Flexibles Arbeiten mit OCR4all

Massenvolltextdigitalisierung von Drucken mithilfe von OCR-D und hochqualitative Transkription von Handschriften

Langhanki, Florian
florian.langhanki@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Wehner, Maximilian
maximilian.wehner@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Baierer, Konstantin
konstantin.baierer@sbb.spk-berlin.de
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

Hinrichsen, Lena
hinrichsen@hab.de
Herzog August Bibliothek Wolfenbüttel

Reul, Christian
christian.reul@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Einleitung

Die automatisierte Texterkennung von historischen Drucken und Handschriften gilt aus geisteswissenschaftlicher wie aus informatischer Perspektive in ganz unterschiedlichen Forschungs- und Anwendungskontexten auch weiterhin als anspruchsvolle und problembehaftete Aufgabe. Während die OCR (Optical Character Recognition) moderner Texte mit ihren zeilenbasierten OCR-Ansätzen (Breuel et al. 2013) weithin als informatisch quasi gelöstes Problem angesehen wird, stellen v. a. die höchst komplexen Layoutstrukturen vormoderner Werke (speziell der vor 1700) und ihr teils schlechter Druck- bzw. Erhaltungszustand immer noch ein großes Problem bei der Herstellung maschinenles- und -verarbeitbarer Texte dar. Verglichen mit der Vielfalt und Varianz der in Drucken verwendeten Typen und Schriftarten, gestaltet sich die Erkennung von Handschriften durch die vielfältigen Ausprägungen einzelner Schriftarten in Kombination mit unterschiedlichen Schreiberhänden noch einmal komplizierter. Selbst der kommerzielle State of the Art der Texterkennungssoftware wie bspw. ABBYY Finereader[1] wird in der Produktion wissenschaftlich nutzbarer Daten hier vor erhebliche Probleme gestellt. Die bereits bekannten Schwierigkeiten einer OCR auf historischen Daten müssen demnach um jene einer HTR (Handwritten Text Recognition) mittelalterlicher und frühneuzeitlicher Werke erweitert werden.

Besonders die neueren Forschungsfelder innerhalb der Geisteswissenschaften und Digital Humanities (Text Mining, Sentiment Analysis etc.) haben diese Schwierigkeiten bei gleichzeitigem Bedarf großer Textmengen zur Anwendung quantitativer Analyseverfahren erkannt. Hier stellt sich zunehmend die Frage nach Möglichkeiten einer Texterkennung historischer Drucke und Handschriften, die sowohl hohen Qualitätsansprüchen als auch einem ebensolchen Automatisierungsgrad genügen.

Es ist unstrittig, dass entsprechende Werkzeuge frei verfügbar sein, kohärente OCR- bzw. HTR-Workflows zur Verfügung stellen müssen und außerdem einfach und selbstständig durch nichtinformatische, geisteswissenschaftliche Nutzer:innen bspw. über eine grafische Benutzeroberfläche nutzbar sein sollten. Hinzu kommen jene spezifischen Anforderungen, die mit der Massenverarbeitung von Texten einhergehen, sowie der Wunsch nach größtmöglicher Flexibilität und nach Vielfalt von Werkzeugen. Den besonderen Anforderungen einer massenhaften Textdigitalisierung wendet sich besonders das DFG-geförderte Projekt OCR-D (Engl et al. 2020) mit dem Ziel zu, die Werke in den Verzeichnissen der deutschsprachigen Drucke (VD 16–18) durch vollautomatische Texterkennung als Forschungsdaten zur Verfügung zu stellen. Während in OCR-D also der Fokus auf Massenverarbeitung, Skalierbarkeit und Flexibilität sowie vielfältigen Anwendungsmöglichkeiten liegt, vereint die an der Universität Würzburg entwickelte Software OCR4all[2] (Reul et al. 2019) die erstgenannten Notwendigkeiten einer einfachen Nutzbarkeit entsprechender Technologien mithilfe einer grafischen Benutzeroberfläche und richtet sich dabei dezidiert an Geisteswissenschaftler:innen.

