großer Quellenwert zu, da die „periodisch erschienenen Schriften die wichtigsten Medien der Aufklärung darstellen", womit „die herausragende Bedeutung der Zeitungen für die Entstehung einer politischen Öffentlichkeit in Deutschland" inzwischen als unumstritten gilt.[4] Zugleich ist jedoch die Zugänglichkeit dieser Quellengattung aufgrund der Fragilität der Originale zunehmend schwierig und eine digitale Bereitstellung dringend notwendig.[5]

Zweitens bieten Zeitungen aufgrund ihrer Struktur und der Anforderung einer Artikelseparierung als Mehrwert für den Forschenden einige besondere Herausforderungen bei der Volltexterschließung, die hier aufgezeigt werden sollen. Genau

---

[3] Deutsche Forschungsgemeinschaft (2013, S. 30).

[4] Böning (2013, S. 36); siehe auch Meier (2013), der ein großes Forschungsdesiderat in der Beschäftigung mit der Sprache und Geschichte deutschsprachiger Zeitungen im Bereich der Sprach- und Literaturwissenschaft sieht und deshalb eine digitale Bereitstellung deutscher historischer Zeitungen an zentraler Stelle sehr begrüßen würde.

[5] DFG-Rahmenantrag zur Digitalisierung historischer Zeitungen (2012, S. 2). Es handelt sich dabei um den Rahmenantrag für die Zeitungsdigitalisierung in Deutschland, ergänzt durch vier Einzelanträge der teilnehmenden Bibliotheken. Für diese Arbeit lag nur der Rahmenantrag vor.

zu diesem Themenbereich wurde Anfang 2016 das DFG-Projekt zur Erstellung eines „Masterplan Zeitungsdigitalisierung“ abgeschlossen. Die Aktualität der Ergebnisse sowie der Empfehlungen des DFG-Projekts, die für diesen Beitrag bereits in unveröffentlichter Form vorlagen, waren ein dritter, ausschlaggebender Grund für die Fokussierung auf digitalisierte Zeitungen.[6]

2. DFG-Projekt: Masterplan Zeitungsdigitalisierung

In zahlreichen Ländern bestehen bereits renommierte, nationale Zeitungsportale, wie etwa in Österreich das ANNO-Projekt der Österreichischen Nationalbibliothek[7], „The British Newspaper Archive“ der British Library (BL)[8] oder auch das „TROVE-Portal“ der Australischen Nationalbibliothek[9], wohingegen in Deutschland bisher nur regional geprägte Digitalisierungsprojekte ohne nationale Koordinierung existieren.[10] Dieser Nachholbedarf war Anlass dafür, dass nach einigen Vorarbeiten 2012 bei der DFG ein Rahmenantrag zur „Digitalisierung historischer Zeitungen“ eingereicht wurde.[11] Beteiligt waren die Deutsche Nationalbibliothek (DNB), die Sächsische Landesbibliothek – Staats- und Universitätsbibliothek (SLUB), die Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB), die Staats- und Universitätsbibliothek Bremen (SuUB), die Universitäts- und Landesbibliothek Sachsen-Anhalt (ULB), sowie die Bayerische Staatsbibliothek (BSB). Die Arbeitspakete des zweijährigen Projekts umfassten eine Verbesserung der Nachweis- und Präsentationsstrukturen (vor allem in der Zeitschriftendatenbank) sowie die Zugänglichmachung ausgewählter Zeitungen als Images und zum Teil als Volltexte, indem unterschiedliche Digitalisierungswege (vom Original, vom Film, u.a.) und Erschließungstiefen (von einer reinen Imagedigitalisierung bis zur Volltextgenerierung und Artikelerschließung) erprobt wurden. Weitere Erschließungstiefen mit einer Normdatenverknüpfung („Named Entity Recognition“, NER) und einer Sach- und Bilderschließung konnten im Rah-

---

[6] Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland (Masterplan Zeitungsdigitalisierung), 2016.
[7] ANNO – AustriaN Newspapers Online: http://anno.onb.ac.at.
[8] The British Newspaper Archive: http://www.britishnewspaperarchive.co.uk.
[9] TROVE – Digitised Newspapers and more: https://trove.nla.gov.au/newspaper.
[10] Siehe zu diesen und weiteren internationalen Projekten sowie zu den verschiedenen in Deutschland laufenden Zeitungsdigitalisierungsprojekten: Masterplan Zeitungsdigitalisierung (2016, S. 6-10).
[11] Siehe Fußnote 4.

men des Projekts nicht getestet werden, sollen aber in einer Hauptphase berücksichtigt werden.[12]

Für die Digitalisierung wählten die Pilotbibliotheken unterschiedliche Zeitungsbestände des 17. bis frühen 20. Jahrhunderts aus: die SuUB digitalisierte ihren einmaligen, vollständigen Bestand deutschsprachiger Zeitungen des 17. Jahrhunderts, die BSB wählte die Allgemeine Zeitung/Cotta'sche Zeitung (19. Jahrhundert) sowie den Illustrierten Sonntag/Der gerade Weg (20. Jahrhundert), an der SLUB wurde die „Leipziger Volkszeitung“ (20. Jahrhundert) digitalisiert und die ULB übernahm das „Hallische Tageblatt“ (17./18. Jahrhundert).[13] Insgesamt wurden 448.000 Seiten vom Original und kleinere Mengen vom Mikrofilm mit OCR bearbeitet.[14] Nur die Ergebnisse der getesteten Volltexterschließungsverfahren, die in verschiedenen Veröffentlichungen ausgewertet wurden, spielen für den vorliegenden Artikel eine Rolle, die Ergebnisse der anderen Arbeitspakete werden außen vor gelassen.[15]

Ziel des Projekts war die Erstellung eines Masterplans mit Mindeststandards bzw. Best Practice-Empfehlungen zur künftigen Zeitungsdigitalisierung in Deutschland.[16] Miteinbezogen in die Zusammenstellung des Masterplans wurden auch die Ergebnisse aus dem von der SBB koordinierten European Newspaper Projects (ENP)[17], das nach einer dreijährigen Laufzeit ebenfalls letztes Jahr abgeschlossen wurde. Deshalb werden die zentralen Auswertungen auch hier genannt.[18]

3. Technischer Prozess der Volltexterschließung

Im Folgenden wird der gesamte Prozess der Volltexterschließung von Retrodigitalisaten vorgestellt, der in verschiedenen Schritten abläuft. Nach der Materialauswahl erfolgt das Pre-Processing der gescannten Seiten, gefolgt von der graphischen Seg-

[12] Masterplan Zeitungsdigitalisierung (2016, S. 36).
[13] Masterplan Zeitungsdigitalisierung (2016, S. 21).
[14] Masterplan Zeitungsdigitalisierung (2016, S. 31).
[15] Folgende Veröffentlichungen zu den Ergebnissen der einzelnen Digitalisierungsprojekte lagen bis Februar 2016 vor: Müller & Hermes (2014a und 2014b) für die SuUB, Sommer et al. (2014) für die ULB, Wernersson (2015) für die BSB. Da von der SLUB keine Veröffentlichung vorliegt, können die Ergebnisse bzgl. einer Volltexterschließung nur insofern rezipiert werden als diese im Masterplan genannt werden. Die Ergebnisse der SBB wurden vor allem im Rahmen des ENP evaluiert und publiziert (siehe Fußnote 17).
[16] DFG-Rahmenantrag (2012, S. 2-3).
[17] Europeana Newspapers-Projekt: http://www.europeana-newspapers.eu.
[18] Im Rahmen des ENP-Projekts entstanden zahlreiche Veröffentlichungen, wovon hier vor allem die Arbeiten von Pletschacher et al. (2014 und 2015) sowie von Clausner et al. (2015) rezipiert wurden.

