Fachhochschule Potsdam

Fachbereich Informationswissenschaften

Archiv (B. A.)

Bachelorarbeit

zum Thema:

Aufwand und Nutzen einer Kooperation mit Transkribus (HTR-Software) für kleine Spezialarchive am Beispiel des Archivs der Max-Planck-Gesellschaft

Vorgelegt von:

Florian Spillert

Fachsemester: 7

Abgabedatum: 29.01.2024

Erstgutachter: Dr. Michael Scholz
Zweitgutachter: Ulf Preuss

Inhalt

1. Einleitung- Bedeutung von Handwritten Text Recognition (HTR) für Archive und Chancen bei der Nutzung der Software Transkribus

Der Einsatz von neuronalen Netzen und der Bereich des Deep Learning gehören noch nicht wirklich zum Methodenapparat der Archivwissenschaft und der archivischen Hilfswissenschaften im Allgemeinen. Eine Auseinandersetzung mit der imitativen Selbstlernstrategie von Computern ist jedoch nicht nur aufgrund des erfolgreichen Einsatzes in den Geisteswissenschaften angebracht, sondern auch aufgrund der Tatsache, dass diese Technologie zunehmend in alltägliche Entscheidungsprozesse integriert wird. Von Suchmaschinen über Online-Shopping bis hin zu Bewerbungsprozessen werden Algorithmen des maschinellen Lernens und vor allem zunehmend aus dem Bereich des Deep Learning mit Trainingsdaten (z.B. vergangenen Entscheidungen) gefüttert, um Modelle zu entwickeln, die Nutzerverhalten vorhersagen und scheinbar effiziente Entscheidungen treffen.[1] Wie der Professor für Digital Humanities, Dr. Tobias Hodel, treffend ausführte, macht dieser Trend auch vor den Archiven nicht halt.

Bei der Aufsicht im Lesesaal des Archivs der Max-Planck-Gesellschaft kam der Autor der vorliegenden Arbeit über Nutzende mit dem Thema der Handwritten Text Recognition (HTR) Ende des Jahres 2022 in Verbindung. Das Thema der algorithmischen künstlichen Intelligenz hat mit der Veröffentlichung des omnipräsent in den Medien behandelten Chatbots ChatGPT seit November 2022 erheblich an Fahrt aufgenommen. Als disruptive Technologie, wie das erwähnte Large Language Model auch schon des Öfteren bezeichnet wurde, könnte man für den Dienstleistungsbereich von Archiven durchaus auch das Aufkommen der OCR- gepaart mit der HTR-Technologie (Handwritten Text Recognition) bezeichnen.[2] Die Möglichkeit alte maschinenschriftliche Texte von Unikaten in Unicode zu übersetzen und so am PC durchsuchbar zu machen, besteht zwar schon lange (OCR existiert als Prototyp-Technologie seit den 1950er Jahren)[3], allerdings wurde die Anpassung der Technologie auf handschriftlichen,

[1] Hodel, 2023, S. 174
[2] Vgl. Heigl & Jörn, 2023, S. 47.
[3] Vgl. Rufenacht, 2020: https://parashift.io/de/die-evolution-der-dokumentenerfassung/ [Zugriff am 23.01.2024].

historische Unterlagen lange Zeit als unüberwindbare Herausforderung angesehen. Dies hat sich zügig verändert, seitdem die HTR-Technologie durch die Plattform Transkribus 2015 mit Mitteln aus dem Rahmenprogramm der Europäischen Kommission als Teil des Projekts tranScriptorium (2013-2015) erhebliche Entwicklungssprünge tat.[4] Erneut von der Europäischen Kommission finanziert, diesmal im Rahmen von Horizon 2020, zielt das Nachfolgeprojekt als Europäische Genossenschaft READ-Coop darauf ab, eine funktionierende Online-Forschungsinfrastruktur zu erhalten, zu entwickeln und zu fördern, in der neue Technologien Innovationen in der Archivforschung ermöglichen.[5]

