Evaluation von Volltextdaten mit Open-Source-Komponenten

Uwe Hartwig, Universitäts-und Landesbibliothek Sachsen-Anhalt, Halle

Zusammenfassung
Im Bereich der Volltexterzeugung stehen heute vollwertige Open-Source Systeme zur Verfügung. Auch bei der Auswertung der Resultate können etablierte Open-Source-Werkzeuge aus den Bereichen Data Science (DS), Information Retrieval (IR) und Natural Language Processing (NLP) eingesetzt werden. Nach einer kurzen Vorstellung üblicher Auswertungsverfahren und Metriken wird exemplarisch über den Einsatz dieser Tools im DFG-Projekt „Digitalisierung Historischer Deutscher Zeitungen I" der Universitäts- und Landesbibliothek Sachsen-Anhalt (ULB) berichtet.

Summary
In the area of full text recognition, several fully-fledged open source systems are available today. Established open source tools stemming from the fields of Data Science (DS), Information Retrieval (IR) and Natural Language Processing (NLP) can also be used to evaluate the results. After a brief discussion of common evaluation procedures and metrics, the application of such tools in the DFG-funded project „Digitisaion of historical German newspapers I (Digitalisierung Historischer Deutscher Zeitungen I)" at the University and State Library Saxony-Anhalt is used as an example.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/5888

Autorenidentifikation: Hartwig, Uwe: ORCID: https://orcid.org/0000-0001-7164-6376

Schlagwörter: Optical character recognition; OCR; Volltext; Evaluation; historische Zeitungen; Zeitung; Digitalisierung; Digitalisat

1. Zeitungsdigitalisierung

Seit 2011 wird bundesweit an der Digitalisierung historischer deutscher Zeitungen unter dem Stichwort „Masterplan Zeitungsdigitalisierung" gearbeitet. In dessen Kontext entsteht eine zentrale Anlaufstelle, um die Zeitungen des deutschen Kulturraums im Internet zu präsentieren.[1]

Zur Universitäts- und Landesbibliothek (ULB) Sachsen-Anhalt gehört eine der größten Sammlungen regionaler Zeitungen in Deutschland. 2014 startete in Halle ein Pilotprojekt, bei dem das Hallische

1 Deutsche Forschungsgemeinschaft: Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland, 2017. Online: <https://www.zeitschriftendatenbank.de/fileadmin/user_upload/ZDB/z/Masterplan.pdf>, Stand: 17.06.2022.

patriotische Wochenblatt (HpW, 1799–1892) inklusive Folgetitel digitalisiert und durch einen externen Dienstleister mit Volltextdaten versehen wurde.[2]

In der anschließenden, ersten Hauptphase wurden ab 2019 weitere regionale Zeitungen digitalisiert. Begonnen wurde mit dem „General-Anzeiger" (GA, 1889–1918) und seinem Nachfolger, den „Hallischen Nachrichten" (HN, 1918-1943). Darauf folgte „Der Bote für das Saalethal" bzw. die „Saale-Zeitung" (SZ, 1867–1933), welche zum Jahreswechsel 1933/34 in „Mitteldeutschland Saale-Zeitung" (MSZ, 1934–1943) umbenannt wurde. Im Rahmen von Pilotierung und erster Hauptphase wurden über 125.000 Ausgaben und Beilagen digitalisiert.

2. Volltexterzeugung in der Retro-Digitalisierung

Unter der Erzeugung von Volltextdaten, der „Optical Character Recognition" (OCR), versteht man einen Prozess, der versucht, auf einem Bild Schriftzeichen (characters) zu lokalisieren und wiederzuerkennen.[3] Für die Erzeugung von Volltexten stellen historische Zeitungen aufgrund des komplexen Layouts eine besondere Herausforderung dar. Dazu kommen inhärente Qualitätsprobleme des Papiers durch die beginnende Massenproduktion in der 2. Hälfte des 19. Jahrhunderts. Eine zusätzliche Unbekannte ergibt sich durch das Mikrofilmmaterial. In der Hauptphase I bestand das Ausgangsmaterial für die OCR aus etwa 450 Mikrofilmen mit über 520.000 Einzelaufnahmen. Diese Masterfilme wurden in den Jahren 1995/96 im Auftrag der ULB Sachsen-Anhalt und des Stadtarchivs Halle von verschiedenen externen Dienstleistern erstellt.

2016 änderte sich aufgrund eines Leitungswechsels die strategische Ausrichtung der ULB Sachsen-Anhalt. Neben offenen Standards treten nun auch Open-Source-Technologien an die Stelle existierender, proprietärer Lösungen. Das betrifft auch den Bereich der Texterkennung, wo in den vergangenen Jahren verschiedene Projekte und Communities gewachsen sind, wie z.B. Tesseract-OCR,[4] OCR4all[5] oder das von der DFG geförderte OCR-D-Framework.[6] Im Gegensatz zur Pilotphase von 2014 sollten nun auch die Volltextdaten intern mit Open-Source-Lösungen erstellt werden.

Zur Vorbereitung wurden aus dem Mikrofilmbestand im Jahr 2018 einige Stichproben ausgewählt und basierend auf den Arbeiten von Sommer et al.[7] und den Empfehlungen der DFG zur Digitalisierung evaluiert.[8] Die Ergebnisse waren vielversprechend und gingen in die Antragstellung ein. Da jedoch

2 Sommer, Dorothea; Heiligenhaus, Kay; Wippermann, Carola; Pankratz, Manfred: Zeitungsdigitalisierung. Eine neue Herausforderung für die ULB Halle, in: ABI Technik 34 (2), 2014, S. 75–85.

3 Rice, Stephen; Jenkins, Frank; Nartker, Thomas: The fifth annual test of OCR accuracy. Information Science Research Institute Los Angeles, 1996.

4 Smith, Ray: An overview of the Tesseract OCR engine, in: Ninth international conference on document analysis and recognition (ICDAR 2007), vol. 2, 2007, S. 629–633.

5 Reul, Christian; Christ, Dennis; Hartelt, Alexander; Balbach, Nico; Wehner, Maximilian; Springmann, Uwe; Wick, Christoph; Grundig, Christine; Büttner, Andreas; Puppe, Frank: OCR4all – An open-source tool providing a (semi-) automatic OCR workflow for historical printings, in: Applied Sciences 9 (22), 2019.

6 Engl, Elisabeth: OCR-D kompakt. Ergebnisse und Stand der Forschung in der Förderinitiative, in: Bibliothek Forschung und Praxis 44 (2), 2020, S. 218–230.

7 Sommer: Zeitungsdigitalisierung, 2014.

8 Deutsche Forschungsgemeinschaft: DFG-Vordruck 12.151–12/16 – Praxisregeln „Digitalisierung". 2016. Online: <https://www.dfg.de/formulare/12_151/12_151_de.pdf>.

nur ein Bruchteil des Materials vorweg gesichtet werden konnte, blieb die Frage nach der Qualität für die Masse der Filme ungeklärt. Angesichts dieser Unsicherheit wurde entschieden, die technischen Möglichkeiten des Microfilmscannerssytems „Zeutschel OM 1800“ auszureizen, um 12bit Graustufenscans mit einer physikalischen Auflösung von 470 DPI zu generieren.

5 Pfennig. Kriegs-Zeitung 5 Pfennig.
Vormittags-Ausgabe. Vormittags-Ausgabe.

General-Anzeiger
für Halle und die Provinz Sachsen

Nummer 167 Halle, Sonntag den 18. Juni 1916

Abflauen der russischen Offensive.

Italienische Angriffe überall zurückgeschlagen.

Ein englischer Zerstörer im Kanal gesunken.

Der amtliche österreichische Heeresbericht.

Die blutigen Kämpfe um Hermanik.

Neue Streitkräfte.

Die hingeopferten Muschiks.

Russischer Bericht.

Abflauen der russischen Offensive.

Ein englischer Zerstörer im Kanal gesunken.

Schwierigkeiten bei der italienischen Kabinettsbildung.

Der russische Einfall in Rumänien.

Eine Kriegsgans an der rumänischen Grenze.

Zunahme der Artillerietätigkeit an der Aisne.

Haig — Joffre.

Die Kulturarbeit in Polen.

Abb. 1: Artikelseite J_0100_0001 General-Anzeiger (GA, 18.06.1916)

2.1 Modelle und Training

Im Rahmen der Antragstellung wurde das Open-Source-System Tesseract-OCR in Version 4.0 eingesetzt, im Projektverlauf in den Versionen 4.1.0 – 4.1.3. Ein wesentlicher Punkt für die Wahl von Tesseract-OCR ist die große Community hinter diesem Projekt. Sie stellt im Internet über 300 Sprachmodelle für das System frei zur Verfügung, auch für historische Fraktur-Schriften.[9] Die Anpassung dieser Modelle auf einen bestimmten Teilbereich, d.h. das werkspezifische Feintraining bzw. Finetuning, ist mit geeigneten Modellen und zusätzlichem Trainingsmaterial mit Komponenten von Tesseract-OCR möglich, d.h. ohne zusätzliche Systeme.