mentierung in einzelne Text- bzw. Zeichenblöcke. Erst danach wird die eigentliche Zeichen- bzw. Worterkennung (Optical Character Recognition, OCR) durchgeführt, die oft missverständlich als Bezeichnung des gesamten Verfahrens gebraucht wird. Eine Kombination der Ergebnisse aus der Segmentierung und der OCR sind schließlich nötig, um eine Strukturerschließung der einzelnen Seite zu erzielen. Gerade bei Zeitungen spielt die Optical Layout Recognition (OLR) eine entscheidende Rolle, da das Ziel hier eine Separierung der einzelnen Artikel ist. Als Software verwendeten alle Projektpartner im Zeitungsdigitalisierungsprojekt die Software ABBYY mit dem FineReader, die den Gesamtablaufprozess abbildet.[19]

3.1 Scan-Vorlage

Die Entscheidung für oder gegen eine OCR-Erkennung beginnt bereits bei der Sichtung des Bestandes, denn verschiedene druckextrinsische Faktoren (zum Beispiel physische Schäden wie Löcher oder Flecken im Druckbild) sowie druckintrinsische Faktoren (zum Beispiel ein verrutschter Druckspiegel oder auch durchscheinende Textseiten) verhindern eine gute Scan-Qualität.[20] Die Qualität des Digitalisats ist jedoch von entscheidender Bedeutung für die weitere OCR-Verarbeitung.[21]

Generell wird für eine automatisierte Texterkennung eine Digitalisierung vom Original bevorzugt. In manchen Fällen existiert bereits ein Mikrofilm, der kostensparend und mit zum Teil sehr guter Qualität digitalisiert werden kann. Jedoch lässt sich das nicht ohne weiteres generalisieren, da bei Mikrofilm nicht nur dieselben Probleme, zum Beispiel mit Buchwölbungen und verzerrten Seiten auftreten können, sondern auch die Qualität der Mikroverfilmung nicht immer gleich gut ist. Außerdem liegt ein Mikrofilm immer in schwarz-weiß vor, womit viele Bildoptimierungsmöglichkeiten im Pre-Processing nicht möglich sind.

---

[19] An der BSB und an der ULB wurde mit ABBYY Version 10 für Frakturschriften gearbeitet, im ENP kam bereits Version 11 zum Einsatz (siehe Masterplan Zeitungsdigitalisierung, S. 37-38). ABBYY wird in der Literatur im Allgemeinen als die „einzige zuverlässige OCR Software für Altbestände“ beschrieben (Kämmerer 2009, S. 633), da sich die Firma mittlerweile seit zehn Jahren auch um historische Schriften und Dokumente bemüht und „mit der Frakturerkennung ein Alleinstellungsmerkmal“ besetzt (Mühlberger 2011, S. 13). Weitere kommerzielle und Open Source-Softwarelösungen sind zum Beispiel OmniPage, Readiris, BIT Alpha oder Tesseract und OCRopus (siehe Kämmerer 2009, S. 626, Mühlberger 2011, S. 11).

[20] Kämmer (2009, S. 637-638).

[21] Mühlberger (2011, S. 12).

Die Erkennungsrate profitiert durchaus von der Binarisierung (siehe Kapitel 3.2), weshalb eine Digitalisierung in Graustufe oder Farbe bevorzugt wird.[22] Vorzugsweise sollten die Images mit einer Auflösung von 300 dpi vorliegen. Eine höhere Auflösung mit zum Beispiel 400 bis 600 dpi führt nicht automatisch zu besseren Ergebnissen, kann sich aber positiv auswirken, wenn vermehrt kleinere Schriften verwendet wurden, wie etwa bei Zeitungen.[23] Bezogen auf die Komprimierung der Images haben Tests ergeben, dass zwischen TIFF unkomprimiert und JPEG mit 90% Komprimierungsfaktor kein signifikanter Unterschied besteht.[24] Die Erfahrungen aus dem Zeitungsdigitalisierungsprojekt bestätigen die in der Fachliteratur genannten Empfehlungen. Bereits in der Antragstellung war man sich darüber einig, dass „die Entscheidung für eine Digitalisierung vom Original oder vom Film in Abhängigkeit von der Qualität des Originals bzw. des Films, der konservatorischen Rahmenbedingungen und der Kosten pro Zeitung zu treffen sind".[25] Dementsprechend verfuhren die Pilotbibliotheken bei der Digitalisierung unterschiedlich: Die SuUB entschied sich für Rückvergrößerungen vom Mikrofilm, die anschließend gescannt wurden. Eine weitere Texterschließung kam aufgrund der „komplexen Materialität" (uneinheitliches Schriftbild und Layout, keine standardisierten Schrifttypen etc.) nicht in Frage.[26] Die ULB, die SLUB sowie die BSB digitalisierten vom Original, die SLUB entschied sich aufgrund der Beschaffenheit der Originalvorlagen für bitonale Scans vom Mikrofilm.[27]

Laut DFG-Antrag sollten alle Projektpartner entsprechend den DFG-Praxisregeln[28] mit einer Auflösung von 300 dpi in Graustufe und in TIFF unkomprimiert scannen.[29] Die ULB entschied sich jedoch für eine Digitalisierung in Farbe und erstellte für Testzwecke zusätzliche Scans mit einer Auflösung in 400 dpi, mit dem Ergebnis, dass die erreichte Qualität bei 300 dpi für eine automatische Texterkennung ausreiche und Speicherplatz einspare.[30] Abgeleitet aus ihren Erfahrungen

[22] Fuchs (2013, S. 198), Mühlberger (2011, S. 12).
[23] Mühlberger (2011, S. 12).
[24] Mühlberger (2011, S. 13).
[25] DFG-Rahmenantrag (2012, S. 11).
[26] Müller & Hermes (2014b, S. 274).
[27] Masterplan Zeitungsdigitalisierung (2016, S. 26).
[28] Deutsche Forschungsgemeinschaft (2013).
[29] DFG-Rahmenantrag (2012, S. 12).
[30] Sommer et al. (2014, S. 78), Masterplan Zeitungsdigitalisierung (2016, S. 31).

empfehlen die Projektpartner, dass besondere Bestände vom Original in Farbe oder Graustufe mit 300 dpi gescannt werden sollten, da dadurch die beste Qualität für die OCR-Bearbeitung gegeben sei. Digitalisierungen vom Mikrofilm sind preisgünstiger und sind dann zu empfehlen, wenn die Filme in sehr guter Qualität vorliegen. Von Reproformen (siehe SuUB) sollte aufgrund der Qualitätseinbußen nur in begründeten Ausnahmen digitalisiert werden.[31]