Seitdem ist die Plattform bei Historikern, Archivaren, Bibliothekaren und Big-Data-Scientists stetig beliebter geworden. Transkribus ist nicht die einzige Software in der Welt der HTR-Software. Aber vor Anbietern wie eScriptorium (seit 2018)[6], ist sie die beliebteste Plattform für die Anwendung an historischen Dokumenten. Zuletzt verzeichnete das Portal, welches viele universitäre Einrichtungen zu seinen Miteigentümern zählt, über 150.000 registrierte Nutzer.[7] Ein Hauptfeature für viele User von Transkribus, ist die Funktion, Digitalisate verknüpft mit eigenen Transkriptionstexten als Groundtruth (GT) in die Plattform einzuspeisen, um so einen Algorithmus auf eine bestimmte Hand- oder Druckschrift zu „trainieren“, damit dieser später mit dem so entstandenen „Modell“ zielgenau für die automatisierte Erkennung bei ähnliche Schriften eingesetzt werden kann.

Auch kleine Spezialarchive wissenschaftlicher Einrichtungen nutzen das Angebot oder müssen sich noch überlegen, ob und in welchem Rahmen ein Einsatz der Software lohnend ist. Der Autor ist sich der Tatsache bewusst, dass bereits einige archivische Prüfungsarbeiten das Thema der praktischen Nutzung der HTR-Software *Transkribus* zum Thema hatten. Jedoch wurde der Fokus bei Arbeiten wie der von Nikolina Milioni von der Universität Uppsala „Automatic Transcription of Historical Documents“ (2020), nicht auf den Aufwand bei der Nutzung der Software und damit einhergehende Prozesse im beteiligten Archiv gerichtet.

---

[4] Mühlberger et al., 2019, S. 957.
[5] Ebd.
[6] Anm. Siehe für Open-Source-Software „eScriptorium”: https://gitlab.com/scripta/escriptorium [Zugriff am 21.01.2024].
[7] Vgl. Transkribus, 2023.

Ziel der Bachelorarbeit ist es, eine Aufwand-Nutzen-Analyse für den Einsatz von Transkribus in kleinen Archiven zu leisten und die Einsatzmöglichkeiten bei eingeschränkten personellen Möglichkeiten auszuloten. Die Arbeit soll anhand eines exemplarischen Workflow-Beispiels aufzeigen, welche Schritte notwendig sind, um einen kleinen Teilbestand, hier die Briefe der berühmten Physikerin Lise Meitner an den Nobelpreisträger Otto Hahn, mit Hilfe der HTR-Software selbständig barrierefrei nutzbar zu machen.[8] Darüber hinaus soll gezeigt werden, ob die Nutzung bereits vorhandener HTR-Modelle ohne die Notwendigkeit, eigene Modelle zu trainieren, für die Erstellung gut lesbarer Transkripte geeignet ist.
HTR-Modelle sind Algorithmen, die auf lernfähigen rekurrenten neuronalen Netzen basieren. Sie lernen aus Regelmäßigkeiten und Wiederholungen und entwickeln somit die Fähigkeit Vorhersagen zu treffen.[9] Transkribus ist nicht die einzige Software-Plattform, die sich diese Mechanismen zunutze macht. Nachfolgend wird der Platz des verwendeten Tools im Markt der HTR-Software eingeordnet, um zu erforschen, warum Transkribus als Plattform mit am ehesten für Kultureinrichtungen wie Archive für die automatisierte Transkription infrage kommt.

1.1 Vergleich verschiedener HTR-Anbieter

Der folgende Exkurs stellt die Eigenschaften der bekanntesten HTR-Anbieter im Vergleich zu Transkribus für einen Überblick dar.

