Projekt OCR-BW

Automatische Texterkennung von Handschriften

Dorothee Huff, Universitätsbibliothek Tübingen
Kristina Stöbener, Universitätsbibliothek Tübingen

Zusammenfassung
Nach der Digitalisierung von historischen Dokumenten ist der nächste konsequente Schritt die Anreicherung der Digitalisate im Präsentationssystem mit einem durchsuchbaren Volltext, um die Zugänglichkeit zu den Texten weiter zu erhöhen und neue Forschungsfragen an das Material zu ermöglichen. Während in vielen Bibliotheken bereits verschiedene Möglichkeiten zur automatischen Texterkennung von Druckwerken genutzt werden, ist die Zurückhaltung bei Handschriften vielfach höher, da handschriftliche Quellen die automatische Texterkennung vor neue Herausforderungen stellen. Mithilfe von Machine Learning wurden auf dem Feld der automatischen Handschriftenerkennung in den letzten Jahren jedoch große Fortschritte gemacht, die von Bibliotheken genutzt werden können, um ihre eigenen Bestände weiter zu erschließen, aber auch, um sich als Servicepartnerin für die Wissenschaft zu etablieren.

Im Rahmen des Projekts OCR-BW (https://ocr-bw.bib.uni-mannheim.de/) werden seit 2019 Transkribus und seit 2021 auch eScriptorium für die Erzeugung von automatischen Volltexten für Handschriften systematisch an ausgewählten Korpora getestet. Die im bisherigen Projektverlauf erzielten Ergebnisse sind sehr positiv und zeigen, dass eine automatische Handschriftenerkennung mit einer Zeichenfehlerrate von unter 5 % möglich und erwartbar ist. Bereits veröffentlichte Volltexte haben die Sichtbarkeit und das Forschungsinteresse an diesen Materialien deutlich erhöht. Das Projekt zielt außerdem darauf ab, die Wissenschaft bei der Vorbereitung und Durchführung von Forschungsvorhaben zu unterstützen. An Beispielen vom mittelalterlichen Gebetbuch über Großbestände wie Juristische Konsilien bis hin zum Expeditionstagebuch des 20. Jahrhunderts soll gezeigt werden, mit welchem Ressourcenaufwand welche Ergebnisse erzielt werden können.

Summary
After the digitization of historical documents, the next logical step is to enrich the digitized material with a searchable full text to further increase the accessibility of the texts and to enable new research questions. While many libraries already use various options for automatic text recognition of printed material, there is much higher reluctance to do so when it comes to manuscripts, since handwritten sources pose new challenges for automatic text recognition. With the help of machine learning, however, great progress has been made in the field of automatic handwritten text recognition in recent years, which libraries can not only use to make their own holdings more accessible, but also to establish themselves as a service partner for science.

As part of the OCR-BW project (https://ocr-bw.bib.uni-mannheim.de/), since 2019 the transcription platforms Transkribus and, from 2021, eScriptorium have been systematically tested on selected corpora to generate automatic full texts for manuscripts. The results achieved during the project so far are very positive and show that automatic handwritten text recognition with a character error

rate of less than 5 % is possible and can be expected. Full texts that have already been published have significantly increased the visibility and research interest in these materials. The project also aims to support science in the preparation and implementation of research projects. Examples ranging from medieval prayer books to large collections such as legal councils to expedition diaries of the 20th century will be used to show which results can be achieved with which resources.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/5885

Autorenidentifikation:
Huff, Dorothee: ORCID: https://orcid.org/0000-0003-0866-9967
Stöbener, Kristina: ORCID: https://orcid.org/0000-0002-3299-974X

Schlagwörter: OCR; HTR; Automatische Texterkennung; Handschrift; Digital Humanities; Künstliche Intelligenz; Digitalisierung

1. Einleitung

Seit 2019 wird das Projekt OCR-BW, eine Kooperation der Universitätsbibliotheken Tübingen und Mannheim, vom Ministerium für Wissenschaft, Kultur und Kunst Baden-Württemberg in zwei Projektphasen (2019–2021 und 2021–2022) gefördert. Ziel des Projekts ist der Aufbau und die Etablierung eines Kompetenzzentrums „Volltexterkennung von handschriftlichen und gedruckten Werken", das Wissenschaftler und Wissenschaftlerinnen, Bibliotheken, Archive und andere Institutionen in Baden-Württemberg bei der Anwendung von automatischer Texterkennungs- und Transkriptionssoftware unterstützt.[1]

Bei einer Volltexterkennung werden textliche Bildinhalte in digitale Textformate übersetzt. Erkannte Texte können durchsucht, kopiert, bearbeitet und für eine Extraktion von Forschungsdaten verwendet werden. Die UB Mannheim stellt mehrere Open-Source-Softwareprodukte, wie z.B. Tesseract, aus dem Bereich OCR als nutzerfreundliche und systemunabhängige Anwendung für historische Drucke zur Verfügung und hat zudem eine Instanz der Transkriptionsplattform eScriptorium aufgesetzt. Die UB Tübingen testet die Einsatzmöglichkeiten der Transkriptionsplattform Transkribus für die automatische Handschriften- und Druckerkennung an eigenen Beständen aus der Handschriftenabteilung und dem Universitätsarchiv. In der gemeinsamen Kooperation können die UB Tübingen und die UB Mannheim ein breites Spektrum an Materialien und Technik abdecken. Das Kompetenzzentrum erarbeitet sich so ein umfassendes Knowhow im Bereich der automatischen Volltexterkennung und vermittelt dieses weiter. Ziel ist, auf Basis der Projektergebnisse Handlungsempfehlungen für unterschiedliche Materialgruppen zu etablieren, um den Zugang zu diesen Texten zu erleichtern und neue wissenschaftliche Fragestellungen und Auswertungsmöglichkeiten zu schaffen.

1 Vgl. die Projekthomepage <https://ocr-bw.bib.uni-mannheim.de/>, Stand 08.09.2022.

Entsprechend den Nutzerbedürfnissen wurde auch die zweite Projektphase geplant, die neben Kooperationen mit wissenschaftlichen Projekten und anderen Unterstützungsleistungen nun die Bearbeitung von kleineren Textkorpora bzw. Einzeldokumenten in den Fokus stellt, um den wissenschaftlichen Praxisfall zu testen, der auch für die bibliothekarische Arbeit relevant ist. Nach der Bearbeitung von Großbeständen in der ersten Phase sollte nun der Frage nachgegangen werden, inwieweit der Einsatz von automatischer Texterkennungssoftware für Handschriften auch bei Einzelbänden sinnvoll ist. Wenn eine Handschrift nur einen Umfang von 200 Seiten hat, ist der Aufwand für die Erstellung eines eigenen Texterkennungsmodells vergleichsweise hoch. Hier lassen sich die in der ersten Projektphase gewonnenen Erkenntnisse nachnutzen.

2. Rahmenbedingungen und Projektdurchführung

Im Projekt OCR-BW[2] wurde für die automatische Texterkennung von Handschriften vornehmlich die Transkriptionsplattform Transkribus eingesetzt. Der Grundstein für Transkribus wurde im Rahmen des Projekts transScriptorium (2013–2015) gelegt. Anschließend wurde Transkribus innerhalb des ebenfalls EU-geförderten Projekts READ (2016–2019) weitergeführt.[3] Mit dem Ende der öffentlichen Förderung wurde die READ-COOP SCE, eine Genossenschaft nach europäischem Recht, gegründet, um den Betrieb und die Weiterentwicklung der Plattform zu sichern und zu fördern.[4] Transkribus ist aktuell in einer Desktop-Version (Expert Client) sowie als im Funktionsumfang reduzierte Browser-Version (Transkribus Lite) nutzbar.[5] Beide Versionen bieten ein User Interface, so dass keine speziellen IT-Kenntnisse benötigt werden. Mit Einführung eines „Freemium-Finanzierungsmodells" im Oktober 2020 ist die automatische Texterkennung nicht mehr kostenlos, sondern wird seitenweise abgerechnet.[6] In der zweiten Phase des Projekts OCR-BW wurden zudem erste Testläufe mit der

2 Wir bedanken uns für Anmerkungen und Hinweise zu diesem Beitrag bei Stefan Weil, Larissa Will und Jan Kamlah (UB Mannheim), Marianne Dörr, Regina Keyler, Annika Timmins und Olaf Brandt (UB Tübingen) sowie Sebastian Colutto (READ-COOP).

3 Siehe: Muehlberger, Guenter; Seaward, Louise; Terras, Melissa; Ares Oliveira, Sofia; Bosch, Vicente; Bryan, Maximilian; Colutto, Sebastian; Déjean, Hervé; Diem, Markus; Fiel, Stefan; Gatos, Basilis; Greinoecker, Albert; Grüning, Tobias; Hackl, Guenter; Haukkovaara, Vili; Heyer, Gerhard; Hirvonen, Lauri; Hodel, Tobias; Jokinen, Matti; Kahle, Philip; Kallio, Mario; Kaplan, Frederic; Kleber, Florian; Labahn, Roger; Lang, Eva Maria; Laube, Sören; Leifert, Gundram; Louloudis, Georgios; McNicholl, Rory; Meunier, Jean-Luc; Michael, Johannes; Mühlbauer, Elena; Philipp, Nathanael; Pratikakis, Ioannis; Puigcerver Pérez, Joan; Putz, Hannelore; Retsinas, George; Romero, Verónica; Sablatnig, Robert; Sánchez, Joan Andreu; Schofield, Philip; Sfikas, Giorgos; Sieber, Christian; Stamatopoulos, Nikolaos; Strauß, Tobias; Terbul, Tamara; Toselli, Alejandro Héctor; Ulreich, Berthold; Villegas, Mauricio; Vidal, Enrique; Walcher, Johanna; Weidemann, Max; Wurster, Herbert; Zagoris, Konstantinos: Transforming scholarship in the archives through handwritten text recognition. Transkribus as a case study, in: Journal of Documentation, 75 (5), 2019, S. 957. Online: <https://doi.org/10.1108/JD-07-2018-0114>, Stand 08.09.2022.