3.2 Pre-Processing

Nach der Digitalisierung müssen die Images für die Text- und Layouterkennung gegebenenfalls einer weiteren Vorverarbeitung bzw. Optimierung unterzogen werden. Dazu kann unter anderem die Anpassung der Auflösung auf mindestens 300 dpi, eine Trennung von Doppelseiten, eine Bildrotation, falls die Scans in einer falschen Ausrichtung vorliegen, oder auch eine Bildbeschneidung, das heißt zum Beispiel das Entfernen von schwarzen Rändern („noise removal“[32]), nötig sein. Typische Scanfehler wie schiefe Textseiten, Wölbungen oder Verzerrungen versucht man ebenfalls in diesem Prozess durch verschiedene technische „Normalisierungs“-Verfahren („De-Skewing“) zu lösen.[33] Der wichtigste Schritt im Pre-Processing ist jedoch die sogenannte Binarisierung, also die Umwandlung des Images in ein Schwarz-Weiß-Bild, da die zur Zeichenerkennung gespeicherten Musterrepräsentationen sich auf schwarz-weiße Zeichen beziehen.[34] Das Image wird dabei mit verschiedenen Algorithmen analysiert und in ein Schwarz-Weiß-Bild überführt, um den Text vom Hintergrund zu trennen und die einzelnen Buchstaben vollständig zu erhalten:

> It is an important and critical stage in the document image analysis and recognition pipeline since it permits less image storage space, enhances the readability of text areas, and allows efficient and quick further processing for page segmentation and recognition.[35]

[31] Masterplan Zeitungsdigitalisierung (2016, S. 26).
[32] Dengel & Shafait (2014, S. 189).
[33] Gatos (2014, S. 104-127).
[34] Fuchs (2013, S. 198-199), Federbusch & Polzin (2013, S. 28-29).
[35] Gatos (2014, S. 77-86, hier S. 77).

Die Sorge, durch eine zu starke Binarisierung könne wiederum die Erkennung der einzelnen Zeichen leiden, wurde sowohl im deutschen Zeitungsdigitalisierungsprojekt als auch im ENP widerlegt.[36]

3.3 Segmentierung

Die Segmentierung oder auch Fragmentierung eines Digitalisats auf Seitenebene ist der zweite Schritt im Prozess der Texterschließung. Die Software unterscheidet dabei nicht nur zwischen Text und Nicht-Text, sondern differenziert weiter in Strukturbereiche wie Textblöcke, Tabellen oder auch Bilder.[37] Es handelt sich dabei um eine Layoutanalyse auf rein graphischer Ebene, die inhaltliche Segmentierung erfolgt erst später (siehe Kapitel 3.5). Sollten in diesem Schritt jedoch bereits Erkennungsfehler entstehen, wird sich das durch die gesamte weitere Texterkennung durchziehen: „If a page segmentation algorithm fails to correctly segment text from image, the character recognition module outputs a lot of garbage characters originating from the image parts.“[38]

Die Segmentierung einzelner Zeilen kann größere Probleme bereiten, da der Textfluss gerade in älteren Drucken nicht immer exakt verläuft. Das bedeutet, dass sich Textzeilen eventuell überlappen, dass Zeilen unterbrochen sind oder auch schief gedruckt wurden. Mit verschiedenen technischen Methoden wird versucht, die einzelnen Textzeilen zu isolieren.[39] Das Vorgehen bei der nachfolgenden Zeichensegmentierung ist ähnlich, denn auch hier hat die Software oft mit sich überlappenden, verzerrten oder auch fragmentierten Zeichen zu kämpfen, die eine Segmentierung der Zeichen erschweren oder unmöglich machen. Häufige Fehlerquellen sind sich berührende Zeichen, wie zum Beispiel „r n“, die in automatisierten Verfahren häufig als „m“ erkannt werden, oder auch gebrochene Zeichen in Folge mangelhafter Binarisierung oder aufgrund einer schlechten Scanvorlage. Diese weitere hierarchische Segmentierung eines Digitalisats bis auf Zeilen-, Wort- und Zeichenebene zählt jedoch bereits zur eigentlichen OCR.[40]

---

[36] Pletschacher et al. (2015, S. 45-46), Masterplan Zeitungsdigitalisierung (2016, S. 31-32).
[37] Fuchs (2013, S. 200), Federbusch & Polzin (2013, S. 29-30).
[38] Dengler & Shafait (2014, S. 190).
[39] Nobile & Suen (2014, S. 263-272).
[40] Nobile & Suen (2014, S. 272-288).

3.4 Optical Character Recognition (OCR)

Ziel der optischen Zeichenerkennung ist es, die isolierten Rastergraphiken zu identifizieren und ihnen ein semantisches Zeichen bzw. einen Buchstaben zuzuweisen, um einen maschinendurchsuchbaren Text zu erhalten. In technischer Hinsicht bezieht sich die OCR somit nur auf den Teilbereich des Mustervergleichs von segmentierten Bildteilen mit einem in der Software hinterlegten Zeichenvorrat. Auf die isolierten Zeichen werden sogenannte Klassifikatoren („classifier"), also zuvor erstellte Muster („patterns"), angewandt, die ein bestimmtes Zeichen repräsentieren und in der Software hinterlegt wurden. Verschiedene Algorithmen zur „pattern recognition" kommen dabei zum Einsatz.[41] Voraussetzung für eine optimale Erkennung ist also, dass das System zuvor auf alle im Text vorkommenden Zeichen, die nicht standardmäßig vorhanden sind, „trainiert" wird. Dadurch können erheblich bessere Ergebnisse erzielt werden.[42] In einem Projekt der SBB zur Volltexterschließung eines ausgewählten Bestandes alter Drucke (Funeralschriften des 17. Jahrhunderts) versuchte man zum Beispiel, der Software eigene Muster für die verwendeten Schrifttypen durch manuelles Klassifizieren und Zuordnen beizubringen, was die Texterkennung deutlich verbesserte.[43] Verschiedene Spracheinstellungen wurden nur im ENP getestet, sollen aber in einem Nachfolgeprojekt des deutschen Zeitungsdigitalisierungsprojekts eine wichtigere Rolle spielen.[44]

In allen für diese Arbeit herangezogenen Projekten ist die Zufriedenheit nicht nur mit der Erkennung von Antiqua-Schrift, sondern gerade auch mit Fraktur-Schrift sehr hoch. So heißt es zum Beispiel in einem Beitrag zum ENP:

> It should be noted that the performance on the very difficult Gothic text is comparatively very good, due to ABBYY FineReader having a new Fraktur module [...]. What used to be near random results for such documents is now dose to 70% which is considered by many the threshold for meaningful full text search. Documents with mixed content (which basical-

[41] Für eine detaillierte Beschreibung der technischen Verfahren und verwendeten Algorithmen zur automatisierten Mustererkennung („pattern recognition") siehe vor allem Cao & Natarajan (2014, S. 332-358).
[42] Mühlberger (2011, S. 14), Fuchs (2013, S. 200-202).
[43] Federbusch & Polzin (2013, S. 23-25).
[44] Masterplan Zeitungsdigitalisierung (2016, S. 34).

