Der Stand der Handschriftenerkennung im ABD-Kontext

MAS-Arbeit

für den

MAS ALIS

an der

Universität Bern

Eingereicht bei Herrn Prof. Dr. T. Hodel

Vorgelegt von Claudia Pfister

Matrikelnummer: 11-722-634
Alpenstrasse 19
8800 Thalwil
044 722 16 75
claudia.pfister@students.unibe.ch

Thalwil, 2022

„Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Art. 36 Abs. 1 Buchst. r des Gesetzes über die Universität Bern und Art. 69 des Universitätsstatuts zum Entzug des aufgrund dieser Arbeiten verliehenen Titels berechtigt ist."

Inhalt

1. Einführung

Die massiven Fortschritte der letzten paar Jahre in der Handschriftenerkennung (HTR) haben dazu geführt, dass sie in immer mehr Bereichen eine praktischer Anwendung erfahren hat und somit ihr grosses Potential aufgezeigt werden konnte. Besonders interessant ist dieses neue Werkzeug für Informationsinstitutionen, die viel handschriftliches Material aufbewahren, wie Archive und wissenschaftliche Projekte, die damit ihr Angebot bedeutend erweitern und verändern können. Die HTR ermöglicht idealerweise den Zugriff auf den Volltext der Archivalien, was den Zugang für den Benutzer ändert und für viele Fragestellungen signifikant verbessert, wodurch neue Möglichkeiten und Ansätze für die Forschung geschaffen werden. Sie stellt dadurch eine willkommene neue Dienstleistung der Anbieterinstitution der Archivalien dar. Eine qualitativ hochwertige maschinelle Tiefenerschliessung ist jedoch sehr zeit- und ressourcenaufwendig und besonders für kleinere Institutionen schwer zu bewältigen. Deshalb wird die zentrale Frage der MAS-Arbeit sein, mit wie geringem Aufwand es möglich ist, einen Volltext zu erstellen, der für Interessierte von genügend hoher Qualität ist, dass er durchsuchbar und maschinell weiterverarbeitbar ist. Als Trainings- und Testmaterial dienen dabei Briefe von Johannes Gessner an Albrecht von Haller, zur Verfügung gestellt von dem SNF-Projekt »Online-Edition der Rezensionen und Briefe Albrecht von Hallers«.

Der erste Abschnitt dieser Arbeit widmet sich den Grundlagen. Dazu gehört eine allgemeine historische und technische Einführung in die HTR und in die Software Transkribus. Das Augenmerk wird dabei darauf sein, wie Archive und verwandte Institutionen von der HTR profitieren können und wie die HTR bereits im ABD-Kontext angewendet wird. Das folgende Kapitel stellt das für das Projekt verwendete Material vor und beschreibt die Transkription eines Briefes. Anschliessend folgt die Evaluation der so erstellten Texte, zuerst im Vergleich zwischen verschiedenen ausgewählten öffentlichen Modellen, dann in Bezug darauf, inwieweit die besten Modelle in unkorrigiertem Zustand zu gebrauchen sind und wie hilfreich die maschinelle Vorarbeit bei der manuellen Transkription ist. Das letzte Kapitel wird sich einem eigenen an den Gessner-Briefen orientierten HTR-System widmen und wird sowohl den Trainingsprozess in Transkribus schildern wie auch die Qualität der so erreichten Transkriptionen analysieren. Zum Schluss sollen die Resultate und Erkenntnisse rekapituliert werden mit Blick auf ihre Bedeutung und Nutzen für ADB-Institutionen.

2. Handschriftenerkennung und Transkribus

2.1 Hintergründe

Die automatische Handschriftenerkennung, meist handwritten text recognition, kurz HTR, genannt, ist lange schnell an ihre Grenzen gestossen, sobald es sich nicht um eine stark standardisierte Schrift gehandelt hat. Anders als die Optical Character Recognition (OCR), bei der der Text in einzelne Buchstaben zerlegt wird, um diese einzeln zu erkennen, ist für die meisten Handschriften ein flexibleres Vorgehen erforderlich, bei dem mehr von der Umgebung der Buchstaben berücksichtigt wird. Dies hat dazu geführt, dass frühere Versuche mit Handschriftenerkennung in der Regel nicht die gewünschte Qualität erreichen konnten. Erst in den letzten Jahren ist dank der massiven Zunahme an digitalisiertem handschriftlichen Material und mit Hilfe immer mächtigerer Computer und neuronaler Netzwerke, die trainiert werden, über ein in Zeilen zerlegtes Bild hinweg Text zu strukturieren und zu annotieren, ein rasanter Fortschritt in der HTR möglich geworden.[1] Mittlerweile ist es realistisch, eine Character Error Rate (CER)[2] von unter 5% zu erreichen, wenn das zu transkribierende Material und das Transkriptionsmodell zueinanderpassen.[3]

Heute gibt es mehrere gut funktionierende HTR-Tools, teils von privaten Anbietern, teils für bestimmte Projekte entwickelt, teils aus der Forschungsgemeinschaft selbst hervorgegangen.[4] Ein Beispiel für ein Digital-Humanities-Projekt, das sein eigenes HTR-System entwickelt, ist *In Codice Ratio*, das mit den Fonds des Archivio Apostolico Vaticano arbeitet.[5] Dieses Projekt arbeitet hauptsächlich mit Convolutional Neural Networks und Crowd Sourcing, wobei die Transkription so konstruiert ist, dass spezifischen Symbole erkannt werden müssen, ähnlich dem Pattern Matching, so dass auch Paläographieunkundige sich beteiligen können.[6] Adam Matthew Digital ist ein Fall von einem wissenschaftlichen Verlag, der HTR in einem

[1] Muehlberger *et al*. (2019), 956. Für eine genauere Beschreibung, wie ein neuronales Netz trainiert wird, siehe Hodel b (2022), 6f.

[2] Die CER gibt an, wie viele aller Zeichen richtig transkribiert worden sind. Ein interessanter Faktor, insbesondere für die Volltextsuche, ist auch die Word Error Rate (WER), also wie viele aller Wörter korrekt wiedergegeben sind. Da innerhalb des Projekts READ aber kein Konsens erreicht werden konnte, wie ein Wort genau zu definieren ist, beziehen sich Transkribus-Modelle in der Regel auf die einzelnen Zeichen und nicht auf die WER, weshalb die Arbeit sich im Folgenden ebenfalls an der CER orientiert, Hodel *et al*. (2021), 4, Anm. 5.

[3] Muehlberger *et al*. (2019), 962.

[4] Terras (2022), 183f.

[5] https://www.inf.uniroma3.it/db/icr/index.html.

[6] Nieddu *et al*. (2021), 2.6.

kommerziellen Kontext benutzt.[7] Bei den HTR-Systeme, die aus der Forschung hervorgegangen sind, sind vor allem drei zu nennen. Die Open Source-Plattform eScriptorium[8] der Université PSL verwendet das OCR-System Kraken[9] und wird bis anhin vor allem in französischen Projekten wie LECTAUREP[10] verwendet. Ein weiteres relativ altes und bekanntes HTR-System ist auch das von der Universität Groningen und dem Nationaal Archief der Niederlanden entwickelte Monk.[11] Und dann gibt es noch Transkribus[12], zur Zeit das Standardprogramm für die HTR.

Transkribus ist ab 2015 im Rahmen des tranScriptorium-Projekts (2013-2015)[13] entstanden und im Anschluss als Herzstück des Recognition and Enrichment of Archival Documents (READ) European Union Horizon 2020 Projekts (2016-2019) weitergeführt worden.[14] Im Anschluss hat sich die Genossenschaft READ-COOP SCE gebildet und ab Juli 2019 die Verantwortung für die Plattform übernommen. Die READ-COOP verfügt über 113 Mitglieder, darunter Privatpersonen und Einrichtungen, in 27 Ländern, darunter Gedächtnisinstitutionen wie das finnische, das schwedische und das norwegische Nationalarchiv, die Universitätsbibliothek Freiburg, die Nationalbibliothek Schottland und die British Library. Als Mitglieder aus der Schweiz sind die Zentralbibliothek Zürich, die Universitätsbibliothek Basel, das Walter Benjamin Kolleg der Universität Bern, das Staatsarchiv des Kantons Zürichs und das Stadtarchiv Zug zu nennen.[15] Die READ-COOP veranstaltet regelmässig Konferenzen mit Präsentationen und Workshops, um ihren Benutzern die Möglichkeit zu geben, sich über die Entwicklungen und Zukunft der HTR auszutauschen.[16] Während die Software Transkribus und die Erstellung eines Nutzerkonto kostenlos angeboten werden, ist mit Abschluss des Projekts die HTR-Funktion seit Oktober 2020 ab mehr als 500 Seiten kostenpflichtig geworden.[17] Eben-

---

7 https://www.amdigital.co.uk/products/handwritten-text-recognition.

8 https://escriptorium.fr/. Für eine ausführliche Vorstellung von eScriptorium siehe Kiessling *et al*. (2019).

9 https://kraken.re/.

10 https://lectaurep.hypotheses.org/.

11 https://www.ai.rug.nl/~lambert/Monk-collections-english.html. Der dort angegebene Link zur Search Engine ist veraltet und entspricht dem aktiven Link https://monk.hpc.rug.nl/cgi-bin/monkweb.

12 https://readcoop.eu/transkribus/.

13 http://www.transkriptorium.com/.

14 Muehlberger *et al*. (2019), 957.

15 https://readcoop.eu/de/mitglieder/.

16 Das Programm für die nächste TUC vom 29.-30. September 2022 an der Universität Innsbruck findet sich auf https://readcoop.eu/de/tuc22/programme/.

17 Terras (2022), 185.

falls kostenpflichtig ist eine Mitgliedschaft, privat oder als Institution, und das Verwenden von »read&search«-Web-Interface, bei dem Transkribus die digitale Edition der Dokumente hostet.[18] Durch das Verwenden dieser Plattform kann ausserhalb des Transkribus GUI die Keyword Spotting-Funktion der allgemeinen Öffentlichkeit angeboten werden. Diese Technologie erlaubt, einen Suchbegriff nicht nur zu finden, wenn er in der ausgegebenen Transkription vorhanden ist, sondern sie bezieht auch andere Transkriptionsvarianten mit geringerer Wahrscheinlichkeit ein; so würde eine Suche nach dem Namen »Muralt« auch die falsche Transkription »Murult« finden, wenn der Algorithmus bei dem fraglichen Vokal beispielsweise mit 60% Wahrscheinlichkeit ein »u« gelesen hat und mit 40% Wahrscheinlichkeit ein »a«.[19]

Die Grundfunktionen von Transkribus erlauben das Hochladen, das automatische Segmentieren, die manuelle und maschinelle Transkription von Dokumenten und den anschliessenden Export der Transkriptionen in verschiedenen Formaten, unter anderem TEI. Das Augenmerk liegt im Folgenden aber auf der maschinellen Transkription. Für diese bietet Transkribus um die 100 frei zugängliche KI-Modelle an, von denen 22 eine CER von weniger als 1% erreichen.[20] Durch die Möglichkeit, eigene Modelle zu trainieren und anschliessend allen Transkribus-Nutzern zur Verfügung zu stellen, steigt die Zahl öffentlicher Modelle stetig und somit auch die die Vielfalt, welche Art Dokumente automatisch transkribiert werden können. Die Bandweite reicht dabei von einem Modell für Russisch-Kirchenslawisch im 11. und 16. Jahrhundert[21] zu Drucken in der Devanagari-Schrift, die um 1900 publiziert worden sind[22]. Der Schwerpunkt der angebotenen Modelle liegt aber auf dem 17. bis 19. Jahrhundert; für die drei Jahrhunderte stehen 61.9% der Modelle zur Verfügung, 21.5% allein für das 18. Jahrhundert. Bei den Sprachen sind für Deutsch, gefolgt von Niederländisch und Französisch, die meisten Modelle vorhanden; abgesehen von zwei Ausnahmen sind nur Modelle für europäische Sprachen momentan öffentlich verfügbar. Das ist aufgrund des europäischen Hintergrunds in der Transkribus-Entwicklung verständlich; so ist auch die Mehrzahl der Transkribus-Nutzer zur

[18] https://readcoop.eu/de/readsearch/.
[19] Muehlberger *et al*. (2019), 962f.
[20] Vgl. https://readcoop.eu/de/transkribus/oeffentliche-modelle/.
[21] https://readcoop.eu/de/modelle/russian-church-slavonic-1/.
[22] https://readcoop.eu/de/modelle/devanagari-mixed-19th-20th-century/.

Zeit aus Europa; 11% allein kommen dabei aus der Schweiz, was sie nach Deutschland zum Land mit den zweitmeisten Transkribus-Nutzern macht.[23]

Für das Erstellen eigener Modelle offeriert Transkribus zwei verschiedene Engines, HTR+ und PyLaia, die beide auf neuronalen Netzwerken basieren, und sich qualitätsmässig wenig unterscheiden.[24] PyLaia ist zudem bei der Anwendung innerhalb von Transkribus günstiger als HTR+ und Open Source, was bedeutet, dass die Netzstruktur der Engine über ihr Github-Repository verändert werden kann.[25] HTR+ und die Serverkomponente von Transkribus sind somit die einzigen Komponenten von Transkribus, deren Aufbau nicht frei einsehbar ist.[26]

Um mit den Engines Modelle zu trainieren, wird empfohlen, sie mit 5000 bis 15'000 Wörtern oder 25 bis 75 Seiten Ground Truth zu trainieren, je nach Schriftqualität; Ground Truth bezeichnet Dokumente, die in guter Qualität vorhanden sind in Bezug auf Digitalisat, Layout und Transkription.[27] Zur weiteren Verbesserung erlaubt Transkribus die Verwendung von Basis- und Sprachmodellen. Ersteres bedeutet die Verwendung eines bereits bestehenden Modells zusätzlich zur Ground Truth, was erlaubt mit einer geringeren Wörterzahl ein Modell zu trainieren. Sprachmodelle erfüllen einen ähnlichen Zweck wie Wörterbücher aber mit grösserer Flexibilität, da sie mit dem Modell und auf ihm aufgebaut erstellt werden;[28] das HTR-Training erstellt also eine Statistik über die Anordnung der Buchstaben im Trainingsset und entscheidet über die Wahrscheinlichkeit von Zeichenfolgen.[29] Sprachmodelle sind besonders wichtig für die Weiterverarbeitung des Texts wie etwa die Named-Entity-Recogniton (NER).[30] Dieser Bereich der HTR ist noch in einer frühen Phase der Entwicklung.[31]

Die automatische Handschriftenerkennung wird immer zuverlässiger und bequemer zu nutzen. Damit stellt sich dann auch die Frage, warum überhaupt diese Technologie von Gedächt-

---

23 Terras (2022), 188f. Anm. 51.

24 Hodel *et al.* (2021), 4. Vgl. auch die Beobachtungen von Alvermann b (2020), wo festgestellt wird, dass PyLaia bei viel Trainigsmaterial besser abschneidet als HTR+; HTR+ ist dagegen besser beim Lesen von »gebogenen« oder senkrechten Textzeilen. Alvermann (2021) belegt, dass PyLaia etwa 1% besser abschneidet und auch zu HTR+ aufholt, was die Leistung bei »gebogenen« Textzeilen betrifft.

25 https://readcoop.eu/de/transkribus/howto/how-to-train-pylaia-models-in-transkribus/.

26 Muehlberger *et al.* (2019), 958.

27 Ebd., 959.

28 https://readcoop.eu/de/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/#elementor-toc__heading-anchor-12.