4 Die UB Tübingen ist seit Januar 2020 Mitglied in der READ-COOP SCE. Als Mitglied der READ-COOP bekommt die Einrichtung nicht nur ein Mitspracherecht bei der weiteren Entwicklung, sondern auch Preisnachlässe für die mittlerweile bepreisten Dienstleistungen der Transkribus-Plattform. Die Zusammenarbeit mit der READ-COOP als Mitglied ist insgesamt als positiv zu werten. Gerade kleinere Feature-Requests werden häufig zeitnah umgesetzt und bei den regelmäßigen Programm-Updates installiert. Durch die Umsetzung von derartigen Nutzerwünschen wird die Benutzbarkeit von Transkribus sowie die Qualität der Ergebnisse der automatischen Texterkennung fortwährend verbessert.

5 Siehe: <https://readcoop.eu/transkribus/>, Stand: 08.09.2022.

6 Alle anderen Funktionen, wie Layouterkennung und Modelltraining, bleiben kostenlos nutzbar.

Open-Source-Alternative eScriptorium durchgeführt, wofür die von der UB Mannheim eingerichtete Instanz genutzt wurde.[7]

In der ersten Projektphase (2019–2021) sind für die Texterkennung vornehmlich größere Textkorpora ausgewählt worden, die über längere Zeiträume entstanden sind, von einem oder mehreren Schreibern verfasst wurden und verschiedene Sprachen, Zeichensysteme und Schriftarten beinhalten. Gerade die mehrere Jahrhunderte und viele Regalmeter umfassenden juristischen Konsilien und Senatsprotokolle der Universität Tübingen decken in dieser Hinsicht ein breites Spektrum ab. Aber auch die Schriftzeugnisse der Einzelautoren Martin Crusius (1526–1607) und Edwin Hennig (1882–1977) sind nur auf den ersten Blick homogen, da das Oeuvre der beiden Schreiber jeweils mehrere Jahrzehnte umspannt und sprachlich divers ist. Jedoch wurde bei der Antragstellung für das OCR-BW-Projekt geplant, aus den zum Teil sehr großen Korpora für die Bearbeitung im Projekt nur einzelne Bände auszuwählen, welche somit jeweils nur einen begrenzten Schreibzeitraum aufwiesen und möglichst nur eine Schreiberhand enthalten sollten. Die Vorannahme zum Zeitpunkt der Projektplanung war, dass das Material für eine erfolgreiche automatische Texterkennung von Handschriften homogen sein müsse. Im Folgenden soll beispielhaft gezeigt werden, ob sich dies bewahrheitet hat und welche Anpassungen bei der Bearbeitung der Textkorpora vorgenommen worden sind.

Für die ausgewählten Dokumente wurden als Grundlage für das Training der Texterkennungsmodelle, die auf neuronalen Netzen beruhen und zur automatischen Transkription angewendet werden können,[8] jeweils Ground-Truth-Daten (GT) erzeugt. Ground-Truth-Daten sind vom Begriff her als hundertprozentig „wahre" Transkriptionen zu verstehen, was jedoch dahingehend relativiert werden muss, dass verschiedene Transkriptionsrichtlinien angewandt werden können und die Richtigkeit somit unterschiedlich definierbar ist. Durch eine Standardisierung der Daten soll zum einen die Nutzbarkeit des automatischen Outputs erhöht, wie auch die Nachnutzung der GT-Daten in anderen Kontexten ermöglicht werden. Für den Bereich der Handschriften gibt es diesbezüglich noch keine anerkannte Normierung, so dass so weit wie möglich die Transkriptionsrichtlinien von OCR-D für Drucke zur Orientierung genutzt werden.[9] Im Projekt OCR-BW wurde entschieden, so dokumentnah und zeichengetreu wie möglich zu transkribieren, um den Nutzerinnen und Nutzern die Informationen des Originals möglichst unverfälscht zur Verfügung zu stellen. Von diesem Ausgangspunkt lässt sich auf Wunsch eine Normalisierung durch Ersetzen von z.B. Sonderzeichen leichter automatisch vornehmen, als dies in umgekehrter Richtung möglich wäre. Trotzdem müssen im Sinne der Praktikabilität auch

7 Siehe: Kiessling, Benjamin; Tissot, Robin; Stökl Ben Ezra, Daniel; Stokes, Peter: eScriptorium. An Open Source Platform for Historical Document Analysis, in: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), 2019, S. 19–24. Online: <https://doi.org/10.1109/ICDARW.2019.10032> sowie <https://gitlab.com/scripta/escriptorium/>, Stand: 08.09.2022.

8 Siehe: Michael, Johannes; Weidemann, Max; Labahn, Roger: HTR Engine Based on NNs P3. Optimizing speed and performance – HTR+, READ-H2020 Project 674943, Deliverable D7.9, 2018. Online: <https://readcoop.eu/wp-content/uploads/2018/12/Del_D7_9.pdf>, Stand: 08.09.2022.

9 Siehe: <https://ocr-d.de/de/gt-guidelines/trans/index.html>, Stand: 08.09.2022. Das OCR-BW-Projekt orientiert sich an Level 2 der OCR-D-Guidelines. Vgl. auch: Boenig, Matthias; Federbusch, Maria; Herrmann, Elisa; Neudecker, Clemens; Würzner, Kay-Michael: Ground Truth. Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities?, in: DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts. Universität zu Köln, 26. Februar bis 2. März 2018, 2018, S. 219–223. Online: <https://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf>, Stand: 08.09.2022.

Abstriche gemacht werden, da nicht alle notwendigen Zeichen im offiziellen Unicode-Zeichensatz zur Verfügung stehen. Dies stellt ein Hindernis sowohl für die Bearbeitung in Transkribus wie auch für die Darstellung der Volltexte dar, da die Zeichen in der Private-Use-Area (PUA)[10] von Unicode von vielen Fonts nicht unterstützt werden. Zudem gibt es für einige Zeichen bislang weder im offiziellen Unicode-Zeichensatz noch in der PUA bzw. bei MUFI[11] überhaupt einen Codepoint, so dass hier nur Annäherungen gefunden werden konnten. Wenn möglich, wurden in diesen Fällen Zeichen gewählt, die von der Bedeutung her dem handschriftlichen Symbol entsprechen, auch wenn die Optik abweicht.

Bewertet werden die Ergebnisse auf Grundlage der Zeichenfehlerrate (engl. Character Error Rate, CER) der entwickelten Texterkennungsmodelle.[12] Angelehnt an die DFG-Richtlinien für Druckerkennung wird ein Ergebnis einer CER von unter 5 % angestrebt.[13] Aktuelle Erfahrungen bei der Texterkennung von Handschriften lassen eine Bewertung in drei Kategorien zu, wobei eine CER von 1.) unter 10 % als gut, 2.) unter 5 % als sehr gut und 3.) unter 2,5 % als exzellent zu bewerten ist.[14] Diese Ergebnisse können jedoch nur als Anhaltspunkt gewertet werden, da sie von der Auswahl des Validation Sets abhängig sind.[15] Je nachdem, ob die hierfür ausgewählten Seiten besonders „leicht“ oder „schwierig“ sind, variiert die CER und ist dementsprechend beeinflussbar. Um möglichst belastbare Aussagen über die zu erwartende CER auf noch unbearbeiteten Textteilen treffen zu können, wurde versucht, das Validation Set möglichst repräsentativ zu gestalten. Auch die Größe des Validation Sets hat einen Einfluss, da den einzelnen Seiten je nachdem eine geringe oder große Gewichtung hinsichtlich des Ergebnisses zukommt. Als gute Größe hat sich ein Wert von ca. 10 % der Ground-Truth-Daten erwiesen, um keinen zu großen Anteil der GT für den Trainingsprozess zu verlieren, aber dennoch ein belastbares Ergebnis zu erhalten.

10 Siehe: <http://www.unicode.org/faq/private_use.html>, Stand: 08.09.2022.

11 Die communitybasierte Medieval Unicode Font Initiative (MUFI) verfolgt das Ziel, für fehlende Zeichen vorrangig aus mittelalterlichen Texten eine Aufnahme in den Unicode-Zeichensatz zu erwirken. Siehe hierzu: <https://mufi.info>, Stand: 08.09.2022. Die dort veröffentlichten Code Charts bieten, wie auch die Beispiele der OCR-D-Guidelines, einen guten Überblick und Einstieg bei der Suche nach einer maschinenlesbaren Wiedergabe von Sonderzeichen, die in historischen Dokumenten des europäischen Kulturraums vorliegen.

12 Die aufgeführten Werte für die Evaluierung der Ergebnisse sind den in Transkribus integrierten Werkzeugen entnommen. Die für ein Modell berechnete CER kann der Modellbeschreibung entnommen werden. Außerdem gibt es noch weitere Vergleichswerkzeuge, mit denen die Fehlerrate eines Modells auf den vorliegenden GT-Daten berechnet werden kann.

13 Siehe: Deutsche Forschungsgemeinschaft: DFG-Praxisregeln „Digitalisierung“. DFG-Vordruck 12.151-12/16, 2016. Online: <https://www.dfg.de/formulare/12_151/>, Stand: 08.09.2022.

14 Vgl. Hodel, Tobias; Schoch, David; Schneider, Christa; Purcell, Jake: 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, S. 2. Online: <https://doi.org/10.5334/johd.46>, Stand: 08.09.2022.

15 Für ein Modelltraining werden die Ground-Truth-Daten manuell oder automatisch in ein Training Set und ein Validation Set aufgeteilt. Anhand des Training Sets lernt die Software, während sie sich gleichzeitig im Trainingsprozess am Validation Set selbst prüft. Das Validation Set entstammt ebenfalls den GT-Daten, aber da es nicht im Trainingsprozess genutzt wird, ist es für das Modell unbekanntes Material. Somit kann anhand der Performance des Modells auf dem Validation Set geschlossen werden, welche Fehlerrate für noch nicht transkribiertes Material zu erwarten ist, das dem Validation Set ähnlich ist.