Bei Tesseract-OCR handelt es sich um ein selbstständig lernendes System, das auf aufbereiteten Input angewiesen ist. Es benötigt zur Optimierung Referenzdaten. Die Standardmodelle von Tesseract-OCR wurden mit sehr großen Mengen von synthetischen, d.h. künstlichen Daten erstellt.[10] Das Tesseract-OCR-Standardmodell frk für Fraktur-Schriften entstand z.B. aus der Kombination von 202.469 Textzeilen und 4.634 Schriftarten.[11] Adaptionen bestehender Modelle sind bereits mit einem Bruchteil dieses Umfangs möglich, aber zusätzlich ist spezifisches Trainingsmaterial notwendig.

Forschungen im Bereich der Volltexterkennung haben empirisch belegt, dass die Erkennungsgenauigkeit eines OCR-Modells durch Verfeinerung bzw. Finetuning auf ein spezifisches Datenset erhöht werden kann.[12] Aktuelle OCR-Systeme „lernen“ auf Zeilenebene, auf der Grundlage von Paaren, bestehend aus Zeilenbildern plus Text. Das Zeilenbild darf nur die entsprechende Zeile beinhalten. Um diese Paare zu erstellen, müssen mindestens Transkriptionen und geometrische Informationen auf Zeilenebene existieren.

Diese zugrundeliegenden Referenzdaten werden als „groundtruth“ bzw. „groundtruth-Daten“ (GT) bezeichnet. Sie sind definiert als fehlerfreie, manuell korrigierte Text- und Strukturdaten, die sowohl für eine fortwährende, automatisierte Bewertung als auch für das Training von OCR-Modellen für Text-und Layouterkennung verwendet werden.[13] In der Praxis sind fehlerfreie Daten der Idealzustand und die Erstellung ist ein aufwendiger Prozess. In der Realität entstehen diese Daten eher iterativ, daher sollten zusätzliche Kapazitäten für Reviews eingeplant werden. Beim Transkribieren steckt der Teufel im (Zeichen-)Detail. Aus diesem Grund muss die Vorgehensweise klar festgelegt werden. Richtlinien, wie z.B. von OCR-D,[14] sind eine wertvolle Hilfe, um eine einheitliche Bearbeitung zu befördern.

In der Massendigitalisierung ist es vom Aufwand her nicht vertretbar, für jedes Buch separat eine Modelladaption in Betracht zu ziehen. Für eine Menge vergleichbarer Drucke, wie z.B. Zeitungen

9 Tesseract-OCR Trained models, <https://github.com/tesseract-ocr/tessdata>, Stand: 07.01.2022.
10 Smith, Ray: History of the Tesseract OCR engine. What worked and what didn't, in: Document Recognition and Retrieval XX, vol. 8658, International Society for Optics and Photonics, 2013.
11 Tesseract-OCR Fraktur, <https://github.com/tesseract-ocr/langdata_lstm/tree/main/frk>, Stand: 08.06.2022.
12 Reul, Christian; Wick, Christoph; Nöth, Maximilian; Büttner, Andreas; Wehner, Maximilian; Springmann, Uwe: Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning, in: The 6th International Workshop on Historical Document Imaging and Processing, 2021, S. 7–12.
13 Engl: OCR-D kompakt, 2020.
14 Ground Truth Richtlinen, <https://ocr-d.de/de/gt-guidelines/trans/trLevels.html>, Stand: 08.06.2022.

mit tausenden Einzelausgaben, erscheint der Aufwand gerechtfertigt. Um den Trainingsprozess mit Tesseract-OCR zu unterstützen, stellt die Open-Source-Community das Modul tesstrain[15] bereit, das die Aufbereitung der Trainingsdaten übernimmt und das eigentliche Training ausführt.

2.2 Re-OCR

Seit der Jahrtausendwende wurden und werden historische Drucke digitalisiert und mit Volltextdaten angereichert. In der Massendigitalisierung wurde diese „Schmutzige OCR" in Kauf genommen – schlicht, weil sie besser war, als gar keine Volltextdaten und wenigstens als Basis für die Suchindizierung verwendet werden konnte.[16] Dazu kommt, dass sich der Bereich der Volltexterzeugung dynamisch entwickelt. Was vor wenigen Jahren als state-of-the-art galt, ist heute überholt. Diese Dynamik erzeugt allerdings auch zusätzlichen Druck, neue Möglichkeiten nutzbar zu machen. Moderne OCR-System können Modelle untereinander austauschen. Somit ist es möglich, dass Spezialisten in der Community ein neues Modell mit verbesserten Erkennungsraten trainieren und publizieren, was an anderen Stellen eingesetzt werden kann. Um von solchen Fortschritten zu profitieren, müssen Volltextdaten kontinuierlich erneuert werden. Unter dem Label „re-running OCR" (Re-OCR) wurden an der Luxemburgischen Nationalbibliothek (BnL) Workflows zusammengefasst, die vorhandene OCR-Daten nach Möglichkeit verbessern wollen.[17]

3. Bewertung der Volltextqualität

Ein zentraler Punkt, um entscheiden zu können, welchen Effekt ein neues Modell oder ein zusätzliches Training auf den Bestand haben kann, ist eine möglichst exakte Bewertung der Volltextqualität. Um die Qualität der Volltextdaten zu beurteilen, müssen diese methodisch nachvollziehbar ausgewählt und reproduzierbar evaluiert werden.

In der Vergangenheit war man an der ULB Sachsen-Anhalt an einer im Rahmen der Kooperation mit externen Dienstleistern stattfindenden Evaluierung interessiert. Die Wiederverwendbarkeit der für solche einmalige Qualitätsbestimmung erzeugten Daten stand nicht im Fokus. Das war keinesfalls nur in Halle der Fall: Aus Mangel an Ressourcen oder wegen fehlender Vorgaben und Richtlinien finden auch in anderen Einrichtungen oft gar keine oder keine systematischen Qualitätskontrollen statt.[18] Zumindest bei den Vorgaben und Empfehlungen ist Besserung in Sicht. Zuletzt wurden einige Übersichten publiziert, die als Hilfestellung dienen können.[19] In der OCR-Community sind seit den

15 tesstrain, <https://github.com/tesseract-ocr/tesstrain>, Stand: 07.01.2022.

16 Nölte, Manfred; Bultmann, Jan-Paul; Schünemann, Maik; Blenkle, Martin: Automatische Qualitätsverbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift Die Grenzboten, in: o-bib. Das offene Bibliotheksjournal 3 (1), 2016, S. 32–55.

17 Maurer, Yves: Improving the quality of the text, a pilot project to assess and correct the OCR in a multilingual environment, 2017; Schneider, Pit: Rerunning OCR. A Machine Learning Approach to Quality Assessment and Enhancement Prediction, arXiv preprint arXiv:2110.01661, 2021.

18 Schink, Manuela: OCR – Evaluierung der Genauigkeit (QM) sowie Tools zur Unterstützung. Online-Konferenz „OCR–Prozesse und Entwicklungen", 1. März 2021. Online: <https://wiki.zbw.eu/pages/viewpage.action?pageId=33620559&preview=/33620559/33620565/2021-02-24 Schink OCR-Evaluierung und Tools.pdf>, Stand 14.07.2022.

19 Neudecker, Clemens; Baierer, Konstantin; Gerber, Maik; Clausner, Christian; Pletschacher, Stefan; Antonacopoulos, Apostolos: A survey of OCR evaluation tools and metrics, 2021. <http://usir.salford.ac.uk/id/eprint/62335/>, Stand:

1990er Jahren verschiedenen Metriken zur Auswertung bzw. Evaluation von Volltextdaten etabliert. Als Grundlage dient meist der Vergleich von speziellen Referenzdaten, d.h. den bereits erwähnten groundtruth-Daten (GT) mit aktuellen Ergebnissen bzw. Kandidaten (candidate, C) des OCR-Systems. Bei diesem Vergleich werden die Unterschiede bzw. Ähnlichkeiten zwischen groundtruth GT und Kandidat C ermittelt.

Abb. 2: Anzeigenseite J_0050_1024 General-Anzeiger (GA, 02.12.1899)

14.07.2022; Neudecker, Clemens; Zaczynska, Karolina; Baierer, Konstantin; Rehm, Georg; Gerber, Mike; Schneider, Julián Moreno: Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten, in: Qualität in der Inhaltserschließung, Berlin; Boston 2021, S. 137–166.