29 Alvermann/Gut (2021), 147. Hodel b (2022), 8.

30 Vgl. Hodel c (2022).

31 Hodel a (2022), 161.

nisinstitutionen benutzt werden sollte und inwieweit sie einen Mehrwert bei der Bestandsvermittlung bildet. Mit dem Angebot von Volltexten online wendet sich der Blick des Kunden mehr dem Inhalt des Bestandes zu und weg von dem eigentlichen Dokument, was die Recherchemöglichkeiten verändert und sie sozusagen der heutigen Zeit anpasst. Vor allem jüngere Forscher haben sich daran gewöhnt, ihre Suchabfragen an Google auszurichten.[32] Eine Volltextsuche kommt dieser Arbeitsweise entgegen. Eine grosse Menge an Computer lesbaren Daten, wie es Volltexte sind, erlaubt auch, die Texte in einem grösseren Zusammenhang zu betrachten, etwa durch Text mining, um ein Corpus genauer zu analysieren oder durch Distant reading, um die Beziehungen innerhalb grosser Textmengen zu untersuchen.[33]

Eine Umfrage unter Transkribus-Nutzern aus dem Frühling 2019 zeigt weitere Vorteile und Probleme auf, die die HTR bringt. Die Umfrage besteht aus 50 Fragen, auf die 19% aller aktiven Nutzer geantwortet haben. Von den Umfrageteilnehmern benützen die meisten Transkribus für persönliche Interessen und 36% die Plattform (auch) für ihre Arbeit innerhalb einer Organisation. Insgesamt haben die Umfrageteilnehmer vor, mehr als 8 Millionen Seiten zu bearbeiten; 43% beziehen das Material, mit dem sie arbeiten, dabei von einer Gedächtniseinrichtung. Besonders interessant sind die Rückmeldungen von Mitarbeitern an grösseren Projekten, die Transkribus als Inspiration bezeichnen sowohl ihren Digitalisierungsprozess zu verbessern, etwa durch die Qualität der Digitalisate, als auch mehr Material zu transkribieren.[34] Dass Transkribus ein Motor in der Digitalisierung ist, zeigt auch, dass 33% zurückmelden, dass ohne die Plattform keine Möglichkeit zur Transkription gegeben wäre, und 40%, dass es ohne Transkribus um einiges zeit- und ressourcenaufwendiger wäre.[35] Die Mehrheit der Umfrageteilnehmer benützt Transkribus für Recherchezwecke, wozu unter anderem gehört: die Texte online zu veröffentlichen, so dass in ihnen gesucht werden kann; die Texte direkt als Primärquelle zu nutzen; die Texte in wissenschaftlichen Ausgaben weiterzuverwenden, oft durch einen TEI-XML-Export; die Texte mit Corpuslinguistik weiterzuverarbeiten.[36]

---

[32] Edmond/Lehmann (2021), ii101.

[33] Vgl. Hodel b (2022), 14-19.

[34] Terras (2022), 187-190. Vgl. besonders die Rückmeldung ebd. 190, »Transkribus is a driver; an impetus to digitizing more textual material [...] It has also highlighted the quality required for digitzation and transcription.«

[35] Ebd, 192f. Eine Rückmeldung kommt zum Schluss, dass durch die HTR 80% an Kosten gespart werden kann.

[36] Ebd., 191.

Als Vorteile von HTR nennen die Umfrageteilnehmer die Beschleunigung des Transkriptionsprozesses, die Zunahme an digitalisierten historischen Quellen, das vereinfachte Teilen von Dokumentinhalten, die Verbesserung von existierenden Transkriptionen oder die Vergrösserung des Umfangs an verfügbaren historischen Dokumenten. Genannt wir auch, dass HTR beweist, dass Technologie einen sozialen und nicht-kommerziellen Nutzen haben kann. Erwähnt wird auch, die Möglichkeit Text Mining und ähnliche Technologien, zu nutzen. Problem bleibt, zu wissen, welche Dokumente anzubieten von Interesse ist. Die nicht digital aufbereiteten oder erst gar nicht digitalisierten Bestände einer Institution werden weiter an den Rand gedrängt und drohen vergessen zu werden, obwohl sie auch wertvolle Informationen enthalten. Daher ist es vor allem wichtig, die Masse an digitalisiertem Material in hoher Bildqualität zu erhöhen, etwa durch die Digitalisierung des Gesamtbestandes eines Archivs und menschliche und digitale Ressourcen darauf zu konzentrieren.[37] Wie die Umfrage zeigt, sind die meisten Transkribus-Nutzer Privatpersonen; diesen Material anzubieten, kann als Citizen Science wertvolle Ergebnisse liefern. Die Nutzer bei der Aufbereitung der Daten einzubeziehen, wie es beispielsweise bei der Plattform e-manuscripta geschieht, auf der die Nutzer ihre Transkriptionen einbringen können[38], wird umso wichtiger, je mehr digitalisierte Texte online gestellt werden. Maschinelle Vortranskriptionen wären dabei eine Möglichkeit, den Einstieg für Anfänger in der Transkription zu vereinfachen und ermutigend zu wirken, da es weniger beängstigt, einen Text zu korrigieren als ihn ganz neu zu erstellen. Es ist in jedem Fall wichtig, im Auge zu behalten, wie das veränderte Angebot sich auf die historischen Bestände und ihre Nutzer auswirkt.[39]

2.2 Transkribus-Projekte in Gedächtniseinrichtungen

Im Folgenden soll auf eine kleine Auswahl von Projekten, die mit Transkribus arbeiten eingegangen werden.[40] Einen besonders wichtigen Anteil haben in diesem Zusammenhang die Projekte des Staatsarchivs Zürich geleistet. Das Staatsarchiv hat bereits 2009, lange vor Transkribus, die Abteilung Editionsprojekte gegründet mit dem Ziel, seine bedeutendsten Dokumente als Volltexte online verfügbar zu machen. Zwischen 2009 und 2017 hat so das Projekt

[37] Terras (2022), 198f.

[38] https://www.e-manuscripta.ch/transcript/home.

[39] Terras (2022), 200.

[40] Vgl. auch die Projekte, die auf der Transkribus-Webseite selbst vorgestellt werden, https://readcoop.eu/success-stories/.

TKR, »Transkription und Digitalisierung der Kantonsratsprotokolle und Regierungsratsbeschlüsse seit 1803« existiert. Die gedruckten Protokolle haben ohne Problem mit OCR bearbeitet werden können, aber für die handschriftlichen Texte, insgesamt rund 270 Millionen Zeichen, ist die manuelle Transkription durch studentische Mitarbeitende nötig gewesen. So ist innerhalb des Projekts langsam ein grosses Corpus von zeichengetreu transkribierten Dokumenten angewachsen, auf das 2014 das sich gerade aufgleisenden Projekt READ aufmerksam geworden ist. Das Staatsarchiv ist zu einem Projektpartner, genannt »Large Scale Demonstrator«, geworden[41] und hat in dieser Rolle seine Daten Transkribus als Trainingsmaterial, als Ground Truth, zur Verfügung gestellt. Durch diese Beteiligung sind sodann weitere grosse Schweizer Gedächtnisinstitutionen wie das Bundesarchiv, die ETH-Bibliothek Zürich und die Universitätsbibliothek Basel auf Transkribus aufmerksam geworden.[42]

Die Zusammenarbeit mit Transkribus hat in der Folge zur Digitalisierung von weiteren Beständen wie Urkundenregesten (1460-1798) und dem Register der Briefbestände des Archivs geführt. Seit 2019 läuft das Projekt »Pilot vormoderne Quellen«, für das die Protokolle des Alten Stadtstaats Zürich zwischen 1484 und 1798 für die Öffentlichkeit digital zugänglich gemacht werden mit Hilfe von Transkribus.[43] Das konkrete Vorgehen dabei besteht darin, zuerst Metadaten zu erheben zum Schriftbild und dazu, wie die Protokolle strukturiert sind. Die Digitalisate, die in Transkribus nach Ratsmanual-Band geordnet importiert werden, kommen von Mikrofilmen. Es folgt die Layouterkennung, die eigentliche Handschriftenerkennung mit aktiviertem Keyword Spotting und, ausserhalb von Transkribus, die Entitätenerkennung (Named Entity Recognition). Für jeden der drei Schritte müssen zuerst manuell Trainingsdaten als Ground Truth erhoben werden, in diesem Fall Textseiten mit bereits ausgezeichneten Textregionen, Transkriptionen und getaggten Elementen. Die so entstandenen Modelle können dann angewendet und deren Output korrigiert werden, was zu neuen Trainingsdaten führt. Dieser Prozess wiederholt sich so lange, bis die neuen Daten zu keiner nennenswerten Verbesserung mehr führen. Für den Fall der Transkription heisst dies konkret, dass zwanzig Textzeilen zufällig ausgewählt und manuell transkribiert werden; das Ziel dabei ist, dass alle verschiedenen im Corpus enthaltenen Hände Teil des Trainingsmaterials ist. Mit 87 Trainingssei-

[41] Für eine Übersicht über die Unterzeichner des Memorandum of Understanding im Rahmen des READ-Projekts und die Begünstigten des Projekts, siehe: https://readcoop.eu/de/netzwerk/.

[42] Plüss/Sieber (2020), 219f.

[43] Ebd., 221.

ten aus acht Bänden ist dann die CER unter 5% gesunken. Das Projekt hat die Erfahrung gemacht, dass Material aus anderen Projekten nicht hilft, ihre Ergebnisse zu verbessern, wohl wegen zu wenig vergleichbarer Texte oder variierender Transkriptionsrichtlinien. Ziel der digitalen Erschliessung ist, die Ratsmanuale in strukturierter Form online zur Verfügung zu stellen und einen hohen Recall zu haben bei der Volltextsuche. Es ist nicht das Ziel, eine wissenschaftliche Edition herauszubringen, nur die Grundlage zu liefern, um dies und ähnliche Projekte Forschenden zu ermöglichen. Für die Zukunft ist geplant, die Öffentlichkeit mehr einzubeziehen und Texterkennung auch im »Projekt zur Evaluierung neuer Erschließungspraktiken« anzuwenden.[44]

Auch universitäre Projekte in der Schweiz benutzen Transkribus. So verwendet »iurisprudentia« Transkribus für die inhaltliche Erschliessung seiner Dokumente, historische Rechtstexte im deutschsprachigen Raum. Das Projekt der Universität Zürich macht das Textcorpus »Recht«, dessen Dokumente über verschiedene Archive und Bibliotheken verteilt sind, digital zugänglich über eine eigene Projektseite, die auf der read&search-Plattform von Transkribus aufbaut.[45] Dieses Corpus ist vor allem für Juristen von grossem Interesse, also Leute, die nicht mit historischen Hilfswissenschaften wie Paläographie vertraut sind. Die HTR kann hier Texte einem Fachkreis zugänglich machen, der sie sonst nur mit grosser Mühe erschliessen könnte.

Ein weiteres wichtiges Projekt, von dem auch das Angebot von Transkribus profitiert, ist »Urkunden und Akten des Klosters und der Hofmeisterei Königsfelden«.[46] Dieses Editionsprojekt hat die Entitäten im Bestand »Urkunden: Kloster und Amt Königsfelden«, um die 1000 Einzelblattdokumente und drei Kopialbücher, des Staatsarchivs Aarau, digitalisiert, transkribiert und getaggt.[47] Die hieraus gewonnenen Transkription sind für die Erstellung drei verschiedener Transkribus-Modelle verwendet worden, eines für deutsche und lateinische gotische Schrift von dem 13. zum 15. Jahrhundert mit 4.92% CER,[48] eines für deutsche Kurrentschrift aus dem 16. bis 18. Jahrhundert mit 8.42%[49] und eines für Chartaschriften mit 6.32%.[50]

[44] Ebd. 225-228.
[45] https://rwi.app/iurisprudentia/de/iurisprudentia.
[46] https://www.koenigsfelden.uzh.ch.
[47] https://www.koenigsfelden.uzh.ch/exist/apps/ssrq/intro.html#philosophie.
[48] https://readcoop.eu/de/modelle/latin-and-german-gothic-book-scripts/.
[49] https://readcoop.eu/de/modelle/german-kurrent-16th-18th/.
[50] https://readcoop.eu/de/modelle/charter-scripts-german-latin-french/.

Ein Vorgängermodell für gotische Schrift von 2017 hat das Kopialbuch von 1336 verwendet, insgesamt etwa 260 Seiten; für das Training sind 26'000 Wörter verwendet worden, womit eine CER von 10% erreicht worden ist. Ein Wörterbuch zur Verbesserung des Outputs ist nicht verwendet worden wegen der vielen abweichenden Schreibarten in dieser Art Text und weil sowohl Latein als auch Mitteldeutsch als Sprachen im Corpus vertreten sind.[51]

Ausserhalb der Schweiz findet sich in der französisch sprachigen Welt etwa das Projekt »Nouvelle-France numérique«. Dieses bemüht sich, die Dokumente von Neufrankreich, Teile des heutigen Kanadas, die über öffentliche und private Archive, Bibliotheken und Museen im ganzen Land und ausserhalb verteilt sind, zusammenzubringen; involvierte Gedächtnisinstitutionen sind die Bibliothèque et Archives nationales du Québec, die Archives nationales d'Outre-Mer, die Library and Archives Canada, das Musée de la civilization de Québec, die Rare Books Library und die Documents and Archives Management Division der Universität of Montreal. Dieses Projekt ist seit seinem Anfang 2018 mit READ-COOP und Transkribus verbunden und nutzt die Plattform, um auf die weit verteilten Dokumente zuzugreifen, sie zu transkribieren und nach TEI-Standards auszuzeichnen. Für Transkribus resultiert dies wiederum in neuen HTR-Modellen, viele mit einer CER von weniger als 5%. Es ist auch geplant, Modelle zu entwickeln, die speziell für indigene Sprachen geeignet sind.[52] Öffentlich verfügbar ist das Modell »New France (17th-18th Century)« mit einer CER von 4.12%, das mit Korrespondenzen und Registern von Kolonialverwaltern in Neufrankreich trainiert ist und aus 296'403 Wörtern besteht; es eignet sich auch als Basismodell für juristische Dokumente aus dieser Epoche.[53]

In Frankreich ist das HIstorical MANuscript Indexing for user-controlled Search (Himanis), ein bekanntes Projekt, das von 2015-2017 gedauert hat. Es befasst sich mit dem Kanzlei-Corpus des Trésor des Chartes mit dem Ziel, es online zugänglich zu machen; im Prozess sollen Indexing- und Suchfunktionen für historische Texte erarbeitet, Volltextsuche als Teil der Studie des historischen Erbes etabliert und daraus heraus eine neue Sicht auf die Bildung von Nationalstaaten in Europa gefunden werden. Mit dem Projekt verbundene Gedächtnisinstitutio-

[51] https://readcoop.eu/medieval-handwriting-and-handwritten-text-recognition/.

[52] https://readcoop.eu/success-stories/nouvelle-france-numerique-collaboration-and-partnership-arising-from-ai/. Durch das Projekt ist die Université du Québec à Rimouski zum ersten nordamerikanischen Transkribusmitglied geworden, https://nouvellefrancenumerique.info/collaboration-et-haute-technologie/. Zur Homepage des Projekts: https://nouvellefrancenumerique.info/.