3. Vorgehensweise bei der Bearbeitung eigener Textkorpora

3.1 Tagebücher von Edwin Hennig

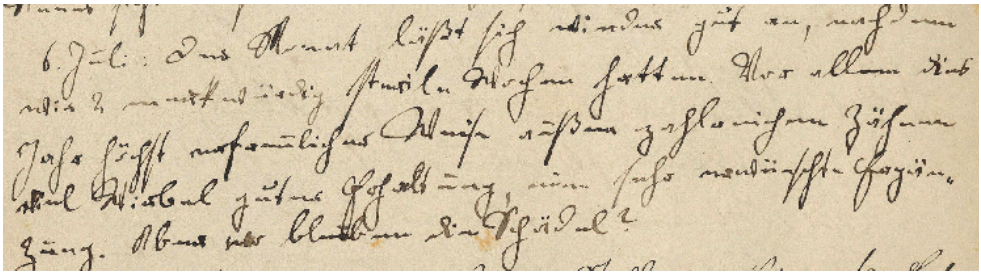

Abbildung 1: Schriftprobe aus den Tagebüchern von Edwin Hennig (Tübingen, Universitätsarchiv, UAT 407/81, S. 22).

Das erste bearbeitete Korpus waren die Tagebücher des Geologen und Paläontologen Edwin Hennig (1882–1977). Aus seinen hinterlassenen Tagebüchern, die er in den Jahren 1897–1973 – von seiner Schulzeit bis ins hohe Alter – abfasste, wurden zunächst drei Bände ausgewählt. Die Tagebücher führte Hennig im Rahmen der Tendaguru-Expeditionen in Afrika (1909–1911).[16] Für diesen Bestand wurden Ground-Truth-Daten erzeugt, indem der Reihe nach Seiten aus den Expeditionstagebüchern transkribiert worden sind. Nach 100 Seiten wurde ein erstes Texterkennungsmodell mit einer CER von 8,88 % trainiert. Um den gewünschten Wert einer CER von unter 5 % zu erreichen, wurden weitere 65 Seiten transkribiert, womit eine CER von 4,32 % erzielt wurde. Bei Anwendung dieses Modells auf frühere oder spätere Jahrgänge zeigte sich jedoch eine spürbar schlechtere Performance, die mit Schriftveränderungen über den langen Schreibzeitraum hinweg erklärt werden kann. Um dies auszugleichen, wurden insgesamt 75 weitere Seiten aus Tagebüchern der Schulzeit Hennigs (1897–1901) sowie den Jahren 1944–46 und 1961–62 transkribiert und dem Modelltraining hinzugefügt.[17] Während die Texterkennung mit den Modellen 1–3 auf diesen Seiten eine deutlich höhere CER aufwies,

16 Siehe: Maier, Gerhard: African dinosaurs unearthed. The Tendaguru expeditions, Bloomington, Ind. 2003 (Life of the Past) und Heumann, Ina; Stoecker, Holger; Tamborini, Marco; Vennen, Mareike: Dinosaurierfragmente. Zur Geschichte der Tendaguru-Expedition und ihrer Objekte, 1906–2018, Göttingen 2018. Die Expeditions-Tagebücher waren bereits vor Beginn des OCR-BW-Projekts von der Wissenschaft stark nachgefragt, da sie für viele Disziplinen der Geistes- und Naturwissenschaften von Interesse sind. Die Online-Stellung der Volltexte hat jedoch weitere Anfragen generiert.

17 Der Trainingsprozess wurde so gestaltet, dass bei weiteren Trainings nach der Erzeugung zusätzlicher GT-Daten für ein Korpus das Vorgängermodell als sogenanntes Base Model, also als ein Grundlagenmodell, verwendet wurde. Auf diese Weise wird das für das Base Model trainierte neuronale Netz als Grundlage für das neue Training genutzt. Hierfür kann neben einem Vorgängermodell mit eigenen GT-Daten auch ein passendes generisches Modell genutzt werden. Oftmals kann auf diese Weise das Ergebnis des neuen Modells ohne zusätzlichen Aufwand verbessert werden bzw. im Fall der Nutzung eines generischen Modells schon mit weniger GT-Daten ein besseres Ergebnis erzielt werden. Zu beachten ist jedoch, dass bei einer Trainingsreihe von ca. sechs konsekutiven Durchläufen eine Art Degenerationseffekt einsetzt und das Modelltraining kein nutzbares Ergebnis erzielt (Dirk Alvermann, Universitätsarchiv Greifswald, pers. Komm.). Siehe zum Einsatz von Base Models: Ströbel, Phillip; Clematide, Simon; Volk, Martin; Schwitter, Raphael; Hodel, Tobias; Schoch, David: Evaluation of HTR models without Ground Truth Material. Preprint 2022, S. 6. Online: <https://www.researchgate.net/publication/357927928_Evaluation_of_HTR_models_without_Ground_Truth_Material>, Stand 08.09.2022.

konnten mit den Modellen 4–6, welche GT-Daten aus den entsprechenden Jahrgängen enthalten,[18] auch für diese Seiten Werte im Bereich der gewünschten Fehlerrate erreicht und so die Gesamtperformance hin zu einer CER von 4,05 % noch einmal verbessert werden.[19] Mit bei der Erkennung hinzugeschaltetem Language Model (LM)[20] sank der Wert auf 3,61 %.

Validation Set	CER für Modell UAT_M1 in %	UAT_M2	UAT_M3	UAT_M4	UAT_M5	UAT_M6	UAT_M6/LM
S. 1 (1910)	2,97	2,03	1,88	1,88	2,03	1,74	1,38
S. 2 (1910)	3,09	3,09	2,7	2,08	2,47	1,85	1,62
S. 3 (1910)	18,7	17,58	17,36	16,69	16,01	15,23	14,39
S. 4 (1910)	8,96	9,3	9,3	8,69	8,96	8,69	8,08
S. 5 (1911)	1,92	1,37	0,89	1,24	0,82	1,1	0,89
S. 6 (1911)	2,2	1,52	1,24	1,24	1,52	1,31	1,17
S. 7 (1911)	3,15	1,8	1,75	1,97	1,69	1,63	1,52
S. 8 (1911)	2,12	1,33	1,27	1,03	1,39	1,15	0,97
S. 9 (1911)	2,07	1,31	0,98	0,92	1,03	0,98	0,92
S. 10 (1944)	18,78	19,5	17,01	8,09	6,74	6,64	5,71
S. 11 (1945)	11,72	8,54	8,39	3,4	3,04	3,4	2,97
S. 12 (1961)	16,72	14,37	12,35	8,05	2,64	2,64	2,64
S. 13 (1961)	20,8	16,79	15,68	11,42	3,52	2,9	2,72
S. 14 (1897)	57,69	53,95	49,64	46,19	35,7	6,54	4,53
S. 15 (1899)	30,64	29,95	23,7	22,64	16,23	4,27	3,35
S. 16 (1901)	14,17	12,31	11,66	10,45	8,26	4,62	4,45
CER Durchschnitt	8,88 (12,94 auf Validation Set UAT_M6)	5,58 (11,6)	4,32 (10,52)	4,31 (8,86)	4,03 (6,81)	4,05	3,61
Seiten Training Set	95	126	156	184	212	224	224
Seiten Validation Set	5	7	9	11	13	16	16
GT insgesamt	100	133	165	195	225	240	240

Abbildung 2: Trainingsreihe für die Hennig-Tagebücher.[21]

18 Das finale Modell M6 beruht auf 240 Seiten GT (47.934 Wörtern).

19 Die ersten Volltexte sind bereits in die Onlinepräsentationsplattform DigiTue der UB Tübingen eingespeist: <http://idb.ub.uni-tuebingen.de/opendigi/UAT_407_080>, <http://idb.ub.uni-tuebingen.de/opendigi/UAT_407_081>, <http://idb.ub.uni-tuebingen.de/opendigi/UAT_407_082> und auch auf GitHub veröffentlicht: <https://github.com/ubtue/Ground-Truth/>, Stand: 08.09.2022.

20 Die in Transkribus einsetzbaren Language Models, also Sprachmodelle, werden im Trainingsprozess auf Grundlage der GT-Daten erstellt. Für die hier eingesetzte Technologie HTR+ wird ein 8-gram zeichenbasiertes Language Model verwendet. Dies bedeutet, dass beim Trainingsprozess Sequenzen aus acht aufeinanderfolgenden Zeichen je nach ihrer Häufigkeit in den Trainingsdaten mit den unterschiedlichen Wahrscheinlichkeiten ihres Vorkommens bewertet werden. Wenn das Language Model nun bei der automatischen Texterkennung hinzugeschaltet wird, wird der Output für die vorliegenden Zeichenfolgen entsprechend dieser Wahrscheinlichkeiten gewichtet, da die durch die Software ausgegebene automatische Transkription nicht die einzige, sondern die wahrscheinlichste Lösung darstellt. Je mehr also die Sprache des zu erkennenden Dokuments mit jener der GT-Daten des eingesetzten Modells übereinstimmt (das bezieht sich nicht nur auf die Sprache an sich, sondern auch auf das verwendete Vokabular und die Grammatik), desto besser ist der Effekt eines Language Models. Dies ist zu unterscheiden vom Einsatz von Wörterbüchern, die den von der Software erkannten Text mit einem festen Bestand von Wörtern abgleichen und entsprechend korrigieren. Vgl. hierzu: Strauß, Tobias; Weidemann, Max; Labahn, Roger: Language Models. Improving transcriptions by external language resources, READ-H2020 Project 674943, Deliverable D7.12, 2018. Online: <https://readcoop.eu/wp-content/uploads/2018/12/D7.12_LMs.pdf>, Stand 08.09.2022.