> ly requires the OCR engine to apply all dassifiers and then to decide which result to use) are naturally harder to recognise.[45]

Als Herausforderung für die Erkennungssoftware wurden jedoch in beiden Projekten „Mischschriften" identifiziert, wenn auf einer Zeitungsseite Antiqua und Fraktur vermischt vorkommen, aber auch Sonderschriften, zum Beispiel für die Auszeichnung von Titeln oder in Werbeblöcken: „Als größeres Problem erwies sich, dass die Überschriften in sehr unterschiedlichen Schriftarten, Schriftstilen oder Schriftmischungen gehalten waren, was deren OCR-Genauigkeit und damit die spezifischen Vorteile einer Suche von bzw. in Artikeln beeinträchtigen kann."[46]

Nach der reinen optischen Zeichenerkennung werden OCR-Korrektur-Methoden auf die OCR-Ergebnisse angewandt, die meistens (wie auch bei ABBYY) Teil des Gesamtprozesses sind und deshalb an dieser Stelle kurz beschrieben werden. Die automatisierte Nachkorrektur erfolgt mit Hilfe lexikalischer und linguistischer Technologien, indem im Hintergrund Wörterbücher und Lexika zur Textanalyse, zum Abgleich und zum Ausbessern fehlerhafter Wörter verwendet werden. Das heißt, alle Wörter werden in einem oder mehreren Lexika nachgeschlagen; wird ein Wort nicht gefunden, ermittelt die Software nahe liegende Verbesserungsvorschläge. Wird zum Beispiel ein „ü" aufgrund einer schlechten Scan-Vorlage nicht korrekt erkannt, kann die Zusatzinformation, dass es sich um einen deutschen Text handelt, helfen, um statt „Munchen" korrekterweise „München" auszugeben.[47]

Bei historischen Texten ist eine derartige Textanalyse ungemein schwieriger, weshalb für bestimmte Themenbereiche spezifische Wortlisten oder eigene Wörterbücher neu angelegt und in der Software hinterlegt werden. Ein Nachteil dieser Korrekturmethode ist, dass die intellektuell erkannten Regeln jedoch im Einzelfall versagen können. Das Original enthält zum Beispiel bereits Druckfehler oder Schreibvari-

---

[45] Clausner et al. (2015, S. 934); zu den positiven Ergebnissen bei der Fraktur-Erkennung siehe auch Wernersson (2015, S. 30) und Fuchs (2013, S. 205). Michael Fuchs betont die großen Fortschritte, die ABBYY in den letzten Jahren durch die enge Zusammenarbeit mit Bibliotheken, Universitäten und Forschungsinstitutionen gerade im Bereich der Erkennung von Fraktur-Schrift erzielt hat. Als maßgebliche Projekte werden IMPACT (= Improving Access to Text), ein EU-Projekt (2008-2011), das die Verbesserung von Volltexterkennung und die Entwicklung neuer Verfahren und Tools zum Ziel hatte, siehe http://www.impact-project.eu, und METAe (Metadata Engine, http://meta-e.aib.uni-linz.ac.at) genannt.

[46] Masterplan Zeitungsdigitalisierung (2016, S. 33). Siehe dazu auch Mühlberger (2011, S. 14).

[47] Fuchs (2013, S. 203), Mühlberger (2011, S. 14).

anten, die nun durch die automatisierte Anwendung bestimmter Korrekturregeln ausgemerzt werden.[48] Bei der Entwicklung geeigneter Tools und Methoden zur Fehler-Korrektur bewegt man sich vor allem im Bereich der Computerlingustik, wozu es zahlreiche Projekte und Forschungsarbeiten mit unterschiedlichen Ansätzen gibt.[49]

Der Masterplan für die Zeitungsdigitalisierung weist auch auf diese Problematik hin und empfiehlt vor allem die Entwicklung von Wörterbüchern mit historischen Varianten (so zum Beispiel Tehyl → Theil → Teil), um den Index für die Suchabfragen damit anreichern zu können: „Das Angebot einer Option 'Suche nach historischen Varianten' würde die Abfrage des Index um die im historischen Wörterbuch verzeichneten validen Varianten erweitern (query expansion) und so auch die Fundstellen dem Benutzer als Treffer anbieten.“[50]

3.5 Optical Layout Recognition (OLR)

Die Volltexterschließung von Retrodigitalisaten beschränkt sich jedoch nicht nur auf eine möglichst akkurate Zeichenerkennung, sondern umfasst auch eine inhaltliche Erschließung des Dokuments. Dadurch werden erweiterte Suchmöglichkeiten innerhalb eines Dokuments (zum Beispiel auf Kapitelebene) möglich sowie auch der Aufbau von Linkstrukturen zur Navigation innerhalb eines digitalisierten Buches:

> The main role of logical labeling in this context is to identify table of contents pages in the books and to link individual entries in the table of content pages to the corresponding chapters/sections. [...] Similarly, when digitizing scholarly material like technical journals, it is often required to extract titles and authors of each artide to facilitate advanced search within the digitized collection.[51]

Eine „logical labeling“-Analyse bedeutet die logische und semantische Auszeichnung eines digitalen Dokuments (Titel, Überschrift, Autor, Seitenzahl, Text etc.), um intellektuell verständliche Informationen zu extrahieren, die wiederum in maschinenles-

---

[48] Federbusch & Polzin (2013, S. 32).
[49] Mühlberger (2011, S. 17), Federbusch & Polzin (2013, S. 33-34). Namhaft in diesem Bereich ist sicherlich das Centrum für Informations- und Sprachverarbeitung (CIS) an der LMU München mit Prof. Klaus Schulze, renommierte Forschungsarbeiten zum Thema OCR-Nachkorrektur stammen unter anderem von Christian M. Strohmaier (2004) und Christoph Ringlstetter (2006, 2009).
[50] Masterplan Zeitungsdigitalisierung (2016, S. 37).
[51] Dengel & Shafait (2014, S. 184).

baren Code umgewandelt werden können. Der Mehrwert für eine wissenschaftliche Nutzung bei historischen Zeitungen liegt vor allem in einer Artikelseparierung. Erschwert wird dies durch die Spaltendarstellung von Zeitungen, da damit nicht unbedingt nachvollziehbar ist, wo auf der Zeitungsseite der Textfluss fortgeführt führt. Um inhaltliche Einheiten einer Zeitung herausfiltern und sogar den Textfluss nachvollziehen zu können, müssen die Ergebnisse der graphischen Layout-Segmentierung mit den Ergebnissen der Mustererkennung zusammenfließen. Außerdem muss der Software auch hier wieder die Bedeutung der verschiedenen Auszeichnungen antrainiert werden: „Generic logical structure information of a document is encoded as a set of layout and typesetting conventions for that document class. These conventions can be regarded as document knowledge for that particular class of documents."[52]

So erhält man über die Segmentierung (siehe Kapitel 3.3) einzelne Textblöcke, und über die Musterkennung ist es möglich, Textformatierungen oder auch syntaktische Zeichen auszuwerten. Bestimmte Formatierungen, wie etwa Fettdruck, große Schriftgröße oder eine besondere Schriftart können auf eine Überschrift hindeuten. Syntaxzeichen, wie zum Beispiel ein Punkt, bedeuten möglicherweise das Ende eines Artikels (verbunden mit einer Einrückung) oder Trennstriche die Fortführung des Textes in einem nächsten Textblock. Jedoch sind diesem Vorgehen durch die hohe Anzahl an Formatierungsmöglichkeiten Grenzen gesetzt. Seit mittlerweile mehr als zwei Jahrzehnten wird an technischen Methoden und Tools zur automatisierten Layout-Analyse gearbeitet, jedoch bringt die verfügbare Software vor allem bezogen auf die Digitalisierung von historischen Zeitungen noch kaum verwertbare Ergebnisse.

