Das Projekt OCR-BW: Automatische Texterkennung auch für Archive

Von Dorothee Huff und Regina Keyler

In der ersten Phase des vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (MWK) geförderten Kooperationsprojekts zwischen den Universitätsbibliotheken Mannheim und Tübingen wurde in Tübingen die Transkriptionsplattform *Transkribus*[1] hinsichtlich der automatischen Texterkennung von Handschriften erprobt. Tests wurden anhand der Tagebücher des Paläontologen Edwin Hennig (1882–1977) sowie der Tagebücher und Predigtnachschriften (in lateinischer und griechischer Sprache) von Martin Crusius (1526–1607) vom Ende des 16. Jahrhunderts durchgeführt. Außerdem bearbeitete das Projektteam einzelne Bände der umfassenden Bestände der juristischen Konsilien und der Senatsprotokolle sowie Texte in der indischen Sprache Malayalam. Für diese Bestände wurden jeweils *Ground-Truth-Daten*, also korrekte Transkriptionen erstellt und auf Grundlage dieser Texterkennungsmodelle, die auf neuronalen Netzen beruhen, trainiert. Das Ziel war, eine Zeichenfehlerrate (CER) von unter 5 % zu erreichen.[2] Die Universitätsbibliothek Mannheim arbeitete in dieser Phase vor allem mit der Software *Tesseract* an der Volltexterkennung von historischen Druckwerken in Fraktur.[3] In diesen zwei Jahren wurden auch – soweit es die Pandemie zuließ – Mitarbeiterinnen und Mitarbeiter von Gedächtnisinstitutionen sowie Wissenschaftlerinnen und Wissenschaftler im Umgang mit den Texterkennungswerkzeugen geschult und beraten.

Die Zeit der Projektverlängerung um ein Jahr nutzte die Universitätsbibliothek Mannheim zur Installation und Evaluation der Plattform *eScriptorium*. In Tübingen wurden nun Tests mit kleineren und heterogeneren Textkorpora wie gemischtes Archivgut, mittelalterliche Handschriften und Inkunabeln durchgeführt und der Workflow umgestellt: Es wurden nun in der Regel keine Texterkennungsmodelle von Grund auf neu erzeugt, sondern bereits veröffentlichte generische Modelle für die Quellen werksspezifisch nachtrainiert. Damit verringerte sich der Aufwand für die Bearbeitung erheblich, da weniger eigenes Trainingsmaterial erstellt werden musste.

Zum Ende des Projektes lagen Online-Schulungsmaterialien für die genutzten Software-Anwendungen vor.[4] Das an den Universitätsbibliotheken Mannheim und Tübingen im Rahmen

---

1 https://readcoop.eu/ (aufgerufen am 10. 04. 2024).

2 Vgl. hierzu: Deutsche Forschungsgemeinschaft: DFG-Praxisregeln „Digitalisierung". DFG-Vordruck 12.151 – 12/16, 2016. Online: https://www.dfg.de/formulare/12_151/ (aufgerufen am 10. 04. 2024).

3 Siehe zur Installation, Anwendung und Modellhinweisen für Tesseract: https://github.com/UB-Mannheim/Tesseract_Dokumentation (aufgerufen am 10. 04. 2024).

4 Verfügbar über ZOERR: http://hdl.handle.net/10900.3/OER_ULGTBJWR (aufgerufen am 10. 04. 2024).

der Projektlaufzeit aufgebaute *Kompetenzzentrum OCR* wird zudem weiterbetrieben, so dass persönliche Beratung sowie Schulungen und Workshops nach wie vor angeboten werden können.[5]

Schwach strukturierte Quellen in Archiven als geeignete Vorlagen für die Volltexterkennung

Bei der Formulierung des Projektantrags 2017 wurden Überlegungen angestellt, welche Quellen aus dem Universitätsarchiv Tübingen (UAT) und der Handschriftenabteilung der Universitätsbibliothek (UB) sich dafür eignen würden, innerhalb des Projekts bearbeitet zu werden. Zum damaligen Zeitpunkt (2017/18) ging man davon aus, dass die Quellen möglichst von einer Schreiberhand stammen sollten, um eine automatische Texterkennung zu ermöglichen. In die nähere Auswahl kamen daher Vorlesungsnachschriften, Tagebücher und Protokolle, die nur von einer Schreiberhand verfasst worden waren.