21 Für die in das jeweilige Modelltraining als Validation Set einbezogenen Seiten sind die CER-Werte schwarz dargestellt, während die grau markierten Werte zeigen, welches Ergebnis die Modelle auf erst im weiteren Trainingsverlauf inkludierten Daten erzielen. Die Abkürzung UAT steht für Universitätsarchiv Tübingen.

Diese ersten Ergebnisse und Erkenntnisse führten zu einer Neuausrichtung der Projektplanung, um das Verhältnis von Kosten und Nutzen zu optimieren. Müssen für jeden einzelnen Band GT-Daten im dreistelligen Bereich erzeugt werden, ist der Aufwand je nach Gesamtseitenzahl des Bandes im Verhältnis zu den Seiten, die noch für eine automatische Texterkennung übrigbleiben, sehr hoch. Die zweite Phase des Modelltrainings der Hennig-Tagebücher hat jedoch gezeigt, dass auch zeitlich heterogenes Material – in diesem Fall Dokumente eines Schreibers, die über einen langen Schreibzeitraum entstanden sind – mit genauso gutem Ergebnis in einem Modell vereinigt werden kann. Ein solches, übergreifendes Modell kann anschließend für die automatische Texterkennung auch eines großen Bestands genutzt werden. Damit ist das Verhältnis von manuell transkribierten zu automatisch zu transkribierenden Seiten deutlich günstiger. Aus der Tatsache, dass dieses Korpus zwar nur eine Schreiberhand aufweist, die jedoch Veränderungen unterworfen ist, lässt sich ableiten, dass dies auch für Dokumente mit unterschiedlichen Schreiberhänden in zumindest einem begrenzten Schreibzeitraum und mit ähnlicher Schrift gelten sollte.

3.2 Griechische Predigtmitschriften und lateinische Tagebücher von Martin Crusius

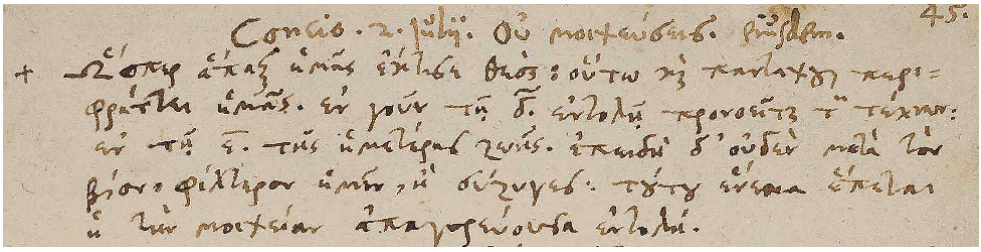

Abbildung 3: Schriftprobe aus den griechischen Predigtmitschriften von Martin Crusius (Tübingen, UB, Mb 19-4, S. 45).

Als Konsequenz aus dieser Erkenntnis wurde für die griechischen Predigtmitschriften von Martin Crusius (1526–1607) nicht nur der bereits ausgewählte, vergleichsweise schön geschriebene Band bearbeitet, sondern ein Korpus von 20 Bänden zusammengestellt, das neben den im Zeitraum von 1563 bis 1604 verfassten Bänden Predigtmitschriften auch die Abschrift eines Martyrologiums durch Crusius sowie einen Band Predigtmitschriften von Samuel Grammer enthält. Aus diesen 20 Bänden wurden jeweils zehn Seiten für die GT-Erzeugung[22] ausgewählt, die die jeweiligen Bände – nach einer kursorischen Durchsicht – in Bezug auf Layout, Beschreibstoff, Beschädigungen, Inhalt, Schrift und die verwendete Tinte so gut wie möglich repräsentieren. Das Ergebnis spiegelt den Erfolg dieser Herangehensweise wider, da die Fehlerquoten auf dem Validation Set, das aus je einer Seite aus jeder Handschrift besteht, für alle Jahrgänge einheitliche Werte ohne große Ausschläge aufweisen. Im Durchschnitt wird für die 18 Bände Predigtmitschriften von Crusius eine Fehlerrate von 3,54 % (3,41 % mit Language Model) erzielt. Betrachtet man allein den Bestand der griechischen Predigtmitschriften von der Hand Martin Crusius', können mit einem Aufwand von 180 Seiten GT (43.020 Wörter) über 10.000 Seiten automatisch erkannt werden. Bei einem Modell, welches die Hand von

22 Die griechischen Transkriptionen wurden von Renate Burri (Bern) erstellt.

Samuel Grammer sowie das Martyrologium miteinbezieht, ist die Erkennung ebenfalls zufriedenstellend, wenn sie auch im Nachkommastellenbereich ein wenig schlechter abschneidet (3,83 % CER).[23] Ein spezialisiertes Modell führt also für einen spezifischen Bestand zum besten Ergebnis. Dennoch macht es Sinn, ein Modell innerhalb eines gewissen Rahmens zu erweitern, um das Kosten-Nutzen-Verhältnis zu optimieren.

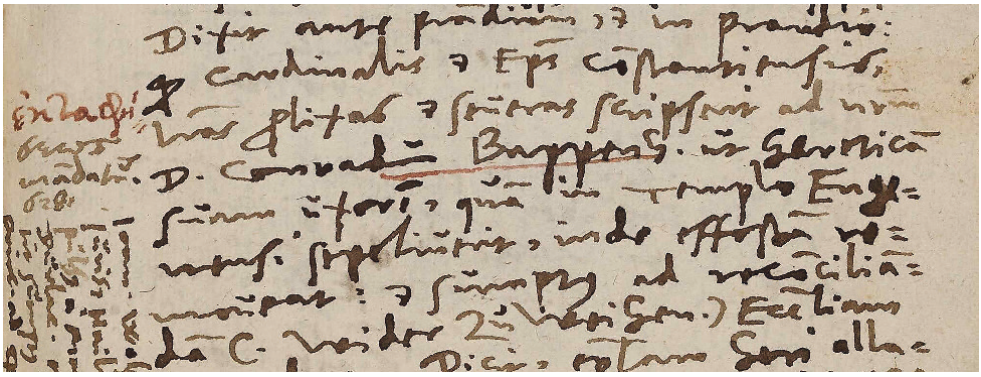

Abbildung 4: Schriftprobe aus den lateinischen Tagebüchern von Martin Crusius (Tübingen, UB, Mh 466-4, S. 627).

Für die Jahre 1596 bis 1605 der zehn lateinischen Tagebücher (1573–1605) von Martin Crusius liegt bereits eine gedruckte Edition vor.[24] Grundsätzlich ist es sinnvoll, bereits vorliegende Transkriptionen nachzunutzen, um den Aufwand der Erzeugung von GT-Daten zu verringern. Transkribus bietet mehrere Optionen, um diese einzufügen. Am einfachsten ist es, wenn die Daten als XML- oder TXT-Datei vorliegen, welche leicht automatisch eingespeist werden können. Dies war bei der vorliegenden Edition nicht der Fall, so dass ein semi-automatisches Verfahren angewandt wurde. Zunächst wurde ein Editionsband digitalisiert und mit ABBYY ein Volltext erstellt. Dieser wurde für die Einspeisung in Transkribus aufbereitet, indem er zeilengenau gegliedert sowie die Lesereihenfolge bearbeitet wurde; auch wurden die Transkriptionsrichtlinien angepasst und einige Fehllesungen korrigiert. Im Anschluss wurde die Transkription per Copy and Paste seitenweise eingefügt. Das Verfahren war recht aufwendig und hat in diesem Fall gegenüber einer manuellen Transkription zu keiner Zeitersparnis geführt, so dass es nach einem ersten Durchlauf wieder aufgegeben wurde. Wenn allerdings eine Edition zur Verfügung steht, die den gewünschten Transkriptionsrichtlinien entspricht und im besten Falle bereits zeilengenau vorliegt, kann diese gewinnbringend für die Erstellung von GT-Daten nachgenutzt werden. Gerade digitale Editionen dürften hierfür eine gute Option darstellen.

Eine Arbeitserleichterung bei der Erstellung der GT-Daten konnte jedoch dadurch erreicht werden, dass im Gegensatz zum ersten bearbeiteten Korpus der Hennig-Tagebücher, für welches erst mit 100 Seiten GT ein erstes Modell trainiert wurde, der Prozess der GT-Erstellung nun von einem steten

23 Dieses Modell beinhaltet die komplette GT von 200 Seiten (47.706 Wörter).

24 Siehe: Crusius, Martin: Diarium Martini Crusii, hrsg. von Wilhelm Goez, Ernst Conrad, Reinhold Stahlecker, Eugen Staiger unter Mitw. von Reinhold Rau und Hans Widmann, 4 Bde., Tübingen 1927–1961.

Modelltraining begleitet wurde. Bei diesem wie auch bei den weiteren bearbeiteten Korpora konnte bereits mit ca. 20–25 Seiten GT ein erstes Modell trainiert werden, das zwar in der Regel noch nicht das gewünschte Ergebnis einer CER von unter 5 % erreichte, aber als Transkriptionsgrundlage für die weitere GT-Erzeugung genutzt werden konnte.[25] Mit regelmäßigem Nachtraining nach Fertigstellung weiterer GT-Daten konnte die GT-Erzeugung durch zunehmend geringeren Korrekturaufwand immer weiter beschleunigt werden.[26] Mit insgesamt 136 Seiten GT (24.548 Wörter) verteilt auf die zehn überlieferten Tagebuchbände (1573–1605) wurde ein Modell für die Textteile in lateinischer Sprache trainiert, das eine CER von 5,13 % erreicht (4,66 % mit Language Model). Neben Latein beinhalten die Tagebücher auch relevante Textteile in deutscher Sprache, die in das Modell hineintrainiert wurden, sowie in griechischer Sprache, die bei diesem Modelltraining ausgespart wurden. Von allen bearbeiteten Textkorpora war das Ergebnis in diesem Fall am schlechtesten. Dies scheint auf eine vergleichsweise unregelmäßige Schriftausprägung, viele Abkürzungen sowie auf das komplexe Layout zurückzuführen zu sein.