Im Zeitungsdigitalisierungsprojekt wurden verschiedene Tests für eine Strukturerschließung durchgeführt, lediglich die BSB testete eine halb-automatisierte Artikelseparierung. Die Ergebnisse bestätigen die genannten Probleme: In den Stichproben ergaben sich große Schwankungen der erkannten Schriftgröße und des fälschlich erkannten Fettdrucks, wodurch die Erkennung und Auszeichnung von Überschriften erschwert wird.[53] Innerhalb der Überschriften beeinträchtigen unterschiedliche Schriftarten, Schriftstile und Schriftmischungen deren Texterkennung. Auch wurden

[52] Dengel & Shafait (2014, S. 197).
[53] Wernersson (2015, S. 13).

zum Teil Überschriften, die mit besonderen Drucken gestaltet sind, nicht als Text, sondern als Bild erkannt.[54]

Die Ergebnisse bzw. Empfehlungen lauten, dass bezüglich der Artikelerschließung eine Layouterkennung mit Artikelseparierung „zum erweiterten Erschließungsstandard für wissenschaftliche Zwecke“ [55] zähle und deshalb besonders bei Zeitungen von überregionaler Bedeutung zu empfehlen sei. Die oft hochgradig komplexe Struktur von Zeitungen bringe die aktuellen Technologien jedoch an ihre Grenzen: „Die Verfügbarkeit valider Werkzeuge sollte zu einer deutlich besseren Erschließung und damit auch Präsentation führen“.[56] Auch im ENP heißt es, „in terms of layout analysis capabilities there is still room for improvement“.[57]

3.6 Nachkorrektur

Eine Korrektur des automatisiert erfassten Volltexts sowie des Layouts kann teilautomatisiert und manuell erfolgen. Auf die Möglichkeiten einer nachträglichen lexikalischen und linguistischen Fehlerkorrektur als Teil des automatisierten OCR-Prozesses wurde bereits in Kapitel 3.4 eingegangen.

Manuelle Postkorrektur-Verfahren wurden lange als zeit- und kostenintensiv eingeschätzt, erfreuen sich jedoch im digitalen Zeitalter mit den Möglichkeiten des Web 2.0 immer größerer Beliebtheit. Als Vorreiter in dieser Hinsicht gilt das „Crowdsourcing-Projekt“[58] der National Library of Australia, die bereits 2008 das Potenzial der sozialen Teilhabe ihrer NutzerInnen am Korrekturprozess mit Hilfe eines Online-Tools erkannte: Innerhalb weniger Wochen korrigierten ca. 1.000 registrierte NutzerInnen über 700.000 Textzeilen in mehr als 50.000 Artikeln. Erfreulich war bei diesem Projekt nicht nur die hohe Teilnahme der Öffentlichkeit, sondern auch die sehr gute Qualität, die erzielt wurde: „... our own users have the potential to achieve a 100% accuracy rate with their knowledge of English, history

[54] Masterplan Zeitungsdigitalisierung (2016, S. 33).
[55] Masterplan Zeitungsdigitalisierung (2016, S. 36).
[56] Masterplan Zeitungsdigitalisierung (2016, S. 36).
[57] Pletschacher (2015, S. 46).
[58] „Crowdsourcing uses social engagement techniques to help a group of people achieve a shared, usually significant, and large goal by working collaboratively together as a group.“ (Holley 2010).

and context, whereas our contractors are only achieving an accuracy of 99.5% in the title headings."[59]

Ein anderes Beispiel für den Einbezug der Öffentlichkeit ist das Digitalisierungsprojekt der New York Times. Die OCR-Korrektur wurde mithilfe sogenannter „Captchas" durchgeführt, das heißt, alle von der OCR-Software nicht automatisch erkannten Wörter wurden in Captchas verwandelt und den NutzerInnen beim Einloggen in passwortgeschützte Websites zur Entzifferung vorgegeben. Diese Methode wird von Google zur Korrektur der Volltexte ihrer digitalisierten Bücher im großen Stil eingesetzt.[60] Auch im Masterplan wird die Einbeziehung der Nutzerinnen und Nutzer bei der OCR-Optimierung als erfolgversprechender Ansatz hervorgehoben.[61]

4. Qualitätsbewertung

Die Bewertung der OCR-Qualität erfolgt anhand der Evaluierung der Erkennungsgenauigkeit, das heißt, der Übereinstimmung zwischen Original und Kopie. Die Ermittlung dieses Wertes ist jedoch nicht trivial und geschieht oft auf Grundlage unterschiedlicher Bemessungsgrundlagen, was einen direkten Vergleich erschwert. Unterschiede liegen in der Festlegung, ob Wort-, Zeichen-, oder auch Layout-Genauigkeit bewertet wird, und in den Überprüfungsmethoden (Ground Truth-Datenset, Stichproben). Nicht zuletzt ist jedoch die Qualität des generierten Volltextes abhängig von den projektspezifischen Rahmenbedingungen.[62]

In der Regel wird die Wort- oder die Zeichengenauigkeit als Bewertungskriterium verwendet. Entsprechend der Levenshtein-Distanzberechnung wird jeder Ersetzungs-, Lösch- und Einfüge-Vorgang als jeweils ein Fehler gezählt.[63] Bei der Bewertung der Zeichengenauigkeit ist ausschlaggebend, wie viele Zeichen korrekt erkannt wurden. Im deutschen Zeitungsdigitalisierungsprojekt entschied man sich für diese Methodik. Beim ENP hingegen entschied man sich für eine Bewertung der Wortgenauigkeit, da zum einen die Auswertung der Zeichengenauigkeit bei umfangreichen Textdokumenten wie Zeitungen zu ressourcenintensiv sei und zum anderen

[59] Holley (2009).
[60] Gugliotta (2011).
[61] Masterplan Zeitungsdigitalisierung (2016, S. 35).
[62] Federbusch & Polzin (2013, S. 130).
[63] Jurafsky (2000), S. 154.

diese eher dem Ziel der Texterkennung entspreche: „scholars use words as search terms, not characters“[64]. Bei der Wortgenauigkeit wird abgeglichen, inwiefern das Wort im Original und im Referenztext übereinstimmt, dabei ist die Wortreihenfolge ausschlaggebend. Für das ENP wählte man aufgrund des komplexen Zeitungslayouts eine „Bag of Words“-Metrik, welche die Wortreihenfolge außen vor lässt und nur die Tatsache bewertet, ob ein Wort korrekt erkannt wurde oder nicht.[65] Im Unterschied zur Bewertung der Zeichengenauigkeit spielen somit zuvor festgelegte Bewertungsregeln, zum Beispiel welches Wort aufgrund seiner „Signifikanz“ (damit sind alle Wörter gemeint, die für eine spätere Suche infrage kommen könnten) überhaupt in die Evaluation miteinbezogen wird, eine zentrale Rolle und erschweren eine Vergleichbarkeit.