Da moderne OCR-Systeme Ergebnisse in verschiedenen Strukturtiefen (z.B. Regionen, Zeilen, Wörter, Glyphen) gliedern, können je nach Auswertungsansatz andere Ebenen zugrunde gelegt werden. Eine Evaluation auf Textbasis interpretiert OCR-Daten als fortlaufende Menge visueller Zeichen bzw. Glyphen. Mit dieser Sichtweise können effiziente Verfahren der Computerlinguistik bzw. Textanalyse zur Berechnung der Unterschiede bzw. Ähnlichkeit von Zeichenketten zur Evaluation genutzt werden. Der Nachteil dieser Sichtweise ist, dass Aspekte der räumlichen Orientierung nicht berücksichtigt werden. Abweichungen bei der Erkennung von Positionen und räumlichen Zusammenhängen können so nicht evaluiert werden. Bei Seiten mit einfachem Layout und fixer Leserichtung ist das unproblematisch, bei komplexen Anzeigenseiten nicht.

Die Verwendung von Verfahren der Textlinguistik nach Levenshtein und Ukkonen zur Qualitätsbestimmung für die Zeichengenauigkeit geht zurück auf die Benchmarks von Steven Rice in den 1990er Jahren.[20] Die Differenz zweier Zeichenketten wird als Summe von Ersetzungs-, Ergänzungs- oder Löschoperationen einer Zeichenketten-Distanzfunktion (edit-distance) bestimmt. Diese Operationen bilden den Aufwand ab, den ein Mensch investieren muss, um vom fehlerhaften Kandidatentext zum groundtruth-Äquivalent zu kommen. Fragestellungen mit Schwerpunkt Modellverbesserung interpretieren diese Unterschiede als Fehler. Zusätzliches Training soll die Differenzen verringern, d.h. es geht um die Minimierung von Unterschieden bzw. Fehlerraten (error rate, err).

Im Kontext von Benchmarks für OCR-Systemen steht die Ähnlichkeit bzw. Erkennungsgenauigkeit (accuracy, acc) im Fokus. Je geringer die Differenz von groundtruth GT und Kandidat C, umso ähnlicher sind sie – im Idealfall zu 100 Prozent. Beide Größen, error rate (err) und accuracy (acc) repräsentieren unterschiedliche Perspektiven auf den gleichen Datenbestand. Hinsichtlich acc und err gilt für die Menge der groundtruth-Daten |GT| und der Menge der Differenzen |Δ| von Groundtruth GT und Evaluationskandidat C:

$$\frac{|GT| - \Delta}{|GT|} = acc$$

$$\frac{\Delta}{|GT|} = err$$

$$acc_C + err_C = 1$$

Abb. 3: Verhältnis accuracy (acc) und error rate (err) für Kandidat C

Die Genauigkeit acc ergibt sich aus der Differenz Δ aller Zeichen des Referenztextes und der edit-distance dieser Zeichenkette für einen Kandidatentext C. Solange die Differenz Δ größer 0 ist, also weniger Fehler als gt-Zeichen existieren, wird das Verhältnis zur Anzahl |GT| der groundtruth-Zeichen bestimmt.

Im Detail sind bei textuellen Metriken einige Fallstricke verborgen. Diskutiert wurde u.A., was als Unterschied bzw. Fehler gezählt wird, ob für spezifische Fehlerklassen zusätzlich eine Wichtung erfolgen soll[21] oder ob, und wenn ja, wie Resultate zu normalisieren sind.[22] Diese Fragen haben dazu geführt, dass grundlegende Verfahren

20 Rice, Stephen; Jenkins, Frank; Nartker, Thomas: The fifth annual test of OCR accuracy, 1996.
21 Neudecker: Methoden und Metriken zur Messung von OCR-Qualität, 2021.
22 Wernersson, Maria: Evaluation von automatisch erzeugten OCR-Daten am Beispiel der Allgemeinen Zeitung, in: ABI Technik 35 (1), 2015, S. 23–35.

zur Differenzbestimmung je nach Zielstellung neu umgesetzt wurden,[23] anstatt existierende Implementationen zu nutzen. An der ULB Sachsen-Anhalt wurde ein derartiges Vorgehen vermieden, um die Vergleichbarkeit der Resultate zu verbessern. Im Rahmen des Projektes wurden sowohl für die fortwährende Qualitätsbestimmung, als auch für das Training verschiedene textbasierte Metriken herangezogen. Als Grundlage diente jeweils ein manuell erstellter Referenztext bzw. groundtruth-Datensatz GT und für die Auswertung der aktuelle Output des OCR-Systems als Kandidatentext C. Vor der Anwendung der Metriken werden Referenz- und Kandidatentext aufbereitet. Nicht sichtbare Zeichen (z.B. Leerzeichen) werden entfernt und alle übrigen Zeichen auf eine kanonische Unicode-Darstellung zurückgeführt, was z.B. nicht sichtbare Unterschiede in der Kodierung diakritischer Zeichen bereinigt. Der Wert der textuellen Metriken wird prozentual angegeben. Die Werte aus dem Bereich Information Retrieval liegen zwischen 0 und 1.

3.1 Character Accuracy (CA)

Die Zeichengenauigkeit bzw. Character Accuracy (CA) ist eine elementare textuelle OCR-Metrik. Ihr Gegenstück in Untersuchungen der OCR-Community mit Schwerpunkt Training ist die Character Error Rate (CER). Sie berücksichtigt alle sichtbaren Zeichen eines Textes und setzt eine fixe Textorientierung voraus. Bei der Massendigitalisierung historischer Drucke gilt eine CA von 90% als gut und über 95% als „excellent".[24] Bezogen auf die Untersuchung der historischen Zeitungsbestände der British Library (BL) ergab sich eine CA von 83.6% („19th Century Newspaper Project") bzw. 75.6% („Burnley Collection"),[25] was im Rahmen dieser Untersuchung zu der Einschätzung führte, dass ... anything pre-1900 will be fortunate to exceed 85% accuracy".[26]

3.2 Letter Accuracy (LA)

Die Buchstabengenauigkeit bzw. Letter Accuracy (LA) ist eine Spezialisierung der Zeichengenauigkeit, welche nur Buchstaben im allgemeinen Sinn berücksichtigt. Auch diese Metrik geht auf die Arbeiten von Stephen Rice und Kollegen aus den 1990er Jahren zurück, die zusätzliche Zeichenklassen für Ziffern, Leerzeichen und Groß- bzw. Kleinbuchstaben eingeführt hatten.[27]

Bei DFG-Projekten erfährt diese Metrik eine besondere Aufmerksamkeit, da bei Kooperationen mit externen Dienstleistern eine Buchstabengenauigkeit von 95% möglichst nicht unterschritten werden soll.[28] Allerdings ist die Darstellung dort nicht eindeutig, da für die Prüfung Interpunktionszeichen wieder zu berücksichtigten sind.[29]

23 Wernersson: Evaluation von automatisch erzeugten OCR-Daten, 2015; Nölte: Automatische Qualitätsverbesserung von Fraktur-Volltexten, 2016.
24 Springmann,Uwe; Florian Fink; Klaus Schulz: Automatic quality evaluation and (semi-)automatic improvement of OCR models for historical printings, arXiv preprint arXiv:1606.05157, 2016.
25 Tanner, Simon; Muñoz, Trevor; Ros, Pich Hemy: Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive, in: D-lib Magazine 15 (7/8), 2009.
26 ebd.
27 Rice, Stephen; Jenkins, Frank; Nartker, Thomas: The fifth annual test of OCR accuracy, 1996.
28 Deutsche Forschungsgemeinschaft: DFG-Vordruck 12.151 – 12/16 – Praxisregeln „Digitalisierung". 2016.
29 ebd.

Im Kontext mehrsprachiger Texte mit hunderten unterschiedlichen Zeichen muss sorgfältig ausgewählt werden. Im Rahmen des Projektes wurde an der ULB Sachsen-Anhalt festgelegt, dass neben den regulären ASCII-Interpunktionen und Ziffern zusätzlich arabische, indische und persische Zahlzeichen sowie Interpunktionszeichen des erweiterten UTF-8-Codebereichs für die Evaluation aus den Daten entfernt werden. Bei der Buchstabengenauigkeit findet durch die Reduktion der Zeichenklassen automatisch eine Reduzierung der Fehlerklassen statt. Daher könnte man annehmen, dass die Buchstabengenauigkeit stets höhere Werte als die Zeichengenauigkeit liefert. Dieser Schluss gilt aber nur, wenn die Bezugsgröße unverändert bleibt, d.h. die Zeichenmenge des Referenztextes. Das ist allerdings zu kurz gedacht, da man für eine faire Auswertung auch die Referenzdaten um jene Zeichen bereinigen muss, die man nicht bei den Kandidaten berücksichtigen möchte. So kommt es in der Realität in seltenen Fällen sogar vor, dass ein Kandidatentext eine höhere Zeichen- als Buchstabengenauigkeit hat. Im arithmetischen Mittel lag die Buchstabengenauigkeit (LA) bei den Evaluationsdaten des Projektes ca. 1% über der entsprechenden Zeichengenauigkeit (CA).