Da die griechischen Predigtmitschriften auch lateinische und deutsche Textteile enthalten, wie ebenfalls in die lateinischen Tagebücher viele griechische Wörter eingefügt sind, wurde der Versuch unternommen, beide Sprach- und Zeichensysteme in einem Modell zu vereinen. Zwar können bei Transkribus verschiedene Texterkennungsmodelle auf unterschiedliche Textregionen angewandt werden, wofür diese jedoch entsprechend ausgezeichnet werden müssten und was bei einem Sprachwechsel in einer Zeile nicht weiterhilft. Bei der Modelltrainingsreihe für die griechischen Tagebücher hatte sich bereits gezeigt, dass bei einer Inklusion der lateinischen und deutschen Anmerkungen in das Modelltraining diese zwar durchaus auch automatisch erkannt wurden, dabei aber größere Schwächen auftraten und die Performance des Modells bezogen auf die CER um ca. 1 % abnahm. Dies ist wohl darauf zurückzuführen, dass das zur Verfügung stehende Trainingsmaterial für die Textteile in lateinischer Sprache vergleichsweise gering war und daher zum einen für eine sichere Erkennung des lateinischen Textes nicht ausreichte und dies zum anderen aber auch zu Unsicherheiten im griechischen Text führte.[27] Eine Zusammenführung der beiden GT-Sätze für die griechischen Predigtmitschriften und die lateinischen Tagebücher von insgesamt 346 Seiten (72.254 Wörter) führte zu einem Modell mit einer stabilen Zeichenfehlerrate von 4,52 % (4,22 % mit Language Model) für beide Zeichensysteme. Dieses Kombinationsmodell kann somit ohne Differenzierung nach sprachlichen Textregionen auf die beiden Bestandsgruppen angewandt werden.

25 Der Zeitaufwand bei der Erstellung der GT-Daten hängt grundsätzlich davon ab, wie gut die vorliegende Schrift gelesen und wie schnell getippt werden kann. Ein weiterer Faktor bei einer zeichengetreuen Transkription ist oftmals das Auffinden eines passenden Codepoints im Unicode-Zeichensatz für Sonderzeichen. Durch den Einsatz einer automatischen Transkription als Grundlage für die GT-Erstellung kann der Zeitaufwand je nach Qualität der automatischen Transkription sowie der Übereinstimmung der Transkriptionsregeln, wenn ein generisches Modell genutzt wird, erheblich verringert werden.

26 Bei einem Korpus wie den Crusius-Tagebüchern ist es bei dieser Herangehensweise am sinnvollsten, die zeitlich aufeinanderfolgenden Bände der Reihe nach zu bearbeiten und nicht vom ersten zum letzten zu springen, damit der Schriftveränderung sukzessive Rechnung getragen werden kann.

27 Ähnliches ließ sich auch für die Tagebücher von Edwin Hennig beobachten, die Einzelwörter in afrikanischen Sprachen enthalten. Diese wurden mit zunehmender GT-Menge immer besser erkannt, zumal wenn die Begriffe wiederholt worden sind; sie blieben aber vergleichsweise fehleranfällig.

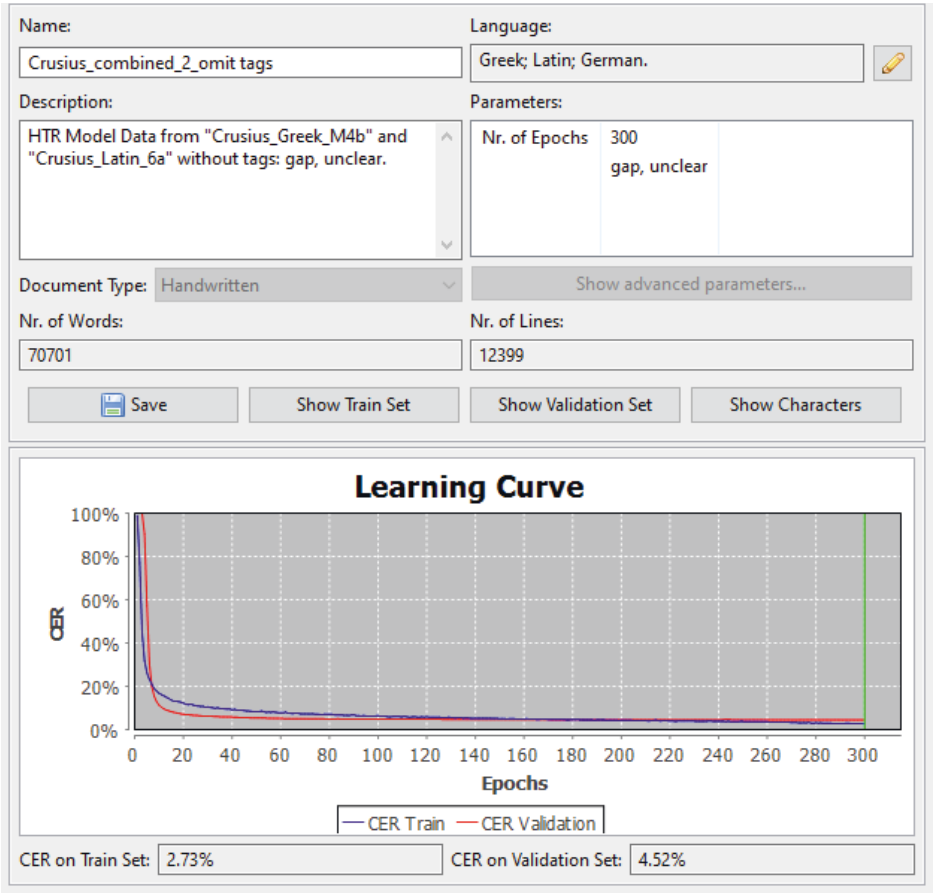

Abbildung 5: Modellbeschreibung und Evaluierung im Transkribus Expert Client. In diesem Fall wurden Zeilen vom Training ausgeschlossen, die die Tags „Unclear" und „Gap" enthalten, mit welchen nicht oder nur uneindeutig lesbare Zeichen oder Worte markiert werden können (Transkribus Expert Client V. 1.20.1).

3.3 Juristische Konsilien

Für das Korpus der juristischen Konsilien wurden keine GT-Daten von Grund auf erstellt. Da zu diesem Zeitpunkt schon einige generische Modelle für deutsche Kurrentschrift öffentlich zugänglich waren, wurde unter deren Verwendung bereits eine erste Transkriptionsgrundlage erzeugt.[28] Neben der eher subjektiven Methode eines Tests verschiedener Modelle auf einer ausgewählten Seite, um die Eignung dieser für das eigene Dokument abzuschätzen, bietet Transkribus ein Tool, mit Hilfe dessen die zu erwartende Performance der verschiedenen Modelle für ein Korpus berechnet werden kann. Hierfür

28 Zur Anwendung von generischen Modellen auf unbekanntem Material siehe bei Hodel: General Models, 2021 (wie Fußnote 14).

wird ein sogenanntes Sample Set erstellt, wobei eine selbst festgelegte Zahl von Zeilen automatisch aus den gewünschten Dokumenten ausgewählt und in einem neuen Dokument zusammengestellt wird. Für das so erzeugte Sample Set, das pro Seite eine Zeile enthält, müssen nun GT-Daten als Vergleichsgrundlage erstellt werden. Im Anschluss können Modelle auf das Sample Set angewendet und deren CER berechnet werden.[29] Hierbei wird ein oberer und ein unterer Grenzwert sowie ein Mittelwert der wahrscheinlich zu erwartenden CER angegeben. Auf dieser Entscheidungsgrundlage wurde für die juristischen Konsilien das von Dirk Alvermann (Universitätsarchiv Greifswald) veröffentlichte Modell „Acta_17 HTR+" ausgewählt, für welches eine CER von 7,198 % als anzunehmender Mittelwert berechnet wurde.[30]

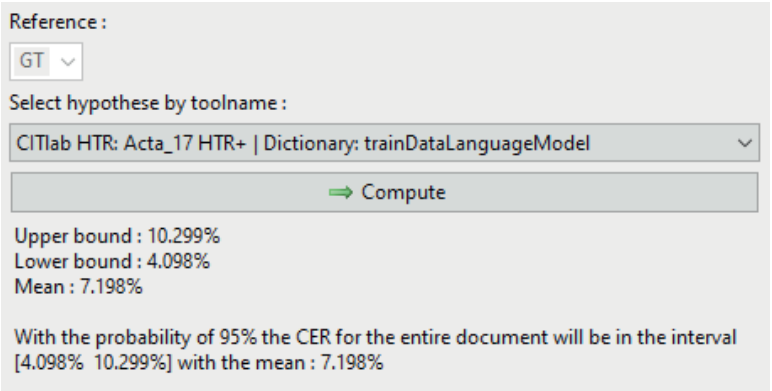

Abbildung 6: Berechnung der zu erwartenden CER mit Hilfe des Werkzeugs „Compare Samples" im Transkribus Expert Client (Transkribus Expert Client V. 1.20.1).

Auf diese Weise wurden zwei Bände der juristischen Konsilien aus dem 17. Jahrhundert bearbeitet, die unter Berücksichtigung von wissenschaftlichem und öffentlichem Interesse ausgewählt worden waren.[31] Es wurden 223 Seiten GT (41.573 Wörter) erstellt, die mehrere Schreiber beinhalten.[32] Tatsächlich konnte bereits mit dem ersten Modell, das auf 26 Seiten GT beruht, eine Fehlerrate von 4,72 % erreicht werden. Letztlich wäre noch nicht einmal dieser Aufwand notwendig gewesen, um das Ziel einer CER von unter 5 % zu erreichen, denn bei einer Akzeptanz der Transkriptionsrichtlinien des Acta_17-Modells – also vor allem der Normalisierung der s-Formen und der Auflösung von Abkürzungen, die im Sample Set als Fehler gewertet worden waren – wurde das entsprechend angepasste Validation

29 Hierfür werden keine Gebühren berechnet.

30 Siehe den im Rahmen des DFG-Projekts „Archivische Findmittel und Quellen. Digitalisierung und Erschließung von Quellen zur gerichtlichen und höchstrichterlichen Entscheidungsfindung im Ostseeraum" geführten Blog: <https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/de/>, Stand: 08.09.2022.