Während der Projektarbeit stellte sich jedoch heraus, dass auch Textkorpora bearbeitet werden können, die von unterschiedlichen Händen stammen oder gemischt handschriftlich und maschinenschriftlich sind – sie müssen nur entsprechend repräsentativ trainiert werden.[6] Zudem können mit generischen Modellen gerade für deutsche Kurrentschrift – abhängig von der Schreiberhand – auch ohne eigenes Training zum Teil schon gute Ergebnisse bei der automatischen Texterkennung erzielt werden.[7] Einen weiteren Schritt in die Richtung automatischer Texterkennung für Handschriften auf Knopfdruck bedeutet der Einsatz von Transformer-Modellen, wobei ein Modell sowohl gedruckten wie auch handschriftlichen Text in unterschiedlichen Sprachen und Schriftarten erkennen kann.[8]

Darum kann nun von neuem die Frage gestellt werden: Für welche Archivalientypen lohnt sich eine Volltexterstellung besonders? Dabei soll auch berücksichtigt werden, dass es unterschiedliche Bedürfnisse gibt: Die der Archive, d.h. der Archivarinnen und Archivare, und die der Nutzenden.

Die nachfolgenden Überlegungen setzen an den Beziehungen zwischen den Erschließungsdaten und dem durchsuchbaren Volltext an. In den Fokus rückt dabei ein Typ von Unterlagen,

---

5 Siehe für aktuelle Informationen: https://ocr-bw.bib.uni-mannheim.de (aufgerufen am 10. 04. 2024).

6 Vgl. für eine ausführliche Darstellung einzelner Modelltrainingsreihen: Dorothee *Huff* und Kristina *Stöbener*. Projekt OCR-BW: Automatische Texterkennung von Handschriften. In: O-Bib. Das Offene Bibliotheksjournal 9/4 (2022) S. 6 – 13. Online: https://doi.org/10.5282/o-bib/5885 (aufgerufen am 10. 04. 2024).

7 Zur Anwendung von generischen Modellen auf unbekanntem Material vgl. Tobias *Hodel* u. a.: General Models for Handwritten Text Recognition. Feasibility and State-of-the Art. German Kurrent as an Example. In: Journal of Open Humanities Data 7/13 (2021). Online: https://doi.org/10.5334/johd.46 (aufgerufen am 10. 04. 2024).

8 Siehe zum Einsatz von Transformer-Modellen in Transkribus: https://readcoop.eu/de/introducing-transkribus-super-models-get-access-to-the-text-titan-i/ (aufgerufen am 10. 04. 2024).