[53] https://readcoop.eu/de/modelle/new-france-17th-18th-centuries/.

nen sind die European Library, die Archives Nationales de France und die Bibliothèque Nationale de France.[54] Für Transkribus ist »HIMANIS Chancery M1+« mit einer CER von 5.33% und basierend auf 666'000 Wörtern aus dem Projekt hervorgegangen.[55] Hier lässt sich auch sehen, wie es Transkribus erlaubt, von einander zu profitieren, da auch das Modell »Charter Scripts XIII-XV_M1«, das bereits beim Königsfeldener Projekt erwähnt worden ist, Dokumente aus dem HIMANIS-Projekt enthält.[56]

Ebenfalls in Frankreich findet sich das Projekt »Foucault Fiches de lecture«, das Notizen von Michel Foucault, die in der BnF aufbewahrt sind, online zugänglich macht. Als Ground Truth sind 200 Seiten aus dem Manuskript *Théories et institutions pénales* zur Verfügung gestanden, für das schon eine Transkription vorgelegen ist, auch wenn es für eine möglichst getreue Abschrift nötig gewesen ist, nochmals korrigierend darüber zu gehen.[57] Probleme, die sich mit den Texten bei der HTR gestellt haben, sind die vielen Abkürzungen und Foucaults Schrift, bei der Buchstaben schwer zu unterscheiden sind, gewesen; dazu kommt, dass das Papier so dünn ist, dass teils die rückseitige Schrift ebenfalls erkannt wird. Um durch schlechte Lesarten die HTR-Modelle nicht zu verwirren, sind TEI-Tags für unleserliche und unsichere Buchstaben verwendet worden.[58] Mit der Segmentation hat die Aufbereitung der Ground Truth etwas über zwei Wochen gedauert; das daraus resultierende Modell weist eine CER von 15% auf. Das ist für genügend befunden worden, um mit den eigentlichen Dokumenten des Projekts weiterzufahren und dank des Modells schneller arbeiten zu können und bereits einen durchsuchbaren Rohtext zu haben. Transkribiert worden sind ungefähr 400 Seiten mit Notizen zu zwei Lesungen.[59] Insgesamt sind jetzt 600 Seiten zum Training zur Verfügung gestanden, was es erlaubt hat eine CER von 8% zu erzielen. Jede dank dieses Models neu transkribierte und korrigierte Seite kann zur Ground Truth hinzugefügt werden, womit wieder ein Modell trainiert werden kann. Ebenfalls werden mehr Möglichkeit zur Zusammenarbeit wie über die Plattform Omeka/eman ins Auge gefasst, damit auch Aussenstehende sich an der Korrektur

[54] https://himanis.hypotheses.org/about. Zur Homepage: http://himanis.huma-num.fr/app/.
[55] https://readcoop.eu/de/modelle/french-and-latin-chancery-documents/.
[56] https://readcoop.eu/de/modelle/charter-scripts-german-latin-french/.
[57] Massot *et al*. (2018), 2f.
[58] Ebd., 5.7.
[59] Ebd., 3f.

der Transkriptionen beteiligen können. Insgesamt wird die Arbeit mit Transkribus innerhalb des Projekts als Erfolg gesehen.[60]

Für Deutschland gibt es ebenfalls mehrere Projekte, die Transkribus nutzen für die Bearbeitung ihrer Corpora. So befasst sich das 2020 begonnene Projekt »Klosterregister und Klosterbuch für Pommern« mit Schriften aus Klöstern des ganzen historischen Pommern aus einem Zeitraum vom 12. bis zum 16. Jahrhundert; beteiligt an dem Projekt sind dem entsprechend alle heutigen Länder, die auf dem Gebiet liegen. Für Transkribus bedeutend ist ein Teilprojekt, das die Urkundenregister der pommerschen Kirchen und Klöster, die 1913-1923 hauptsächlich von Hermann Hoogeweg, damals Direktor des Staatsarchivs Stettin, erarbeitet worden sind, digitalisiert. Dieser Bestand, insgesamt 34 Bände mit Regesten zu 7346 Dokumenten, liegt im preussischen Staatsarchiv Stettin, heute in Polen, das innerhalb des Projekts die Digitalisierung der Bände übernommen hat. Das Universitätsarchiv Greifswald hat sodann die HTR der Regestenbände durchgeführt und die Universitätsbibliothek Greifswald die Bände mit Volltext online gestellt.[61] Das Universitätsarchiv hat also alle Transkribus direkt betreffenden Aufgaben übernommen. Dazu gehört, die Digitalisate, die eine gute Auflösung von 400 dpi aufweisen, nach Band geordnet bei Transkribus hochzuladen und zu segmentieren. Da die Regestenbücher als Tabellen aufgebaut sind, wird ihr Layout auch als solches gehandhabt, um die Daten besser ordnen zu können und die Nachnutzung zu erleichtern. Für die HTR ist ein eigenes Modell entwickelt worden, da nur drei Hände in den Dokumenten vertreten sind und daher ein konsistentes Schriftbild vorhanden ist. Zu beachten ist dabei, dass alle drei Handschriften im Trainigscorpus und auch alle drei Sprachen, Deutsch, Mittelniederdeutsch und Latein, vertreten sind. Die Erstellung einer ersten Ground Truth, etwa 30'000 Wörtern, ist im Universitätsarchiv Greifswald und an der Universität Kiel erfolgt. Mit dieser ist sodann mit der Engine HTR+ ein Modell trainiert worden; mit Hilfe dieses Modells ist in der Transkription fortgefahren worden bis etwa 100'000 Wörter Ground Truth erreicht worden sind und das

---

[60] Ebd., 7f.

[61] Alvermann/Gut (2021), 130-136.140. Alle 34 Regestenbände sind inzwischen einerseits konsultierbar über den Webkatalog der Universitätsbiblbiothek Greifswald: https://www.digitale-bibliothek-mv.de/viewer/toc/PPNAPSzczecinie_65_78_0_3_1/, andererseits liegt für eine bequemere Textübersicht und eine Suche mit Keyword Spotting ein read&search-Portal vor https://transkribus.eu/r/regestapomeraniae/#/. Allerdings sind dort die Namen der Regesten vertauscht (zum Beispiel sind die »Regesten zu den Urkunden des Cisterzienser-Nonnenklosters zu Bergen auf Rügen« als »Regesten zu den vom Marienstift zu Stettin deponierten Urkunden« abgelegt), weshalb bei der Verwendung auf die jeweiligen Titel der Bände aufgepasst werden muss.

mit diesen Daten trainiertes Modell eine CER von 1.75% aufgewiesen hat. Mit der Verwendung eines Sprachmodells ist es möglich das Resultat noch weiter zu verbessern.[62]

Erwähnenswert ist auch das Projekt »Rechtsprechung im Ostseeraum« von 2019-2021, an dem sich bei den Gedächtniseinrichtungen das Universitätsarchiv Greifswald, das Archiv der Hansestadt Wismar, das Landesarchiv Mecklenburg-Vorpommern und die Universitätsbibliothek Greifswald beteiligen. Das zu digitalisierende Corpus enthält die Spruchakten der Greifswalder Juristenfakultät zwischen 1580-1893, die Urteilsbegründungen der Assessoren am Wismarer Tribunal zwischen 1746-1845 und des Wismarer Ratsgerichts zwischen 1701-1879, was in insgesamt 257'000 Seiten resultiert. Der Corpus ist über Transkribus' read&search-Angebot durchsuchbar.[63] Eines der grössten Modelle in Transkribus, trainiert mit 1'840'000 Wörtern, basiert auf der Arbeit dieses Projekts, die German_Kurrent_17th-18th mit einer CER von 5.5%.[64] Zwei weitere Modelle sind die Acta_17 PyLaia mit über 594'000 Wörtern und einer CER 5,8 %[65] und die Acta (extended) mit mehr als 1'500'000 Wörtern und einer CER von 7.1%.[66] Die Modelle werden in Kapitel 3 für die eigene Transkriptionsarbeit getestet werden.

Ein ganz anderer aber nicht weniger interessanter Anwendungsbereich von Transkribus ist, die Software für pädagogische Zwecke zu verwenden, um etwa Studenten Paläographie näher zu bringen. Ein Vorteil dabei ist auch, dass Transkribus Transkriptionen in der Cloud und somit Zusammenarbeit erlaubt. In einem konkreten Fall haben Studenten der Universität Zürich mit 2200 Sendschreiben aus dem 15. Jahrhundert von Bern an die Landvögte von Thun gearbeitet; dieses Corpus besteht aus vielen verschiedenen Händen und eine über die Jahrzehnte sich wandelnde Schrift. Jeder und jede der 11 Studenten in dem dokumentierten Fall hat zwei Sendschreiben transkribiert. Die Zusammenarbeit hat dabei dank Transkribus besser funktioniert, auch wenn kein zeitlicher Vorteil zu erkennen gewesen ist. Die Wichtigkeit von Transkriptionsregeln und allgemeine Grundsätze bei wissenschaftlichen Editionen sind aber so hervorgehoben worden durch mehr Diskussionen wegen der Zusammenarbeit. Am Ende ist eine Ground Truth von 21'682 Wörter aus der Lehrveranstaltung herausgekommen, was eigentlich mehr als genug sein sollte für ein Modell. Dennoch hat es nur eine CER von 26% auf

---

62 Ebd., 143-148.

63 https://transkribus.eu/r/jurisdiction/info.

64 https://readcoop.eu/de/modelle/german-kurrent-17th-18th-century/.

65 https://readcoop.eu/de/modelle/german-low-german-and-latin/.

66 https://readcoop.eu/de/modelle/german-lower-german-latin-17th-century/.

dem Validierungsset erreicht, obwohl das Resultat mit einem Sprachmodell um 7-9% verbessert werden konnte. Allerdings hat es sich bei den Verbesserungen vor allem standardisierte Phrasen gehandelt, weniger um Entitäten, also die Begriffe wie Namen oder Orte, die hauptsächlich bei Suchanfragen an ein Corpus interessieren. Der Hauptgrund für das schlechte Abschneiden der HTR in diesem Fall liegt an den zu verschiedenen Schriften. Nichtsdestotrotz hat Transkribus die Studenten und ihre Arbeit, die zu dem Modell geführt hat, inspiriert und lässt sich daher als pädagogisches Hilfsmittel empfehlen.[67]

Die obig beschriebenen Projekte haben alle das grosse Potenzial von HTR aufgezeigt und auch wie unerlässlich Gedächtniseinrichtungen sind bei der Bereitstellung und Verarbeitung von Material bei universitären Projekte. Nicht alle Digitalisierungsprojekte in Gedächtnisinstitutionen können aber innerhalb grösserer wissenschaftlicher Projekte untergebracht werden und speziell für kleinere Institutionen bedeuten Transkriptionen einen hohen Arbeits- und Zeitaufwand, für die schnell die Mittel knapp werden. Transkribus verspricht hier Abhilfe. Wie gross diese sein kann, soll in den folgenden zwei Kapitel nachempfunden werden.[68]

3. Trainieren und Evaluation

3.1 Das Trainingsmaterial

Den Nutzen für kleinere Projekte zu testen, die etwa auch von einem Archiv ohne viele Ressourcen bewältigt werden können, soll also in diesem und im nächsten Kapitel simuliert werden. Viele Archive verfügen über Briefsammlungen, daher wird auch im folgenden mit einer solchen gearbeitet, indem sie automatisch transkribiert und die ausgegebenen Texte auf ihre Nützlichkeit geprüft wird. Als Trainingscorpus gewählt werden die Briefe von Johannes Gessner (1709-1790) aus Zürich an den Berner Albrecht von Haller (1708-1777), die in der Burgerbibliothek Bern liegen.[69] Der umfangreiche Briefwechsel der beiden Gelehrten ist auf Latein verfasst und enthält Briefe aus einem Zeitraum von beinahe 50 Jahren, von 1728 bis zu Hallers Tod; auf der Plattform *hallerNet* sind 660 Briefe verzeichnet, von denen 639 mit ei-

[67] Hodel (2017).

[68] Rückmeldungen bei Umfragen sind vielversprechend, vgl. etwa in der Umfrage bei Terras (2022), 192, »we are a small workforce so having HTR complete even a portion of the transcription process is helpful to us«.

[69] Die Briefe füllen drei Schachteln, die Bestände N Albrecht von Haller 105.20-22 Korrespondenz: Briefe an Haller: Gessner, Johannes, 1-3 (Konvolut/Codices/Bände): http://katalog.burgerbib.ch/detail.aspx?ID=54931, http://katalog.burgerbib.ch/detail.aspx?ID=54932, http://katalog.burgerbib.ch/detail.aspx?ID=54933.

nem Digitalisat verbunden sind.[70] Verwendet wird der Übersicht halber nur eine Auswahl der ältesten 50 Briefe, die zwischen 1728 und 1734 entstanden sind. Somit muss das System nur eine einzige Hand innerhalb eines stark beschränkten Zeitraums erkennen, was einerseits für das Erstellen eines eigenen Modells von Vorteil ist, andererseits ist fraglich, wie gut ein allgemeiner trainiertes Modell mit einer so individualisierten Schrift zurechtkommt, speziell da es sich um Alltagsbriefe handelt, in denen kein grosser Wert auf ein einheitliches Schriftbild oder eine saubere Darstellung gelegt worden ist. Die Briefe liegen bereits als TIF-Dateien mit einer Auflösung von 300 dpi auf dem Bildserver des Projekts »Online-Edition der Rezensionen und Briefe Albrecht von Hallers« vor, in dessen Rahmen sie digitalisiert worden sind; öffentlich einsehbar sind sie über das Projektportal *hallerNet*.[71] Nur für eine sehr kleine Auswahl liegen die Briefe in modernen Editionen vor; Haller selbst hat jedoch 1773-1775 eine grössere Zahl veröffentlicht als *Epistolarum ab eruditis viris*.[72] Zwar sind dies keine wissenschaftlichen Transkriptionen und teils werden Wörter oder Satzkonstruktionen leicht verändert. Haller lässt auch immer wieder ganze Abschnitte aus, wenn er sie als nicht interessant genug für eine weite Öffentlichkeit erachtet hat.[73] In vielen Fällen ermöglichen die Briefe aber eine schnellere Entzifferung, auch wenn ein kritischer Blick auf das Digitalisat nötig bleibt.

Jeder der zur Verfügung gestellten Briefe ist als ein Ordner vorhanden, in dem die Digitalisate der Seiten gespeichert sind. Das ist ideal für Transkribus, da für das Hochladen alle Bilder in einem Ordner sein müssen. Jeder Brief, egal aus wie vielen Seiten er besteht, wird also als ein Dokument Teil einer eigens dafür eingerichteten »Gessner-Collection«

Zuerst muss jetzt die Layouterkennung erfolgen und es zeigt sich, dass hierbei die Resultate nicht die saubersten sind. Die Korrektur des Layouts ist nicht nötig für die eigentliche HTR und eine etwaige Volltextsuche, aber da es ein Ziel ist, die Briefe möglichst sauber zu transkribieren, wird das Layout so gut wie möglich bereinigt. Ein Problem besteht darin, dass die meisten Briefe auf einem grossen Blatt Papier geschrieben sind, das in der Mitte gefaltet ist,

---

[70] Gründe für das Fehlen eines Digitalisats sind ein anderer Archivierungsort als die Burgerbibliothek oder, hauptsächlich, der Verlust des Briefes, sodass seine Existenz nur über das handschriftliche Briefverzeichnis von G.E. Haller belegt ist, das ebenfalls in der Burgerbibliothek liegt: http://katalog.burgerbib.ch/detail.aspx?ID=54856.

[71] https://www.hallernet.org/. Für die Arbeit stellte mir Christian Forney freundlicherweise Kopien zur Verfügung.

[72] Für ein Digitalisat des Werkes siehe: https://doi.org/10.3931/e-rara-24956.

[73] Vgl. bsp. den Originalbrief vom 30.6.1732, https://hallernet.org/data/letter/02424/facs, mit Hallers Version, https://www.e-rara.ch/zuz/content/zoom/7850882.

so dass mit Vorder- und Rückseite vier Seiten entstehen, die alle beschrieben werden können. Diese vier Seiten sind einzeln bei aufgefaltetem Papier gescannt, wobei bei Seite 1 ein Teil von Seite 4 zu sehen ist und bei Seite 2 ein Teil von Seite 3 und umgekehrt. Die Layouterkennung erkennt nun auch Blöcke für einen Text, der auf einer anderen Seite vollständig vorhanden ist. Diese Textregionen müssen gelöscht werden. Dazu kommen noch Kleinigkeiten, wie dass die Signatur als Textregion oder einfach nur als Zeile erkannt werden; dies alles muss ebenfalls gelöscht werden. Jeder Brief beginnt mit einer deutlich abgesetzten Begrüssung, di als separate Textregion betrachtet werden könnte, der Einfachheit halber wird aber nur ein Textblock pro Seite definiert. Es wird kein Tagging der Briefstruktur innerhalb von Transkribus vorgenommen.[74] Was überhaupt nicht als Text erkannt wird, sind die Adressen, die meistens auf der letzten Seite eines Briefes stehen und die vertikal geschrieben sind, dazu noch in Französisch. Hier kann eine Textregion selber eingefügt und die Zeilen gezeichnet werden; da aber für die Versuche im Lauf der Arbeit die Adressen in der Transkription nicht berücksichtigt werden, ist das nicht nötig. Auf der Ebene der Zeilen, Baselines genannt, wird es noch umständlicher. Gessner korrigiert sich oft, streicht etwa Wörter durch oder fügt eines oberhalb der Zeile ein. Das bringt auch die Nummerierung der Baselines durcheinander; da es für die Korrektur der Transkriptionen einfacher ist, einen zusammenhängenden Text zu haben, wird auch das korrigiert. Das heisst auch, wenn in der Mitte einer Zeile ein Wort eingefügt ist, das nicht oder als einzelne Zeile erkannt worden ist, muss die eigentliche Zeile aufgebrochen und neu nummeriert werden mit Berücksichtigung des eingefügten Worts als Baseline. Da die Briefe aber relativ unübersichtlich sind, werden viele Kleinigkeiten nicht vor der HTR bemerkt, so dass im Nachhinein das Layout nochmals korrigiert werden muss; das ist insofern ein Problem, als dass Transkribus die CER nicht mehr berechnen kann, wenn die Baselines umgestellt werden. Für ein wirklich sauberes Layout, in dem alle Baselines in der richtigen Reihenfolge wiedergegeben werden, sind pro Seite an die drei Minuten einzuberechnen. Weniger Aufwand ist nur möglich, wenn allein ein Volltext zur Durchsuchung der Briefe generiert werden soll und die Navigation des Lesers im Digitalisat stattfindet, was nicht der ideale Kundenservice ist. Es ist möglich ein Baseline-Modell in Transkribus zu trainieren,[75] aber da

[74] Für das Taggen der Textstruktur und das Trainieren eines entsprechenden Struktur-Modells mit Transkribus siehe den entsprechenden How-To-Guide: https://readcoop.eu/de/transkribus/howto/how-to-use-the-structural-tagging-feature-and-how-to-train-it/.