OCR4all ist eine Software die OCR- und HTR-Lösungen miteinander verbinden möchte.[10] Sie wurde von der Universität Würzburg im Zentrum für Philologie und Digitalität (ZDP) entwickelt und ist seit 2019 frei verfügbar. Der Fokus der Anwendungsmöglichkeiten lag zuletzt bei historischen Drucken der frühen Neuzeit. Es erfolgte 2021 ein Zusammenschluss mit dem OCR-D-Projekt, welches das Ziel hat, historischer Drucke aus dem 16. Und 17. Jhd. einer Volltexttransformation zu unterziehen. Die von der Universität Würzburg als anwenderfreundlich dargestellte Software hat Ihren Wandel hin zur Implementierung von HTR-Funktionen noch nicht

[8] Anm. Briefe Lise Meiterns in: Archiv der Max-Planck-Gesellschaft, III. Abt., Rep. 14.
[9] Alvermann, 2022, S. 189.
[10] Universität Würzburg, 2022.

abgeschlossen. Der Fokus liegt zum jetzigen Stand noch bei der Erkennung von Druckerzeugnissen.[11]

*tranSkriptorium i*st ein Dienstleister, der verschiedene KI-HTR-Softwarelösungen für seine Kunden anbietet, Ihnen aber selbst nicht frei zur Anwendung zur Verfügung stellt. Der spanische Anbieter verwendet keine eigene Softwarelösung, sondern bedient sich auch dem Dokumentenanalysetool „PyLaia“, welches Transkribus nutzt. [12]

Eine Stärke sieht der Anbieter in der Textwahrscheinlichkeitsberechnung und Indexierung von Entitäten zum Aufbau eines Thesaurus. Daraus geht die „Named Entity Recognition/Information Extraction“ hervor, die z. B. die Indexierung von Personennamen erleichtern soll. Für die Auslagerung größerer Projekte erscheint eine Kooperation mit Archiven unter Umständen lohnend, jedoch weniger für kleinere Projekte, über die man die volle (finanzielle und inhaltliche) Kontrolle behalten will.

Loghi ist eine noch sehr junge (2023) niederländische HTR-Software (HTR), welche bislang für Developer und nicht direkt für Anwender geeignet ist. Bislang ist die Open-Source-Anwendung nur auf niederländische Manuskripte beschränkt trainiert und bietet für Außenstehende keine Trainingsfunktion für eigene HTR-Modelle. Ein wichtiger Fortschritt für das Projekt war die Implementierung einer Website, die automatisiert transkribierte Scans durchsuchbar macht und unter dem Namen „Verleden Tekst“ bekannt ist. Die Zeichenfehlerrate - Character Error Rate (CER) weist einen niedrigen Wert auf, sodass die Durchsuchbarkeit der eingespeisten Texte weitgehend gewährleistet ist. Dies legt nahe, dass die Software bereits in der Lage ist, eine effektive und benutzerfreundliche Umgebung für die Verarbeitung und Suche von historischen Texten zu bieten.[13]

Die HTR-Software *eScriptorium* wurde erstmals 2018 eingeführt und hat seitdem mehrere Versionen durchlaufen. Sie kommt Transkribus in Bezug auf Oberfläche und Einsatzgebiet am nächsten. Die Software ermöglicht die Vorverarbeitung von Bildern sowie manuelle und automatische Layoutanalysen. Nutzer können Ground Truth (GT) sammeln und Modelle trainieren, wobei ein Webbrowser-Interface für die Editierung

[11] Universität Würzburg, 2022.
[12] Transkriptorium (o. D) – Website.
[13] Hoitink, 2023.

von Text genutzt werden kann. Entwickelt im Rahmen eines israelisch-französischen Projekts an der Université Paris, hat eScriptorium neben Schrift- und Sprachmodellen für Hebräisch und Französisch mittlerweile auch deutschsprachige Modelle implementiert. Die Software arbeitet mit dem Algorithmus "Kraken" und erfordert Docker oder Linux. Sie ist nur nach Einladung für wissenschaftliche Projekte verfügbar, wobei Open-Source-Installationen auf eigenen Servern möglich sind. Im Vergleich zu Transkribus, wie von der Universität Heidelberg in einem Vergleichsversuch an mehrsprachigen Wörterbüchern des 17. Jhd. festgestellt, überzeugt eScriptorium durch eine verbesserte und leichter trainierbare Layoutanalyse mit der Engine „Kraken“. [14]