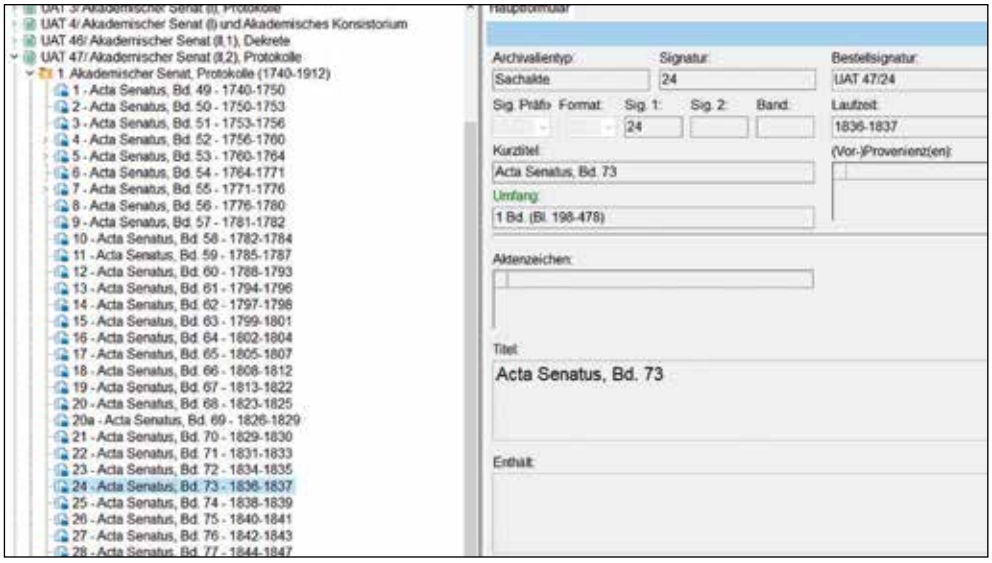

Abb. 1: Erschließungsdaten „schwach strukturierter“ Archivalieneinheiten (Bestand UAT 47, Senatsprotokolle).

von dem bislang vor allem im Kontext der digitalen Unterlagen die Rede war: die *schwach strukturierten Unterlagen*. In der archivwissenschaftlichen Diskussion sind damit meist digital entstandene Unterlagen, Dateisammlungen oder der Inhalt von E-Mail-Postfächern gemeint, die nicht systematisch klassifiziert, sondern z. B. nur chronologisch abgelegt sind.[9] Hier soll der Begriff jedoch auf analoge Archivalien angewandt werden, die in sich *schwach strukturiert* sind und deren Titel aus einer formalen oder materiellen Beschreibung gebildet wird und damit wenig Rückschlüsse auf den Inhalt zulässt. Inwieweit können Volltexte als Ersatz für eine ressourcenaufwändige Tiefenerschließung dienen?

Einige Beispiele: Die Einträge in Tagebüchern sind chronologisch angeordnet, der Titel eines Bandes kann lauten: *Reisetagebuch 1903*. Inhaltlich erschließbar wäre das Tagebuch nur durch Regesten oder eine Verschlagwortung – ein Aufwand, der im archivischen Alltag kaum zu leisten ist.

[9] Vgl. Bewertungskriterien für schwach strukturiertes Schriftgut. Empfehlungen des Arbeitskreises „Archivische Bewertung“ im VdA. In: Bewertung schwach strukturierter Unterlagen (Mitteilungen aus dem Stadtarchiv von Köln 107). Köln 2021. S. 15–26. Auch in diesen Empfehlungen wird der Begriff auf analoge Unterlagen übertragen, z. B. auf Handakten oder Korrespondenzen im Nachlassbereich.

Ähnlich sieht es bei der Erschließung von Protokollen aus: Auch hier werden in den Metadaten für die Erschließung meist nur das Gremium und die Laufzeit des Bandes angegeben. Erst die Aufnahme der Tagesordnungspunkte der Sitzungen würde das Auffinden einzelner behandelter Themen ermöglichen.

Ein weiteres Beispiel sind Korrespondenzserien: Eine Erschließung auf Ebene der Verzeichnungseinheit würde lauten: *Korrespondenz A–G, 1975*. Mit viel Aufwand könnten die Korrespondenzpartnerinnen und Korrespondenzpartner – am besten mit einer Normdaten-Verknüpfung – erhoben werden. Eine fünfbändige Korrespondenz zwischen zwei wichtigen Persönlichkeiten sollte jedoch tiefer erschlossen werden als nur über die beiden Namen und die Laufzeit. Dies wäre allerdings wieder nur mit aufwändigen Regesten oder gar einer Edition möglich.