Mit dem im Juli 2021, im Rahmen der dritten Projektphase von OCR-D[3] , gestarteten Würzburger Forschungsprojekt OCR4all-libraries[4] rückt mit der geplanten Integration der OCR-D-Lösungen in die dort entwickelte Software nun noch einmal verstärkt eine notwendige Vereinfachung und Individualisierung komplexer und projektspezifisch flexibel anwendbarer OCR- und HTR-Workflows in den Fokus. Die Anwendung der Software im Spannungsfeld einer Massenvolltextdigitalisierung wie jener der VD16 – 18[5] und einer hochqualitativen Erfassung mittelalterlicher Handschriften erfährt hier einen neuen wie nachhaltigen Rückenwind.

OCR4all

Die im Workshop verwendete Software orientiert sich in seinem Aufbau an den Hauptkomponenten eines OCR-Workflows (s. u.), gliedert diesen jedoch noch einmal in unterschiedliche Teilmodule. Dieser modulare Aufbau erlaubt eine Einbindung und Verwendung bereits bestehender Softwarelösungen, die gemäß ihre n Stärken zu einem kohärenten OCR-Workflow kombiniert werden. Im Allgemeinen umfasst der typische Ablauf einer OCR bzw. HTR die **Vorverarbeitung** (Preprocessing), die **Regionen- und Zeilensegmentierung** (Region-, Line-Segmentation), die **Texterkennung** (Recognition) und die **Nachkorrektur** (Post Correction) (s. Abb. 1).

Abb. 1: Hauptkomponenten eines typischen OCR-Workflows. Von links nach rechts: Originalbild, Vorverarbeitung, Segmentierung, Texterkennung, Nachkorrektur.

Im Preprocessing werden die Einzelbilder gerade gestellt und binarisiert oder in Graustufen umgewandelt (s. Abb. 1). Dabei werden alle gängigen Eingabeformate für Bilddateien unterstützt. Dem schließt sich die Layouttypisierung mithilfe des Segmentierungstools LAREX[6] (s. Abb. 2) an. Hier können werkspezifische Parameter zur Text- und Bildtypisierung festgelegt sowie zu erkennende Layoutregionen (Haupttext, Überschriften, Marginalien, Seitenzahlen, Anstreichungen, Randnotizen etc.) definiert werden. Je nach Komplexität des vorliegenden Seitenlayouts ist nach einer automatischen Layouterkennung ein Eingriff in das vorliegende Ergebnis mittels unterschiedlicher Korrekturwerkzeuge möglich. Weiterhin kann in LAREX die Lesereihenfolge der Layoutbestandteile markiert werden, um den Lesefluss des Originals später vorlagengetreu nachbilden zu können. Vor allem für die Verwendung des maschinenverarbeitbaren Textes in digitalen Editionen sind viele der beschriebenen Funktionen unverzichtbar.

Der Layouttypisierung folgt die Zeilensegmentierung. In dieser werden die Text beinhaltenden Layoutbestandteile in einzelne Zeilenbilder zerteilt (OCRopus[7]), um damit die eigentliche OCR vorzubereiten.

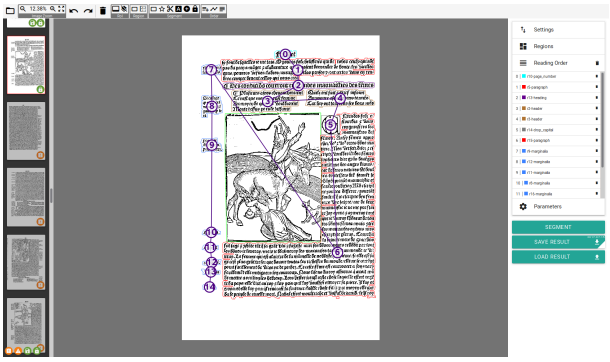

Abb. 2: Im Teilmodul der Segmentierung erfolgen die Typisierung der Layoutelemente sowie die Festlegung der Lesereihenfolge entweder von Grund auf oder in Form der Korrektur eines automatisch generierten Ergebnisses.

Anschließend wird bei der Recognition aus den vorliegenden Einzelzeilenbildern (mittels der OCR-Engine Calamari[8]) maschinenverarbeitbarer Text generiert. Dazu können in OCR4all bereits standardmäßig integrierte gemischte Modelle für Schriftarten unterschiedlicher Epochen genutzt werden. Als ‘gemischt’ werden Modelle bezeichnet, deren Trainingsgrundlage aus einer Vielzahl verschiedener Drucktypen und Schriftarten besteht. Nach der Recognition können die entstandenen Texte in einem Editor komfortabel korrigiert werden (s. Abb. 3).