31 Die Bearbeitung eines breiteren Spektrums war aufgrund fehlender Kapazitäten bei der Digitalisierung nicht möglich. Im Gegensatz dazu konnten für die Senatsprotokolle sieben für den Zeitraum von 1799–1847 repräsentativ ausgewählte Bände digitalisiert und bearbeitet werden. Auch bei dieser größeren Zeitspanne sowie der Vielzahl an Schreiberhänden (sowohl Berufsschreiber wie auch in Vertretung Professoren) wurde auf Grundlage von 214 Seiten GT eine CER von 4,60 % (4,24 % mit Language Model) erreicht.

32 Die hohe Anzahl an GT-Daten war hierbei nicht der Notwendigkeit für die Modellentwicklung, sondern dem Projektantrag geschuldet.

Set bereits mit der gewünschten Fehlerrate erkannt. Das eigene spezialisierte Modell konnte die CER noch einmal auf einen Wert von 2,09 % (1,95 % mit Language Model) senken.

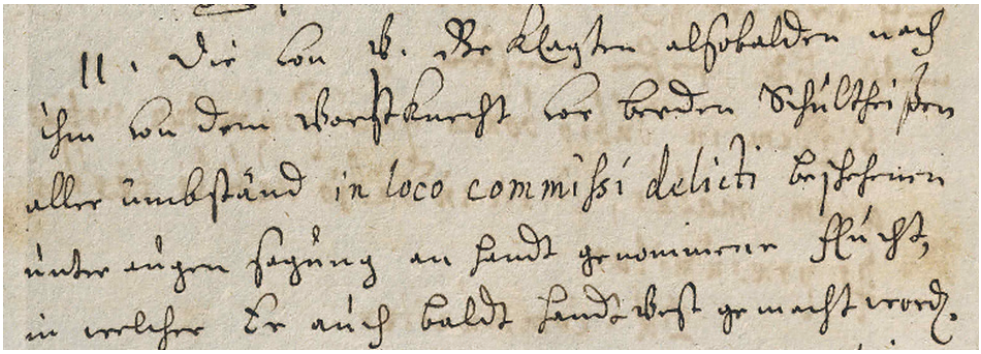

Abbildung 7: Schriftprobe aus den Juristischen Konsilien (Tübingen, Universitätsarchiv, UAT 84/13, Bl. 208v).

Zum Vergleich wurde unter anderem für die juristischen Konsilien auch ein in eScriptorium nutzbares Modell erzeugt. eScriptorium ist eine kostenlose Open-Source-Alternative zu Transkribus und bietet im Gegensatz zu Transkribus den Vorteil, dass Texterkennungsmodelle exportiert und importiert werden können. Im Rahmen des OCR-BW-Projekts hat die UB Mannheim eine Instanz aufgesetzt und seitdem auch zur Weiterentwicklung der Plattform beigetragen. Transkribus erlaubt keinen Export der trainierten Modelle, jedoch ist es möglich, die erstellten GT-Daten im PAGE-XML-Format aus Transkribus zu exportieren und nach eScriptorium zu importieren.[33] Diese exportierten Daten enthalten alle notwendigen Informationen – insbesondere die Grundlinien der Textzeilen und die Zeilentexte – für ein Training mit anderer OCR-Software. Konkret konnte Stefan Weil (UB Mannheim) damit ein Training mit der Texterkennungsengine Kraken, die auch in eScriptorium eingesetzt ist, durchführen und dabei die gleiche Aufteilung der Daten in Training und Validation Set wie in Transkribus verwenden, um eine Vergleichbarkeit der beiden Plattformen zu gewährleisten.[34] Ergebnis dieses Trainings war ein Modell, das auf dem Validation Set eine CER von 5,27 % erreicht und damit schlechter abschneidet als das mit Transkribus erzielte Ergebnis.[35] Die Unterschiede der Fehlerraten lassen sich nicht mit den unterschiedlichen plattformimmanenten Messsystemen erklären, wie ein Vergleich außerhalb beider Systeme mit dem externen Tool Dinglehopper gezeigt hat,[36] so dass momentan davon auszugehen ist, dass mit Transkribus bessere Ergebnisse erzielt werden können.[37]

33 Siehe für eine Anleitung zur Datenübertragung: <https://ub-mannheim.github.io/eScriptorium_Dokumentation/Modell%C3%BCbertragung_Transkribus_nach_eScriptorium>, Stand: 08.09.2022.

34 Auch über die Oberfläche von eScriptorium kann ein Modelltraining angestoßen werden, aber dort ist eine individuelle Zuweisung des Training und Validation Sets aktuell nicht möglich.

35 Siehe: <https://github.com/UB-Mannheim/kraken/wiki/Training#juristische-konsilien-t%C3%BCbingen>, Stand: 08.09.2022. Hierbei ist zu bedenken, dass keine Nachkorrektur der importierten Daten vorgenommen wurde und Übertragungsfehler das Ergebnis möglicherweise beeinflussen können.

36 Siehe für einen Vergleich mit Dinglehopper: <https://ub-backup.bib.uni-mannheim.de/~stweil/Juristische_Konsilien_Tuebingen/report-transkribus.html> und <https://ub-backup.bib.uni-mannheim.de/~stweil/Juristische_Konsilien_Tuebingen/report-kraken.html>, Stand: 08.09.2022.

37 Die hier vorgestellten Modelle wurden mit HTR+ trainiert, da die mit dieser Engine erzeugten Modelle bei der vorhandenen GT-Datenmenge stets besser abschnitten als Modelle, die mit der Engine PyLaia trainiert wurden. Dies

4. Unterstützung von Forschung und Lehre

Das Interesse am Thema Volltexterkennung für Handschriften und historische Drucke war im universitären Kontext seit Projektbeginn sehr groß.[38] Der Bedarf an automatischer Texterkennung besteht nicht nur in den klassisch textverarbeitenden Disziplinen, sondern, wie sich gezeigt hat, auch durchaus in den Naturwissenschaften.[39] So wurde die Bedienung entsprechender Bedarfe zu einem zunehmend größeren Anteil der Projektarbeit.[40] Das Kompetenzzentrum informiert grundsätzlich über die aktuellen technischen Möglichkeiten, bietet eine Einschätzung für die Bearbeitungsmöglichkeiten und nötigen Aufwände für verschiedene Materialgruppen und Verwertungszwecke, gibt Einführungen in die Techniken und unterstützt entsprechende Projektplanungen und -antragstellungen sowie betreut laufende Projekte. Während gerade im Bereich der Lehre ein räumlicher Schwerpunkt auf die Universität Tübingen gesetzt wird, wo Schulungsveranstaltungen für angemeldete Gruppen oder als offenes Angebot in den Räumlichkeiten der UB oder direkt im Rahmen von Lehrveranstaltungen abgehalten werden,[41] wenden sich mittlerweile Wissenschaftlerinnen und Wissenschaftler anderer Institutionen wie auch Mitarbeiterinnen und Mitarbeiter anderer Bibliotheken und Archive aus dem ganzen Bundesgebiet an das OCR-BW-Projekt.

Bei der Zusammenarbeit mit Forschungsprojekten ist zu beachten, dass sich die Tätigkeit der UB Tübingen aufgrund beschränkter Personalressourcen hauptsächlich auf Unterstützungsleistungen beschränken muss. Das bedeutet vor allem, dass der zeitaufwendige Teil der Erstellung von Ground-Truth-Daten nicht übernommen werden kann. Eine Ausnahme stellt im Einzelfall die Bearbeitung eigener UB-Bestände dar. Die Beständeauswahl für die zweite Projektphase von OCR-BW wurde so konzipiert, dass Materialien mit konkretem Forschungsinteresse ausgewählt worden sind. So wurden im Rahmen von OCR-BW für einen Briefnachlass des 19. Jahrhunderts, der sich zu seinem größeren Teil in der UB Tübingen befindet und der innerhalb eines wissenschaftlichen Projekts ediert werden soll, Digitalisate für einzelne Jahrgänge – aufgeteilt über den Gesamtüberlieferungszeitraum – erstellt, GT-Daten erzeugt und Modelle trainiert. Für das Editionsprojekt wird die Digitalisierung wie auch die Kostenübernahme der automatischen Texterkennung für die übrigen Jahrgänge beantragt.[42] Die

lässt sich zum Teil dadurch erklären, dass es bei der Entwicklung der hier vorgestellten Modelle technisch noch nicht möglich war, ein Base Model zu nutzen. Mittlerweile wurde dieses Problem gelöst; vorläufige Tests zeigen eine deutliche Verbesserung der PyLaia-Modelle, auch wenn die Fehlerrate immer noch leicht höher ist als bei HTR+-Modellen. Bei sehr großen Modellen sind die Fehlerraten auch ohne Base Model in der Regel ähnlich, so dass sich hier aufgrund der geringeren Kosten oft der Einsatz von PyLaia-Modellen lohnt. – Nachtrag: Seit November 2022 ist das Training und der Einsatz von HTR+-Modellen nicht mehr möglich.

38 Zum ersten Projektworkshop im Februar 2020, der über das Thema automatische Texterkennung von historischen Drucken und Handschriften informierte und erste Projektergebnisse sowie die eingesetzten Programme vorstellte, kamen über 50 Studierende und Wissenschaftlerinnen und Wissenschaftler aus verschiedenen Disziplinen: <https://ocr-bw.bib.uni-mannheim.de/2020/02/26/1-workshop-zu-ocr-und-handschriftenerkennung-an-der-ub-tuebingen/>, Stand 08.09.2022.