Eine ausführliche OCR-Evaluation des Zeitungsdigitalisierungsprojekts der BL von 2009 ergab, dass die Zeichengenauigkeit immer höher lag als die Wortgenauigkeit, aber dass die Erkennungsrate signifikanter Wörter immer schlechter abschnitt als die gesamte Worterkennungsrate.[66] Da also keine Klarheit in der Wahl der Evaluierungsmethodik besteht, empfiehlt der Masterplan Zeitungsdigitalisierung, „eine differenzierte fachwissenschaftliche Bewertung der Qualitätseinstufungen von Zeichen- und Wortgenauigkeit“ im Rahmen des aktuellen OCR-D-Projektes anzustreben.[67]

Die Überprüfung der Layout-Erkennungsrate ist ungemein schwieriger, was nicht zuletzt daran liegt, dass die Entwicklungen in diesem Bereich noch sehr am Anfang stehen und allgemein definierte Bewertungskriterien noch nicht vorliegen (siehe Kapitel 3.5). Der Fachbeitrag der BSB liefert eine ausführliche Evaluierung, der als Kriterien der Lesefluss sowie die Formaterkennung zugrunde lagen, woraus

---

64 Pletschacher (2015, S. 39), siehe auch Tanner et al. (2009).
65 Pletschacher et al. (2014, S. 7), Pletschacher (2015, S. 40-41).
66 Tanner et al. (2009) schlussfolgert weiter daraus: „Considering that the average word accuracy across the whole 19th Century Newspaper Project is 78% with the significant word accuracy at 68.4% then it is dear that searching the resource will not be as satisfactory for the end user as might be desired. It is also inescapable that this is due to a combination of two major factors: the OCR process itself and the content + physical quality of the newspapers.“
67 Masterplan Zeitungsdigitalisierung (2016, S. 40). OCR-D ist ein aktuelles Koordinierungsprojekt, welches auf die Weiterentwicklung von Verfahren der Optical Character Recognition (OCR) für historische Drucke des 16. bis 18. Jahrhunderts im deutschsprachigen ausgerichtet ist. Zu den Hauptzielen zählt die Weiterentwicklung der Optical Layout Recognition sowie die Erstellung von Verfahren der Qualitätssicherung. Da das Projekt jedoch erst Ende 2015 begann, lagen für diese Arbeit noch keine verwertbaren Ergebnisse vor. Weitere Informationen siehe unter: http://www.ocr-d.de.

Rückschlüsse bezüglich der Erkennung von Überschriften gezogen werden könnten. Eine statistische Auswertung der Formaterkennung liegt nicht vor, sondern lediglich Kommentierungen, deren Ergebnisse keine abschließenden Schlussfolgerungen zulassen.[68]

Für die Ermittlung der Erkennungsrate ist ein Referenztext notwendig, der je nach den Rahmenbedingungen des Projektes unterschiedlich gewonnen wird. Möchte man eine 100%ige Genauigkeit für einen ausgewählten Textkorpus als Bemessungsgrundlage anwenden können, so muss ein sogenannter Ground Truth erstellt werden. Die Zusammenstellung eines Ground Truth-Datensets ist sehr aufwändig und kostspielig, nicht nur weil der Volltext fehlerfrei sein muss, sondern auch weil die Auswahl der Images ein möglichst „realistisches" und „repräsentatives" Abbild der gesamten Kollektion darstellen sollte. Somit beinhaltet das Datenset auch problematische Images, zum Beispiel mit Verzerrungen und Seitenwölbungen, mit unterschiedlichem Layout und natürlich unterschiedlichen Schrifttypen.[69] Im ENP wurde ein umfangreiches Datenset aus ca. 600 Images aufgebaut, um damit aussagekräftige OCR-Evaluationstests durchführen zu können.[70] Der Aufwand wurde wie folgt beziffert: „Reflecting on the creation of the dataset, it required several hundreds of person-hours from a variety of people in different organisations to select, pre-process, manually correct (this is the most easily quantifiable cost: €15,000), verify, ingest and categorise the page images and ground truth content."[71]

Alternativ zu Ground Truth-Daten wird mit Stichproben gearbeitet, die mit Hilfe statistischer Verfahren die Ermittlung von Genauigkeitswahrscheinlichkeiten ermöglichen. Durch die Angabe einer sogenannten Konfidenzzahl kann die unbekannte Erkennungsquote einer Software besser eingeschätzt werden, sie gibt die Güte der Schätzung an. Bei einer Konfidenzzahl von 95% liegt die tatsächliche Erkennungsrate mit einer Wahrscheinlichkeit von 95% im Intervall [Mittelwert-a, Mittelwert+a]. Je höher die Konfidenzzahl, desto wahrscheinlicher ist die korrekte Bewertung des OCR-Ergebnisses. Dieses Verfahren verhilft zur besseren Beurteilung einer

[68] Wernersson (2015, S. 30 und S. 35).
[69] Valvery (2014, S. 986). Vorreiter bei der Performanz-Evalution auf Basis von Ground Truth-Datensets war die BL (siehe Tanner et al. (2009)).
[70] Pletschacher et al. (2014, S. 11-13).
[71] Clausner et al. (2015, S. 935).

zu verwendenden Software.[72] Eine andere Möglichkeit zur Überprüfung der Güte eines Textes, wenn es darum geht, die Arbeitsergebnisse eines Dienstleisters zu überprüfen, ist das sogenannte Bernoulli-Experiment. Ziel ist es, anhand einer Stichprobe die vom Dienstleister behauptete Erkennungsquote zu überprüfen, wobei die Wahrscheinlichkeit für einen Irrtum möglichst niedrig gehalten werden soll, während zugleich die Stichprobengröße noch praktisch anwendbar ist. Die DFG-Praxisregeln empfehlen dazu, dass in einer Stichprobe von 500 Zeichen bei einer Irrtumswahrscheinlichkeit von 2,5% mindestens 489 Zeichen erkannt werden müssen, wenn der Dienstleister eine Genauigkeit von 96% behauptet.[73] Im Rahmen des Zeitungsdigitalisierungsprojekts entschied man sich gegen die Erstellung und Evaluierung mit Hilfe von Ground Truth-Daten, stattdessen wurde mit Stichproben gearbeitet. Seitens ULB und BSB wurden umfangreiche OCR-Tests durchgeführt.

