OCR-D – Koordinierte Förderinitiative zur Weiterentwicklung von OCR-Verfahren

Elisa Herrmann, Herzog August Bibliothek Wolfenbüttel
Thomas Stäcker, Universitäts- und Landesbibliothek Darmstadt

Zusammenfassung:
Das Projekt OCR-D hat zum Ziel, das Verfahren der automatischen Texterkennung historischer Texte weiterzuentwickeln. Nach einer primären Phase der Bedarfsanalyse folgt 2018 die Modulprojektphase. Der vorliegende Artikel beschreibt in Kürze das in der ersten Projektphase erarbeitete Funktionsmodell von OCR-D und geht auf die Herausforderungen der einzelnen Prozessschritte ein. Für diese sollen die Modulprojekte zukünftig Lösungen erarbeiten.

Summary:
The OCR-D Project aims to refine the process of automatic text recognition especially for historical texts. After an initial analysis of the requirements in the first phase of the project, the second project phase – the module project phase – will start in 2018. The article describes the function model of OCR-D and addresses the challenges which have to be met at different process steps. These are supposed to be solved in the module projects.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2017H4S199-203
Autorenidentifikation: Herrmann, Elisa: GND 1143232690;
Stäcker, Thomas: GND 141905573,
ORCID: http://orcid.org/0000-0002-1509-6960
Schlagwörter: OCR, Digitalisierung

1. Stand der Digitalisierung und Texterkennung in Deutschland

In den vergangenen 30 Jahren ist ein beträchtlicher Teil des im deutschen Sprachraum erschienenen schriftlichen kulturellen Erbes in mehreren, durch die Deutsche Forschungsgemeinschaft (DFG) geförderten Kampagnen in den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke des 16.-18. Jahrhunderts (VD16, VD17, VD18) zunächst nachgewiesen und seit 2006 digitalisiert worden. Auch wenn damit viel erreicht ist und die Forschungsbedingungen erheblich verbessert worden sind, ist dies doch nur eine unabdingbare Voraussetzung für den zweiten Schritt: die vollständige Umwandlung der Images in eine maschinenlesbare Form. Das gesamte schriftliche Kulturerbe als Volltext für Recherche und weitere Bearbeitung, etwa für digitale Editionen, zur Verfügung zu stellen ist nicht nur eine verwegene Vision, sondern angesichts der jüngsten technischen Entwicklungen im Bereich der Optical Character Recognition (OCR) ein realistisches Szenario.

Auf dem DFG-Workshop „Verfahren zur Verbesserung von OCR-Ergebnissen“ im März 2014 kamen Expertinnen und Experten zu dem Entschluss, dass eine dringende Notwendigkeit für freien Zugang zu historischen Textkorpora und lexikalischen Ressourcen zum Training von vorhandener Software

zur Texterkennung bestehe.[1] Ebenso müssen Open-Source-OCR-Engines zur Verbesserung der Textgenauigkeit weiterentwickelt werden wie auch Anwendungen für die Nachkorrektur der automatisch erstellten Texte. Daneben sollten Workflows, Standards und Verfahren der Langzeitarchivierung mit Blick auf zukünftige Anforderungen an den OCR-Prozess optimiert werden.

Als zentrales Ergebnis dieses Workshops stand fest, dass eine koordinierte Fördermaßnahme der DFG notwendig ist.

2. Das Projekt OCR-D

Die „Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)", kurz OCR-D, begann im September 2015 und versucht seitdem einen Lückenschluss zwischen Forschung und Praxiseinsatz, indem für die Entwicklungsbedarfe Lösungen erarbeitet und der aktuelle Forschungsstand zur OCR mit den Anforderungen aus der Praxis zusammengebracht werden.

OCR-D versteht sich dabei als Koordinierungsgremium und Netzwerk zugleich, bringt Entwickler/innen, Forscher/innen und Anwender/innen zusammen, um aktuelle Erkenntnisse aus der Forschung mit den Anforderungen aus der Praxis in einer praktikablen Lösung zu vereinen. Das DFG-geförderte Projekt wird federführend von der Herzog August Bibliothek Wolfenbüttel (HAB) sowie der Berlin-Brandenburgischen Akademie der Wissenschaften in Berlin (BBAW), der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB) und dem Karlsruher Institut für Technologie (KIT) durchgeführt.