Auch bei einem weiteren Rechercheszenario wäre ein Volltext hilfreich: Das Universitätsarchiv Tübingen verwahrt Aufnahmebücher der Tübinger Universitätskliniken seit ihrer Entstehung. Darin lassen sich die Aufenthalte eigentlich aller Patientinnen und Patienten nachweisen. Die dazugehörigen Akten sind dagegen nur in Auswahl überliefert und außerdem häufig nicht nach Namen erschlossen. Schwierig wird die Suche nach Personen, wenn die Anfragenden nicht wissen, in welcher Klinik die Patientin oder der Patient war oder wann sie oder er aufgenommen wurde. Günstig für das Modelltraining ist es, wenn die Einträge von wenigen (diensthabenden) Personen stammen, so dass genügend Material pro Schreiberhand vorliegt. Ungünstig ist allerdings, dass es sich bei den zu lesenden Texten hauptsächlich um Eigennamen (Personennamen und Ortsnamen) handelt, Wörter also, die auch die KI häufig vor große Herausforderungen stellen.

Für all diese Unterlagen, deren Inhalt durch die Erschließungsdaten nur schwer darstellbar ist, würde sich die Erstellung eines durchsuchbaren Volltextes sehr lohnen.[10] Der Text wäre bei der Recherche nach Sachbegriffen oder Namen durchsuchbar und womöglich könnte man den Volltext auch mit Hilfe weiterer KI-Anwendungen zur automatischen Vergabe von Schlagworten und damit zur Ergänzung der Erschließungsdaten nutzen.[11] Auch in *Transkribus* selbst können Tags gesetzt und mit Wikidata-Einträgen verknüpft werden. Automatische Texterkennung und die Zurverfügungstellung von Volltexten kann also in gewissem Maße die Erschließungsarbeit des Archivs ergänzen, unterstützen oder zum Teil ersetzen.

---

[10] Der an der Universitätsbibliothek Mannheim entwickelte OCR-Recommender bietet Empfehlungen, welches OCR-Werkzeug für welches spezifische Anliegen am besten geeignet ist: https://www.berd-nfdi.de/limesurvey/index.php/996387 (aufgerufen am 10.04.2024).

[11] In einem Versuch wurde nach Anregung durch den Beitrag von Tobias Hodel der Volltext eines Bandes der Juristischen Konsilien mit der KI-Anwendung ChatPDF zusammengefasst. Das Ergebnis der Zusammenfassung ist von ausreichender Qualität und könnte als Inhaltsangabe den Erschließungsangaben beigegeben werden.

Bearbeitungsaufwand und Grenzen

Zwar lässt sich je nach Schrift bei der Volltexterkennung auch ohne großen Eigenaufwand bereits ein gutes Ergebnis erzielen – insofern ein passendes Texterkennungsmodell vorliegt. Für dessen Optimierung ist je nach Zielsetzung jedoch zumindest ein gewisser Arbeitsaufwand nötig. Nicht unterschätzt werden sollte dabei der Aufwand für die Erstellung des zugrunde liegenden Layouts, also die Festlegung von Zeilen und Textregionen, auf deren Grundlage die Texterkennung vorgenommen wird. Zwar gibt es automatische Tools zur Layout-Analyse, die bei einfachen Layouts mit nur einem Textblock bereits gut funktionieren. Bei einem mehrspaltigen Text, Tabellen, interlinearen Einfügungen, Marginalien etc. stoßen diese jedoch an ihre Grenzen und es ist eine manuelle Korrektur notwendig, die je nach Dokumentlänge einige Zeit in Anspruch nehmen kann. Bei einem größeren Bestand mit sich wiederholenden Layoutelementen kann ein entsprechendes Strukturtraining in Betracht gezogen werden. Hier bietet *Transkribus* seit kurzer Zeit die Möglichkeit sowohl des Trainings von Tabellen wie auch des Trainings sogenannter Field-Modelle, mit denen sich wiederholende Strukturelemente z. B. auf Karteikarten trainiert werden können.[12]