In Halle wurde die Zeichengenauigkeit basierend auf einer zufällig ausgewählten Stichprobe von 40 Probewerten zu je 1000 Zeichen gemessen. Bei einer Wahrscheinlichkeit von 99% wurde eine Erkennungsrate von 98,36% (bei 300 dpi) erzielt, „was gemeinhin als gut gilt (Gut: 97-99,5 Prozent)“[74]. Die Ergebnisse beziehen sich jedoch nur auf die Erkennung der Frakturschrift und nicht auf die Gesamtheit der Mischschriften. In München erfolgte die Evaluierung auf Basis von 50 Stichproben mit jeweils 925-992 Zeichen (1.000 mit Zeilenumbrüchen). Bei einer Konfidenzzahl von 95% wurde eine Zeichengenauigkeit von 94,70-97,65% erzielt, was als „gutes, zufriedenstellendes Resultat“ gilt.[75] Es wird darauf hingewiesen, dass die Ergebnisse beider Evaluationen nicht direkt vergleichbar sind, weil mit unterschiedlichen Fehlerklassifikationen gearbeitet wurde. An der BSB wurde eine differenziertere Klassifikation entwickelt, die bezüglich der Operationen und ihrer Bewertung über die klassische Levenshtein-Distanzberechnung hinausgeht.[76] Im Masterplan heißt es zusammenfassend, dass bei Frakturschrift in Zeitungen über 95% Zeichengenauigkeit erreicht werden könne, was einer anzustrebenden Wortgenauigkeit von ca. 80% ent-

[72] Federbusch & Polzin (2013, S. 132).
[73] Deutsche Forschungsgemeinschaft (2013, S. 31-32), siehe auch Federbusch & Polzin (2013, S. 133-135).
[74] Sommer et al. (2014, S. 80), Masterplan Zeitungsdigitalisierung (2016, S. 37).
[75] Wernersson (2015, S. 23), Masterplan Zeitungsdigitalisierung (2016, S. 37-38).
[76] Zur Fehlerklassifikation der ULB siehe Sommer et al. (2014, S. 79-80), zur Fehlerklassifikation der BSB siehe Wernersson (2015, S. 24-25).

spricht.[77] Die Ergebnisse des ENP ergaben bei Antiqua eine Wortgenauigkeit von 81,4%, bei Fraktur 67,3% und bei Mischschriften 64%.[78]

Die tatsächliche Qualität eines OCR-Textes kann jedoch nur unter Einbeziehung der projektspezifischen Einflussfaktoren, wie etwa der Beschaffenheit der Scan-Vorlage, der Scan-Parameter, der verwendeten OCR-Software sowie der Korrekturmöglichkeiten, abschließend beurteilt werden: „Je nach Scanbeschaffenheit und Vorlage erreicht man bei einem Buch 99%, beim nächsten nur noch 95%, selbst wenn alle Software-, Trainings- und Wörterbuchparameter optimiert wurden. Nötig wäre hier eine Art Qualitätsindex für das Digitalisat und die Vorlage.“[79] Auch ökonomische Fragestellungen spielen eine entscheidende Rolle, insofern Mittel für manuelle Nachkorrekturen zur Verfügung stehen oder nicht.

5. Fazit

Einig ist man sich in den genannten Evaluationen, dass OCR-Daten ohne die Angabe, zu welchem Zwecke sie benötigt werden, was also das wissenschaftliche Ziel ist, nicht zu beurteilen sind: „It is of crucial importance to objectively evaluate the results of digitisation projects not only in terms of apparent accuracy (e.g. percentage of correct words) but more importantly in the context of their different intended use scenarios also.“[80] Zu unterscheiden sind hier verschiedene Nutzungsszenarien, wie etwa eine Stichwort- oder auch eine Phrasensuche, aber auch zunehmend quantitative Analyseverfahren und eine semantische Verarbeitung innerhalb des Forschungsfeldes der Digital Humanities. Gerade bei historischen Zeitungen ist eine Suche auf Artikelebene essenziell, die eine sehr gute Layout-Erkennung voraussetzt. Je nach wissenschaftlichem Ziel ist eine unterschiedlich hohe Qualitätsstufe des Volltextes notwendig. Inwiefern bereits durch die Bereitstellung von „schmutzigem“ OCR (also eine Erkennungsrate < 90%) ein wissenschaftlicher Nutzen erfüllt wird, weil immerhin „Orientierung im Text“ geboten würde, ist in der Fachcommunity umstritten.[81] In den DFG-Praxisregeln heißt es, dass auf Basis einer Stichprobenauswertung „erst

[77] Masterplan Zeitungsdigitalisierung (2016, S. 40).
[78] Pletschacher et al. (2015, S. 44).
[79] Federbusch & Polzin (2013, S. 131).
[80] Pletschacher (2015, S. 39), siehe auch Wernerssnon (2015, S. 34), Federbusch & Polzin (2013, S. 137) und Deutsche Forschungsgemeinschaft (2013, S. 30).
[81] Federbusch & Polzin (2013, S. 137), Kämmerer (2009, S. 631 und 634).

Texte ab einer Genauigkeit von 99,5% als wissenschaftlich zuverlässig gelten können“ […]. Unterhalb von 80% scheint der Gesamtnutzen einer Konversion eher fragwürdig, im Bereich zwischen hochwertigem und schlechtem Text kommt es aber immer auf die Art des Projekts an wie auch auf die damit verbundenen Kosten.“[82]

Wie in diesem Artikel aufgezeigt werden sollte, hat die Entwicklung geeigneter Verfahren und Tools für eine automatisierte Volltexterschließung in den letzten Jahren große Fortschritte gemacht. Dies bestätigt auch der Masterplan zur Zeitungsdigitalisierung in Deutschland: “Die Ergebnisse dieser Pilotphase sind ermutigend und zeigen, dass auch Zeitungen mit einer – gemessen an Drucken des 18. bis 20. Jahrhunderts – viel komplexeren Layoutstruktur sowohl mit Antiqua- als auch Fraktur-Lettern über OCR als Volltexte aufbereitet werden können und sollten.“[83] Eine weitere Volltexterschließung mit Layouterkennung und Artikelseparierung (OLR) entspreche jedoch noch nicht den wissenschaftlichen Bedürfnissen und muss noch stark verbessert werden: „Die Verfügbarkeit valider Werkzeuge sollte zu einer deutlich besseren Erschließung und damit auch Präsentation führen.“[84]

Bei der Planung eines Digitalisierungsprojektes mit Volltexterschließung muss also genau abgewogen werden, inwiefern die betreffende Kollektion mit den bestehenden Werkzeugen hinreichend bearbeitet werden kann, welches wissenschaftliche Ziel angestrebt wird und ob gegebenenfalls genügend finanzielle Mittel eingesetzt werden sollten, um durch manuelle Korrekturen das gewünschte Ergebnis zu erzielen. Aus dem bestehenden Content also *sinnvollen* und *brauchbaren* Context zu erzeugen, bleibt eine herausfordernde Aufgabe, die in Abhängigkeit vom jeweiligen Bestand und Projektziel entschieden werden muss.

[82] Deutsche Forschungsgemeinschaft (2013, S. 32).
[83] Masterplan Zeitungsdigitalisierung (2016, S. 35).
[84] Masterplan Zeitungsdigitalisierung (2016, S. 36).

Literatur

ANNO - Austria Newspapers Online. http://anno.onb.ac.at (abgerufen am 02.02.2016).

Böning, H. (2013). Vom Wert der Quellen für die Rekonstruktion historischer Ereignisse und publizistischer Unternehmungen. Einige Gedanken zum Sinn von biobibliographischer Grundlagenforschung. In H. Böning (Hrsg.), *Deutsche Presseforschung. Geschichte und Forschungsprojekte des ältesten historischen Instituts der Universität Bremen* (S. 27–53). Bremen: Ed. Lumière.

Cao, H. & Natarajan, P. (2014). Machine-Printed Character Recognition. In D. Doermann & K. Tombre (Hrsg.), *Handbook of document image processing and recognition* (S. 331–358). London: Springer Reference.

Centrum für Informations- und Sprachverarbeitung - LMU Munich. http://www.cis.uni-muenchen.de (abgerufen am 10.02.2016).