[75] Vgl. den entsprechenden How-To-Guide: https://readcoop.eu/de/transkribus/howto/how-to-train-baseline-models-in-transkribus/.

Gessners Briefe keine konsistente Struktur aufweisen, ist darauf verzichtet und die Standard-Layoutanalyse verwendet worden.

Was die Handhabung von Transkribus als Software angeht, ist sie zum grossen Teil recht intuitiv; wenn aber nicht aufgepasst wird, kann es aus Versehen passieren, dass eine Textregion oder eine Baseline verschoben wird. Nach den ersten ein, zwei Seiten sind die meisten Werkzeuge aber leicht einsetzbar. Ein weiterer Nachteil ist, dass nur eine Seite auf ein Mal offen sein kann.[76]

Bevor nun transkribiert werden kann, muss bestimmt werden, welche Regeln dafür gelten sollen. Dafür dient wegen der Natur der Briefe das Editionsmodell der hallerNet-Plattform als Orientierung, das heisst, es wird sich nahe an die Originaltexte gehalten. Dazu gehört beispielsweise Zirkumflexe zu berücksichtigen, die Gessner manchmal braucht für Ablativendungen im Singular der a- und o-Deklination; gleiches gilt für den Gravis bei Adverbien. Vokalische Ligaturen wie »æ« und »œ« sind ebenfalls beizubehalten sowie die Wiedergabe von »s« und »ſ« «, so gut Gessners Buchstaben zu unterscheiden sind. Ebenfalls befolgt werden soll die Gross- und Kleinschreibung, die allerdings oft schwer zu erkennen ist, speziell bei »s« und »a«. Es geht letztendlich mehr darum, konsistent zu sein; so scheint Gessner vor allem Eigennamen und Pronomina der zweiten Person grosszuschreiben. Satzanfänge sind manchmal deutlich gross geschrieben, daher wird dies in all diesen Fällen so gehandhabt. Was Abkürzungen betrifft, werden Endsilbenverschleifungen wie etwa beim Genitiv Plural, der 3. Singular Passiv oder beim enklitischen -que aufgelöst ebenso wie die Abkürzungen für »non« und »enim«; »i« und »j« werden so wiedergegeben wie im Schullatein, also der Genitiv Singular »-ij« wird »-ii«; das sollte für Transkribus kein Problem sein, da es sich immer um die gleiche Buchstabenreihenfolge am Satzende handelt; ein »j« in einem anderen Kontext kann immer noch als solches erkannt werden. Die Verdoppelung von Konsonanten durch Überstreichung (m̄) wird aufgelöst. Der Umgang mit durchgestrichenen Passagen, von denen es zahlreiche gibt und von verschiedener Qualität, teils ein sauberer Strick, teils durchgekringelt, ist schwierig. Wenn der Text absolut unleserlich ist, werden sie weggelassen und die Baselines gelöscht, um die HTR nicht zu verwirren; allerdings kann Transkribus bei durchgestriche-

[76] Vgl. die Kritiken bei Massot *et al*. (2018), 6.

nen Passagen oft mehr entziffern als der Mensch,[77] daher sollte, wann immer möglich, versucht werden, den Text richtig wiederzugeben inklusive der Streichungen – wenn eine korrekte Version direkt über eine falsche geschrieben wird, beispielsweise ein »o« wird ausgefüllt und ein Punkt oberhalb ergänzt, um ein »i« zu schreiben, dann wird nur die korrekte wiedergegeben. Unterstreichungen, die Gessner etwa für Pflanzennamen verwendet, werden berücksichtigt. Gessner zeichnet oftmals weitere Linien in den Text hinein, um Bezüge herzustellen oder Einfügungen genau zu platzieren, so dass eine Menge Striche vorkommen, die keine Bedeutung für die Transkription haben. Diese bilden möglicherweise ein Problem bei der HTR.

Nun gilt es nur noch zu sichten, welche Briefe sich für einen ersten Versuch der HTR mit öffentlichen Modellen eignen und dann kann transkribiert werden.

3.2 Transkribieren mit Transkribus: Vorauswahl der Modelle

Als Nächstes müssen jetzt die Briefe ausgewählt werden, welche sich am besten eignen, um an ihnen das Transkriptionskönnen der öffentlichen Modelle zu testen. Feststeht, dass sie mit einer Transkription versehen sein müssen, um sicher zu gehen, dass die Lesart jeweils korrekt ist. Weiter sollte es ein Brief sein, der sich in gutem Zustand befindet, um nicht schon anfangs mit Problemen wie durchscheinender Tinte, verblasster Schrift oder Siegelflecken zu kämpfen zu haben.

Das weitere Vorgehen besteht nun darin, die öffentlichen KI-Modelle von Transkribus auszuwählen, die zeitlich, räumlich und sprachlich zum Testmaterial passen. Die Wahl fällt dabei auf die Modelle (mit der verwendeten Engine in Klammern) Pylaia_NeoLatin_Ravenstein (PyLaia)[78], NeoLatin_Ravenstein_1643-1772 (HTR+)[79], Acta 17 (extended) (HTR+)[80], Acta_17 PyLaia (PyLaia)[81], German_Kurrent_XVI-XVIII_M1 (HTR+)[82] und German_Kurrent_17th-18th (PyLaia)[83]. Die ersten beiden sind gewählt worden, weil die Dokumente, mit denen sie trainiert wurden, den Zeitraum 1643-1772 abdecken und sie in Latein verfasst sind. Allerdings sind sie aus den Niederlanden und ihr Kontext ist völlig verschieden. Die nächsten

[77] Vgl. Alvermann a (2020), »[HTR-Modelle] lesen auch da noch sinnvolle Inhalte heraus, wo ein Transcriber längst aufgegeben hätte.«

[78] https://readcoop.eu/model/neo-latin-17th-18th-century/.

[79] https://readcoop.eu/model/neo-latin/.

[80] https://readcoop.eu/model/german-lower-german-latin-17th-century/.

[81] https://readcoop.eu/model/german-low-german-and-latin/.

[82] https://readcoop.eu/model/german-kurrent-16th-18th/.

[83] https://readcoop.eu/model/german-kurrent-17th-18th-century/.

beiden sind diejenigen aus dem Projekt »Rechtsprechung im Ostseeraum«, deren Trainingsmaterial einen längeren Zeitraum abbilden, 1580-1705 beziehungsweise 1750, und lateinische Texte unter anderen enthält. Ihre geographische Herkunft und ihr Kontext sind wiederum sehr verschieden von Gessners informellen Briefen. Das zweitletzte Modell hat den Vorteil hauptsächlich aus Schweizer Dokumenten zu bestehen, darunter die aus dem Staatsarchiv Zürich und aus Königsfelden. Wie bei den vorigen Modellen wird eine relativ lange Zeitspanne vom 16. zum 18. Jahrhundert abgedeckt und eine grössere Art Dokumente; als Sprache angegeben ist nur deutsch, was von Nachteil sein dürfte. Das letzte Modell enthält ebenfalls Dokumente aus dem »Rechtsprechung im Ostseeraum«-Projekt und ist zeitlich am engsten begrenzt, hauptsächlich bestehend aus Dokumente des 17. und 18. Jahrhundert. Diese sind allerdings wieder aus Norddeutschland und nur in deutscher Sprache verfasst. Welches Modell in der Praxis am besten abschneidet, muss durch Versuchstranskriptionen getestet werden.

Zuerst sollen alle Modelle auf dem gleichen Text getestet werden, damit ein Parallelvergleich einfach möglich ist. Dafür soll nur ein sehr kurzer Textauszug analysiert werden, um die Modelle auszusortieren, die am wenigsten für Gessners Schrift geeignet sind. Dazu dient der kürzeste Brief.[84] Um das Abschneiden der Modelle zu demonstrieren, wird die letzte Zeile des Briefes wiedergegeben:

Korrekt Version: »eorum mecum plurimum gratias agunt«.

Pylaia_NeoLatin_Ravenstein: »Copex. Paecute. plecgiturum Nortias, Agnate«.

NeoLatin_Ravenstein_1643-1772: »Dovmm. Naceee precsiuiti groctrus aguat«.

Acta 17 (extended): »corm nacum plurimum gratis agnat«.

Acta_17 PyLaia: »conum nocum plarimum gratius agnat«.

German_Kurrent_XVI-XVIII_M1: »corar nacuu plurimuer gratres aguat«.

German_Kurrent_17th-18th: »comm num plusimam gratios aguit«.

Es ist sofort klar, dass dies nicht die ermutigendsten Resultate sind; ein einziges Modell, Acta 17 (extended), hat es geschafft ein einziges Wort, »plurimum«, korrekt erkennen. Keines der sechs Modelle hat das »e« in »eorum« korrekt lesen können, obwohl es nicht besonders unleserlich aussieht; immerhin das folgende »o« hat keine Problem bereitet. Die Verwirrung

[84] https://hallernet.org/data/letter/02438/core.

der »rum«-Buchstabenfolge ist etwas verständlicher, da das »u« ohne Kontext kaum als solches erkannt werden kann und dadurch auch die Lesbarkeit des »u« beeinträchtigt; die meisten Modelle haben zumindest das »-m« ausgemacht. Bei dem »mecum«, das für das Menschenauge leicht lesbar ist, können nur die »Acta«-Modelle zumindest das Ende des Wortes lesen, wahrscheinlich weil das Training mit lateinischen Texten sie für das Erkennen solcher Endungen trainiert hat. »plurimum« bereitete noch am wenigsten Probleme; Gessners »p-« scheint für keines der Modelle ein Problem. »gratias« ist auch nicht zu schlecht, wenn bedacht wird, wie Gessner das letzte »a« hingekleckst hat. Es ist recht erstaunlich, dass nicht einmal ein mit lateinischen Texten trainiertes Modell ein so gewöhnliches Wort wie »agunt« lesen kann; es ist verständlich, warum das »u« als »n« gelesen wird, da in Gessners Schrift der Unterschied zwischen den beiden Zeichen gering ist, etwas weniger, wieso kein einziges Modell das »n« als solches erkannt hat.

Schauen wir Pylaia_NeoLatin_Ravenstein über das ganze Dokument genauer an, sieht die Situation nicht viel besser aus. In der Adresse, die etwas klarer geschrieben ist, kann das Modell immerhin ein »atque« entziffern. Aber wie schon in der letzten Zeile sehen wir beispielsweise ein »r« das als »p« gelesen wird. Auch das »Frater« zu »fater« und »tamen« zu »fraltrem« wird, ist ein Zeichen, dass die Schriften, auf die dieses Modell aufgebaut ist, nicht zu Gessner passen; ansonsten würde an einem Ort nicht ein »r« ergänzt und es an einem anderen wegfallen. Insgesamt ist eine CER von 52.36% weit davon entfernt hilfreich zu sein, weder für ein Überfliegen des Textinhalts noch als Hilfsmittel bei der Transkription. Auch Keyword Spotting funktioniert in einem solchen Fall nicht zuverlässig, nicht bei einer WER von 98%. Es ist schon nach diesem kleinen Textabschnitt klar, dass es sich nicht lohnt, dieses Modell weiter anzuschauen.

Ähnlich verhält es sich mit NeoLatin_Ravenstein_1643-1772, obwohl schon eine deutliche Verbesserung festzustellen ist; die CER liegt bei 39.21%, ein Rückgang von über 13% verglichen mit Pylaia_NeoLatin_Ravenstein. Dieses Modell lässt die »r« zwar auch nicht immer »r« sein, aber immerhin werden sie nur zu einem »n« oder einem »s«, was eine weniger extreme Abweichung ist; »s« und »r« sind bei Gessner kaum auseinanderzuhalten. Immerhin ist hier »Frater« »fratr« und »tamen« »Tacien«; beides ist nicht korrekt, aber es sind erklärliche Fehler, wenn auf das Digitalisat geschaut wird. Wenn das menschliche Auge jedoch Mühe

hat, »precsiuiti« zu lesen, wo »plurimum« steht, verschmiertes mittleres »m« hin oder her, ist die Verwirrung doch zu gross, um hilfreich zu sein. Die Fehler sind eher nachvollziehbar, aber zum damit Arbeiten eignet sich auch dieses Modell nicht.

Die anderen Modelle versprechen bessere Resultate. Bei Acta 17 (extended) sinkt die CER nochmals um 17% auf 22.22%. Das ist immer noch nicht an dem Punkt, an dem ein Text einfach verständlich ist, das Modell erreicht aber ein Resultat, das für das Keyword Spotting genügen sollte und damit für eine Volltextsuche mit dieser Technologie.[85] Auch wenn ein »a« für ein »u« oder ein »o« für ein »e« gehalten werden, sind das Fehler, die auch einem Menschen ohne Kontext passieren könnten. Es sind relativ kleine Fehler und teilweise gibt es perfekt gelesene Abschnitte bis wieder alles unleserlich wird und ein »Frater« zum »tater« wird, »tamen« zu »tacen«.

Acta_17 PyLaia schneidet etwas schlechter ab, was nicht erstaunt, wenn bedacht wird, dass es weniger Material zur Verfügung hat als Acta_17 (extended) und auch keine Dokumente enthält aus der Zeit, in der diese Briefe verfasst worden sind. Die CER steigt leicht um 2% auf 24.14%. Die Transkriptionen der Modelle sind sich auch untereinander so ähnlich wie dem korrekten Text und in den meisten Fällen ist ersteres Modell korrekter; so wird »Frater« zu »tator«, um einen Buchstaben weniger korrekt; »tacen« bleibt als Transkription. Es gibt einzelne Ausnahmefälle, wo dieses Modell besser abschneidet, insbesondere die Zeile, die in korrektem Latein als »iam spes mihi quam olim apparet« transkribiert wird und bei Acta_17 (extended) » f spec nehi quam obm apperet« lautet. Besonders interessant ist hier, dass Gessner das »jam« durchgestrichen hat. Möglicherweise kann Acta_17 PyLaia besser mit schlecht leserlichem Material zurechtkommen.

German_Kurrent_XVI-XVIII_M1 erreicht eine CER von 26.56%, wiederum etwas über 2% schlechter als das vorige Modell, was umso enttäuschender ist, als dass es Schweizer Dokumente enthält und daher regional besser passen sollte. Wir sehen hier den »Frater« als »Feater«, immerhin die einzige Lesart, bei der der Grossbuchstabe erkannt wird, und »tamen« als »tacer«. Wieso so viele Modele das »-m-« als »-c-« lesen wollen ist verblüffend, speziell, da sie nicht noch einen Buchstaben wie etwa »n« ergänzen – es wird nur der erste Strich des »m«

[85] Gemäss dem How-To-Guide von Transkribus ist ab einer CER von 20-30% Keyword Spotting einsetzbar, https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/#elementor-toc__heading-anchor-10.

gelesen und dann zum »e« gesprungen. Insgesamt bietet dies Modell fast nirgends bessere Lesarten als die beiden Acta-Modelle.