Auch aufgrund der individuellen Ausprägung der Schrift kann das Ergebnis der automatischen Texterkennung nur eingeschränkt hilfreich sein – falls man es nicht sogar mit einer Schriftart und/oder Sprache zu tun hat, für die noch kein passendes Texterkennungsmodell zugänglich ist. Wenn in einem solchen Fall ein eigenes Texterkennungsmodell erzeugt werden soll, ist zwar der Trainingsprozess schnell gestartet und läuft im Hintergrund, jedoch müssen als Grundlage dafür zunächst *Ground-Truth-Daten* erstellt werden. Dafür kann entweder das Ergebnis einer automatischen Transkription korrigiert oder eine eigene Transkription von Grund auf neu erstellt werden. Wenn ein vorhandenes Texterkennungsmodell werksspezifisch auf eine bestimmte Schriftausprägung nachtrainiert werden soll, reichen dafür wenige Seiten, während der für ein gutes Ergebnis benötigte Umfang für ein von Grund auf neu trainiertes Modell größer ausfällt. Hier ist das Kosten-Nutzen-Verhältnis zu beachten. Soll nur ein Band von z. B. 100 Seiten bearbeitet werden, wird bereits ein verhältnismäßig großer Anteil dieser Seiten für ein eigenes Modelltraining benötigt werden. Bei einem Bestand wie den Tübinger Senatsprotokollen hingegen wirkt der Aufwand der Erstellung von ca. 200 Seiten *Ground-Truth-Daten* auf den ersten Blick sehr hoch. Das erzeugte Modell kann dann jedoch auf mehrere Tausend Seiten der Jahrgänge 1799 bis 1847 mit unterschiedlichen Schreiberhänden angewandt werden. Für weitere Bände bzw. neue Schreiberhände muss das vorhandene Modell wiederum mit einigen Seiten entsprechend nachtrainiert werden.

Bei dem im Projektverlauf bearbeiteten Bestand des gemischten losen Schriftguts, der aus Akten mit einzelnen Seiten unterschiedlicher Schreiber sowie gedrucktem und maschinenschriftlichen Material besteht, hat sich ein eigenes Modelltraining hingegen als unwirtschaftlich erwiesen. Für das Modelltraining hätte in diesem Fall zumindest ein Großteil, wenn nicht sogar der

---

[12] Siehe https://readcoop.eu/introducing-table-models-trainable-layout-ai-in-transkribus/ sowie https://readcoop.eu/introducing-field-models-trainable-layout-ai-in-transkribus/ (aufgerufen am 10. 04. 2024).

Validation Set	CER
insgesamt	3,72
UAT 47/15, S. 8 (1799-1801)	6,39
UAT 47/15, S. 465 (1799-1801)	11,27
UAT 47/19, S. 17 (1813-1822)	3,87
UAT 47/19, S. 504 (1813-1822)	7,14
UAT 47/20a, S. 113 (1826-1829)	2,53
UAT 47/20a, S. 476 (1826-1829)	1,31
UAT 47/22, S. 233 (1831-1833)	0,71
UAT 47/22, S. 353 (1831-1833)	2,66
UAT 47/22, S. 510 (1831-1833)	5,06
UAT 47/22, S. 595 (1831-1833)	4,82
UAT 47/24, S. 5 (1836-1837)	3,3
UAT 47/24, S. 250 (1836-1837)	5,75
UAT 47/25, S. 36 (1838-1839)	3,06
UAT 47/25, S. 247 (1838-1839)	1,06
UAT 47/28, S. 33 (1844-1847)	5,15
UAT 47/28, S. 44 (1844-1847)	2,87
UAT 47/28, S. 55 (1844-1847)	2,93
UAT 47/28, S. 66 (1844-1847)	1,08
UAT 47/28, S. 92 (1844-1847)	4,61
UAT 47/28, S. 108 (1844-1847)	0,96
UAT 47/28, S. 120 (1844-1847)	4,44
UAT 47/28, S. 130 (1844-1847)	1,43