3.3 Word Accuracy (WA)

Rice et al. 1996 definierten neben den bereits erwähnten Metriken auch die Wortgenauigkeit bzw. „Word Accuracy“ (WA). Ein Wort im Sinne der Evaluation ist eine Abfolge von einem oder mehreren Zeichen, getrennt durch ein Leerzeichen.[30] Jede Wortform zählt separat. Ebenso gelten Abkürzungen, durch Zeilenumbrüche entstandene Trennungs-Artefakte oder Jahreszahlen ebenso als „Wörter“. Die Charakteristik als Zeichenfolge, als Bruchstücke bzw. „Token“ überträgt den Ansatz der Ähnlichkeit vom Glyph auf die nächsthöhere Ebene, also das Wort. Die Fehlerzahl ergibt sich analog aus den Operationen, um von der Wortfolge des Kandidaten (C) zum korrekten Referenztext (GT) zu kommen. Dabei spielt es keine Rolle, ob mehrere Fehler in einem konkreten Wort enthalten sind. Im Rahmen der BL-Studien ermittelten Tanner et al. 2009 für das „19th Century Newspaper Project“ eine Wortgenauigkeit von 78% und für die „Burnley Collection“ 65%.[31] Entsprechend resümierte Mühlenberger 2011, dass für die Wortgenauigkeit historischer Zeitungen nicht mehr als 80–90% erwartet werden könne.[32]

3.4 Bag of Words (BoW)

Grundlegend für die „Bag of Words“-Metrik (BoW) sind die Arbeiten von Clausner et al. 2015[33] und Pletschacher et al. 2015[34] im Kontext des European Newspaper Project (ENP). Sie wird ebenfalls in den DFG-Empfehlungen zur Zeitungsdigitalisierung erwähnt.[35]

30 Rice: The fifth annual test of OCR accuracy, 1996.
31 Tanner: Measuring Mass Text Digitization Quality, 2009.
32 Mühlberger, Günter: Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Text-und Strukturerkennung (OCR), in: Zeitschrift für Bibliothekswesen und Bibliographie 58 (1), 2011, S. 10–18.
33 Clausner, Christian; Papadopoulos, Christos; Pletschacher, Stefan; Antonacopoulos, Apostolos: The ENP image and ground truth dataset of historical newspapers, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 931–935.
34 Pletschacher, Stefan; Clausner, Chrisian; Antonacopoulos, Apostolos: Europeana newspapers OCR workflow evaluation, in: Proceedings of the 3rd international workshop on historical document imaging and processing, 2015, pp. 39–46.
35 Deutsche Forschungsgemeinschaft: Empfehlungen zur Digitalisierung historischer Zeitungen, 2017.

Fehler in der Layouterkennung führen dazu, dass die Leserichtung gestört ist, obwohl die Wörter an sich korrekt erkannt wurden. Obwohl also der Lesefluss z.T. stark gestört sein kann, bleibt die Verbindung zwischen einer digitalisierten Seite und einem Suchbegriff weiterhin bestehen und das Wort ist über einen entsprechenden Index immer noch zu finden.

Die BoW-Metrik ist unabhängig vom Layout einer Seite bzw. eines Seitenbereichs. Es handelt sich um Wortgenauigkeit ohne inhärente Ordnung. Zur Ermittlung dieser Metrik werden alle Wortformen im Sinne der Wortgenauigkeit des Kandidaten- und Referenztextes gelistet und daraus die Differenz gebildet. Diese Differenz Δ wird in Relation zur Anzahl der GT-Wörter gesetzt. Grundsätzlich gilt hier das gleiche Verhältnis von Ähnlichkeit bzw. Unterschieden als Summe der Fehler analog zu den bisherigen Metriken. Es gilt: WA ≤ BoW, d.h. bei einem Text ohne Layoutprobleme sind WA und BoW identisch.

Bei Zeitungen sind komplexe Layouts allgegenwärtig. Bei den Analysen zum ENP wurde für reine Gotik- bzw. Frakturschriften ein BoW-Wert von 67.3% erreicht.[36] Für Mischungen von Schriftarten, welche sehr häufig im Kontext von Anzeigenseiten auftreten, sank dieser Wert auf 64.0%.[37]

3.5 Dictionary Metric (DM)

Ein erster Vorschlag, Wörterbücher systematisch für die Qualitätseinschätzung bzw. Quality Prediction zu verwenden, stammt von Alex et al. 2014 unter dem Begriff „Simple Quality".[38] Dieser Ansatz ist seither unter verschiedenen Bezeichnungen aufgetaucht, z.B. als „Words in Dictionary"[39], „Dictionary Method"[40] oder „Dictionary Mapping".[41] Allen Bezeichnungen gemein ist der Abgleich von Wortformen eines Kandidatentextes C mit einer vorhandenen Wortliste bzw. einem Wörterbuch. Im Rahmen der Evaluation des Zeitungsprojektes der ULB Sachsen-Anhalt wird diese Kennzahl als „Dictionary Metric" (DM) bezeichnet. Dieser Ansatz benötigt keine zusätzlichen Referenzdaten. Es sind keine spezifischen groundtruth-Daten erforderlich, stattdessen ein passendes Wörterbuch oder eine Wortliste. Diese Metrik behandelt einen Text wie eine Wortmenge und untersucht, wie viele dieser Wörter bzw. Wortbestandteile in einem Wörterbuch auffindbar sind. Nonsens-Kombinationen durch schlechte OCR bilden keine sinnvollen Wörter. Sie sind weder im Wörterbuch zu finden noch werden sie von einer Rechtschreibprüfung akzeptiert. Eigennamen, Abkürzungen, historische und regionale Schreibvarianten stellen allerdings eine konzeptionelle Hypothek dieser Metrik dar, da sie in einem normalen Wörterbuch bzw. Nachschlagewerk fehlen. Aus einem Text entsteht eine Menge von Wortformen |W|, die Dopplungen enthalten kann.

36 Pletschacher: Europeana newspapers OCR workflow evaluation, 2015.
37 ebd.
38 Alex, Beatrice; Burns, John: 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, pp. 97–102.
39 Clausner, Christian; Papadopoulos, Christos; Pletschacher, Stefan; Antonacopoulos, Apostolos: Quality prediction system for large-scale digitisation workflows, in: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), IEEE, 2016, pp. 138–143.
40 Maurer: Improving the quality of the text, 2017.
41 Schneider: Rerunning OCR, 2021.

$$DM_C = \begin{cases} 0, & \text{if } (|GT| - \Delta) < 0 \\ \frac{|GT| - \Delta}{|GT|} * 100, & otherwise \end{cases}$$

Abb. 4: Dictionary Metric (DM) für Kandidate C

Jeder nicht gefundene Eintrag bzw. Rechtschreibfehler wird gezählt. Diese Summe wird von der Gesamtzahl gefundener Wortformen abgezogen und das Ergebnis anschließend in Relation zu dieser Gesamtzahl gesetzt, analog wie bei den anderen Metriken. Der wesentliche Nachteil dieser Metrik ist, dass aufgrund des fehlenden Referenztextes nicht mit Sicherheit davon ausgegangen werden kann, dass die gefundenen Wörter tatsächlich im Text vorkommen. Ebenso bedeutet ein schlechter Wert nicht automatisch, dass die Seite schlecht erkannt wurde. Z.B. kann ein Anzeigentext mit vielen Abkürzungen und Eigennamen einen niedrigen DM-Wert erhalten, auch wenn die „Wörter" an sich korrekt erkannt wurden.

3.6 Information Retrieval (IR)

In der modernen, vernetzten Wissenschaftswelt sind Stichwortsuchen in großen Datenbeständen ein längst nicht mehr wegzudenkendes Hilfsmittel. Entsprechend erscheint es sinnvoll, die Qualität von Volltextdaten auch mit Kennzahlen aus dem Bereich des klassischen Information Retrieval (IR) mit Metriken wie Precision (PRE) oder Recall (REC) zu bewerten.[42] Es existieren verschiedene Vorschläge, die bereits vorgestellte Wortgenauigkeit so zu spezialisieren, dass sie auf Stichwörter einer Online-Recherche zielt. Auch hier stammen erste Anregungen unter der Bezeichnung „Non-stopword accuracy" von Rice et al. 1996.[43] In den Untersuchungen der British Library (BL) versuchten später Tanner et al., diese begrifflich mit den Labeln „Significant word accuracy" bzw. „Significant words with capital letter start accuracy" zu erfassen.[44]

Im Rahmen von Massendigitalisierungsprojekten ist es nicht praktikabel, für jede einzelne Seite Stichwörter manuell zu annotieren. Im Englischen kann der Umstand ausgenutzt werden, dass Eigennamen innerhalb eines Satzes mit einem Großbuchstaben beginnen. Im Deutschen trifft das aufgrund der im Vergleich zum Englischen komplizierteren Regeln für Groß- u. Kleinschreibung nicht zu. Im Rahmen der Evaluation von Zeitungsdaten wird aus allen Wörtern automatisch ein Set erstellt. Jedes Wort bzw. Token kommt in diesem Set genau einmal vor – egal, wie häufig es tatsächlich auf der Seite enthalten ist. Dieses Set wird vor der Auswertung bereinigt. Dabei werden häufig auftretende Wortformen mit geringem Informationswert entfernt, sogenannte „Stopwörter" (stop words). Dabei handelt es sich um Ausdrücke wie z.B. Artikel, Präpositionen und Pronomen, die sehr wahrscheinlich nicht den Gegenstand einer Suchanfrage bilden oder aufgrund ihres häufigen Vorkommens nicht beim Auffinden relevanter Informationen helfen können. Dieser Schritt wird mit dem Referenztext GT und dem Kandidatentext C durchgeführt.