Die Projektlaufzeit ist in zwei Phasen unterteilt: In der ersten Projektphase wurden (Entwicklungs-)Bedarfe analysiert und eine Koordinierungsinfrastruktur für die zweite Projektphase aufgebaut, in welcher Modulprojekte Lösungen für die erkannten Bedarfe umsetzen sollen.

2.1. Phase I: Das Funktionsmodell

Das wesentliche Arbeitsergebnis der ersten Projektphase ist das Funktionsmodell als beispielhafter OCR-Durchlauf. Das Modell untergliedert die Bearbeitung des Images bzw. Textes in vier Bearbeitungsebenen, auf denen die algorithmischen Verfahren angewandt werden: Dokument, Seite, Absatz (bzw. Textzone) und Zeile. Einzelne Operationen wirken sich dabei auf verschiedenen Ebenen aus.

Der OCR-Prozess beginnt bereits einige Schritte vor der eigentlichen Texterkennung. Zunächst wird das Bild-Digitalisat im „Preprocessing" vorbereitet. Prozessschritte dieser Vorverarbeitung können dabei das Zuschneiden (*Cropping*), das Begradigen (*Deskewing*) und Entzerren (*Dewarping*) der Seiten sowie das automatische Bereinigen (*Despeckling*) sein. Das Preprocessing wird meist durch die Binarisierung abgeschlossen, bei der das Bild in eine Grafik mit lediglich schwarzen und weißen Pixeln umgewandelt wird. Zu beachten ist, dass je nach Material und verwendeter Texterkennungsmethode auf einzelne Schritte des Preprocessings verzichtet werden kann.

1 Deutsche Forschungsgemeinschaft, *Protokoll zum Workshop „Verfahren zur Verbesserung von OCR-Ergebnissen"* (2014), zuletzt geprüft am 02.11.2017, http://www.dfg.de/download/pdf/foerderung/programme/lis/140522_ergebnisprotokoll_ocr_workshop.pdf.

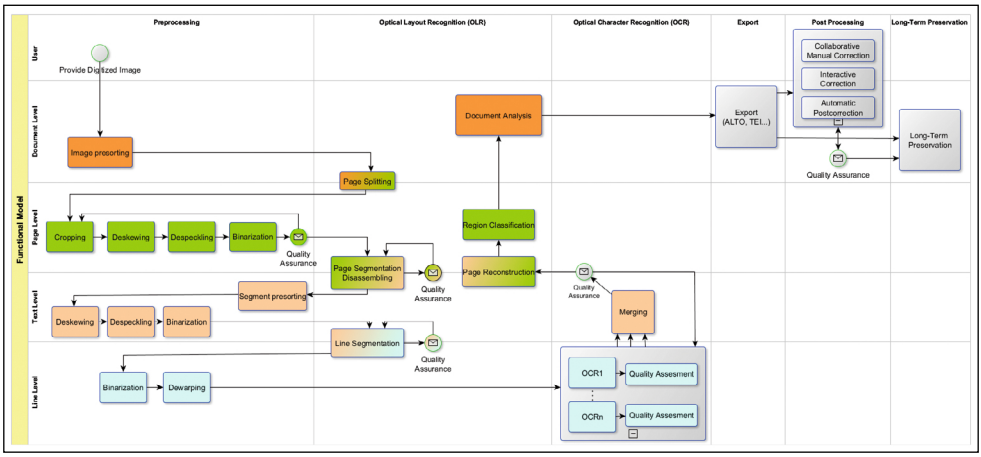

Abb. 1: Das OCR-D-Funktionsmodell

In der folgenden „Layout-Erkennung“ wird die Seite in Text- und Nicht-Textzonen und weiter bis auf einzelne Texteilen unterteilt. Dabei können die einzelnen Schritte des Preprocessing nochmals auf Absatz- und/oder Zeilenebene wiederholt werden. Komplexe Layouts, etwa mehrspaltige Texte oder Marginalien stellen eine der größeren Herausforderungen dieses Prozessschrittes dar.

Im Anschluss erfolgt die eigentliche „Texterkennung“ mittels OCR-Software. Sind derzeit vor allem noch klassische Verfahren der Optical Character Recognition bei vielen Digitalisierungsvorhaben verbreitet, drängen neuere Ansätze mittels Neuronaler Netze stärker in den Fokus, deren Praxistauglichkeit für Massenprozessierungen es jedoch noch zu prüfen gilt. Eine Schwierigkeit der Texterkennung stellt vor allem der Schriftypen- und Sprachmix, teilweise auf Wortebene, dar.

