

KI statt Paläographie: Automatische Transkription von Handschriften und Drucken Einführung in Transkribus und eScriptorium

Will, Larissa

larissa.will[at]uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID-iD: 0009-0004-6220-8939

Huff, Dorothee

dorothee.huff[at]uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID-iD: 0000-0003-0866-9967

Zusammenfassung: Texterkennungs- und Transkriptionsplattformen wie Transkribus und eScriptorium können bei der Erschließung historischer Quellen unterstützen und bieten gegenüber kommandozeilenbasierter Texterkennungssoftware den Vorteil einer einfachen Anwendung über eine graphische Oberfläche. Dabei spielt es keine Rolle, ob es sich um eine Postkarte, eine historische Zeitung oder eine mittelalterliche Handschrift handelt. Von der Layouterkennung und -korrektur über die automatisierte Texterkennung bis hin zur manuellen Korrektur der OCR-Ergebnisse und dem Training eigener Modelle werden die grundlegenden Funktionen von Transkribus und eScriptorium vorgestellt und können von den Teilnehmenden selbst an Beispielen getestet werden.

Abstract

Egal, ob es sich um Editionsprojekte, ein Seminar zur Handschriftenkunde oder Citizen Science-Projekte handelt: Schriftstücke aus vergangenen Zeiten lassen sich ohne paläographische Kenntnisse oft nur mit hohem Aufwand entziffern und maschinell weiterverarbeiten. Texterkennungs- und Transkriptionsplattformen wie Transkribus und eScriptorium können hier bei der Erschließung von historischen Quellen unterstützen bzw. diese bei großen Korpora überhaupt erst ermöglichen. Sie bieten gegenüber kommandozeilenbasierter Texterkennungssoftware den Vorteil einer einfachen Anwendung über eine graphische Oberfläche. Dabei spielt es keine Rolle, ob es sich um eine Postkarte, eine historische Zeitung oder eine mittelalterliche Handschrift handelt.

Transkribus ist eine Plattform zur Transkription handschriftlicher und gedruckter Dokumente, die verschiedene Tools wie die automatische

Layout- und Texterkennung auf Grundlage von neuronalen Netzen vereint und daneben unter anderem auch das Tagging von Entitäten und Strukturelementen sowie das Training von eigenen Modellen ermöglicht. Die erzeugten Daten können in vielen unterschiedlichen Formaten für die Weiterverarbeitung exportiert werden. Transkribus war 2016 bis 2019 Teil des EU geförderten Projekts READ (Recognition and Enrichment of Archival Documents) und wird seitdem von der READ-COOP SCE weiterbetrieben.¹ Aktuell kann das Programm in einer Desktop-Version (Expert Client) und Browser-Version (Transkribus Lite) nach einer einfachen Registrierung ohne Lizenz genutzt werden. Abgesehen von der automatischen Texterkennung sind alle Funktionen gebührenfrei nutzbar. Aufgrund der großen Usercommunity stehen viele öffentlich zugängliche Texterkennungsmodelle für Drucke und Handschriften in unterschiedlichen Sprachen und Schriftarten zur Verfügung, die oftmals schon bei geringem Eigenaufwand mit gutem Ergebnis auf eigene Texte angewandt werden können. Während eine automatische Transkription auf Knopfdruck sehr bequem ist, so muss dabei jedoch beachtet werden, dass der Output auf diese Weise unter Umständen Richtlinien entspricht, die sich nicht mit den eigenen decken, uneinheitlich und/oder nicht nachvollziehbar sind. Sollen diese Probleme mit einem eigenen Modelltraining umgangen werden, kann jedoch auch ein nichtstandardisierter Output als Ausgangspunkt für selbst aufbereitete Ground-Truth-Daten dienen und deren Erzeugung erleichtern.

eScriptorium ist eine freie Open-Source-Transkriptionsplattform, die als Teil des scripta-Projektes² an der Université Paris Sciences et Lettres entwickelt wird.³ Für die Texterkennung und die Segmentierung verwendet eScriptorium die OCR-Engine Kraken.⁴ Die Quelloffenheit von eScriptorium ermöglicht es der Anwenderschaft einen Beitrag zur Entwicklung zu leisten oder an eigene Anforderungen zu adaptieren. eScriptorium lässt sich auf dem eigenen PC oder Server betreiben. Handschriftliche und gedruckte Texte können automatisiert oder manuell segmentiert und transkribiert werden. Da die Daten hierfür nicht auf Fremd-Server hochgeladen werden müssen, ist es damit auch möglich, Digitalisate mit restriktiven Auflagen für die Weitergabe zu bearbeiten.

¹ READ-COOP, 2023.

² PSL, 2023.

³ Scripta/escriptorium, 2023.