42 Manning, Christopher; Raghavan, Prabhakar; Schütze, Heinrich: Introduction to Information Retrieval, Cambridge 2008.
43 Rice: The fifth annual test of OCR accuracy, 1996.
44 Tanner: Measuring Mass Text Digitization Quality and Usefulness, 2009.

4. Anwendung

4.1 OCR-Pipeline

Für die Inhouse-Volltexterzeugung wurde zum Start der Hauptphase I an der ULB Sachsen-Anhalt auf einen hochgradig automatisierten Workflow gesetzt. Die OCR-Pipeline verwaltet einen Pool von Tesseract-OCR-Instanzen und verteilte die über ein Austauschlaufwerk eintreffenden Bilddatenpakete nacheinander auf diese Instanzen. Ein einzelnes Paket entsprach im Projekt einem kompletten Mikrofilm mit durchschnittlich 1.100 Aufnahmen.

Für jedes zu verarbeitende Bild wird parallel ein separater Durchgang gestartet. Bei schwerwiegenden Fehlern, die zum Fehlschlag des Prozesses führen, ist somit nur das aktuelle Bild betroffen und nicht das komplette Paket. Auf einer einzelnen Maschine mit 14 CPUs erzeugte dieses System beispielsweise mit 12 Tesseract-OCR-Instanzen Volltext für etwa 4.500 Zeitungsseiten pro Tag. Optional kann als Teilschritt dieser Pipeline ein Container mit dem Textkorrektur-Service Language Tool[45] ausgeführt werden, der eine Rechtschreibprüfung gegen ein aktuelles deutsches Wörterbuch durchführt. Dieser Schritt ermittelt für jede Seite in der Pipeline die Dictionary Metric (DM) auf die oben dargestellte Art und Weise.

Die Implementation wurde mittlerweile publiziert.[46] Für den metrischen Teil inklusive der stop words und den statistischen Teil der Evaluation kommen die Open-Source-Frameworks „rapidfuzz“[47], „nltk“[48] und „numpy“[49] aus dem Python-Ökosystem zum Einsatz.

4.2 Erstellung von Referenzdaten

Die Erstellung von groundtruth-Daten ist aufwendig, aber als Basis für die Erstellung kann der aktuelle Output des OCR-Systems genutzt werden. Sofern die Erkennung nicht dramatisch versagt hat, ist das Korrigieren solcher Daten schneller, als von Null zu beginnen. Ein weiterer, sehr wichtiger Punkt sind die zur Verfügung stehenden Werkzeuge. Ein Tool, das Bild und OCR-Daten kombiniert darstellen kann, stellt bei komplexen Layouts eine erhebliche Vereinfachung der Arbeitsabläufe dar. Für die Korrektur der groundtruth-Daten wurde im Verlauf des Projektes ab 2020 mit einer modifizierten Version des Open-Source-Expert-Clients „Transkribus-SWT“[50] gearbeitet, welche zusätzlich den Import des Tesseract-OCR-Formates mit verschachtelten Regionen ermöglicht.[51] Nach kurzer Einarbeitung können auch Nicht-Expert*innen damit arbeiten.

45 Language-Tool, <https://github.com/languagetool-org/languagetool>, Stand: 03.05.2022.
46 ocr-pipeline, <https://github.com/ulb-sachsen-anhalt/ocr-pipeline>, Stand: 08.08.2022.
47 rapidfuzz, <https://github.com/maxbachmann/RapidFuzz>, Stand: 07.11.2022.
48 Natural Language Toolkit, <https://www.nltk.org/>, Stand: 08.08.2022.
49 Numeric Python, <https://numpy.org/>, Stand: 07.11.2022.
50 Kahle, Philip; Colutto, Sebastian; Hackl, Günter; Mühlberger, Günter: Transkribus – a service platform for transcription, recognition and retrieval of historical documents, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 4, IEEE, 2017, pp. 19–24.
51 Transkribus-SWT, <https://github.com/ulb-sachsen-anhalt/transkribus-swt-gui>, Stand: 15.12.2022.

4.2.1 Auswahlkriterien

Vorrangiges Ziel der Evaluation war neben der Überwachung des Trainings die fortwährende Prüfung der Qualität des erzeugten Volltextes. Die Einschätzung der Pilotphase 2014 beruhte auf einer repräsentativen Stichprobe von 40 Samples aus einer Vorauswahl der Jahrgänge 1800, 1801, 1832, 1833, 1856, 1868, 1872 und 1876.[52] Repräsentativ bedeutet, dass alle vorhandenen Formate in die Stichprobe integriert waren.

Repräsentativität ist nicht das einzige Auswahlkriterium für Referenzdaten. In der British Library (BL) wurde 2009 grundsätzlich ebenfalls mit Stichproben gearbeitet, allerdings in anderen Dimensionen. Dort wurde ein Sample erstellt, welches ca. 1% der insgesamt 2 Millionen Seiten umfasste, also insgesamt ca. 20.000 Seiten. Auf diesen Seiten wurden anschließend zwei kleine Regionen mit ca. 100 Wörtern ausgewählt. Dabei wurden gezielt Bereiche mit gehobener Qualität ausgesucht. Den Autoren ging es nicht um Repräsentativität, sondern um eine Abschätzung der bestmöglichen Qualität nach oben: „Segments were selected on the basis of being amongst the clearest on the page image".[53] Die Datengrundlage wurde mit Bedacht nicht zufällig zusammengestellt. Die Untersuchungen zum European Newspaper Project (ENP) basieren auf einem Sample von 528 Images (bzw. 46.889 Regionen oder 202.524 Zeilen) aus einem Pool von ca. 10 Millionen Zeitungsseiten.[54]

In der Hauptphase des Zeitungsprojektes der ULB Sachsen-Anhalt mit über einer halben Million Seiten konnten ebenfalls nur Stichproben erhoben werden. Um trotzdem eine möglichst objektive und repräsentative Auswahl zu treffen, wurde ein Intervall von 25 Filmen festgelegt, welches etwa 10 Jahrgänge einer Zeitung umfasste. Von jedem 25. Film wurden anschließend die erste und letzte Seite mit Inhalt und dazu jede 256. Seite gewählt, unabhängig vom Materialzustand.

Zu Projektbeginn waren weder Erfahrungswerte über die erforderliche Zeitdauer zur Generierung der Volltextdaten vorhanden, noch für die Erstellung von Referenzdaten. Darum stand die Befürchtung im Raum, dass die Korrektur der groundtruth-Daten nicht mit dem Scannen der Mikrofilme Schritt halten könne.

Um dem zu begegnen, wurde entschieden, nur Ausschnitte von einzelnen Seiten als groundtruth-Daten zu transkribieren, um die Samplegröße einzuhalten. Es entstanden zusammenhängende, rechteckige Teilbereiche (1/2 Seite, 1–4 Spalten) mit 2.700 bis 10.000 Zeichen (ohne Leerzeichen). In der Tendenz gilt, dass ältere Ausgaben aus dem 19. Jahrhundert kleinere Formate mit weniger Umfang bilden.

52 Sommer: Zeitungsdigitalisierung, 2014.
53 Tanner: Measuring Mass Text Digitization Quality and Usefulness, 2009.
54 Clausner: The ENP image and ground truth dataset of historical newspapers, 2015.

4.3 Evaluation der Projektdaten

Grundlage für die Evaluierung der Datensätze bildet das Tool „digital-eval“, das ebenfalls von der ULB Sachsen-Anhalt für die Allgemeinheit publiziert wurde.[55] Es implementiert die vorgestellten Metriken, aggregiert Datensätze nach verschiedenen Kriterien und nutzt die Möglichkeiten der oben genannten Open-Source-Komponenten für die Datenaufbereitung, Kalkulation und zur Ermittlung deskriptiver statistischer Größen wie Median und Standardabweichung.