Clausner, C. et al. (2015). The ENP image and ground truth dataset of historical newspapers. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (S. 931–935).

Dengel, A. & Shafait, F. (2014). Analysis of the Logical Layout of Documents. In D. Doermann & K. Tombre (Hrsg.), *Handbook of document image processing and recognition* (S. 177–222). London: Springer Reference.

Deutsche Forschungsgemeinschaft (2013). *DFG-Praxisregeln „Digitalisierung"*. http://www.dfg.de/formulare/12_151/12_151_de.pdf (abgerufen am 02.02.2016).

DFG-Rahmenantrag (2012). Digitalisierung historischer Zeitungen. Unveröffentlichtes Dokument zum internen Gebrauch. o.O.

Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland (Masterplan Zeitungsdigitalisierung). Ergebnisse des DFG-Projektes „Digitalisierung historischer Zeitungen" Pilotphase 2013-2015 (2016). http://www.zeitschriftendatenbank.de/fileadmin/user_upload/ZDB/z/Masterplan.pdf (abgerufen am 01.02.2018).

Europeana Newspapers. http://www.europeana-newspapers.eu (abgerufen am 02.02.2016).

Federbusch, M. & Polzin, C. (2013). Volltext via OCR: Möglichkeiten und Grenzen ; Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz. Berlin: Staatsbibliothek zu Berlin - Preußischer Kulturbesitz.

Fuchs, M. (2013). Grundlagen und Herausforderungen bei der Verarbeitung historischer Dokumente mit Fraktur-OCR - historische Dokumente in einer „digitalen" Welt. In J. Meier, F. Kopp & J. Schrastetter (Hrsg.), Digitale Quellensammlungen. Erstellung – Archivierung – Präsentation – Nutzung (S. 197–210). Berlin: Weidler.

Gatos, G. B. (2014). Imaging Techniques in Document Analysis Processes. In D. Doermann & K. Tombre (Hrsg.), *Handbook of document image processing and recognition* (S. 73–132). London: Springer Reference.

Gugliotta, G. (2011). Deciphering Old Texts, One Woozy, Curvy Word at a Time. http://www.nytimes.com/2011/03/29/science/29recaptcha.html?_r=0 (abgerufen am 10.02.2016).

Holley, R. (2009). How Good Can It Get? *D-Lib Magazine*, 15(3/4). doi:10.1045/march2009-holley.

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine*, 16(3/4). doi:10.1045/march2010-holley.

Impact. Improving access to text. http://www.impact-project.eu (abgerufen am 08.02.2016).

Jurafsky, D. S. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Kämmerer, C. (2009). Technik. Vom Image zum Volltext – Möglichkeiten und Grenzen des Einsatzes von OCR beim alten Buch. *Bibliotheksdienst,* 43(6). doi:10.1515/bd.2009.43.6.626.

Lübbers, B. (2015). „Content is King, Context is Queen". *B.I.T. Online*, (2), 182–185. http://www.b-i-t-online.de/heft/2015-02-interview-ceynowa.pdf (abgerufen am 02.02.2016).

Meier, J. (2013). Die Relevanz digitaler Quellensammlungen für die linguistische Forschung. In J. Meier, F. Kopp & J. Schrastetter (Hrsg.), *Digitale Quellensammlungen. Erstellung - Archivierung - Präsentation - Nutzung* (S. 29–47). Berlin: Weidler.

METAe - Meta Data Engine. http://meta-e.aib.uni-linz.ac.at (abgerufen am 08.02.2016).

Mühlberger, G. (2011). Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Tex tund Strukturerkennung (OCR). In Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB) 58 (1), 10–18.

Müller, M. E. & Hermes, M. (2014a). Die digitale Transformation eines ganzen Jahrhunderts: Digitalisierung der Zeitungen des 17. Jahrhunderts an der SuUB Bremen. *Bibliotheksdienst*, 48(12). doi:10.1515/bd-2014-0123.

Müller, M. E. & Hermes, M. (2014b). Digitalisierung der vollständigen deutschsprachigen Zeitungen des 17. Jahrhunderts in der SuUB Bremen: Ein Werkstattbericht. *o-bib. Das offene Bibliotheksjournal*, 1(1), 265–279. doi:10.5282/o-bib/2014H1S265-279.

Neudecker, C., Lieder, H.-J. & Kobel, S. (2015). D2.2 Specification of requirements of OCR and structural refinement-services for digitised newspapers in Europeana. http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/Final_Report.pdf (abgerufen am 03.02.2016).

Nobile, N. & Suen, C. Y. (2014). Text Segmentation for Document Recognition. In D. Doermann & K. Tombre (Hrsg.), *Handbook of document image processing and recognition* (S. 258–290). London: Springer Reference.

OCR-D. Koordinierungsprojekt zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR). http://www.ocr-d.de (abgerufen am 02.02.2016).

Pletschacher, S., Clausner, C. & Antonacopoulos, A. (2014). D3.5_Performance_Evaluation_Report_1.0. http://www.europeana-newspapers.eu/wp-content/uploads/2015/05/D3.5_Performance_Evaluation_Report_1.0.pdf (abgerufen am 09.02.2016).

Pletschacher, S., Clausner, S. & Antonacopoulos, A. (2015). Europeana Newspapers OCR Workflow Evaluation: Proceedings of the 2015 Workshop on Historical Document Imaging and Processing (HIP2015), Nancy, France, August 2015, 39–46. http://primaresearch.org/www/assets/papers/HIP2015_Pletschacher_OCRWorkflowEvaluation.pdf (abgerufen am 16.02.2016).

Ringlstetter, C. et al. (2009). Successfully detecting and correcting false friends using channel profiles. *International Journal on Document Analysis and Recognition (IJDAR)*, 12(3), 165–174. doi:10.1007/s10032-009-0091-y.

Ringlstetter, C.. (2006). Fehlererkennung und Fehlerkorrektur (Inaugural-Dissertation). Ludwigs-Maximilians-Universität, München.

Sommer, D. et al. (2014). Zeitungsdigitalisierung: eine neue Herausforderung für die ULB Halle. *ABI Technik,* 34(2), 75–85. doi:10.1515/abitech-2014-0013.

Strohmaier, C. M. (2004). Methoden der lexikalischen Nachkorrektur OCR-erfasster Dokumente. https://edoc.ub.uni-muenchen.de/3674/1/Strohmaier_Christian.pdf (abgerufen am 12.02.2016).

Tanner, S., Muñoz, T. & Ros, P. H. (2009). Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive. http://www.dlib.org/dlib/july09/munoz/07munoz.html (abgerufen am 08.02.2016).

The British Newspaper Archive. http://www.britishnewspaperarchive.co.uk (abgerufen am 02.02.2016).

TROVE – Digitised Newspapers and more. https://trove.nla.gov.au/newspaper. (abgerufen am 02.02.2016).

Valveny, E. (2014). Datasets and Annotations for Document Analysis and Recognition. In D. Doermann & K. Tombre (Hrsg.), *Handbook of document image processing and recognition* (S. 983–1009). London: Springer Reference.

Wernersson, M. (2015). Evaluation von automatisch erzeugten OCR-Daten am Beispiel der Allgemeinen Zeitung. *ABI Technik*, 35(1), 23–35. doi:10.1515/abitech-2015-0014.