Abb. 3: Im Editor kann generierter Text mithilfe einer virtuellen Tastatur (rechts) zeichengetreu korrigiert werden.

Für die Berechnung der Fehlerrate einer Zeichenerkennung kann im Evaluationsmodul der ursprünglich erkannte Text mit der durch die Nutzer:innen vorgenommenen Korrektur verglichen werden.

Darüber hinaus bietet OCR4all die Möglichkeit, unter Verwendung vorgenommener Textkorrekturen, selbstständig werkspezifische Modelle zu trainieren, anzuwenden und iterativ zu verbessern. Besonders für Werke mit großer Typenvielfalt und -varianz, auf denen bereits bestehende gemischte Modelle keine hinreichende Erkennungsergebnisse erzielen, werden auf diese Weise dennoch sehr hohe Erkennungsraten erreicht.

Im abschließenden Modul zur Nachkorrektur können die während des Workflows generierten Texte editionsreif korrigiert und anschließend als Plain Text und im Kontext weiterer Strukturdaten als PAGE-XML[9] ausgegeben werden. Letzteres Format beinhaltet neben dem erkannten und ggf. nachkorrigierten Text so auch die Koordinaten aller ausgezeichneten Layoutelemente der Scanseite sowie deren semantische Funktion innerhalb des originalen Seitenlayouts.

Derzeit ist der Workflow auf die hier erläuterten Methoden beschränkt. Im Verlauf des OCR4all-libraries Projekts werden bis zum Workshop jedoch auch die im Rahmen des OCR-D-Projekts erarbeiteten Lösungen verfügbar gemacht werden, wodurch die Nutzer:innen den Workflow eigenständig um Einiges flexibler gestalten und präziser auf den eigenen Anwendungsfall abstimmen können.

In Abhängigkeit des Ausgangsmaterials variiert der zum Erreichen einer sehr hohen Genauigkeit benötigte Arbeitsaufwand zwischen wenigen Minuten bei Werken mit einfachen Layoutstrukturen und einigen Stunden bei sehr komplexen Werken, für die spezifische Erkennungsmodelle erst noch trainiert werden müssen (Reul et al. 2019).

Workshopkonzeption

Der ganztägige Workshop soll einem informatisch wie technisch nicht speziell geschulten Nutzer:innenkreis einen einfachen und verständlichen Einstieg in das Themen- und Problemfeld der OCR und HTR historischer Materialien bieten. Er wird dazu befähigen, mithilfe der vorgestellten Software eigenständig und innerhalb kurzer Zeit qualitativ hochwertige Texte aus ganz unterschiedlichen Ausgangsdaten zu generieren. Die Workshopkonzeption erfolgt deshalb besonders praxisbezogen. Dies bedeutet einen angeleiteten und stets individuell anpassbaren Durchgang des oben vorgestellten OCR- bzw. HTR-Workflows anhand verschiedener, nach Layoutkomplexität, Typographie und Schriftart, Erhaltungszustand und Entstehungszeitraum gruppierter Drucke und Handschriften. Dabei sollen anwendungsbezogen u. a. die folgenden Grundfragen der OCR und HTR beantwortet werden:

- Auf welchen Daten ist OCR4all anwendbar? Was ist OCR-D und welchen Mehrwert bringt die Integration von OCR-D-Lösungen?
- Wie verändert sich entsprechend des Ausgangsmaterials die Anwendung des in OCR4all integrierten OCR- bzw. HTR-Workflows und der in ihm enthaltenen Submodule?
- Mit welchem (manuellen) Aufwand ist in unterschiedlichen Bearbeitungsphasen des Materials zu rechnen?
- Wie stark lässt sich der Workflow in Abhängigkeit des vorliegenden Materials automatisieren?

- Wie und nach welchen Maßgaben können (im Rahmen eines iterativen Ansatzes) projekt- und werksspezifische Texterkennungsmodelle trainiert werden? Welche Erkennungsgenauigkeiten sind zu erwarten?
- Welcher Aufwand ist mit Blick auf die spätere Verwendung der produzierten Texte überhaupt sinnvoll?