Im Projektverlauf zeichnete sich ab, dass einige Seiten bereits mit dem Tesseract-Standardmodell für Fraktur frk respektable Resultate erzielen, andere jedoch nicht. Hier zeigte sich der Einfluss des Materials durch die strikte Auswahlmethode. Um solche Probleme einzuordnen, wurde für die Evaluation eine an Data Science-Verfahren angelehnte Ausreißer-Ermittlung (Outlier-Detection) mittels Interquartilsabstand implementiert.[56] Wenn Ausreißer identifiziert werden, wird eine zweite Evaluation ohne diese Werte durchgeführt. Das Bereinigen führt dazu, dass die Standardabweichung der Messwerte zurückgeht. Diese Größe gibt an, wie weit Datenwerte im Messbereich verteilt sind. Je geringer die Streuung, umso homogener die Werte. Aussagen zur bereinigten Qualität beziehen sich nicht mehr auf alle Daten, sondern auf solche, die aufgrund ihrer Eigenschaften entsprechende Ergebnisse liefern können. Bei den durchgeführten Evaluationen wurden über den Interquartilsabstand fast immer Ausreißer im unteren Bereich ermittelt und ausgeschlossen. Das Entfernen dieser Outlier führt in allen Fällen zu einem deutlichen Absinken der Standardabweichung und zu einem Anstieg der Mittel- und Medianwerte.

4.3.1 Referenzcorpus ZD1_25

Als Resultat der Bemühungen zur groundtruth-Datenerstellung entstand das Referenzcorpus „ZD1_25“. Die Daten folgen der Auswahlmethodik, aus jedem 25.Film Samples zu entnehmen, durchgängig für die Zeitungen „General-Anzeiger“ (GA, 1889–1918) und den Nachfolger „Hallische Nachrichten“ (HN, 1918–1943). Zu diesem Set gehören 39 Datensätze mit insgesamt 196.000 Zeichen. Zwei Drittel sind Artikelseiten (124.000 Zeichen ohne Leerzeichen) und ein Drittel Anzeigenseiten (72.000). Jedes Item geht gleichwertig in die Berechnung ein.

Da der Schwerpunkt auf Artikeln lag, wurden an den Dateinamen zusätzliche Informationen annotiert, um Artikelseiten automatisch zu identifizieren. Es entstanden schließlich zwei Klassen: Artikel- (art) und Anzeigen- und Ankündigungsseiten bzw. Announcementseiten (ann).

Jeder groundtruth-Datensatz GT wird anschließend bei der Evaluation einer dieser Klassen zugeordnet. Durch das Auswahlverfahren war von jedem Film mindestens eine Titelseite mit Artikeln art und eine Rückseite mit Anzeigen und Ankündigungen ann enthalten. Die Werte der Metriken wurden mit der im Projektrahmen entwickelten Evaluation automatisch nach den Klassen art und ann und hierarchisch nach Jahrgängen und der übergeordneten Zeitung zusammengefasst.

55 digital-eval, <https://github.com/ulb-sachsen-anhalt/digital-eval>, Stand: 08.08.2022.
56 Interquartile Range, <https://en.wikipedia.org/wiki/Interquartile_range>, Stand: 05.01.2022.

4.3.2 Metriken und Modelle

Für die Qualitätsbestimmung und das fortlaufende Training wurden mit unterschiedlichen Tesseract-OCR Modellkonfigurationen Volltextdaten erzeugt und mit dem Referenz-Corpus ZD1_25 verglichen.

item	Typ	$CA(N_{Cs})$	$LA(N_{Ls})$	$WA/BoW(N_{Ws})$	$PRE/REC(N_{Ts})$
001_0002	art	26.56(4662)	27.64(4500)	3.08/71.72(778)	0.66/0.70(413)
001_0256	ann	75.84(3530)	80.95(3280)	24.54/80.26(542)	0.66/0.75(333)
001_0512	art	98.49(5640)	99.10(5358)	93.09/94.82(926)	0.93/0.92(525)
001_0768	art	15.36(5105)	19.46(4877)	0.00/46.92(876)	0.17/0.31(475)
001_0938	ann	89.15(2737)	90.22(2445)	75.88/90.00(510)	0.82/0.87(301)

Tab. 1: Detail General-Anzeiger (GA), Tesseract 4.1.3, Modell ulbhalzd1_22k

Fällt eine Metrik ab, dann verhalten sich die anderen ebenso. Eine Ausnahme bilden Wortgenauigkeit (WA) und Bag-of-Words (BoW). Sie können über viele Prozentpunkte auseinander liegen. Dieser Unterschied resultiert aus einer fehlerhaften Layouterkennung. Nur bei einem von 16 Datensätzen des Typs ann beträgt der Unterschied von WA zu BoW weniger als 10 Prozentpunkte, für 10 Datensätze sogar mehr als 20 Punkte. Somit schlägt sich in mehr als 90% der Fälle das komplexe Anzeigenlayout auf Wortebene unmittelbar in der Erkennungsqualität nieder.

Bei Artikelseiten bzw. Seitenfragmenten mit verhältnismäßig homogenen Spalten zeigt sich dieser Effekt ebenfalls. Hier existieren nur 3 Datensätze, bei denen WA und BoW in ähnlichem Grad wie bei Anzeigen divergieren (≈13%). Anzeigenseiten leiden an Layoutproblemen und werden insgesamt deutlich schlechter erkannt als Artikel (–22.9 Prozentpunkte). Die Ergebnisse können mit dem im Projektrahmen entwickelten Evaluationstool inhaltlich nach Artikel- und Anzeigensamples und hierarchisch nach Jahrgängen und Zeitungen zusammengefasst werden.

	$\overline{LA}$	
LA@zd1_25	78.20	39 items, 183297 refs, std: 25.95, median: 93.04
LA@zd1_25	88.50	36 items (-3), 170821 refs, std: 21.48, median: 93.87
LA@zd1_25/0001	63.47	5 items, 20460 refs, std: 33.20, median: 80.95
LA@zd1_25/0025	85.39	5 items, 23636 refs, std: 8.58, median: 86.70
LA@zd1_25/0050	72.35	5 items, 18253 refs, std: 16.64, median: 69.79
LA@zd1_25/0075	71.96	4 items, 22846 refs, std: 29.00, median: 81.44
LA@zd1_25/0100	95.36	5 items, 27615 refs, std: 1.69, median: 95.55
LA@zd1_25/0125	64.73	6 items, 27113 refs, std: 32.78, median: 67.11
LA@zd1_25/0150	91.87	4 items, 17361 refs, std: 10.46, median: 97.09
LA@zd1_25/0175	84.67	5 items, 26013 refs, std: 26.68, median: 97.37
LA@zd1_25@ann	65.13	16 items, 63903 refs, std: 22.92, median: 69.87
LA@zd1_25@art	87.30	23 items, 119394 refs, std: 23.98, median: 97.01
LA@zd1_25@art	97.17	19 items (-4), 102031 refs, std: 1.88, median: 97.37

Tab. 2: Buchstabengenauigkeit (LA), Tesseract 4.1.3, Modell ulbhalzd1_22k

Die Streuung Buchstabengenauigkeit (LA) über die Stichprobe ist für Anzeigen so groß, dass dort über den Interquartilsabstand keine Ausreißer ausgeschlossen werden. Die Standardabweichung für Artikel sinkt durch das Nicht-Berücksichtigen der 4 Extremwerte deutlich von 23.67 auf 2.09 Prozentpunkte.

Für die Dictionary Metric (DM) gilt, dass sie nur eine Abschätzung und keine exakte Quantifizierung ermöglicht. Deutlich wird das bei einer Evaluation der groundtruth-Daten. Es gilt, dass Bezeichnungen, die dem System unbekannt sind, als Fehler gewertet werden. Besonders häufig tritt dieses Phänomen bei Anzeigenseiten auf, welche ein Konzentrat von Eigennamen (Personen, Ortschaften, Adressen) und Abkürzungen bilden.

item	DM	t
1667522809_J_0001_0002.gt.art	91.327	22.95s
1667522809_J_0001_0256.gt.ann	78.656	26.39s
1667522809_J_0001_0512.gt.art	89.848	24.76s
1667522809_J_0001_0768.gt.art	90.717	24.20s
1667522809_J_0001_0938.gt.ann	78.759	23.80s

Tab. 3: Ausschnitt Dictionary Metric (DM) und Laufzeit t für GT, Tesseract 4.1.3

Wie der Tabelle 3 zu entnehmen ist, besteht ein signifikanter Unterschied zwischen Artikeln und Anzeigen. Über den kompletten Corpus „ZD1_25“ mit 39 Samples betrachtet bleibt festzustellen, dass die 10 schlechtesten Dictionary Metric-Werte allesamt Anzeigenseiten erzielen. Der niedrigste Wert im Set (57.95) wird von einer Anzeigenseite markiert, der beste (97.16) von einer Artikelseite.

Ein hoher DM-Wert korrespondiert mit einer guten OCR. Umgekehrt gilt diese Beobachtung nicht, da es speziell im Zeitungsbereich Seiten gibt, die mit Wörterbüchern oder Wortlisten schwer zu fassen sind (z.B. Anzeigen, Börsenlisten, Fahrpläne).