Die lokale Installation erleichtert den Zugriff auf Dateien, eliminiert die Notwendigkeit des Uploads und ermöglicht sogar den Import von Daten aus Transkribus. Trotz dieser Vorteile weist eScriptorium jedoch einige Nachteile auf, darunter eine als umständlich und "unergonomisch" empfundene manuelle Segmentierung sowie eine technisch anspruchsvolle Ersteinrichtung und komplexe Bedienung, die Kenntnisse in Git, Python und SQL erfordert.[15]

Transkribus kann, basierend auf der im Projekt vorherrschenden Erfahrung, der eindeutigen Ausrichtung auf historische Dokumente, der Benutzerfreundlichkeit, dem Funktionsumfang und der ständigen Weiterentwicklung, als eine der führenden Plattformen im Bereich der Handwritten Text Recognition (HTR)-Software betrachtet werden.

Es ist die erste einsteigerfreundliche Software, die auf die Transkription historischer Handschriften spezialisiert ist. Transkribus (insbes. Transkribus Lite – Webanwendung) wird von Teilen der Community, dank seines selbsterklärenden Interfaces, als besonders nutzerfreundlich wahrgenommen.[16]

Sie integriert Tools verschiedener Forschungsgruppen, darunter PRHLT der Technischen Universität Valencia mit der „PyLaia“- Software und die CITlab-Gruppe der Universität Rostock mit „HTR+“. Die Software verfügt über eine zeilenbasierte, automatisierte

[14] Vgl. Website der Plattform „eScriptorium“: https://sofer.info/ [Zugriff am 24.01.2024].
[15] Vgl. Folien der Präsentation am Heidelberg Forum Digital Humanities, 09.02.2022: https://backend.uni-heidelberg.de/en/documents/ocr-technologies-in-comparison-from-manuscripts-and-old-prints-to-database-structures-and-htr-models/download (Download) [Zugriff am 20.12.2023].
[16] Milioni, 2020, S. 13.

Layoutanalyse und ermöglicht, wie bereits dargestellt, die Speicherung von Ground Truth (GT) zum Modelltraining. Auch zahlreiche öffentliche Modelle für verschiedene Schrifttyparten, sogenannte generische Modelle (z.Zt. 168 - Stand Januar 2024), können zur Texterkennung genutzt werden. Die Plattform wird per Browser als Web-Interface oder als eigenständige Java-basierte Software verwendet. 2020 wurde ein Subskriptionssystem nach Einstellung der EU-Förderung eingeführt. Jedoch bietet Transkribus zur Unterstützung der Forschungsgemeinschaft Stipendien für nicht kommerzielle Forschungsprojekte an.[17]
Zusammenfassend ist Transkribus aufgrund seiner langen Entwicklungsgeschichte, Spezialisierung auf historische Handschriften, Benutzerfreundlichkeit, vielseitigen Funktionen und Zusammenarbeit mit verschiedenen Forschungsgruppen eine führende HTR-Plattform, die für kleine oder größere Projekte in kleinen Archiven gedacht ist (und innerhalb dieser auch rege genutzt wird), weshalb sich diese Arbeit mit dessen Anwendung befasst.

Im Rahmen der Arbeit nicht ausführlich darstellbar, aber erwähnenswert, sind die beiden Firmen und Anbieter Parashift[18] und MyScript[19], deren Schwerpunkt auf der Echtzeiterkennung von modernen Handschriften liegt. Die „Read“ OCR-Engine von Microsoft[20] und die Cloud Vision API von Google[21] fokussieren sich bei der Weiterentwicklung von Texterkennungstechnologie auf Alltagsanwendungen und die Verschmelzung von Bild- und Texterkennung.

1.2 Transkribus-Workflow allgemein

Im Folgenden wird die von der READ-Coop und dem Gründervater von Transkribus, Günther Mühlberger generell empfohlene Vorgehensweise für den Workflow der Texterkennung skizziert und dessen Umsetzung bzw. Abwandlung in ein paar deutschen Archiven besprochen.[22] Dies soll den im Kapitel 3 dargestellten Versuchen

[17] Mühlberger, 2018,S. 146f.
[18] Rufenacht, 2020.
[19] MyScript, 2023.
[20] Microsoft Corporation, 2024.
[21] Google Ireland Limited, 2024.
[22] Anm. Hauptquelle für den Abschnitt ist: Mühlberger et al., 2019, S. 958-964.

am eigenen Bestand und den Überlegungen zu Arbeitsabläufen im Archiv der Max-Planck-Gesellschaft eine Grundlage geben.