39 Im Bereich der Biodiversitätsforschung besteht zum Beispiel aktuell großes Interesse an den Aufzeichnungen früherer Fachvertreter, die handschriftlich unter anderem in Form von Briefen und Tagebüchern vorliegen. Das Projekt OCR-BW plant diesbezüglich eine Zusammenarbeit mit dem Fachinformationsdienst Biodiversitätsforschung (BIOfid).

40 Mittlerweile wurde über die Projektlaufzeit hinaus ein Service für OCR-Dienste eingerichtet. Siehe <https://uni-tuebingen.de/einrichtungen/universitaetsbibliothek/publizieren-forschen/ocr/>, Stand 08.09.2022.

41 Für Lehrveranstaltungen bietet sich die Verwendung von Transkribus Lite an. Transkribus Lite kann im Browser genutzt werden; die Anwendung erfordert zudem einen geringeren Einarbeitungsaufwand als der Expert Client.

42 Diese Kosten sind oftmals im Vergleich zum Gesamtvolumen eines Projekts sehr gering, aber es ist wichtig, sie schon bei der Antragstellung einzuberechnen.

in diesem Editionsprojekt korrigierten Transkriptionen sollen wiederum an die UB Tübingen zurückgespielt und gemeinsam mit den Digitalisaten in deren Präsentationsplattform DigiTue eingefügt werden, wo sie somit allen Nutzerinnen und Nutzern als Volltexte zur Verfügung stehen. Durch eine derartige Kooperation wird gleichzeitig für die Wissenschaft der Arbeitsaufwand verringert[43] und für die bestandshaltende Bibliothek ein geprüfter Volltext erzeugt.

Ein solcher zeitlicher Aufwand im Rahmen einer Projektvorbereitung oder -betreuung kann jedoch nicht immer betrieben werden. In den meisten Fällen werden wissenschaftliche Projekte bei der eigenständigen Anwendung von automatischer Texterkennung unterstützt. Das Kompetenzzentrum fungiert dabei als Anlaufstelle und Multiplikator der notwendigen Kenntnisse. Anhand von Testseiten wird zunächst ausgelotet, welche Ergebnisse mit den zur Verfügung stehenden Techniken mit welchem Aufwand erzielt werden können, um das gewünschte Ergebnis zu erreichen. Mitunter bedeutet dies nur die Empfehlung eines geeigneten Modells, da die automatische Transkription bereits als Lesehilfe ausreicht und die automatische Layouterkennung kaum einer Nachkorrektur bedarf.[44] Gerade für Kurrentschriften stehen in Transkribus generische Modelle zur Verfügung, die ohne Eigenaufwand gute automatische Ergebnisse erzielen. Darüber hinaus werden die Wissenschaftlerinnen und Wissenschaftler bei der Nutzung von Transkribus angeleitet und wenn nötig z.B. bei der GT-Erstellung und dem Modelltraining betreut.

Vielfach sind in der Wissenschaft bereits Transkriptionen vorhanden, die zu GT-Daten aufbereitet und für Modelltrainings nachgenutzt werden können. Ein Beispiel hierfür ist die Kooperation mit dem DFG-geförderten Projekt „Narrative Vermittlung religiösen Wissens: Edition und Kommentierung geistlicher Vers- und Prosatexte des 13. bis 16. Jahrhunderts“[45] der Universitäten Tübingen und Köln, das mit Textzeugen aus ca. 600 Handschriften aus verschiedenen Institutionen arbeitet. Von den Projektmitarbeiterinnen wurden Digitalisate und Transkriptionen für bislang 22 volkssprachliche Handschriften des 15. und 16. Jahrhunderts übermittelt, die im Rahmen von OCR-BW in Transkribus eingespeist wurden. Auf dieser Grundlage wurde ein generisches Modell für kursive Handschriften und Bastarden entwickelt, das entsprechend der zugrunde liegenden Transkriptionen auf die Projektbedürfnisse zugeschnitten ist. Das Modell ist somit an die projekteigenen Transkriptionsrichtlinien angepasst und hat etwa gelernt, Abkürzungen aufzulösen und die ergänzten Buchstaben in Klammern zu setzen, wodurch dieser Schritt in der weiteren Bearbeitung erspart bleibt.[46] Für eine Nutzung in anderen Kontexten können diese Zusatzinformationen durch die Search/Replace-Funktion herausgelöscht werden, so dass das Modell flexibel bleibt.[47] Mithilfe des Modells werden im Projekt automatische

43 Im Rahmen eines geplanten Forschungsprojektes wurde von den Mitarbeitern der Aufwand bei der Korrektur von automatischen Transkripten, die mit dem Modell für die lateinischen Tagebücher von Martin Crusius erzeugt worden sind, im Vergleich zu einer rein manuellen Transkription kalkuliert. Inklusive der Anpassung der Transkriptionsrichtlinien von einer zeichengetreuen zu einer normalisierten Transkription wurde eine Zeitersparnis von 50 % berechnet. Bei einem entsprechend angepassten Modell dürfte der Wert entsprechend höher ausfallen.

44 Im Rahmen von OCR-BW wurde die automatische Texterkennung für einzelne Werke übernommen, wenn dies ohne großen Aufwand bei der Layoutkorrektur möglich war und ein passendes Modell zur Verfügung stand.

45 <https://religioese-kurzerzaehlungen.uni-koeln.de/>, Stand 08.09.2022.

46 Die Daten werden im TEI-Format exportiert und in Oxygen weiterverarbeitet.

47 Um die Anpassungsfähigkeit eines Modells zu erhöhen, ist es in der Regel sinnvoll, in den GT-Daten zunächst so viele Informationen wie möglich zu bewahren, die im weiteren Verlauf durch automatische Skripte an verschiedene Bedürfnisse angepasst werden können.

Transkripte für weitere Texte erzeugt und so die weitere Transkriptionsarbeit beschleunigt. Parallel wird das Ergebnis durch sukzessives Nachtraining mit neuen Transkripten erweitert und verbessert.[48] Die hier erzielten Erkenntnisse lassen sich wiederum auf die zukünftig geplante Bearbeitung des UB-eigenen mittelalterlichen Handschriftenbestands übertragen.

Ein weiteres Feld ist umgekehrt die Aufbereitung von Druckeditionen für z.B. die Erstellung einer Onlineausgabe oder für die Weiterverarbeitung in Datenbanken. Obwohl es für Drucke viele weitere Optionen zur Erstellung eines Volltextes gibt, hat der Einsatz von Transkribus in mehreren bisher betreuten Projekten einige Vorteile geboten. So offeriert Transkribus auch Modelle für verschiedene Druckschriften, die durch ein werkspezifisches Training auf Sonderzeichen für z.B. historische skandinavische und altorientalische Sprachen angepasst wurden. Aufgrund der zur Verfügung stehenden Nutzeroberfläche, die keine tiefgehenden IT-Kenntnisse verlangt, wurde die GT-Erstellung wie auch die manuelle Nachkorrektur nach einer Einweisung von den Projektmitarbeiterinnen und -mitarbeitern übernommen, während das Modelltraining durch das Tübinger Kompetenzzentrum erfolgte. Weiteres Ziel der Arbeit des Kompetenzzentrums ist es also, Studierende und Forschende in die Lage zu versetzen, Texterkennungssoftware eigenständig anzuwenden, während das Kompetenzzentrum beratend und unterstützend zur Seite steht.

5. Fazit

Grundsätzlich lässt sich zusammenfassen, dass die Evaluation von Transkribus erfolgreich verlaufen ist und es sich als Werkzeug zur automatischen Texterkennung von Handschriften und Drucken sowohl für institutionelle als auch für wissenschaftliche Zwecke gut eignet. Dies hat sich sowohl bei der Bearbeitung eigener Quellenkorpora als auch bei der Unterstützung von Institutionen, Studierenden sowie von Wissenschaftlerinnen und Wissenschaftlern gezeigt. Eine automatische Handschriftenerkennung mit einer Zeichenfehlerrate von unter 5 % ist unabhängig von Sprache, Schriftart und Schreibzeitraum möglich und erwartbar. Für die automatische Texterkennung von großen homogenen Beständen, wie sie in der ersten Projektphase als Evaluationsgrundlage dienten, ist Transkribus besonders gut geeignet. Je formalisierter eine Schrift ist, desto besser sind die Ergebnisse bereits mit generischen Modellen ohne eigenes Spezialtraining. Gerade Großbestände wie die juristischen Konsilien, die oftmals nur eingeschränkt erschlossen sind, jedoch über weite Strecken ein einheitliches Layout aufweisen und von professionellen Schreibern verfasst worden sind, können also durch automatische Volltexterkennung mit gutem Ergebnis zugänglich gemacht werden. Aber auch für heterogene Bestände haben die bisherigen Versuche gezeigt, dass hier großes Potential liegt und Aufwände durch werkspezifisches Training reduziert werden können. Der Einsatz unterschiedlicher Parameter im Trainingsprozess und die Verwendung des im Trainingsprozess erzeugten Language Models verbessern die Ergebnisse oftmals signifikant. Für den Bereich der Unterstützung

48 Für besonders lange Texte wird mit einem werkspezifischen Training auf Grundlage weniger Seiten das Modell so weit angepasst, dass sich für den übrigen Text oftmals noch eine deutliche Verbesserung ergibt. Für kürzere Texte haben Stichproben ergeben, dass das Modell ihm unbekannte Schreiberhände aus dem Projektkorpus mit einer CER von ca. 8–10 % liest, während sich durch werkspezifisches Training eine Verbesserung auf ca. 3–5 % ergibt. Die Varietät von Digitalisaten, die in unterschiedlichen Qualitäten vorliegen, scheint nur einen geringen Einfluss auf das Ergebnis zu haben. Auch schwarz-weiße Mikrofilmaufnahmen werden gut erkannt, insofern das Hintergrundrauschen nicht zu stark ist.

von Wissenschaft und Forschung, der mit Anfragen zu unterschiedlichen Materialgruppen einhergeht, hat sich die Kooperation mit der UB Mannheim bei der Etablierung des gemeinsamen Kompetenzzentrums für Volltexterkennung von handschriftlichen und gedruckten Werken bewährt. Die an beiden Standorten einlaufenden Anfragen werden gemäß der jeweiligen Expertise verteilt, so dass ein breites Wissens- und Beratungsspektrum mit verteiltem Personalaufwand bedient werden kann.