4.4 Training

Tesseract-OCR ermöglicht das Weitertraining vorhandener Modelle. Zunächst wurden im Rahmen des Projektes Volltextdaten auf Basis eines vorhandenen Standard-Modells generiert. Dieses Modell sollte anschließend mit zusätzlichen Trainingsdaten adaptiert werden, um schließlich mit einer an realen Zeitungsdaten angepassten Version eine komplette Re-OCR über alle Seiten des Projektes durchzuführen. Für das Training lag der Fokus auf Artikelseiten art. Die beste Performance zeigte ein Modell, welches auf Basis des Modells Fraktur_5000k[57] mit 16.000 Bild-Text-Zeilenpaaren für 22.000 Iterationen verfeinert wurde. Dieses angepasste Modell konnte die bereits hohen Resultate inklusive Outlier des Ausgangsmodells für Artikel noch einmal um 1.69 (CA) bzw. 1.44 (LA) Prozentpunkte steigern. Die erstellten Trainingsdaten und das Tesseract-OCR Modell, welches für die Re-OCR zum Einsatz kam, stehen über das GitHub-Repository der ULB Sachsen-Anhalt zur Verfügung.[58]

57 Fraktur_5000k, <https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/Fraktur_5000000/tessdata_best/>, Stand: 28.06.2022.

58 Training Data Zeitungsdigitalisierung HP I, <https://github.com/ulb-sachsen-anhalt/ulb-zeitungsprojekt-hp1>, Stand: 07.01.2022.

Modell	$\overline{CA_0}/\overline{CA_1}$	$\sigma_{CA_0}/\sigma_{CA_1}$	$\overline{LA_0}/\overline{LA_1}$	$\sigma_{LA_0}/\sigma_{LA_1}$
frk	83.92 / 94.47	26.03 / 2.17	85.04 / 94.76	23.66 / 2.09
Fraktur_5000k	84.91 / 94.77	24.11 / 2.39	85.86 / 95.54	23.52 / 2.37
frak2021	85.68 / 95.80	24.77 / 2.05	86.85 / 96.60	23.68 / 2.01
ulbdhz1_22k	86.60 / 96.66	24.51 / 1.88	87.30 / 97.10	23.98 / 1.88

Tab. 4: Modellvergleich CA und LA für Klasse art, Tesseract 4.1.3

Bemerkenswert ist die Performance des Modells frak2021, welches im Verlauf des Jahres 2021 von Stephan Weil an der UB Mannheim publiziert wurde.[59] Damit können bessere Ergebnisse als mit dem Vorgängermodell Fraktur_5000k erzielt werden wobei es nur die Hälfte der Zeit benötig.

4.5 Re-OCR

Für den initialen Workflow im Zeitungsdigitalisierungsprojekt I, bei dem Filme nacheinander im Mikrofilmscanner eingelesen wurden, war die Geschwindigkeit aufgrund der Projektlaufzeit mit einem einzigen OCR-System völlig ausreichend. Für die Re-OCR im Rahmen des Projektes wurden drei virtuelle Maschinen mit 36 Tesseract-OCR-Instanzen eingesetzt. Somit war es möglich, über 500.000 Seiten in nur 2 Monaten mit dem verbesserten Modell ulbdhz1_22k zu verarbeiten.

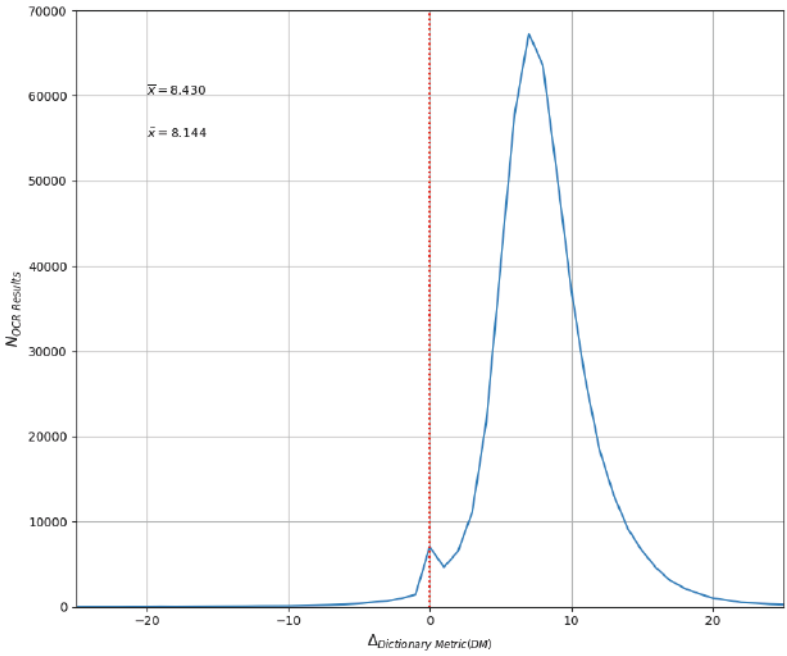

Abb. 5: Steigerung Dictionary Metric (DM) durch Re-OCR

59 Frak2021, <https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/frak2021/tessdata_best/>, Stand: 28.06.2022.

Im Zuge des zweiten Durchlaufs wurde abermals die Dictionary Metric (DM) für jede Seite erhoben. Deren Werte lassen auf eine signifikante Verbesserung des erzeugten Volltextes schließen. Der Anstieg übertrifft die Werte, welche nach den fortwährenden Evaluationen gegen das „ZD1_25"-Set des trainierten Modells erwartbar waren. Ein möglicher Erklärungsansatz berücksichtigt die unterschiedlichen Bild-Formate. Die ersten Volltextdaten wurden direkt im laufenden Digitalisierungs-Workflow mit unkomprimierten TIFF-Daten erzeugt. Mit diesem Material wurden das Finetuning des Modells und die fortwährenden Evaluationen der angepassten Modelle durchgeführt. Für die neuen OCR-Daten wurden dagegen JPG-Images über die Webschnittstelle geladen. Diese Bilder haben identische Abmessungen und belegen nur noch etwa 1/3 des Speicherplatzes bei minimal reduzierter Qualität (95%). Diese Reduktion wird über Komprimierungs-Algorithmen erreicht, die vermutlich eine ähnliche Wirkung wie Verfahren zur Entfernung von Bildstörungen beim Aufbereiten der Bilddaten (Denoising) haben.

5. Zusammenfassung und Ausblick

Eine hochwertige OCR ist bei der Massendigitalisierung historischer deutscher Zeitungen des 19. und 20. Jahrhunderts mit Open-Source-Komponenten möglich. Eine Voraussetzung dafür sind Workflows mit hohem Automatisierungsgrad. Dazu kommt, dass die Generierung von Volltextdaten mit Prozessen des Digitalisierungs-Workflows verknüpft sein muss, damit der neue Volltext in den Bestand zurück fließt. Es entsteht zusätzlicher Aufwand, um Daten über verschiedene Bereiche hinweg aktuell zu halten. PDFs für Digitalisate müssen erneuert werden, ebenso wie die entsprechenden Einträge in Such-, Präsentations- und Archivierungssystemen. Obliegt die Erstellung von Volltext externen Dienstleistern, fallen für eine kontinuierliche OCR-Generierung von Beständen, die Millionen von Seiten umfassen, auch erhebliche Kosten an. Kommerzielle Anbieter wie ABBYY rufen z.B. für ein reguläres A4-Format 14 Cent ab bzw. das 4fache für eine großformatige Zeitungsseite.[60]

Durch die Vereinigung von Plattform und Organisation eröffnen sich durch die digitale Transformation für Bibliotheken neue Möglichkeiten. Mit den entsprechenden Kompetenzen können die Einrichtungen Adaptionen von OCR-Modellen vornehmen. Die Erstellung zusätzlichen Trainingsmaterials ist sehr aufwendig und kann von kleineren Einrichtungen nur in bestimmten Anwendungsszenarien oder Kooperationen geleistet werden. Allerdings bedeutet eine Adaption immer zusätzlichen Aufwand. Wer diesen scheut oder nicht über entsprechende Kapazitäten verfügt, kann durch ein anderes Erkennungsmodell den Volltextbestand verbessern.

Sehr viel einfacher gestaltet sich die erneute Generierung von OCR-Daten, sofern entsprechende Workflows existieren. Das Beispiel mit dem neuen Tesseract-OCR-Modell frak2021 illustriert, wie eine Bibliothek an aktuellen Weiterentwicklungen der OCR-Community partizipieren könnte.

Um den Effekt eines Modellwechsels beurteilen zu können, bilden die hier vorgestellten Metriken eine wichtige Grundlage. Für detaillierte Messungen sind allerdings groundtruth-Daten erforderlich. Ohne diese Daten fehlt die Grundlage einer reproduzierbaren Auswertung.