Bestandswahl

Am Anfang eines jeden Projekts stehen die Bestandswahl und die Zieldefinition. Kriterien für die Bestandswahl sind:

- hohe Nachfrage bei Nutzenden zu bestimmten Beständen und Themen, die den Bestand oder den Urheber der Handschrift tangieren
- der Umfang von laufenden und geplanten Digitalisierungsprojekten, d. h. die Möglichkeit von einfachem Zugang zu Digitalisaten für die nähere Auswahl
- Bestände, deren Zustand so mangelhaft ist, dass eine Digitalisierung in absehbarer Zeit notwendig ist, um Informationsverlust vorzubeugen[23]

Upload der Digitalisate/ Bildbearbeitung

Eine Bildbearbeitung bzw. ein „Pre-Processing“ ist bereits in der Software automatisiert möglich. Nötig kann sie sein, wenn die Digitalisate zu geringe Kontrastwerte aufweisen. Hierbei können schwarz-weiß-Versionen erzeugt werde. Außerdem wird „De-Noising“ eingesetzt, um eventuelles Bildrauschen annähernd zu neutralisieren. Dieses Verfahren soll auch zur Ausblendung von Verunreinigungen und irrelevanten Punkten genutzt werden. „De-Skewing“ wird für die bestmögliche Ausrichtung des Textes verwendet. Wenn möglich sollten TIF-Dateien verwendet werden, um die höchstmögliche dpi (300 oder mehr) für den größten Detailgrad zu verwenden. [24]

Layoutanalyse und Segmentierung

Das Vorgehen nach dem Import der Dokumente im Schritt der Layoutanalyse hängt davon ab, ob bereits vorhandene Transkriptionen in die Erstellung der Groundtruth (GT) einbezogen werden können. Ohne vorhandene Transkriptionen sollte eine Segmentierung halbautomatisiert erfolgen.[25] Die Layoutanalyse dient der späteren automatisierte Texterkennung zur Orientierung über Zeilen und Wörter im Image.

[23] Vgl. Milioni 2020, S.29f.
[24] Milioni, 2020, S. 13f.
[25] Milioni , S. 19

Hierfür wird das Strukturanalyse-Tool „P2PaLA“ verwendet, welches Regionenebenen und Grundlinien in Transkribus auf der Basis von vortrainierten Modellen erkennt. Es wird eins von vielen Layoutanalyse-Modellen ausgewählt, dass die Eigenschaften der Quelle am besten berücksichtigen kann. Das Ergebnis wird einer Kontrolle unterzogen. Dabei werden im Editor händisch überschüssige Grundlinien entfernt, verschobene Linien begradigt oder zusätzliche Linien händisch ergänzt.
Der nächste Schritt wäre bei dieser Anwendungsmöglichkeit ohne bereits vorhandene GT eine manuelle Transkription innerhalb der Transkribus-Oberfläche. Bei bereits vorhandenen Daten kann das „Text2Image“ verwendet werden, um die Layoutanalyse durchzuführen. Hierbei werden die vorhandenen Zeichen, die in einem txt-Dokument gespeichert werden müssen, Zeile für Zeile mit dem Image der Seite abgeglichen.[26] Zeilenumbrüche sollten bei dem txt-Dokument allerdings eingefügt werden, um die Korrektur von nicht übereinstimmenden Zeilen im Nachgang möglichst wenig aufwendig zu gestalten.[27]

Erstellung von einem Training-Data-Set

Es sollten nach den Empfehlungen der Transkribus-Schöpfer mindestens 25-75 Seiten GT eingepflegt und in einer Collection (Explorer-Ordnersystem in Transkribus) gespeichert werden, bevor ein Modelltraining gestartet werden kann und der Algorithmus genug Daten für das Training (zwischen 5.000 und 15.000 Wörtern) hat.[28]

[26] Vgl. Muehlberger et. al, 2019, S. 960.
[27] Milioni, 2020, S. 19.
[28] Vgl. Mühlberger et. Al, 2019, S. 959.