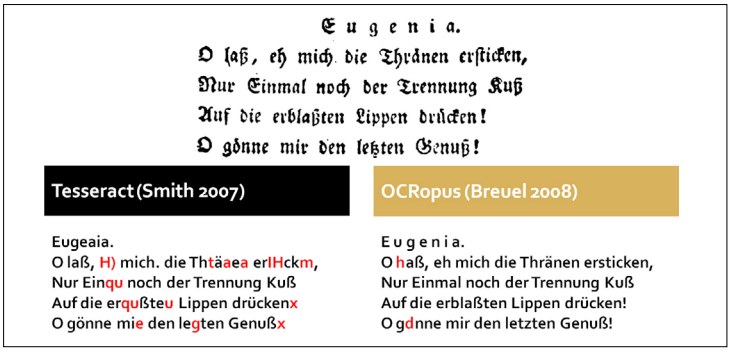

Abb 2: Vergleich der Software Tesseract[2] mit den klassischen Verfahren und OCRopus[3], das bereits moderne Ansätze verfolgt. Alle Fehler wurden rot markiert, Vokale mit hochgestelltem e wurden in diesem Fall als richtig angesehen, wenn sie als Umlaut ausgegeben wurden.

2 Tesseract Open Source OCR Engine, https://github.com/tesseract-ocr/tesseract.
3 The OCRopus OCR System, https://github.com/tmbdev/ocropus.

Nach der Texterkennung erfolgt die automatische „Document Analysis“, bei der das Dokument auf seine Struktur analysiert wird. Durch die „Region Classification“ als Teil der Document Analysis werden die layout-semantische Funktion der einzelnen Textregionen, etwa Überschrift, Seitenzahl oder Marginalie, bestimmt. Im zweiten Schritt wird die Dokumentenstruktur aus den zuvor erkannten Strukturelementen erfasst, etwa um automatisch Inhaltsverzeichnisse zu generieren.

Jedoch erreichen auch moderne Verfahren bei historischen Vorlagen kaum eine gewünschte Qualität von >99% Textgenauigkeit, wie bspw. in der Fallstudie zur historischen OCR im Rahmen des RIDGES-Projekt der Humboldt Universität zu Berlin nachgewiesen wurde.[4] Diese Umstände erfordern in vielen Fällen eine „Nachkorrektur“ der OCR-Ergebnisse. Die entsprechende Nachkorrektur kann manuell, oft auch in Form von Crowdsourcing-Projekten, oder halbautomatisch mittels entsprechender Software durchgeführt werden. Solche Tools bieten die Möglichkeit, potentiell falsch erkannte Wörter hervorzuheben, und bieten ggf. mit Wörterbucheinträgen Vorschläge zur Korrektur an. Der Einsatz vollautomatischer Nachkorrekturverfahren wird auf Grund der uneinheitlichen und auch innerhalb eines Dokumentes variierenden Schreibweise sowie zum Teil starker dialektaler Einflüsse in den historischen Materialien derzeit nicht in der Praxis angewandt.

Im finalen Schritt wird der fertige OCR-Text unter einer freien Lizenz in leicht zugänglichen Repositories zur Verfügung gestellt und langzeitarchiviert werden, wobei eine besondere Anforderung darin bestehen wird, laufend stattfindende Textverbesserungen zentral nachzuweisen, um den jeweils besten Text anbieten zu können.

Charakteristika dieses Funktionsmodells sind zum einen die maximale Adaptivität bzgl. spezifischer Herausforderungen auf Bild- und Textebene, zum anderen soll durch Qualitätssicherungsmethoden an geeigneter Stelle schon frühzeitig in den Prozess eingegriffen werden können, um die Nachbearbeitung zu minimieren. Bisher erfolgt die Qualitätsmessung am Prozessende. Bei dem derzeitig in den DFG-Praxisregeln zur Digitalisierung empfohlenen Bernoulli-Verfahren zur Qualitätsermessung werden stichprobenartig Fehler im OCR-Text erfasst und auf den gesamten Text hochgerechnet, um die Gesamtfehlerquote zu bestimmen.[5] Dies ist bereits eine Erleichterung, da nicht jeder einzelne Fehler gezählt werden muss, um eine Einschätzung der Textgenauigkeit zu erhalten, jedoch ist auch dieses Verfahren noch zu zeitintensiv für den Einsatz in der Massendigitalisierung. Erstrebenswert und im Fokus dieses Projekts ist daher eine Methode zur automatischen Qualitätsbestimmung, ohne Ground-Truth-Abgleich. Ground-Truth bezeichnet hierbei die originalgetreue, maschinenlesbare Darstellung des digitalisierten Dokuments, mit deren Hilfe die Qualität des OCR-Prozessergebnisses mittels Abgleich bestimmt werden kann. Die Erstellung dieses dokumentspezifischen Ground-Truths erfolgt derzeit teilweise oder in Gänze durch manuelle Transkription des Textes. Diese Vorgehensweise ist zum einen zeit- und kostenaufwändig, zum anderen wäre die OCR obsolet, wenn jedes Dokument nochmals transkribiert werden würde. Das Projekt untersucht deshalb Methoden, wie die