Das letzte Modell German_Kurrent_17th-18th hat mit Acta_17 PyLaia gemein, dass es den »spes«-Abschnitt richtig wiedergibt. Es neigt dazu, »s« als »c« zu lesen oder Buchstaben zu kürzen; so bleibt von »-rum« nur »-mm«, von »mecum« »num«. Es hat in ein paar Fällen die beste Lesart, aber an den Stellen, wo die anderen Modelle Probleme gehabt haben, scheitert es auch. Insgesamt ist es Acta_17 PyLaia am ähnlichsten, zwischen einander haben die durch sie erzeugten Texte eine CER von 20.1%. Mit einer CER von 22.86% schneidet German_Kurrent_17th-18th aber über 3% besser ab als Acta_17 PyLaia, nur übertroffen von Acta_17 (extended).

Vorerst sollen im Folgenden die beiden Modelle weggelassen werden, die bei weitem am schlechtesten abgeschnitten haben, also Pylaia_NeoLatin_Ravenstein1 und NeoLatin_Ravenstein_1643-1772; die anderen sind alle innerhalb der 20%-30% CER, wo Keyword Spotting einsetzbar ist, und alle haben am einen oder anderen Ort die beste Lesart gehabt.

	Pylaia_NeoLatin_Ravenstein	NeoLatin_Ravenstein_1643-1772	Acta_17 (extended)	Acta_17 PyLaia	German_Kurrent_XVI-XVIII_M1	German_Kurrent_17th-18th
CER	52.36%	39.21%	22.22%	24.14%	26.56%	22.86%

Deswegen soll beispielsweise German_Kurrent_XVI-XVIII_M1 nicht vorzeitig aussortiert werden, denn dieser eine Brief, der hier verwendet worden ist, entspricht schliesslich nicht dem ganzen Corpus und es kann nicht ausgeschlossen werden, dass die Resultate für einen anderen Brief abweichen werden und die Modelle verschiedene Stärken und Schwächen zeigen werden. Die vier Modelle werden im Folgenden weiter getestet werden, diesmal mit dem Fokus nicht nur darauf, wie korrekt die Transkription ist, sondern, wie viel effizienter die manuelle Transkription wird durch die maschinelle Vorarbeit.

3.3 Transkribieren mit Transkribus: Testen der Modelle

In diesem Abschnitt sollen die obig ausgewählten vier Modelle darauf getestet werden, inwieweit sie bei der Transkription helfen und die Arbeitsdauer verkürzen. Anders als oben kann hier also nicht der gleiche Abschnitt von allen Modellen übersetzt werden, sondern es

braucht einen neuen für insgesamt sechs Transkriptionen, diejenigen mit den vier Modellen und diejenigen ohne HTR-Unterstützung zur Kontrolle, um die Arbeit mit und ohne HTR-Modelle zu vergleichen. Dafür müssen gefunden werden, die sich untereinander möglichst ähnlich sind in Schriftbild und Inhalt. Am einfachsten ist es, diese im gleichen Brief, wo Schrift und Papierqualität konsistent sein sollten, zu suchen; verwendet wird daher der längste Brief, der im Corpus existiert.[86] Elemente wie die Begrüssungs- und die Abschiedsformel werden ausgelassen, da der Inhalt in diesen Fällen vorhersehbar ist, was die Transkription vereinfacht und das Resultat verfälschen könnte.

Es ist zu beachten, dass im Lauf der Zeit, die eigene Fähigkeit, den Text zu transkribieren, zunimmt, da sich das Auge an Gessners Schrift gewöhnt. Daher wird für die ungestützte Transkription die Zeit zweimal gemessen, vor und nach den Transkriptionen mit den vier Modellen. Der erste Abschnitt auf Seite 1 besteht dabei aus 19 Zeilen mit insgesamt 164 Wörter von »Octavo Iduum« bis »facta erit«. Die Transkription benötigt 35 Minuten, also dauert es ungefähr 12.8 Sekunden um ein Wort zu transkribieren; die eigene CER beträgt dabei 9.18%, an dem Punkt, an dem es schwierig geworden ist, ohne Hilfsmittel weiter zu transkribieren. Auf Seite 6 geht der gewählte Abschnitt vom Beginn der Seite bis »petieram misit«, insgesamt 20 Zeilen und 192 Wörtern. Bei einer Transkriptionsdauer von 34 Minuten ergibt dies 10.6 Sekunden pro Wort bei einer CER von 10.22%.

Die HTR sollte dort als Stütze dienen, wo am meisten Probleme bei der Transkription bestehen. Dazu gehört etwa, dass «ss« als »p« wie etwa »prope« statt »posse« gelesen wird; das liegt daran, dass Gessner das erste »s« »̓ « schreibt, was mit dem folgenden kleinen »s« wie ein »p« mit einem Schlenker nach oben daherkommt; da »pope« kein Wort ist, ist ein »r« schnell ergänzt. »s«, »r« und »i« sind leicht zu verwechseln, ebenso »a«, »e« und »o«. Schnell ist auch ein »s« am Ende eines Wortes übersehen. oder das Abkürzungszeichen des Genitiv Plurals. Die Ligatur »æ« ist teils auch schwer zu erkennen. Ein weiteres Problem, das auch für die Modelle schwierig werden könnte, sind die Striche, mit denen Gessner aufzeigt, wo ein ausserhalb der Zeile geschriebenes Wort hingehört; diese Striche nicht auf den nächsten Buchstaben zu beziehen könnte für eine Maschine schwierig sein, da die nicht versteht, dass es ein Wort ausserhalb der aktuellen Baseline gibt, das in den Text gehört. Was nichts mit der

[86] https://hallernet.org/data/letter/02422/core.

HTR zu tun hat, aber wichtig ist für die Erarbeitung einer guten Transkription, ist bei der Korrektur auf Abweichungen zwischen Gessners Hand und Hallers Textwiedergabe achtzugeben. So wird in Hallers Edition »singillatim« zu »sigillatim«, doch der Text und die lateinische Sprache sprechen für Gessner; das gleiche gilt für »liberalissime«, bei Haller »liberatissime«. Weitere Unterschiede sind etwa Gessners »Airol«, bei Haller »Airolum«, und Gessners »alloquio«, das von Haller zu dem gewöhnlicheren »colloquio« umgeschrieben wird.

Die zweite Seite wird nun mit German_Kurrent_17th-18th, dem Modell, das vorher am zweitbesten abgeschnitten hat, maschinell transkribiert und dann korrigiert, zuerst ohne die Hilfe von Hallers Edition. Für einen Ausschnitt mit 190 Wörtern, von »Cerinthe major« bis »speciebus quas«, wird eine CER von 22.58% erreicht, was dem Resultat bei dem vorigen Brief sehr ähnlich ist und zeigt, dass das Modell recht zuverlässig ist für die Transkription von Gessners Hand. Benötigt worden für eine Bereinigung des Textes sind 37 Minuten, das heisst 11.7 Sekunden pro Wort, was zwischen den beiden Werten der Transkription ohne maschinelle Hilfe ist; zeitlich ist also keine grosse Unterstützung festzustellen. Die Transkription hat sich gefühlsmässig zwar hilfreich angefühlt, aber der Text weist nach der Korrektur immer noch eine CER von 11.77% auf, höher als bei dem Text ohne Transkriptionshilfe. Allerdings gilt es festzustellen, dass dies zwar ein Abschnitt aus dem selben Brief ist, die Thematik aber eine andere. Es geht in viel mehr Detail um die Bestimmung von Pflanzen, weshalb sehr viel mehr Fachbegriffe, vor allem Pflanzennamen und botanische Quellenangaben, als im ersten Abschnitt erwähnt werden. Diese sind logischerweise in keinem Trainingscorpus und daher für ein auf Juristenlatein trainiertes Modell nicht leicht zu erkennen und auch für einen Nicht-Botaniker bereiten diese unbekannten Begriffe Probleme.

Für die dritte Seite wird Acta_17 PyLaia verwendet, das vorher am drittbesten abgeschnitten hat. Für einen Abschnitt mit 172 Wörtern, von »Myosotis verò« bis »sicca Tigurum«, dauert die Transkription diesmal ganze 44 Minuten oder 15.3 Sekunden pro Wort. Das ist eine seltsame Steigerung; die HTR-Transkription hat sich zwar etwas weniger hilfreich angefühlt als mit dem vorherigen Modell, aber doch besser als mit nichts. Die CER beträgt auch nur 25.19%, 1% schlechter als im ersten Brief, und nach der Korrektur liegt die CER noch bei 9.98%; erstaunlicherweise ist das besser als vorher bei German_Kurrent_17th-18th, aber immer noch schlechter als ohne Modell. Dieser Abschnitt ist ähnlich geartet wie der vorige, also

mit vielen Pflanzennamen und Fachbegriffen. »g« und »y« werden oft als »q« wiedergegeben; »q« ist ein häufiger Buchstabe in Relativpronomina im Latein, was ein Grund sein mag, dass das Modell diesen Buchstaben bevorzugt, während »y« nur in griechischen Fremdwörtern vorkommt. Römische Zahlen und Zahlen allgemein bereiten Mühe. Doch das sind alles keine grossen Unterschiede von dem vorherigen Modell; es ist einfach alles noch ein bisschen weiter weg von dem erstrebten Resultat.

Für Seite 4 kommt nun German_Kurrent_XVI-XVIII_M1 an die Reihe, dasjenige Modell, das bei dem vorherigen Brief am schlechtesten abgeschnitten hat. Für 185 Wörter auf 19 Zeilen, von »Euphrasia lutea« bis »quærendum ė se«, dauert die Transkription diesmal 36 Minuten, was 11.7 Sekunden entspricht und etwa gleich lang ist wie bei German_Kurrent_17th-18th. Die CER liegt bei 22.83%, was auch nur wenig schlechter ist als German_Kurrent_17th-18th und eine massive Verbesserung im Vergleich mit dem ersten Brief; nach der Bereinigung der maschinellen Transkription sinkt die CER diesmal auf 7.45%, ist also das erste Modell, das bessere Ergebnisse liefert als die ungestützte Transkription. Es ist allerdings zu bedenken, dass an diesem Punkt schon einiges gesichtet worden ist, das Auge sich also an Gessners Schrift gewöhnt haben kann. Auch kommen etwas weniger Pflanzennamen vor; auf der anderen Seite zeigt sich eine weitere Komplikation, die Verwendung eines anderen Alphabets. Auf der 15. Zeile ist ακριβεια (ohne Spiritus oder Akzent) geschrieben und das Modell kann das andere Alphabet nicht erkennen, da es nicht dafür trainiert worden ist. Bei Zahlen hat das Modell auch einige Probleme, kommt aber mit »q« und »p« gut zurecht. Eine seltsam Lesung, speziell da das Modell dazu neigt, Buchstaben zu überspringen, ist in der 17. Zeile »Semprer«, wo deutlich »Sempe« steht.

Zum Schluss kommt noch das Siegermodell des vorherigen Briefs zum Zug, Acta 17 (extended), das diesmal um einiges weniger gut abschneidet und nur 24.26% CER vorweisen kann. Nur das nahe verwandte Acta_17 PyLaia hat schlechter abgeschnitten und das ist, wie schon erwähnt, vom Zeitrahmen und dem Trainingsmaterial her logisch. Trotz der hohen Fehlerzahl benötigt die Transkription nur 38 Minuten für 218 Wörter (von »Quidam in distinctis« bis »angusto instructo«), was 10.5 Sekunden pro Wort entspricht; das ist der schnellste Transkriptionsprozess. Dabei muss aber die Genauigkeit der Geschwindigkeit zum Opfer gefallen sein, denn die CER erreicht danach nur 12.81% und schneidet somit am schlechtesten ab. Die

Transkription des Modells fühlte sich auch weniger hilfreich an, als bei den anderen Modellen. Was es jedoch um einiges besser gekonnt hat als die vorherigen Modelle, ist Zahlen zu lesen. Bei Fachwörtern hilft es jedoch so wenig wie die anderen Systeme: »Muscus trichoides« wird einmal zum »Mscum Insaenden«, ein andermal zu »Msus inherdes«. Die Endungen sind zwar richtig, aber die Bedeutung der Wörter zu erraten ist unmöglich.

Zusammenfassen lassen sich die Resultate in der folgenden Tabelle:

	Eigentranskription Seite 1	Eigentranskription Seite 6	German_Kurrent_17th-18th	Acta_17 PyLaia	German_Kurrent_XVI-XVIII_M1	Acta 17 (extended)
Sec/Wort	12.8	10.6	11.7	15.3	11.7	10.5
CER HTR-Transkription	-	-	22.58%	25.19%	22.83%	24.26%
CER eigene Transkription	9.18%	10.22%	11.77%	9.98%	7.45%	12.81%

Es zeigt sich, dass die Modelle nicht ganz gleich abgeschnitten haben wie zuvor. Die Leistung von Acta_17 (extended) bei dem ersten Brief kann nicht bestätigt werden und wird vom 1. auf den 3. Platz verdrängt. Stattdessen hat German_Kurrent_XVI-XVIII_M1 viel besser abgeschnitten als nach dem ersten Experiment zu erwarten gewesen ist, und hat beinahe am besten abgeschnitten. Allein German_Kurrent_17th-18th steht ungefähr gleich gut da wie bei dem vorigen Brief. Insgesamt haben alle Modelle ähnliche Probleme, nur verschieden stark ausgeprägt, aber es lassen sich keine übergeordneten Muster ableiten, die es erlauben mit Sicherheit ein Modell dem anderen vorzuziehen. So kann nur Acta_17 PyLaia ohne Probleme weggelassen werden, da es nichts leistet, was Acta_17 (extended) nicht besser kann.

Für einen dritten Versuch, die Qualität der Transkriptionen durch die drei verbliebenen Modelle zu testen, wird wieder ein anderer Brief[87] ausgewählt und jeweils eine ganze Seite transkribiert mit den drei vorher bestimmten drei Modellen. Die erste Seite wird wieder ohne HTR-Unterstützung, die zweite mit German_Kurrent_XVI-XVIII_M1, die dritte mit Acta 17 (extended) und die letzte mit German_Kurrent_17th-18th transkribiert. Diesmal ähnelt die

[87] https://hallernet.org/data/letter/02420/core.

Leistung der einzelnen Modelle eher jener bei dem ersten Brief, wie aus der Tabellenzusammenfassung der im Verlauf dieses Experiments erreichten Resultate hervorgeht:

	Eigentranskription	German_Kurrent_ XVI-XVIII_M1	Acta 17 (extended)	German_Kurrent_1 7th-18th
Sec/Wort	11.7	10.6	10.2	8.8
CER HTR-Transkription	-	26.21%	20.43%	24.78%
CER korrigierter Text	6.25%	3.62%	4.94%	4.09%

Es zeigt sich jetzt auch, dass die HTR-Modelle doch die Transkription leicht beschleunigen. Auf das Wort heruntergerechnet ist die Zeiteinsparung nicht besonders eindrücklich, aber über eine längere Zeit wäre die positive Auswirkung ersichtlicher. Auch qualitativ ist eine Verbesserung deutlich zu erkennen dank der Anwendung eines HTR-Modells; da an diesem Punkt die Transkriptionsarbeit um einiges vertrauter ist, kann angenommen werden, das mehr Vertrauen in die Konsistenz der gewonnenen Daten möglich ist. Leider überschneiden sich die besten Aspekte der drei Modelle nicht – jedes hat seine eigenen Stärken. Acta_17 (extended) liest am genauesten, German_Kurrent_17th-18th erlaubt die schnellste Transkription und German_Kurrent_XVI-XVIII_M1 liefert das beste Endresultat.