Abb. 2: Modelltraining für die Tübinger Senatsprotokolle im Zeitraum 1799–1847 mit einer Aufschlüsselung der Zeichenfehlerrate (CER) auf dem Validation Set.

komplette Bestand des handschriftlichen Materials herangezogen werden müssen, um dem Modell die entsprechende Trainingsgrundlage zu bieten. Hier schien der Einsatz eines generischen Modells am sinnvollsten, das mit einer manuellen Nachkorrektur der am schlechtesten erkannten Seiten kombiniert werden könnte.

Bei der Nutzung des Ergebnisses der automatischen Texterkennung für eine Volltextsuche ist grundsätzlich zu beachten, dass dieses in den wenigsten Fällen hundertprozentig korrekt ist. Dieser Umstand wirkt sich natürlich auf das Suchergebnis aus. Gerade Eigennamen, die in der Regel nicht Teil des gewöhnlich gebrauchten Wortschatzes sind, werden oftmals schlechter erkannt, da sie in den einem Modell zugrunde liegenden Trainingsdaten nur selten, wenn nicht sogar überhaupt nicht vorkommen.[13] Auch die Transkriptionsrichtlinien können sich auf den Sucherfolg auswirken. Wurden Abkürzungen zeichengetreu wiedergegeben und Sonderzeichen verwendet, muss dies bei der Durchsuchung eines Textes entsprechend beachtet werden.

[13] In der Umgebung von Transkribus kann die sogenannte Smart Search angewandt werden, um den Sucherfolg auch bei fehlerhaften Ergebnissen der Texterkennung zu erhöhen. Dabei wird nicht nur der tatsächlich ausgegebene Text durchsucht, sondern auch im Hintergrund gespeicherte alternative Lesungen.

Präsentation von Volltexten

Automatisch erzeugte Volltexte sollten jedoch nicht nur dem internen Dienstgebrauch dienen, sondern auch der Öffentlichkeit zur Verfügung gestellt werden. Idealerweise geschieht dies zusammen mit den Metadaten und den Digitalisaten, damit in einem Suchvorgang Erschließungsdaten und Volltext durchsucht werden können.

In der Archivwelt sind die unterschiedlichsten Präsentationsmodule im Einsatz. Unabhängig vom Archivinformationssystem arbeitet das Universitätsarchiv Tübingen mit der von der UB Heidelberg entwickelten Anwendung *DWork*. Neben jeder als Digitalisat vorliegenden Seite wird der Volltext präsentiert. Der Export des erkannten Textes aus *Transkribus* im Ausgabeformat TEI mit den Koordinaten der einzelnen Zeilen erlaubt es, dass die jeweilige Textstelle im Digitalisat mit einem Rahmen markiert wird. Auf diese Weise wird die Orientierung erleichtert, wenn die Nutzerinnen und Nutzer das Ergebnis der automatischen Transkription mit dem Originaltext abgleichen wollen.