4 Uwe Springmann und Anke Lüdeling, „OCR of Historical Printings with an Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus,“ *Digital Humanities Quarterly* 11, Nr. 2 (2017), zuletzt geprüft am 02.11.2017, http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html.

5 Deutsche Forschungsgemeinschaft, *DFG-Praxisregeln „Digitalisierung“* (Bonn, 2016), 35, zuletzt geprüft am 02.11.2017, www.dfg.de/formulare/12_151/12_151_de.pdf.

Qualität ohne Verwendung dokumentspezifischem Ground-Truth erfolgen kann. Bei der Lösung dieses Problems ergeben sich weitere anforderungsspezifische Fragen, etwa was als Fehler betrachtet wird und was ggf. nicht. Ist ein nicht erkanntes Satzzeichen in der Fußnote genauso problematisch für die spätere Forschungsarbeit wie ein nicht erkanntes Zeichen in einem Eigenwort? Generell wird im Projekt stärker auf das spätere Nutzungsszenario eingegangen. So muss die Qualitätsbestimmung nicht unbedingt in „sehr gut", „gut" und „schlecht" eingeteilt werden, vielmehr soll den Nutzenden vermittelt werden, wofür ein Text mit einer Genauigkeit von 85 % benutzt werden kann und wofür eher nicht.

2.2. Phase II: Die Modulprojekte

Im März 2017 veröffentlichte die DFG eine Ausschreibung für Modulprojekte, die für einzelne Problemfelder Lösungen erarbeiten sollen. Die sechs Module sind: Bildvorverarbeitung (Modul 1), Layouterkennung (Modul 2), Textoptimierung (Modul 3), Modelltraining (Modul 4), Langzeitarchivierung und Persistenz (Modul 5) und Qualitätssicherung (Modul 6).

Derzeit durchlaufen die Modulprojektanträge den Begutachtungsprozess bei der DFG, die Projektstarts werden ab 2018 erfolgen.

2.3. Ausblick

Für zukünftige Digitalisierungsprojekte werden die Ergebnisse aus OCR-D weitreichende Veränderungen mit sich bringen. Zum einen soll die Transformation der Titel aus den VD-Projekten in maschinenlesbare Form vorbereitet werden, zum anderen werden auch Vorschläge für die Aktualisierung der DFG-Praxisregeln „Digitalisierung" auf der Grundlage der neuen Erkenntnisse erarbeitet. Nicht zuletzt soll so im Geiste europäischer und nationaler Agenden die mit der Imagedigitalisierung begonnene und derzeit noch andauernde Medienkonversion des gesamten im deutschen Sprachraum erschienenen schriftlichen kulturellen Erbes mittel- bis langfristig durch eine Wandlung in qualitativ hochwertigen Volltext vollendet werden.

Literaturverzeichnis

- Technology Watch des Projekts OCR-D mit weiterführender Literatur: https://www.zotero.org/groups/ocr-d.

- Deutsche Forschungsgemeinschaft. *DFG-Praxisregeln „Digitalisierung"*. Bonn, 2016. Zuletzt geprüft am 02.11.2017. www.dfg.de/formulare/12_151/12_151_de.pdf.

- Deutsche Forschungsgemeinschaft. *Protokoll zum Workshop „Verfahren zur Verbesserung von OCR-Ergebnissen"* (2014). Zuletzt geprüft am 02.11.2017. http://www.dfg.de/download/pdf/foerderung/programme/lis/140522_ergebnisprotokoll_ocr_workshop.pdf.

- Springmann, Uwe und Anke Lüdeling. „OCR of Historical Printings with an Application to Building Diachronic Corpora: A Case Study Using the RIDGES Herbal Corpus." *Digital Humanities Quarterly* 11 Nr. 2 (2017). Zuletzt geprüft am 02.11.2017. http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html.