Der Trend, das German_Kurrent_XVI-XVIII_M1 am meisten bei der Transkription hilft, lässt sich also bestätigen. Es liest also dort am besten, wo es für den Menschen schwierig ist. Acta 17 (extended) und German_Kurrent_17th-18th hingegen haben die Ränge getauscht für Geschwindigkeit und Ausgangs-CER. German_Kurrent_XVI-XVIII_M1 und Acta 17 (extended) scheinen am meisten in der Qualität zu schwanken, was nicht etwas ist, das in einem HTR-Modell wünschenswert ist. German_Kurrent_17th-18th dagegen hat über alle drei Briefe hinweg, auf denen es getestet worden ist, relativ konsistent abgeschnitten.

Wird darauf geschaut welche Probleme in der menschlichen Transkription durch die HTR verhindert worden sind, gibt es nur wenige Stellen, auf die explizit verwiesen werden kann. Es werden sehr viele Buchstaben vertauscht, sowohl da, wo ein Mensch es auch täte, als auch anderswo. Es gibt einige Fälle, wo die Modelle helfen, aber ebenso oft lenkt die HTR-Tran-

skription mehr ab von dem richtigen Begriff als auf ihn zu verweisen. Ein gutes Beispiel für eine Stütze bei der Übersetzung findet sich bei dem langen Brief vom 25. Mai 1732; German_Kurrent_17th-18th erkennt die Zeichenfolge »ſ s«, die so leicht als »p« gelesen werden kann, korrekt als »ss«. German_Kurrent_XVI-XVIII_M1 liest zumindest ein »s«; zwar weniger korrekt, lenkt es doch die Gedanken des Transkribierenden in die richtige Bahn. Es zeigt sich auch wie erwartet, dass die Striche, die Ergänzungen über dem Text auf der unteren Zeile einordnen, die Systeme verwirren. So wird auf Seite 2 desselben Briefes »junctam« zu »punctam« bei German_Kurrent_17th-18th; diese Verwirrung ist dem Strich, der vor dem Wort »D. Magnol« einfügt, geschuldet, der das »j« durchschneidet. Es gibt aber auch Stellen, wo diese Linien die Transkription nicht beeinflussen, so mit Acta_17 PyLaia auf der nächsten Seite vor »quas memoras« oder auf der fünften Seite vor »distinctius« mit Acta 17 (extended).

Es ist ebenfalls festzuhalten, dass die Textgattungen der Corpora nicht unwichtig sind. Das mag etwas verwirren, da neuronale Netze sich bekanntlich nicht um die Bedeutung scheren, die hinter den Daten steckt, mit denen sie lernen. Aber verschiedene Sprachen haben verschieden häufige Buchstabenreihenfolgen und das Modell merkt sich das ganze Umfeld eines Buchstabens; es weiss schliesslich nicht, was genau alles »Buchstabe« ist im Bild. Ein ungewohntes Umfeld, wie es in einer anderen Sprache auftaucht, verwirrt, so dass auch ohne Sprachmodell die Modelle nicht ganz sprachunabhängig sind. Weiter tauchen Inkonsistenzen auf, die auf die Trainingscorpora zurückzuführen sind. Besonders auffällig ist dies bei der Abkürzung »&«, die Acta (extended) sowohl als »&« als auch als »et« wiedergibt; die Transkription als »e«, wie sie bei mehreren Modellen vorkommt, hängt möglicherweise auch damit zusammen. Für die Genitiv-Plural-Endungen sind die Transkriptionen auch nicht ideal. So wird »caulicorum« auf Seite 3 des Briefs vom 25. Mai 1732 mit Acta_17 PyLaia zu »conliruterid« und wenige Zeilen unterhalbe »florum« zu »ferud«; der Abkürzungsschlenker wird als »d« gelesen wird. Auf Seite 5 liest Acta_17 (extended) wiederum »muscorum« als »musciens«. Zusätzlich kommt, dass keines der getesteten Modelle Ligaturen wie »æ« oder Akzente verwendet. Da der Referenztext für den Vergleich der CER diese aber enthält, steigt die CER automatisch, auch wenn es für die Lesbarkeit nicht nötig ist, das Auflösen einer Ligatur oder das Weglassen eines Akzents als Fehler zu werten.[88]

[88] Vgl. die Beobachtung bei Plüss/Sieber (2020), 227, »Variierende Transkriptionsrichtlinien können ebenfalls eine Rolle spielen [d.h. daran, dass Trainingsdaten aus anderen Projekten die Ergebnisse nicht verbessern]«.

Auf der reinen Navigationsebene muss zu dem Transkriptionserlebnis mit Transkribus gesagt werden, dass es nicht immer ideal ist. Ein allgemeines Designprobleme von Transkribus ist die Unterstreichung der Baseline, die zwar nicht sehr dick ist, aber doch dick genug, um in manchen Fällen im Weg zu sein.[89] Es ist auch nicht möglich im Transkriptionstext auf eine andere Zeile zu springen, um die Baseline loszuwerden, denn dann ändert sich die Bildansicht und muss wieder angepasst werden. Es lässt sich recht problemlos darum herum arbeiten, aber ständige darum herum Navigieren kann ermüden. Auch mit der Tastatur Sonderzeichen einzugeben ist nicht ideal. Beispielsweise bei einem »â« oder »ò« fügt der Texteditor automatisch den zuvor gelöschten Buchstaben wieder hinzu, was zu Fehlern im Text führt, wird nicht darauf geachtet.

Transkribus und seine frei verfügbaren Modelle sind also nur bedingt geeignet, um ungefähre Transkriptionen zu erstellen, wobei die Effizienz natürlich textabhängig ist und stark variieren kann. Die meisten Modelle sind auf offiziellen Dokumenten trainiert, in denen oft Wert darauf gelegt wird, gleichmässig und sauber zu schreiben. Bei solchen Unterlagen ist es möglich, dass das Layout und die Lesarten bei der Transkription etwas regulärer sind, was möglicherweise die Differenz unter den verschiedenen Modellen mehr hervorstechen liesse. Gessners Briefe sind dagegen hauptsächlich für seinen Freund bestimmt, auf den Informationsaustausch ausgerichtet und nicht im Voraus geplant oder ins Reine geschrieben. So kann die Schrift am Ende sehr gedrängt werden, was die Transkription auch verschlechtert. Die HTR bietet für diese Art Text nur eine mässige Unterstützung bei der Transkription. Mit der Wahl des richtigen Modells ist sie allerdings gut genug für Keyword Spotting, was doch den Zugang zum Text verbessert, wenn auch zuerst nur für diejenige Person, die darauf zurückgreifen kann. Es müsste also genau bedacht werden, was für ein Endzustand der Transkriptionen angestrebt wird und wie sie der Öffentlichkeit anschliessend zur Verfügung gestellt werden soll. Ansonsten lohnt es sich, doch an der HTR-Transkription selbst nochmals anzusetzen.

[89] Vgl. bereits Massot *et al*. (2018), 6, »la ligne en cours de saisie est surlignée en bleu sur l'image, ce qui peut nuire à la lisibilité.«

4. Das Gessner-Modell

4.1 Das Erstellen eines Modells

Ein wesentlicher Grund, warum es sinnvoll ist im Fall der Gessner-Briefe daranzugehen ein eigenes Transkriptionsmodell zu bilden, ist der einfache Umstand, dass es ein grosses Textcorpus ist, in dem nur eine einzige Hand vorkommt. Wenn von etwa 500 Wörtern pro Brief ausgegangen wird, kann mit fast 320'000 Wörtern insgesamt gerechnet werden; das heisst, es braucht theoretisch weniger als 5% allen Materials vortranskribiert zu werden, um ein sehr gut funktionierendes Modell zu trainieren.[90] Es gibt einige Einschränkungen im Gedächtnis zu behalten, wie dass sich eine Handschrift im Lauf der Zeit verändert. In der Praxis wäre es also wichtig, Briefe aus allen Lebensphasen des Autors in ein Trainingscorpus einzubeziehen, was wie anfangs von Kapitel 3 erklärt hier der Übersichtlichkeit halber nicht der Fall ist. Auch wird das System nicht mit 15'000 Wörtern trainiert werden; es geht hier schliesslich auch darum zu testen, wie gering der Transkriptionsaufwand sein kann, um zumindest ein bereits bestehendes Modell überflügeln zu können und ab wann es einen bedeutenden Vorteil bietet für die Transkription. Angestrebt wird in Anbetracht der Leistungen der Modelle in Kapitel 3 eine CER von unter 20%.

In einem ersten Schritt wird nur angestrebt, ein Modell zu erstellen, das besser ist als ein öffentliches. In jedem Fall ist jedoch zu betonen, dass es nie das Ziel ist, maschinell einen druckreifen Text zu erhalten; der wissenschaftliche Anspruch an die Transkription dieser Art Archivalien ist zu hoch, um sie einzig der Maschine zu überlassen. Es geht einzig darum, wie viel Aufwand nötig ist, bis ein Vorteil durch die HTR erlangt werden kann.

In Kapitel 3 sind drei Briefe transkribiert worden mit insgesamt 13 Seiten, 501 Zeilen und 3624 Wörtern.[91] Das ist deutlich unter der empfohlenen Mindestzahl von 5000 Wörtern und ein Teil dieser Daten muss noch zur Seite gelegt werden, um als Validierungsset zu dienen. Schlussendlich enthält das Trainingsset für dieses erste eigene Modell nur 12 Seiten, 455 Zeilen und 3251 Wörter, das Validierungsset 1 Seite mit 46 Zeilen und 373 Wörter.

---

[90] Vgl. Muehlberger *et al*. (2018), 959, »A ground truth data set of 15,000 transcribed words (or around 75 pages) is generally sufficient for training an HTR engine to recognise text written in one hand«.

[91] Nicht gerechnet sind unbeschriebene oder nur die Adresse enthaltende Seiten. Eingeschlossen sind dagegen die zwei letzten Seiten des Briefs vom 25. Mai 1732, die zur Zeitgewinnung mit Hilfe von Hallers Edition rein manuell transkribiert worden sind.

Um zu sehen, welche Engine besser funktioniert bei einem so kleinen Trainingsset, wird sowohl ein Modell mit HTR+ trainiert und mit den genau gleichen Trainingsdaten eines mit PyLaia. Für das HTR+-Modell wird die Grundeinstellung von 50 Epochen beibehalten; bei PyLaia werden ebenfalls die Standardeinstellungen verwendet, also maximal 250 Epochen und eine Lernrate von 0.0003; einzig der Mindestanhaltspunkt wird zur Sicherheit auf 50 erhöht, da das Trainingsset sehr klein ist.[92] Das Training dauert mit HTR+ eine Stunde und fünfzig Minuten, mit PyLaia nur 28 Minuten. Die CER auf dem Validierungsset beträgt für HTR+ 12.55%, für PyLaia 19.6%.

Beide Modelle haben es also unter 20% geschafft und sind beide schon jetzt besser als die öffentlichen Modelle. Das HTR+-Modell ist sogar sehr nahe an der 12%-Grenze, ab der ein Modell als nützlich betrachtet werden kann.[93] Die Lernkurven zeigen weiter, dass mit der HTR+-Engine nach ungefähr acht Epochen die endgültige CER erreicht werden kann; die PyLaia-Engine braucht etwas über 150 Epochen. Die Leistungen auf dem Trainingsset sind besser bei HTR+ und das ab der ersten Epoche, während sie bei PyLaia ungefähr gleich der auf dem Validierungsset bleiben bis zur 80. Epoche und erst dann langsam weiter sinken. Insgesamt lässt die Lernkurve schliessen, dass beide Modelle an einem guten Punkt ihren Abschluss gefunden haben.

Es gilt jetzt noch, das Vermögen der Modelle an einem Brief Gessners zu testen; der Brief vom 14. Oktober 1732[94] ist auf zwei Seiten beschrieben. Die erste Seite mit dem HTR+-Modell transkribiert erreicht eine CER von 20.82%. Das ist nur knapp besser als die besten öffentlichen Modellen und viel schlechter als die Leistung des Modells auf dem Validierungsset. Ein möglicher negativer Faktor, der zu dem schlechten Resultat geführt hat, kann die Tatsache sein, dass die Schrift in diesem Brief stärker verblasst ist als in dem Trainigsmaterial und das Modell daher auch mehr Mühe gehabt hat, alle Buchstaben zu erkennen. Ein gutes

[92] Vgl. den Hinweis im Tutorial: »Wenn es keine oder nur wenig Variation im Validation Set gibt, stoppt das Modell möglicherweise zu früh. Wenn Ihr Validation Set also eher klein ist, erhöhen Sie bitte den "Early Stopping"-Wert, um zu vermeiden, dass das Training stoppt, bevor es alle Trainingsdaten gesehen hat.«, https://readcoop.eu/de/transkribus/howto/how-to-train-pylaia-models-in-transkribus/#h-early-stopping. In der Praxis hat sich das aber nicht als Problem erwiesen und alle Modelle haben problemlos 250 Epochen durchlaufen.

[93] Vgl. Hodel (2020), 84, »Ab der Schwelle um 12% wird die Korrektur von erkanntem Text gegenüber von händisch erstellten Transkriptionen ökonomisch sinnvoll. Gleichzeitig sind die Resultate ab 12% für Menschen insofern nützlich, da die Navigation im Text, insbesondere für Personen mit Kenntnissen der Dokumente, rasch und zielsicher möglich ist.«

[94] https://hallernet.org/data/letter/02426/core.

Zeichen ist, dass die Korrektur der 162 Wörter nur 16 Minuten, also 5.9 Sekunden pro Wort, gedauert hat, was deutlich besser als bei den Modellen in Kapitel 3 ist; das Modell halbiert hier den Zeitaufwand für die Transkription. Ein speziell auf das eigene Corpus angepasstes Modell hilft also schon mit ganz wenig Trainingsdaten deutlich mehr als ein allgemeines. Auch das Keyword Spotting funktioniert gut.

Für die zweite Seite mit PyLaia dauert die Transkription von 153 Wörtern 17 Minuten, das heisst 6.7 Sekunden dauert die Transkription eines Wortes. Die CER liegt bei 21.97%, was nur etwas schlechter ist als bei dem HTR+-Modell, und keine signifikante Verschlechterung im Vergleich mit dem Validierungsset. PyLaia erlaubt kein Keyword Spotting. Der Unterschied zwischen den beiden Modellen ist für die eigentliche Transkription um einiges geringer als die CER des Validierungsset suggeriert, aber die Beobachtungen deuten doch darauf hin, dass das HTR+-Modell bei einer ganz geringen Menge Trainingsmaterial vorzuziehen ist. Die Reduktion der Transkriptionszeit zeigt auch, dass diese Modelle mehr bei der Arbeit helfen. Probleme haben beide mit Elementen wie Buchstaben, die teils über oder unter der Linie geschrieben sind; diese werden häufig nicht vollständig gelesen. Beispielsweise wird ein »h« als »n« wiedergegeben. Ebenfalls Mühe bereiten Grossbuchstaben, die spärlicher im Trainingscorpus vorhanden sind.