60 Deutsche Forschungsgemeinschaft: Empfehlungen zur Digitalisierung historischer Zeitungen, 2017.

Ebenso ist der Einsatz von Open-Source-Komponenten bei der Evaluation möglich. Die Verwendung offener Komponenten mit tausendfach erprobten Standardimplementationen verringert den Aufwand eigener Evaluationsverfahren erheblich.

Um die Vergleichbarkeit von Maßzahlen zu verbessern, müssen für Metriken und die Erstellung der groundtruth-Daten klare Vorgaben existieren und umgesetzt werden. Punktuelle statistische Qualitätsprüfungen, die mit sehr elaborierten, aber oft auch sehr spezifischen Auswertungsverfahren aufwarten, erschweren die Vergleichbarkeit über Projekt- und Einrichtungsgrenzen hinweg. Hier gilt es, den Datenaspekt stärker in den Fokus zu stellen. Die Erstellung strukturierter groundtruth-Daten bleibt aufwendig, aber ihre Nachnutzbarkeit macht diese Anstrengungen wieder wett.

Konforme Referenzdaten können leichter in anderen Kontexten genutzt werden: Nicht nur in Folgeprojekten der Bibliothek, wo sie entstanden sind, sondern z.B. auch für die Optimierung und Verbesserung der eingesetzten OCR-Systeme innerhalb der Open-Source-Community.

Auf diese Weise entsteht ein Mehrwert für alle Anwender dieser offenen Systeme.

Literaturverzeichnis

- Alex, Beatrice; Burns, John: Estimating and rating the quality of optically character recognised text, in: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 97–102, 2014. Online: <https://dl.acm.org/doi/pdf/10.1145/2595188.2595214>.
- Clausner, Christian; Papadopoulos, Christos; Pletschacher, Stefan; Antonacopoulos, Apostolos: The ENP image and ground truth dataset of historical newspapers, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 931–935, IEEE, 2015. Online: <https://dl.acm.org/doi/10.1109/ICDAR.2015.7333898>.
- Clausner, Christian; Papadopoulos, Christos; Pletschacher, Stefan; Antonacopoulos, Apostolos: Quality prediction system for large-scale digitisation workflows, in: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 138–143, IEEE, 2016. Online: <https://www.primaresearch.org/www/assets/papers/DAS2016_Clausner_QualityPrediction.pdf>, Stand: 10.11.2022.
- Deutsche Forschungsgemeinschaft: DFG-Vordruck 12.151 – 12/16 – Praxisregeln „Digitalisierung“. 2016. Online: <https://www.dfg.de/formulare/12_151/12_151_de.pdf>, Stand: 10.11.2022.
- Deutsche Forschungsgemeinschaft: Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland, 2017. Online: <https://zeitschriftendatenbank.de/fileadmin/user_upload/ZDB/z/Masterplan.pdf>, Stand: 10.11.2022.
- Engl, Elisabeth: OCR-D kompakt: Ergebnisse und Stand der Forschung in der Förderinitiative, in: Bibliothek Forschung und Praxis 44 (2), 2020, S. 218–230. Online: <https://www.degruyter.com/document/doi/10.1515/bfp-2020-0024/pdf>, Stand: 10.11.2022.
- Kahle, Philip; Colutto, Sebastian; Hackl, Günter; Mühlberger, Günter: Transkribus – a service platform for transcription, recognition and retrieval of historical documents, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 4, pp. 19–24, IEEE 2017.

- Manning, Christopher; Raghavan, Prabhakar; Schütze, Heinrich: Introduction to Information Retrival, Cambridge 2008.
- Maurer, Yves: Improving the quality of the text, a pilot project to assess and correct the OCR in a multilingual environment, 2017. Online: <https://slub.qucosa.de/api/qucosa%3A16445/attachment/ATT-0/>, Stand: 10.11.2022.
- Mühlberger, Günter: Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Text- und Strukturerkennung (OCR), in: Zeitschrift für Bibliothekswesen und Bibliographie 58 (1), 2011, S. 10–18.
- Neudecker, Clemens; Zaczynska, Karolina; Baierer, Konstantin; Rehm, Georg; Gerber, Mike; Schneider, Julián Moreno: Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten, in: Qualität in der Inhaltserschließung, Berlin; Boston 2021, S. 137–166. Online: <https://pdfs.semanticscholar.org/09ce/7181d7751cfc05365039475b7432f89afcfd.pdf>, Stand: 10.11.2022.
- Neudecker, Clemens; Baierer, Konstantin; Gerber, Maik; Clausner, Christian; Pletschacher, Stefan; Antonacopoulos, Apostolos: A survey of OCR evaluation tools and metrics, 2021. Online: <https://dl.acm.org/doi/pdf/10.1145/3476887.3476888>.
- Nölte, Manfred; Bultmann, Jan-Paul; Schünemann, Maik; Blenkle, Martin: Automatische Qualitätsverbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift Die Grenzboten, in: o-bib. Das offene Bibliotheksjournal 3 (1), 2016, S. 32–55. Online: <https://doi.org/10.5282/o-bib/2016H1S32-55>.
- Pletschacher, Stefan; Clausner, Chrisian; Antonacopoulos, Apostolos: Europeana newspapers OCR workflow evaluation, in: Proceedings of the 3rd international workshop on historical document imaging and processing, 2015, pp. 39–46. Online: <https://dl.acm.org/doi/pdf/10.1145/2809544.2809554>.
- Reul, Christian; Christ, Dennis; Hartelt, Alexander; Balbach, Nico; Wehner, Maximilian; Springmann, Uwe; Wick, Christoph; Grundig, Christine; Büttner ,Andreas; Puppe, Frank: OCR4all–An open-source tool providing a (semi-)automatic OCR workflow for historical printings, in: Applied Sciences 9 (22), 2019. Online: <https://www.mdpi.com/2076-3417/9/22/4853/htm>, Stand: 10.11.2022.
- Reul, Christian; Wick, Christoph; Nöth, Maximilian; Büttner, Andreas; Wehner, Maximilian; Springmann, Uwe: Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning, in: The 6th International Workshop on Historical Document Imaging and Processing, 2021, S. 7–12. Online: <https://dl.acm.org/doi/pdf/10.1145/3476887.3476910>.
- Rice, Stephen; Jenkins, Frank; Nartker, Thomas: The fifth annual test of OCR accuracy. Information Science Research Institute Los Angeles, 1996. Online: <https://www.stephenvrice.com/images/AT-1996.pdf>, Stand: 10.11.2022.
- Schink, Manuela: OCR – Evaluierung der Genauigkeit (QM) sowie Tools zur Unterstützung. Online-Konferenz „OCR–Prozesse und Entwicklungen“, 1. März 2021. Online: <https://wiki.zbw.eu/pages/viewpage.action?pageId=33620559&preview=/33620559/33620565/2021-02-24 Schink OCR-Evaluierung und Tools.pdf>, Stand: 10.11.2022.

- Schneider, Pit: Rerunning OCR. A Machine Learning Approach to Quality Assessment and Enhancement Prediction, arXiv preprint arXiv:2110.01661, 2021. Online: <https://arxiv.org/pdf/2110.01661>, Stand: 10.11.2022.
- Smith, Ray: An overview of the Tesseract OCR engine, in: Ninth international conference on document analysis and recognition (ICDAR 2007), vol. 2, 2007, S. 629–633. Online: <https://research.google/pubs/pub33418.pdf>, Stand: 10.11.2022.
- Smith, Ray: History of the Tesseract OCR engine. What worked and what didn't, in: Document Recognition and Retrieval XX, vol. 8658, International Society for Optics and Photonics, 2013. Online: <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8658/865802/History-of-the-Tesseract-OCR-engine—what-worked-and/10.1117/12.2010051.pdf>, Stand: 10.11.2022.
- Sommer, Dorothea; Heiligenhaus, Kay; Wippermann, Carola; Pankratz, Manfred: Zeitungsdigitalisierung. Eine neue Herausforderung für die ULB Halle, in: ABI Technik34 (2), 2014, S. 75–85.
- Springmann,Uwe; Florian Fink; Klaus Schulz: Automatic quality evaluation and (semi-)automatic improvement of OCR models for historical printings, arXiv preprint arXiv:1606.05157, 2016. Online: <https://arxiv.org/pdf/1606.05157>, Stand: 10.11.2022.
- Tanner, Simon; Muñoz, Trevor; Ros, Pich Hemy: Measuring Mass Text Digitization Quality and Usefulness. Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive, in: D-lib Magazine 15 (7/8), 2009. Online: <http://www.dlib.org/dlib/july09/munoz/07munoz.html>, Stand: 10.11.2022.
- Wernersson, Maria: Evaluation von automatisch erzeugten OCR-Daten am Beispiel der Allgemeinen Zeitung, in: ABI Technik 35 (1), 2015, S. 23–35.