Trotz der guten Ergebnisse bei der Modellentwicklung wurden bisher nur die Volltexte für einzelne Dokumente aus den bearbeiteten Quellenkorpora auf der Präsentationsplattform DigiTue der UB Tübingen eingebunden, da eine umfassendere Layoutbearbeitung weiterer Bände im Projektkontext nicht leistbar ist. Zusätzlich zu den Metadaten zum Dokument sowie einer Inhaltsübersicht findet sich bei den betreffenden Digitalisaten nun auch ein Reiter mit dem OCR-Volltext. Die Daten werden von Transkribus im TEI-Format mit den Koordinaten für die Bounding-Boxen exportiert. Die Bounding-Boxen ermöglichen den Nutzerinnen und Nutzern ein komfortables Navigieren im Text, da beim Scrollen durch den Volltext die entsprechenden Zeilen im Digitalisat hervorgehoben werden.[49]

Zu problematisieren ist bei der Zurverfügungstellung von automatischen Transkriptionen, dass diese mit den aktuellen technischen Möglichkeiten für Handschriften nur in Einzelfällen ein hundertprozentig korrektes Ergebnis bieten. Auch die im OCR-BW-Projekt erzielten guten Ergebnisse im Rahmen von 2–5 % CER zeitigen Fehler, die die Durchsuchbarkeit der Texte einschränken.[50] Bei Handschriften sind zudem nicht nur die Prozentzahlen der richtig erkannten Zeichen bzw. Wörter ein Kriterium für die vollständige Auffindbarkeit aller gewünschten Suchbegriffe, sondern auch die Transkriptionsrichtlinien. Hier bieten sich zwei grundsätzliche Möglichkeiten: Zum einen können die Zeichen entsprechend der Vorlage wiedergegeben werden, zum anderen kann die Transkription entsprechend den heutigen Konventionen normalisiert werden. Aktuell stellt die UB Tübingen die Volltexte direkt in der mit Transkribus erkannten und ausgegebenen Version, also ohne Ersetzung der Sonderzeichen, zur Verfügung. Beide Optionen bieten Vor- und Nachteile und zeigen, dass die sogenannten Ground-Truth-Daten nicht unbedingt die absolute Wahrheit abbilden, sondern vielmehr eine Interpretation darstellen, die im automatischen Output fortgeführt wird. Es muss also gefragt werden, was eine Bibliothek, die OCR-Erkennung als Dienst anbietet, hinsichtlich des Ergebnisses leisten kann und in welcher Form die Nutzbarkeit am besten gewährleistet ist. Es lässt sich sowohl für einen normalisierten wie für einen diplomatischen Output argumentieren. Sollte eine Bibliothek ihren Nutzerinnen und Nutzern das Werk im Volltext so originalgetreu wie möglich zur Verfügung stellen und dessen Informationen auch in dieser Darstellung bewahren oder sollte im Prozess der Umwandlung größtmöglicher Nutzungskomfort durch eine Wiedergabe nach heutigen Konventionen geschaffen werden? Diese Frage konnte im OCR-BW-Projekt bisher nicht zufriedenstellend gelöst werden. Sie ist für historische Dokumente jedoch auch abgesehen von den Problematiken einer automatischen Texterkennung werkimmanent, da die Nutzerinnen und Nutzer abgesehen von Transkriptionsfehlern mit uneinheitlicher Orthografie rechnen und diese bei ihren Suchanfragen

49 Hierfür sollte darauf geachtet werden, dass in Transkribus dieselbe Bildgröße eingespeist wird wie im Präsentationssystem, da sonst die Koordinaten umgerechnet werden müssen.

50 Innerhalb von Transkribus besteht die Möglichkeit, Dokumente mit dem sogenannten Keyword Spotting zu durchsuchen. Einige Institutionen wie z.B. das finnische Nationalarchiv haben bereits mit einer Einbindung dieser Suchfunktion experimentiert: <https://tuomiokirjat.narc.fi/en/info>, Stand: 08.09.2022.

berücksichtigen müssen. Sinnvoll erscheint von daher vor allem Transparenz hinsichtlich der angewandten Technik und Transkriptionsrichtlinien, auf die die angezeigten Volltexte zurückzuführen sind. Auch wenn sich noch Schwächen zeigen, so ist die Technik im Bereich der automatischen Texterkennung für Handschriften mittlerweile schon so weit fortgeschritten, dass auf diesem Wege der Zugang zu historischen Dokumenten vereinfacht wird. Während das Angebot eines Volltextes für Bibliotheken der nächste konsequente Schritt in der Bestandserschließung nach der Digitalisierung sein muss, wird damit zugleich für Forschung und Lehre eine Grundlage für weitere Auswertungen geschaffen und ein Bereich eröffnet, in welchem sich Bibliotheken als Partnerinnen der Wissenschaft etablieren können.

Literatur

- Boenig, Matthias; Federbusch, Maria; Herrmann, Elisa; Neudecker, Clemens; Würzner, Kay-Michael: Ground Truth. Grundwahrheit oder Ad-Hoc-Lösung? Wo stehen die Digital Humanities?, in: DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts. Universität zu Köln, 26. Februar bis 2. März 2018, 2018, S. 219–223. Online: <https://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf>, Stand: 08.09.2022.
- Crusius, Martin: Diarium Martini Crusii, hrsg. von Wilhelm Goez, Ernst Conrad, Reinhold Stahlecker, Eugen Staiger unter Mitw. von Reinhold Rau und Hans Widmann, 4 Bde., Tübingen 1927–1961.
- Deutsche Forschungsgemeinschaft: DFG-Praxisregeln „Digitalisierung“. DFG-Vordruck 12.151-12/16, 2016. Online: <https://www.dfg.de/formulare/12_151/>, Stand: 08.09.2022.
- Heumann, Ina; Stoecker, Holger; Tamborini, Marco; Vennen, Mareike: Dinosaurierfragmente. Zur Geschichte der Tendaguru-Expedition und ihrer Objekte, 1906–2018, Göttingen 2018.
- Hodel, Tobias; Schoch, David; Schneider, Christa; Purcell, Jake: 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, S. 1–10. Online: <https://doi.org/10.5334/johd.46>, Stand: 08.09.2022.
- Kiessling, Benjamin; Tissot, Robin; Stökl Ben Ezra, Daniel; Stokes, Peter: eScriptorium. An Open Source Platform for Historical Document Analysis, in: 2019 International Conference on Document Analysis and Recognition Workshops *(ICDARW)*, 2019, S. 19–24. Online: <https://doi.org/10.1109/ICDARW.2019.10032>, Stand: 08.09.2022.
- Maier, Gerhard: African dinosaurs unearthed. The Tendaguru expeditions. Bloomington, Ind. 2003 (Life of the Past).
- Michael, Johannes; Weidemann, Max; Labahn, Roger: HTR Engine Based on NNs P3. Optimizing speed and performance - HTR+, READ-H2020 Project 674943, Deliverable D7.9, 2018. Online: <https://readcoop.eu/wp-content/uploads/2018/12/Del_D7_9.pdf>, Stand: 08.09.2022.
- Muehlberger, Guenter; Seaward, Louise; Terras, Melissa; Ares Oliveira, Sofia; Bosch, Vicente; Bryan, Maximilian; Colutto, Sebastian; Déjean, Hervé; Diem, Markus; Fiel, Stefan; Gatos, Basilis; Greinoecker, Albert; Grüning, Tobias; Hackl, Guenter; Haukkovaara, Vili; Heyer, Gerhard; Hirvonen, Lauri; Hodel, Tobias; Jokinen, Matti; Kahle, Philip; Kallio, Mario; Kaplan, Frederic; Kleber, Florian; Labahn, Roger; Lang, Eva Maria; Laube, Sören; Leifert, Gundram; Louloudis, Georgios; McNicholl, Rory; Meunier, Jean-Luc; Michael, Johannes; Mühlbauer,

Elena; Philipp, Nathanael; Pratikakis, Ioannis; Puigcerver Pérez, Joan; Putz, Hannelore; Retsinas, George; Romero, Verónica; Sablatnig, Robert; Sánchez, Joan Andreu; Schofield, Philip; Sfikas, Giorgos; Sieber, Christian; Stamatopoulos, Nikolaos; Strauß, Tobias; Terbul, Tamara; Toselli, Alejandro Héctor; Ulreich, Berthold; Villegas, Mauricio; Vidal, Enrique; Walcher, Johanna; Weidemann, Max; Wurster, Herbert; Zagoris, Konstantinos: Transforming scholarship in the archives through handwritten text recognition. Transkribus as a case study, in: Journal of Documentation, 75 (5), 2019, S. 954–976. Online: <https://doi.org/10.1108/JD-07-2018-0114>, Stand 29.07.2022.

- Strauß, Tobias; Weidemann, Max; Labahn, Roger: Language Models. Improving transcriptions by external language resources, READ-H2020 Project 674943, Deliverable D7.12, 2018. Online: <https://readcoop.eu/wp-content/uploads/2018/12/D7.12_LMs.pdf>, Stand 08.09.2022.
- Ströbel, Phillip; Clematide, Simon; Volk, Martin; Schwitter, Raphael; Hodel, Tobias; Schoch, David: Evaluation of HTR models without Ground Truth Material. Preprint 2022. Online: <https://www.researchgate.net/publication/357927928_Evaluation_of_HTR_models_without_Ground_Truth_Material>, Stand 08.09.2022.