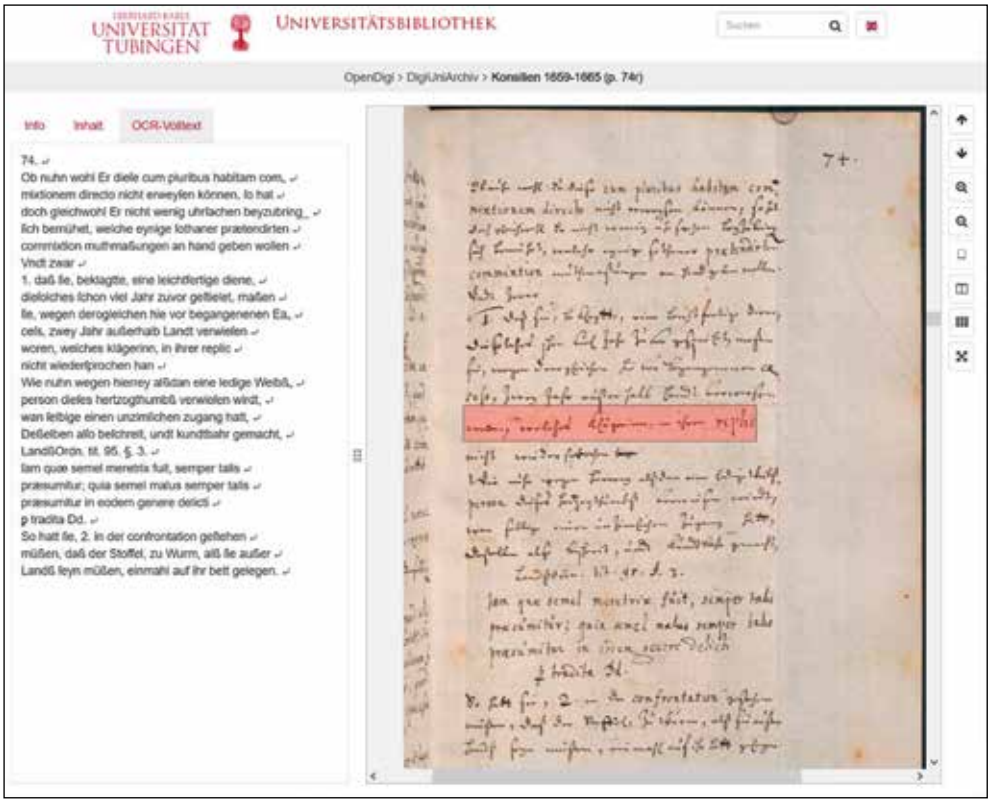

Abb. 3: Präsentation von Volltexten (OpenDigi UB Tübingen). UAT 84/13, 74r.

Diese Lösung ist jedoch nicht optimal, da die Metadaten zwischen dem Archivinformationssystem und *DWork* nicht direkt ausgetauscht werden können. Ein weiterer Nachteil ist, dass in zwei unterschiedlichen Systemen gesucht werden muss: Einmal im internen Archivinformationssystem oder in den exportierten Daten im Archivportal-D und einmal im Präsentationssystem der Digitalisate.

Unterstützung für Archivnutzerinnen und Archivnutzer

Neben der online-Bereitstellung von Volltexten kann es auch sinnvoll sein, von Seiten der Archive Benutzerinnen und Benutzern ohne paläographische Kenntnisse Hilfestellungen im Umgang mit Transkriptionssoftware zu geben. In einem Universitätsarchiv können das z. B. Fachwissenschaftlerinnen und Fachwissenschaftler sein, die über die Geschichte ihres Faches forschen möchten. Auch Wissenschaftlerinnen und Wissenschaftler, die große Textmengen bearbeiten oder durchsuchen möchten, können so Forschungsfragen angehen, die ohne technische Hilfe nicht zu bearbeiten wären.

Für den Ersteinstieg werden Nutzerinnen und Nutzer auf die Seite https://transkribus.ai/ verwiesen. Dort können einzelne digitalisierte Seiten hochgeladen werden, die direkt nach Einstellung der Sprache und der Angabe, ob es sich um ein Druckwerk oder eine Handschrift handelt, von einem vorgegebenen Modell erkannt werden. Das Ergebnis hilft oftmals zumindest bei einer ersten Einschätzung, wovon der Text handelt, und ist ohne Einarbeitung in das Programm möglich.