⁴ Kiessling, 2015 und Kiessling, 2021, S. 87. (Bei Kraken handelt es sich um einen Quelltext-Fork der OCR-Lösung OCRopus, siehe dazu: OCRopus, 2023).

Ein Vorteil von eScriptorium ist, dass nicht nur die Daten, die für das Modelltraining genutzt wurden gemäß der FAIR-Prinzipien bereitgestellt werden können, sondern auch die Segmentierungs- und Texterkennungsmodelle selbst. Dadurch wird die Interoperabilität und Wiederverwendbarkeit der Modelle gefördert. Die Auffindbarkeit wird durch die Bereitstellung von mit einer DOI versehenen Modellen z. B. auf Zenodo gewährleistet. Außerdem können die Modelle außerhalb des Systems für Massen-OCR mit Kraken verwendet werden. Dieser offene Umgang mit den erarbeiteten Daten fördert den Wissenstransfer und ermöglicht es auch kleineren lokalhistorischen Projekten, Studierenden und Forschungsprojekten mit wenig finanziellen Mitteln eine kooperative und komfortable Texterkennungs- und Transkriptionsumgebung zu nutzen. Die Möglichkeit, Modelle auszutauschen und zu teilen, fördert die Zusammenarbeit in der Community und ermöglicht es den Benutzern, auf bereits existierende Ressourcen zurückzugreifen. Dies trägt dazu bei, die Effizienz und den Fortschritt bei der Transkription von historischen Dokumenten zu steigern.

Die Universitätsbibliothek Mannheim betreibt seit Oktober 2021 eine eigene Instanz von eScriptorium⁵ und hat mittlerweile schon einige Projekte u. a. in Lehrveranstaltungen damit umgesetzt sowie einen Transcribathon.⁶ Mittlerweile hat die eScriptorium-Community ein breites Portfolio an Texterkennungsmodellen trainiert.⁷ Die UB Mannheim hat dazu einen wichtigen Teil beigetragen, dazu gehören diverse Frakturmodelle, ein generisches Handschriftenmodell sowie ein Modell zur Erkennung von Behördenschriftgut, welches mit Schreibmaschine erstellt wurde.⁸

Für die Teilnahme am Workshop sollten alle Beteiligten Zugriff auf einen eigenen PC oder Laptop mit Internetanschluss haben. Anhand eigener oder von uns bereitgestellter Dokumente kann der gesamte OCR-Workflow durchlaufen werden. Von der Layouterkennung und -korrektur über die automatisierte Texterkennung bis hin zur manuellen Korrektur der OCR-Ergebnisse und dem Training eigener Modelle bietet der Workshop einen spannenden Einstieg in die Welt der modernen Transkriptionsarbeit. Die Teilnehmenden können sich beim Workshop einen Einblick in die Vor- und Nachteile von Transkribus und eScriptorium verschaffen und deren Nutzen für ihre Forschungsarbeit prüfen.

⁵ Universitätsbibliothek Mannheim/eScriptorium, 2023.

⁶ z. B. Kamlah et. al, 2022.

⁷ OCR/HTR model repository, 2023.

⁸ Weil, 2023.

Bibliografie

„Home - READ-COOP“. READ-COOP. Abgerufen am 27.04.2023.
<https://readcoop.eu/>.

Kamlah, Jan, Thomas Schmidt und Renat Shigapov. 2022. “Extracting Research Data from Historical Documents with EScriptorium and Python.”, präsentiert bei Focused Tutorial on Capturing, Enriching, Disseminating Research Data Objects, Use Cases from Text+, NFDI4Culture und BERD@NFDI, Mannheim, Deutschland.
<https://doi.org/10.5281/zenodo.7373134>.

Kiessling, B. (2021). Advances in Optical Character Recognition for Historical Arabic. Université Paris Sciences et Lettres.

Kiessling, Benjamin. „Kraken Home“. Kraken, 2015. <http://kraken.re/>.

„OCR/HTR model repository | Zenodo“. Zenodo - Research. Shared., 21.03.2023. https://zenodo.org/communities/ocr_models?page=1&size=20.

„OCROPUS“. GitHub. Abgerufen am 27.04.2023.
<https://github.com/ocropus>.

„Scripta | PSL“. Université PSL (Paris Sciences & Lettres) | PSL. Abgerufen am 27.04.2023. <https://www.psl.eu/en/scripta>.

„Scripta / escriptorium - GitLab“. GitLab, 27.04.2023. <https://gitlab.com/scripta/escriptorium/>.

Universitätsbibliothek Mannheim, „eScriptorium - Homepage“, OCR-BW | Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen, abgerufen am 12.05.2023, <https://ocr-bw.bib.uni-mannheim.de/escriptorium/>.

Weil, Stefan. „Index of /~stweil/tesstrain/kraken“. 11.05.2023.
<https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/kraken/>.