Nun soll versucht werden, wie sehr das Resultat verbessert werden kann durch das Hinzufügen eines Basismodells; alle anderen Faktoren, Epochen und Grösse von Trainings- und Validierungsset, bleiben die gleichen. In Kapitel 3 haben drei Modelle sich vernünftig zurechtgefunden in Gessners Briefen; diese bieten sich somit dafür an, als Unterstützung zu den eigenen Daten hinzugefügt zu werden. Ein Basismodell ist explizit dafür da, bei wenig Trainingsmaterial Verbesserungen zu bewirken,[95] was genau auf dieses Szenario zutrifft. Allerdings könnten andere Hände bei einem so spezifisch auf eine Person zugeschriebenen Modell möglicherweise mehr verwirren als helfen, somit ist eine Verbesserung nicht garantiert. Ein Basismodell funktioniert nur, wenn es mit der gleichen Engine trainiert worden ist wie das neue Modell. German_Kurrent_17th-18th kann also mit PyLaia verwendet werden und Acta 17 (extended) und German_Kurrent_XVI-XVIII_M1 mit HTR+. Die Resultate auf dem Validie-

[95] Vgl. die Beschreibung im Tutorial zu HTR+, https://readcoop.eu/de/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/#elementor-toc__heading-anchor-5, »Ein großer Vorteil der Arbeit mit Basismodellen ist, dass sie es ermöglichen, mit einer geringeren Anzahl von Trainingsseiten zu beginnen, was bedeutet, dass der Transkriptionsaufwand reduziert wird.«

rungsset sind mit einer Ausnahme vielversprechend und bereits an einem Punkt angelangt, an dem die Transkription in der Regel als effizient betrachtet werden kann:

	PyLaia-Modell mit German_Kurrent_17th-18th	HTR+-Modell mit Acta 17 (extended)	HTR+-Modell mit German_Kurrent_XVI-XVIII_M1
CER Validation	9.10%	98.19%	9.46%

Das Modell mit Acta 17(extended), um zuerst den Abweichler anzusprechen, ist nicht nur sofort als unbrauchbar erkennbar, es kann überhaupt nichts transkribieren, obwohl das Character Set alle benötigten Zeichen enthält. Wird das Modell aber auf eine Seite angewandt, kommt kein einziger Buchstabe zurück. An der Lernkurve ist ersichtlich, dass die Validationslinie nach der ersten Epoche erwartungsgemäss sinkt und in der zweiten Epoche bei 92.3% liegt, aber in der dritten sofort wieder hinaufgeht auf 98.8 und danach sich fast nicht mehr bewegt; die Trainingslinie sinkt bis zu 89.3% in der vierten Epoche und pendelt sich dann in der sechsten Epoche bei 98.19% ein. Es wird keine Fehlermeldung ausgegeben. Acta 17 (extended) ist aus irgendeinem Grund absolut inkompatibel mit dem eigenen Trainingsmaterial, obwohl nichts an den Hintergrundinformationen zum Modell oder dem Character Set auf die Ursache hindeutet, weder in Bezug auf die Gessner-Briefe noch im Vergleich mit German_Kurrent_XVI-XVIII_M1. Dieses andere Basismodell hat aber die CER im Vergleich zu dem rein mit Gessners Schrift trainierten Modell um über 3% verbessert; das gleiche ist der Fall bei PyLaia, wo sich die CER sogar mehr als halbiert hat, obwohl bei dieser Engine vor dem Training jeweils gewarnt wird, dass Basismodell und eigenes Modell das gleiche Zeichenset benutzen müssen und daher, wenn irgendwo, hier ein Problem zu befürchten gewesen wäre.[96]

Für den Praxistest wird der Brief vom 15. Mai 1733 verwendet.[97] Die erste Seite, transkribiert mit dem PyLaia-Modell, erreicht eine CER von 8.06%, nochmals signifikant besser als

[96] Der genaue Wortlaut der Warnung, die vor dem Trainieren eines PyLaia-Modells mit Basismodell erscheint, lautet: »Training with base models for PyLaia requires the exact same character set. Elsewise, the training will produce an error or a model that outputs only the characters from the base model and is unable to use a language model. Only use base models if you are really sure that the training data contains the exact same characters as the base model.«. In der Praxis hat sich aber gezeigt, dass die Modelle, die mit PyLaia trainiert worden sind, durchaus die Buchstaben, beispielsweise ein » «, die nur in den Gessner-Briefen und nicht im Character Set des Basismodells vorhanden sind, wiedergeben können. Auch das Sprachmodell steht jeweils zur Verfügung.

[97] https://hallernet.org/data/letter/02436/core.

auf dem Validierungsset; für 275 Wörter dauert die Transkription 20 Minuten, 4.4 Sekunden pro Wort, ebenfalls eine beachtliche Verbesserung zu vorhin. Ein genauerer Blick auf die Transkription zeigt, dass die Verbindung der Gessner-Transkriptionen mit einem grösseren Modell genau dort verbessert, wo es nötig ist. Ligaturen wie »æ« und Akzente werden jetzt recht zuverlässig wiedergegeben, ebenso die »-que«-Abkürzung. Die zweite Seite, transkribiert mit dem HTR+Modell, das German_Kurrent_XVI-XVIII_M1 als Basismodell nützt, erreicht nur eine CER von 10.7% für 274 Wörter, die 19 Minuten Transkriptionszeit benötigen, kurz 4.2 Sekunden pro Wort, also ungefähr gleich schnell wie das PyLaia-Modell. Bei der CER muss angemerkt werden, dass bei der Korrektur einige Buchstaben, die nicht sicher entziffert werden konnten, gelöscht worden sind, so dass die CER leicht zu hoch ist. Das Basismodell hilft in diesem Fall bei PyLaia etwas mehr, wenn auch offen bleibt, ob dies an der Engine liegt oder an den Vorzügen des Basismodells, das gemäss den Beobachtungen in Kapitel 3 besser zu Gessners Schrift passt.

4.2 Die Erhöhung des Trainingssets

Diese Modelle mit extrem wenig Trainingsmaterial sind allen getesteten öffentlichen Modellen weit überlegen und bereits äusserst wirksam, sowohl für die Transkriptionsarbeit als auch, im Fall des HTR+-Modells, für die Ausgabe eines mit Keyword Spotting vernünftig durchsuchbaren Texts; selbst eine einfache Textsuche wird viele korrekte Passagen finden. Nach diesen vielversprechenden ersten Ergebnissen kann geschaut werden, wie bedeutend sich die Modelle verbessern können durch Hinzufügung von neuem Trainingsmaterial. Erhöht wird langsam in einem ersten Schritt um zwei Briefe auf fünf mit insgesamt 18 Seiten, 647 Zeilen und 4541 Wörtern im Trainingsset und zwei Seiten, 96 Zeilen und 731 Wörtern im Validierungsset.

Für ein mit HTR+ und ohne Basismodell trainiertes Modell sinkt die CER auf dem Validierungsset auf 11.98%, ist also um .57% niedriger als bei dem ersten Modell aber nicht ganz so gut wie dieses erste mit einem Basismodell. Es ist kein grosser Sprung, aber es ist ein Schritt in die richtige Richtung und zeigt, dass das Trainingsmaterial konsistent ist; auch ist das Trainingsset immer noch unter der empfohlenen Mindestgrösse, so dass kein überragendes Resultat zu erwarten gewesen ist. Wird das Modell auf dem Brief vom 26. November 1731, ein

Brief mit gut sichtbarer Tinte aber recht kleiner Schrift,[98] getestet, dann erreicht es auf der ersten Seite eine CER von 11.42%; für Seite 2 des Briefes vom 24. Januar 1732, der ähnlich ist in Tintenqualität aber etwas grösser geschrieben,[99] liegt die CER bei 11.64%, für Seite 3 bei 9.88%; benötigt werden für die Transkriptionen dabei bei 7.3, 6.9 und 7.7 Sekunden pro Wort, was etwas länger ist als zuvor. Das Modell ist tendenziell genauer als das Validierungsset suggeriert, aber nicht das hilfreichste bei der Transkription. Ein Fortschritt ist, dass das »m̄« nun auch als »mm« transkribiert wird. Insgesamt sind die Fehler, die die Modelle begehen, viel weniger störend als bei den öffentlichen Modellen. Es fällt dabei auf, wie viel häufiger spezifisch auf die Gessner-Hallersche Konversation zugeschnittene Begriffe, die ein grosses Problem für die allgemeinen Modelle gewesen sind, wie Namen oder Orte, korrekt wiedergegeben werden; beispielsweise werden »Scheuchzer«, »Muralt«, »Lavater«, »Dillenius« oder »Eltham«, erkannt; allerdings gibt es auch andere Fälle, wo das Modell weniger Glück hat wie bei »Füssli« oder »Lochmann«. Teils sind die erkannten Namen Teil des Trainingscorpus', so etwa Scheuchzers, was die Wiedererkennung für das Modell erleichtert. Die häufigsten Namen sind auch die, nach denen am ehesten gesucht wird, was es erfreulich macht, dass sie recht zuverlässig erkannt werden.

Wird zu einem so gearteten Modell wieder German_Kurrent_XVI-XVIII_M1 als Basismodell hinzugefügt, sinkt die Fehlerzahl wie schon vorher beobachtet, diesmal auf 9.73% CER auf dem Validierungsset, also über 2% besser als ohne Basismodell. Zugleich ist dies aber schlechter als bei dem Modell mit weniger Trainingsdaten und Basismodell, das auf eine CER von 9.46 gekommen ist. Dieses Resultat ist gegen die Erwartung; es scheint, dass das Basismodell nicht besonders gut zu Gessners Briefen passt, so dass die Erhöhung der Briefzahl möglicherweise zu Verwirrung geführt hat. Leichter verständlich ist dagegen, warum die CER auf dem Trainingsset schlechter ist als ohne Basismodell, nämlich 1.11% entgegen .72%. Das lässt sich damit erklären, dass das Basismodell auf dem Trainingsset dem Modell mehr alternative Lesarten vorschlägt, als wenn das Modell sich allein auf das Trainingsset spezialisieren kann. Für den Praxistest mit der zweiten Seite des Briefs vom 26. November 1731 ergibt sich eine CER von 8.66% bei einer Geschwindigkeit von 6.8 Sekunden pro Wort, auf der dritten eine CER von 7.23 und 5.4 Sekunden pro Wort und auf der vierten Seite sogar eine CER von

[98] https://hallernet.org/data/letter/02417/core.
[99] https://hallernet.org/data/letter/02419/core.

3.84, obwohl die anschliessende Korrektur etwas länger benötigt hat, nämlich 8.2 Sekunden pro Wort. Das ist alles um mindestens 2% besser, wenn auch etwas weniger zeitlich effizient, als die CER, die für die Transkription eines Briefs bei dem Modell mit weniger Trainingsmaterial erreicht worden ist. Möglicherweise ist daher die CER des Validierungsset mehr Zufall; es ist schliesslich, genau wie das Trainingscorpus, sehr klein und damit fehleranfällig.

Wird bei HTR+ noch etwas an der Anzahl Epochen geschraubt, lassen sich keine grossen Veränderungen feststellen. Wird deren Anzahl auf 100 erhöht, mit Beibehaltung des Basismodells, beträgt die CER jetzt 10.21% auf dem Validierungsset, ist also .5% schlechter als mit 50 Epochen; das einzige, was sich verbessert, ist die CER des Trainingssets. Das bedeutet nur, dass das Modell sich zu sehr auf die ganz spezifischen Merkmale des Trainingssets spezialisiert, sogenanntes »overfitting«.[100] Die Lernkurve zeigt, dass nach ungefähr der 12. Epoche die CER auf dem Validierungsset sich kaum mehr verändert. Für die drei Seiten des Briefs vom 30. September 1731,[101] dessen Tinte etwas verblasst ist, liegt die CER bei 8.79%, 7.17% und bei 8.81%, was sich nicht gross unterscheidet von den Resultaten mit der Epochengrundeinstellung. Das Einzige, was sich etwas erhöht ist die Geschwindigkeit, die bei 6.1, 5.6 und 4 Sekunden pro Wort liegt.

Das mit PyLaia trainierte Modell, trainiert mit dem oben beschriebenen erweiterten Trainingsmaterial und den gleichen Einstellungen wie in den oberen PyLaia-Modellen, erreicht eine CER von 15.2%. Das ist eine starke Verbesserung von über 4% zum Zustand mit 1000 Wörtern weniger; mit PyLaia zeigen sich zumindest in diesem Stadium mit noch wenig Trainingsmaterial schneller Verbesserungen durch die Hinzufügung von mehr Trainingsmaterial als mit HTR+; an diesem Punkt liegt das Ergebnis aber noch unter der Leistung des HTR+-Modells. Getestet auf dem Brief vom 24. März 1733[102] werden CERs von 31.55, 14.35 und 14.07 erreicht bei einer Transkriptionsgeschwindigkeit von 3.5, 5.6 und 4.2 Sekunden pro Wort. Nur in einem Einzelfall hat sich die CER im Vergleich mit dem ersten Modell nicht verbessert.

Wird sodann wieder German_Kurrent_17th-18th als Basismodell hinzugefügt, sinkt die CER auf 8.3% auf dem Validierungsset. Der Unterschied zu dem Modell mit weniger Trainingsma-

[100] Vgl. Hodel (2020), 86.
[101] https://hallernet.org/data/letter/02416/core.
[102] https://hallernet.org/data/letter/02434/core.

terial ist jetzt geringer, nur noch eine Senkung der CER um 1.1%. Es ist aber wiederum besser als das entsprechende HTR-Modell, sogar bedeutend deutlicher als die Differenz mit einem kleineren Trainingsset gewesen ist. In der Praxis heisst dies für den Brief vom 9. März 1733,[103] dass die CER der Seiten bei 8.98%, 6.88% und 8.43% bei 4.2, 3.8 und 3.5 Sekunden pro Wort liegt. In diesem Text finden sich griechische Begriffe, darunter auch solche, die im Trainingsset sind; aber das vereinzelte Vorkommen reicht nicht, um dem System nochmals ein ganz anderes Alphabet zu lehren und die Zeichen werden nicht als griechische erkannt. Die Unterschiede zu dem HTR+-Modell sind in der Anwendung minimal; PyLaia scheint die Transkription aber stärker zu beschleunigen..

Mittlerweile haben diese Transkriptionen zu noch mehr neuen Trainingsdaten geführt, insgesamt sind jetzt 26 Seiten mit 916 Zeilen und 6820 Wörtern als Training Set vorhanden. Das ist jetzt innerhalb des unteren Bereichs, der empfohlen wird für das Trainieren von Modellen. Das Validierungsset bleibt gleich wie zuvor. Nun können wieder Modelle nach dem gleichen Muster wie vorher gebaut werden. Es wird darauf verzichtet, die Leistungen ohne Basismodell zu testen, da auf diese Weise trainierte Modelle, selbst bei denjenigen mit dem meisten Trainingsmaterial, auf dem Validierungsset schlechter abschneiden und es sich auch in der Praxis gezeigt hat, dass die Basismodelle einen klaren Vorteil bieten.

Das HTR+-Modell erreicht mit den Grundeinstellungen und dem gleichen Basismodell wie zuvor eine CER auf dem Validierungsset von 9.86%; das ist ein weiterer Anstieg der CER, wie er schon zuvor beobachtet worden ist. Mit 3251 Wörtern Trainingsmaterial hat die CER noch 9.46% betragen. Wird das Modell auf den Brief vom 23. Mai 1731[104] angewandt, zeigt sich eine CER von 7.46%, wobei für jedes Wort 3.9 Sekunden benötigt werden, was im Rahmen der vorherigen Resultate der besten Modelle ist. Ein weiterer Versuch auf 100 Epochen zu erhöhen führt zu ähnlichen Resultaten wie beim ersten Versuch. Die CER liegt wieder etwas höher als mit 50 Epochen, diesmal bei 9.9% auf dem Validierungsset. Für den Brief vom 28. September 1731[105] heisst dies für Seite 1, dass die CER bei 5.82% liegt, und für Seite 2 bei 6.19% bei Transkriptionszeiten von 4.4 und 4.5 Wörter pro Sekunde. Das sind recht gute Resultate und zeigen, dass die Erhöhung des Trainingsmaterials sich doch positiv auswirkt.

[103] https://hallernet.org/data/letter/02433/core.
[104] https://hallernet.org/data/letter/02414/core.
[105] https://hallernet.org/data/letter/02415/core.

Die Anzahl Epochen ist insignifikant, so lange sie mindestens bei etwa 20 liegen. Die Modelle wirken gleich gut, egal ob sie 250 Epochen (CER auf dem Validierungsset von 9.77) oder 40 (CER auf dem Validierungsset von 9.96) durchlaufen haben.

Eine weitere Möglichkeit das Modell zu verbessern ist das Hinzufügen eines Sprachmodells, bevor die Transkription einer Briefseite gestartet wird. Verwendet wird hierfür das Sprachmodell, das während des Modelltrainings gebildet worden ist; 231 andere Wörterbücher stehen für HTR+ daneben zur Verfügung, davon auch einige für Latein, aber da sie nicht effizient mit dem eigenen Corpus verglichen werden können, wird auf sie verzichtet. Die so erstellte Transkription weist beim Brief vom 23. Mai 1731 auf der zweiten Seite eine CER 5.96 auf mit einer Transkriptionsgeschwindigkeit von 3.6 Sekunden pro Wort; mit 250 Epochen kommt die die dritte Seite auf 7.39% CER und 4.2 Sekunden pro Wort. Das Modell mit 100 Epochen erreicht für die dritte Seite des Briefes vom 28. September 1731 6.14% CER bei 4 Sekunden pro Wort. Die Unterschiede zu den Ergebnissen ohne Sprachmodell sind also minimal, auch wenn es tendenziell zu einer Verbesserung führt. Sprachmodelle sollte eigentlich recht effizient sein, bei Modellen mit einer CER zwischen 5 und 10%.[106] Die hier verwendete Messmethode ist bei so geringen Unterschieden aber fehleranfällig.