Sollen größere Dokumente bearbeitet, andere als die voreingestellten Texterkennungsmodelle genutzt oder sogar eigene trainiert werden, ist die Arbeit in der Umgebung einer Transkriptionsplattform zu empfehlen. *Transkribus* bietet aktuell zwei interoperable Varianten, nämlich den *Expert Client*[14] mit dem größten Funktionsspektrum als Desktop-Version sowie *Transkribus Lite*,[15] das im Browser läuft und eine modernere Nutzeroberfläche bietet.[16] Ohne vertiefte IT-Kenntnisse können Tools zur Layouterkennung und Texterkennung angewandt, eigene Transkriptionen angefertigt oder das Ergebnis automatischer Texterkennung korrigiert, Textinhalte sowie Strukturelemente getaggt und die Ergebnisse in verschiedenen Formaten exportiert werden. Der Zugang zu den Dokumenten kann geöffnet und damit ein kollaboratives Arbeiten im beschränkten Kreis oder auch ein öffentliches Crowd-Sourcing-Projekt ermöglicht werden. Nach einem ähn-

---

14 https://readcoop.eu/transkribus/download/ (aufgerufen am 10. 04. 2024).

15 https://readcoop.eu/transkribus/ (aufgerufen am 10. 04. 2024).

16 Die READ-COOP plant, den Betrieb in Zukunft gänzlich auf Transkribus Lite zu verlagern. Der Expert Client wird zwar zunächst aufrechterhalten, erhält jedoch keine Updates mehr. Beide Systeme sind interoperabel, so dass bei einem Wechsel keine Daten verlorengehen bzw. die Versionen parallel genutzt werden können. Aktuell fallen für einzelne Funktionen Gebühren an, wobei jeder Account pro Monat ein Freikontingent von 100 Credits erhält.

lichen Prinzip funktioniert die Transkriptionsplattform *eScriptorium*, die lokal auf dem eigenen Rechner installiert werden kann und mit zunehmender Weiterentwicklung eine gute Alternative darstellt.[17]

Fazit

Gerade für deutsche Kurrentschriften, die einen großen Teil der historischen Archivbestände ausmachen, lassen sich bereits *Out-of-the-box* oftmals gute bis sehr gute Ergebnisse mit generischen Texterkennungsmodellen erzielen. Wenn ein eigenes Texterkennungsmodell für einen bestimmten Bestand erzeugt werden soll, beeinträchtigt auch heterogenes Material wie z. B. unterschiedliche Schreiberhände und/oder lange Schreibzeiträume das Ergebnis nicht wesentlich und verlangt bei entsprechender Planung nicht unbedingt einen höheren Ressourcenaufwand. Dabei stellen auch unterschiedliche Sprachen und Schriftsysteme kein Problem dar und können bei Bedarf in einem Modell vereinigt werden. Liegt bereits ein einigermaßen passendes Modell vor, kann dieses in der Regel schon mit wenigen Seiten durch ein werksspezifisches Training auf ein Dokument angepasst und die Fehlerrate für dieses erheblich gesenkt werden. Soll das Ergebnis jedoch fehlerfrei sein, bedarf es einer manuellen Nachkorrektur.

Die Erzeugung von Volltexten aus handschriftlichen Archivalien ist somit auch für kleinere Einrichtungen mit vertretbarem Aufwand möglich. Mit Transkriptionsplattformen wie *Transkribus* und *eScriptorium*, die über eine Benutzeroberfläche verfügen, bedarf es keiner vertieften IT-Kenntnisse für den Einsatz von Texterkennungssoftware. Gegenüber einer personalaufwändigen Tiefenerschließung gerade auch von schwach strukturierten Quellen, die eine intensive Beschäftigung mit dem Dokument erfordern würde, kann die Erzeugung von automatischen Volltexten mit geringerem Ressourceneinsatz den Zugang zu den Materialien bereits deutlich erhöhen bzw. es den Nutzerinnen und Nutzern ermöglichen, mit großen Textmengen zu arbeiten.

[17] Siehe zu Nutzungs- und Installationshinweisen für eScriptorium: https://ub-mannheim.github.io/eScriptorium_Dokumentation/ (aufgerufen am 10. 04. 2024).