Da sich neben den Spezifika des Ausgangsmaterials auch eine grundlegende technische Expertise der Anwender:innen im Bereich der OCR und HTR als Grundbedingung der Produktion hochwertiger maschinenlesbaer Texte herausgestellt hat, strebt der Workshop neben einer praktischen Handlungsanleitung auch die Vermittlung der wichtigsten Funktionskonzepte der in OCR4all integrierten Submodule an.

Darüber hinaus umfasst die Veranstaltung auch Fragen der Einrichtung und Installation der Software, um den Teilnehmer:innen eine stabile und nachhaltige Anwendung von OCR4all über den Workshopkontext hinaus zu ermöglichen. Um einen reibungslosen Ablauf des Workshops selbst zu garantieren, wird durch die Antragsteller:innen eine Serverversion der Software zur Verfügung gestellt. Die max. 25 Teilnehmer:innen benötigen für die Teilnahme deshalb lediglich einen internetfähigen Laptop. Die Verwendung einer Maus wird empfohlen. Digitalisate werden zur Verfügung gestellt, gerne darf aber auch eigenes Material mitgebracht und im Workshop bearbeitet werden.

Forschungsinteressen der Beitragenden

Florian Langhanki ist Wissenschaftlicher Mitarbeiter am ‘Zentrum für Philologie und Digitalität’ der Universität Würzburg. Seine Forschungsinteressen sind Übersetzungsliteratur und Zweisprachigkeit in Mittelalter und Früher Neuzeit sowie die OCR und HTR frühneuzeitlicher Werke und Sammelhandschriften.

Maximilian Wehner ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für ältere deutsche Philologie der Universität Würzburg. Seine Forschungsinteressen sind die wissensvermittelnde Literatur der Frühen Neuzeit, die OCR bzw. HTR mittelalterlicher und frühneuzeitlicher Drucke und Handschriften sowie deren Nutzung in universitärer und schulischer Lehre.

Konstantin Baierer ist Wissenschaftlicher Mitarbeiter an der Staatsbibliothek zu Berlin und betreut dort seit 2018 das OCR-D-Projekt.

Lena Hinrichsen ist Wissenschaftliche Mitarbeiterin an der Herzog August Bibliothek Wolfenbüttel und Projektkoordinatorin von OCR-D. Ihre Forschungsinteressen sind OCR und Objekterkennung sowie Bild-Text-Verhältnisse.

Dr. Christian Reul leitet die Digitalisierungseinheit des ‘Zentrum für Philologie und Digitalität’ der Universität Würzburg. Seine Forschungsschwerpunkte sind die OCR/HTR auf historischem Material sowie die Neu- und Weiterentwicklung der entsprechenden Software.

Fußnoten

1. https://www.abbyy.com/de-de/finereader/
2. http://ocr4all.de/
3. https://ocr-d.de/de/phase3/
4. https://www.uni-wuerzburg.de/zpd/news/single/news/ocr4all-libraries-genehmigt/
5. https://ocr-d.de/de/about/
6. https://github.com/OCR4all/LAREX
7. https://github.com/tmbdev/ocropy
8. https://github.com/Calamari-OCR/calamari
9. https://www.primaresearch.org/tools/PAGELibraries

Bibliographie

Breuel, Thomas M. / Ul-Hasan, Adnan / Al-Azawi, Mayce Ali / Shafait, Faisal (2013): High-Performance OCR for Printed English and Fraktur Using LSTM Networks, in: *12th International Conference on Document Analysis and Recognition:* 683-687.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): OCR4all - An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings, in: Applied Sciences 2019. (9) 22. URL: https://www.mdpi.com/2076-3417/9/22/4853

Engl, Elisabeth / Boenig, Matthias / Baierer, Konstantin / Neudecker, Clemens / Hartmann, Volker (2020): Volltexte für die Frühe Neuzeit. Der Beitrag des OCR-D-Projekts zur Volltexterkennung frühneuzeitlicher Drucke, in: Zeitschrift für Historische Forschung 47 (2), 2020, S. 223–250. URL: https://elibrary.duncker-humblot.com/journals/id/28/vol/47/iss/5737/art/58179