Die Leistung mit PyLaia als Engine und dem Basismodell führt zu ähnlichen Erkenntnissen wie in den früheren Versuchen. Die CER senkt sich auf 7.1% für das Validierungsset, also über 2% besser als HTR+ wie schon bei den vorherigen Modellen mit weniger Trainingsmaterial und Basismodell. Im Gegensatz zu HTR+ sinkt die CER auch im Vergleich mit den Modellen mit weniger Trainingsmaterial; mit 2000 Wörtern mehr hat die CER sich um über 1% verbessert. Angewandt auf einen Gessner-Brief[107] erreicht das Modell eine CER von 7.05% bei 3.6 Sekunden pro Wort. Das ist nicht signifikant besser als die Resultate mit HTR+. Mit Sprachmodell auf einem Brief[108] getestet wird jetzt eine CER von 4.89%, 2.16% und 6.95% erreicht bei einer Geschwindigkeit von 3.3, 3.7 und 3.1 Sekunden pro Wort. Es ist also eine leichte Verbesserung dank des Sprachmodells wahrscheinlich.

Mit einer weiteren Erhöhung auf 41 Seiten, 1306 Zeilen und 9889 Wörter und 3 Seiten im Trainingsset, 138 Zeilen und 1120 Wörter als Testset lassen sich wiederum zwei Modelle mit

[106] Vgl. Hodel b (2022), 8.
[107] https://hallernet.org/data/letter/02432/core.
[108] https://hallernet.org/data/letter/02431/core und https://hallernet.org/data/letter/02441/core.

den beiden Engines erstellen, beide unterstützt von ihrem Basismodell. HTR+ kommt diesmal auf eine CER von 7.08% auf dem Validierungsset; es wird hier endlich eine Verbesserung deutlich und die CER sinkt bei über 2% im Vergleich mit den vorherigen Modellen. Noch besser schneidet aber wiederum das PyLaia-Modell ab, das diesmal eine CER von 6.81% erreicht. Der Abstand zu der Leistung des HTR+-Modells ist allerdings kleiner geworden.

Das HTR+-Modell erreicht auf dem Brief vom 23. April 1731[109] eine CER von 5.06% mit einer Transkriptionsgeschwindigkeit von 2.9 Sekunden pro Wort. Mit Sprachmodell wird auf der nächsten Seite des Briefes eine CER von 7.41% erreicht bei 3.6 Sekunden pro Wort. Es scheint, das Sprachmodell ist in diesem Fall also nicht besonders hilfreich; teils ist dies wahrscheinlich auch den mathematischen Überlegungen und Gleichungen geschuldet, die in diesem Brief diskutiert werden und die nicht alle sauber mit dem Transkribuseditor dargestellt werden können. Das PyLaia-Modell erreicht ohne Sprachmodell auf Seite 3 bei einer Geschwindigkeit von 4 Sekunden pro Wort eine CER von 7.77%; das gleiche bezüglich mathematischen Formeln wie bei der vorherigen Seite gilt hier. Mit Sprachmodell ist mit 3.3 Sekunden pro Wort eine CER von 6.22% erreichbar. Die Resultate haben sich nicht offensichtlich verbessert im Vergleich mit dem Modell mit 3000 Wörtern weniger Trainingsdaten.

Insgesamt hat die Verwendung von Sprachmodellen keinen eindeutigen Vorteil ergeben. Einfach zu messen ist das wegen der Vielfalt im Erscheinungsbild und im Inhalt der Gessner-Briefe nicht, besonders da die Verbesserungen durch Sprachmodelle in einem Bereich von unter 1% liegen.[110] Zwischen einzelnen Textseiten sind höhere Schwankungen unter den genau gleichen Bedingungen nicht auszuschliessen; mit noch mehr Experimenten liesse sich möglicherweise ein Trend ausmachen. Ein Grund, warum Sprachmodelle keinen konkreten Vorteil gebracht haben, mag sein, dass Schwierigkeiten in der Transkription am meisten dort auftreten, wo die Briefe von der Norm abweichen, beispielsweise durch einen griechischen Begriff, und für diese Fälle sind Sprachmodelle nutzlos. So findet sich beispielsweise auf der vierten Seite des Briefs vom 23. April 1731 anstelle des Eigennamens »Gronovius« ein »honorius«; »honor« ist ein Wort, das mehrmals im Trainingsset erscheint. Möglicherweise liesse sich durch den gezielten Einbezug spezieller Passagen die Leistung in diesen Bereichen erhöhen, aber es besteht dann sogleich die Gefahr, dass die Transkription dann an anderer Stelle

[109] https://hallernet.org/data/letter/02413/core.
[110] Hodel b (2022), 8.

schlechter wird. Spezielle Tags könnten ebenfalls helfen, aber dass erhöht den Aufwand in der Textaufbereitung ohne signifikante Einsparungen an anderer Stelle.

Basismodelle haben sich dagegen als unverzichtbar erwiesen und es erlaubt bei sehr geringer Menge Trainingsmaterial effiziente Modelle zu trainieren. Während Modelle mit HTR+ ohne Basismodell bei diesen geringen Mengen Trainingsmaterial bessere Resultate erzielen als PyLaia, hat sich dieses mit Basismodell als bessere Engine für die Gessner-Briefe empfohlen, was wahrscheinlich eher an der im Vergleich mit German_Kurrent_XVI-XVIII_M1 besseren Kompatibilität von German_Kurrent_17th-18th und Gessners Schrift liegt als an der Engine. Anders lässt sich der Rückgang in der Qualität trotz Hinzufügung von Trainingsmaterial bei den HTR+-Modellen nicht erklären:

Wörter im Trainingsset	3251	4541	6820	9889
CER auf Validierungsset - PyLaia mit Basismodell	9.1	8.3	7.1	6.81
CER auf Validierungsset - HTR+ mit Basismodell	9.46	9.73	9.86	7.08

Die eigens für Gessners Briefe erstellten Modelle haben alle schon mit sehr wenig Trainingsdaten massiv bei der Transkription geholfen und sich dabei deutlich von den öffentlichen Modellen abgesetzt. Durch sie ist eine schnelle Erhöhung des zur Verfügung stehenden Trainingsmaterials möglich geworden, was zu guter Letzt zu Modellen geführt hat, die mit recht grosser Zuverlässigkeit durchsuchbar sind. Allerdings hat es keines der Modelle unter 5% geschafft, was nötig wäre um mit Hilfe von einer Fuzzy-Search ein Corpus zuverlässig durchsuchen zu können.[111] An diesem Punkt lohnt es sich, weiter Modelle zu bauen mit noch mehr Trainingsmaterial, das mit Hilfe der bis jetzt erstellten Modelle leicht zu gewinnen ist, bis die CER nicht mehr weiter sinkt.

[111] Vgl. Hodel b (2022), 14f.

5. Zusammenfassung

Die automatische Handschriftenerkennung erleichtert die wissenschaftliche Arbeit und erlaubt neue Fragestellungen, weswegen es wünschenswert ist, dass sich Gedächtniseinrichtungen mit ihr beschäftigen, um dieses neue Angebot zur Verfügung stellen zu können. Transkribus bietet dabei den leichtesten Einstieg, besonders da auch eine reiche Menge an Tutorials zur Verfügung stehen. Der Einstieg in Transkribus ist somit relativ einfach ohne viel technische Vorkenntnisse und es werden keine Spezialisten benötigt, die meist in kleinen Betrieben nicht vorhanden sind. Da bis zu 500 Seiten gratis transkribiert werden können, ist es möglich, mit der Software vertraut zu werden und sogar einen kleineren Bestand zu transkribieren ohne finanziellen Mehraufwand.

Es ist jedoch unerlässlich genau zu wissen, von welcher Art die Bestände sind und wie sie später zugänglich gemacht werden sollen. Es muss entschieden werden, wie hoch die Qualität des Textes sein sollte, ob etwa die Zeilenreihenfolge stimmen sollte. Die Schrift und die Anzahl Hände wiederum entscheiden, ob es einfacher ist, ein vorhandenes Modell zu verwenden, falls ein passendes vorhanden ist, oder ein eigenes zu erstellen. In dem Fall der Gessner-Briefe haben sich die vorhandenen Modelle als nicht besonders hilfreich erwiesen, auch wenn sie gefühlsmässig die Transkription unterstützt haben. Mittels Keyword Spotting lassen sie sich zwar einigermassen zuverlässig untersuchen, aber wenn diese Technologie verwendet werden soll, braucht es die read&search-Plattform und es ist in den wenigsten Fällen sinnvoll eine solche für einen noch kaum leserlichen Text zu betreiben.

Was sich dagegen lohnt, ist ein eigenes Modell zu erstellen, wenn die Handschriften in einem Corpus einheitlich sind oder, wie im Fall der Gessner-Briefe, nur von einer Person stammen; Nachlässe sind da ein Beispiel für Bestände, bei denen ein solches Vorgehen zu prüfen sich empfiehlt. Die durch ein solches Modell erhaltenen Transkriptionen sind über weite Strecken lesbar, auch wenn es sich nicht um wissenschaftlich edierte Texte handelt, deren Erstellung auch in der Regel nicht das Ziel eines Archivs oder Bibliothek sein sollte.[112] Diese Art Text kann für eine Volltextsuche online verfügbar gemacht werden; die Benutzer sollten aber

[112] Vgl. dazu die Rückmeldung bei Milioni (2020) 32, wo Gedächtnisinstitute zitiert werden mit der Ansicht »transcriptions are scholar's responsibility and not the curator's«. Mit der heute vorhandenen Masse an digitalisiertem Material ist ein so radikaler Standpunkt aber nur noch schwer haltbar.

darauf hingewiesen werden, dass die Volltexte nicht von einem Menschen überprüft worden sind.

Hier bietet sich auch die Möglichkeit, den Nutzenden zu erlauben, die Transkriptionen selbst zu verbessern; für solche Angebote ist allerdings eine entsprechend übersichtliche Darstellung online nötig, was für eine einzelne Institution nicht einfach aufzubauen ist. Andere mögliche Verbesserungen könnte durch NER erfolgen, falls diese in der Lage ist falsch geschriebene Namen dennoch als Entität zu erkennen; diese könnten dann einzeln korrigiert werden, was für viele Fragestellungen an das Corpus eine entscheidende Verbesserung wäre.

Wie die vielen Beispiele erfolgreicher Anwendungen und die eigenen Erfahrungen gezeigt haben, lohnt es sich auf jeden Fall auch für kleinere Institutionen zu überlegen, für welche Bestände HTR effizient eingesetzt werden kann, auch wenn viele Facetten zu berücksichtigen sind, an die rechtzeitig gedacht werden muss, von der angestrebte Qualität bis zur anschliessenden Präsentation, von der eigenen Arbeit bis zum Kontakt mit anderen Institutionen oder Projekten. Die Digitalisierung in Bibliotheken und Archiven weitet sich aus und ändert sich in Richtung von immer mehr Vernetzung. Es gilt auch vorauszuplanen, wie Transkriptionen, als Volltexte oder angereichert mit Metadaten, am besten in diesem sich rasch ändernden Umfeld zu präsentieren sind.

6. Literaturhinweise

Transkribus-Plattform: https://readcoop.eu/de/transkribus/.

hallerNet-Plattform: https://hallernet.org/.

(a) Alvermann, D., »Behandlung von Streichungen und Schwärzungen« in: Blog des Projekts *Rechtsprechung im Ostseeraum. Digitization & Handwritten Text Recognition*, (17.3.2020); https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/de/treatment-of-erasures-and-blackenings/.

(b) Alvermann, D., »HTR+ oder Pylaia« in: Blog des Projekts *Rechtsprechung im Ostseeraum. Digitization & Handwritten Text Recognition*, (18.12.2020); https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/de/htr-versus-pylaia/.

Alvermann, D., »HTR+ oder Pylaia Teil 2« in: Blog des Projekts *Rechtsprechung im Ostseeraum. Digitization & Handwritten Text Recognition,* (22.2.2021); https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/de/htr-versus-pylaia-part-2/.

Alvermann, D., Gut, P., »Transkribus im Archiv – Ein polnisch-deutsches Projekt zur Handschriftentexterkennung an historischen Dokumenten« in: *Archeion*, (2021), Bd. 122, 129-153.

Edmond, J., Lehmann, J., »Digital humanities, knowledge complexity, and the five 'aporias' of digital research« in: *Digital Scholarship in the Humanities* (2021), Bd. 36, Suppl. 2, ii95-i-i108.

Hodel, T., »IMC Leeds Paper: Sending 15th century missives through algorithms. Testing and evaluating HTR with 2200 documents« in: *Schrift im Kloster. Ein Blog zum Kloster Königsfelden, klösterlicher Verwaltung im Mittelalter und digitaler Geschichtswissenschaft* (11. Juli 2017); https://solascriptum.wordpress.com/2017/07/11/imc-leeds-paper-sending-15th-century-missives-through-algorithms-testing-and-evaluating-htr-with-2200-documents/.

Hodel, T., »Best-practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale«, in Schöch, Ch. (Hrsg.), *DHd 2020. Spielräume Digital Humanities zwischen Modellierung und Interpretation*, (Paderborn 2020), 84–87.

Hodel, T., Schoch, D., Schneider, C., Purcell, J., »General Models for Handwritten Text Recognition: Feasibility and State of the Art. German Kurrent as an Example« in: *Journal of Open Humanities Data* (2021), Bd. 17, Nr. 13, 1-10.

(a) Hodel, T., »Chapter 6: Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities« in: Jaillant, L. (Hrsg.), *Archives, Access and Artificial Intelligence. Working with Born-Digital and Digitized Archival Collections* (Bielefeld 2022), 157-177.

(b) Hodel, T., »Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft – Anzeichen einer Revolution der Geisteswissenschaften?« in: *Historische Zeitschrift* (2022) [Vorabdruck].

(c) Hodel, T., »Die Maschine und die Geschichtswissenschaft: der Einfluss von deep learning auf eine Disziplin« (2022) [Vorabdruck].

Kiessling, B., Tissot, R., Stokes, P., Stökl Ben Ezra, D., »eScriptorium: An Open Source Platform for Historical Document Analysis« in: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, (2019), 2.Bd., 19-24.

Massot, M.-L., Sforzini, A., Ventresque, V., »Transcrire les fiches de lecture de Michel Foucault avec le logiciel Transkribus: compte rendu des tests« in *HAL* (2018).

Milioni, N., *Automatic Transcription of Hisorical Dokuments. Transkribus as a Tool for Libraries, Archives and Scholars* (Uppsala 2020): http://www.diva-portal.org/smash/get/diva2:1437985/FULLTEXT01.pdf.

Muehlberger, G., et al., »Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study« in: *Journal of Documentation*, 75 (5/2019), 954-976.

Nieddu, E., Firmani, D., Merialdo, P., Maiorino, M., »In Codice Ratio: A crowd-enabled solution for low resource machine transcription of the Vatican Registers« in: *Information Processing and Management* (2021) Bd. 58, Nr. 5, 1-20.

Plüss, R., Sieber, Ch., »Digitalisierungsprojekte des Staatsarchivs Zürich mit Einsatz von Machine-Learning-Verfahren« in: *ABI Technik* (2020), Nr. 40(3), 218-228.

Terras, M., »Chapter 7: Inviting AI into the Archives: The Reception of Handwritten Technology into Historical Manuscript Transcription« in: Jaillant, L. (Hrsg.), *Archives, Access and AI: Working with Born-Digital and Digitised Archival Collections* (Berlin 2022), 179-204.

Alle Links und Zahlenangaben sind zuletzt am 30.7.2022 überprüft worden.