Modelltraining

Für das Modelltraining ist zur Feststellung der Genauigkeit der Fähigkeiten des zu trainierenden Modells ein Trainingsset und ein Validierungsset, welches nicht beim eigentlichen Training des HTR-Modells berücksichtigt wird, von Nöten. Für die Validierung wird 10% aller GT empfohlen, um möglichst aussagekräftige Ergebnisse zu erzielen. Das Validierungsset kann individuell zusammengestellt werden, um eine große Varianz innerhalb der GT zu testen. Meistens sind mehrere Trainingsdurchläufe und die Erstellung von Untermodellen notwendig, um die Fehlerrate pro Zeichen - Character Error Rate (CER) einschätzen zu können und die Verbesserung eines Modells zu beurteilen. Basismodelle, die mit ähnlichen Schreibstilen trainiert wurden, werden verwendet, um die Trainingsdaten mit bereits vorhandenen Ergebnissen (innerhalb ihrer Muster-Erkennung) anzureichern. Dabei hängt die Sinnhaftigkeit deren Verwendung vom Grad der Ähnlichkeit des Schriftstils des eigenen Bestandes ab.
Die Anzahl der „Epochs“ (Anzahl der Trainingsdurchläufe) wird je nach Material eingestellt, um die bestmögliche CER und Word Error Rate (WER) zu erreichen.[29]
Mehr Epochs führen dabei nicht zwangsläufig zu einer höheren Genauigkeit, da ggf. ungewünschte Eigenschaften zu stark trainiert werden und später Ergebnisse verfälschen.[30]

Anwendung an Digitalisaten des gesamten Bestandes

Nach erfolgreicher Anwendung der Modelle bzw. des zuverlässigsten Modells auf eine ausreichende Anzahl von Testdaten kann das Modell auf einen vollständigen Bestand angewendet werden.

Qualitätskontrolle

Im Vorfeld einer Veröffentlichung muss eine Korrektur der HTR-generierten Texte erfolgen, wenn eine hundertprozentige Genauigkeit gewünscht ist.

Keyword-Spotting

Wenn dieser Grad an Genauigkeit nicht angestrebt wird oder aus Aufwandsgründen nicht erreicht werden kann, steht dem Anwender oder später den Nutzenden die Funktion des „Keyword-Spotting“ zur Verfügung. Dieses Verfahren ermöglicht es, bei

[29] Vgl. Mühlberger, 2018, S. 151f.
[30] Vgl. Milioni, 2020, S. 21.

einer Recherche im Text, trotz fehlerhafter Transkription, viele der falsch erkannten Wörter trotzdem zu finden. Anstatt auf die genaue Erkennung des gesamten Textes zu setzen, nutzt diese Methode „confidence matrices“, die während der Texterkennung mit neuronalen Netzen erstellt werden. Auch bei Modellen mit niedriger Erkennungsrate können korrekte Zeichen identifiziert, jedoch mit geringerer Sicherheit versehen werden. Durch den Einsatz von Variantentabellen können mögliche Variablen einer gesuchten Zeichenfolge effektiv gefunden werden. Dies ermöglicht eine erfolgreiche Suche nach bestimmten Wörtern oder Begriffen, die das Modell nicht als am wahrscheinlichsten berechnet hat. Vorreiterprojekte setzen diese Methode bereits ein, indem sie sie durch leistungsfähige Datenbanksysteme indexieren.[31]

1.3 Beispiele für Workflows und Arbeitsergebnisse in Archiven

Die obigen Empfehlungen zur Nutzung der Software wurden in den Arbeitsabläufe der anwendenden Archive zum Teil abgewandelt umgesetzt. Im Folgenden wird der Workflow einiger Archive exemplarisch wiedergegeben, um damit auch die Einsatzvielfalt und die damit verbundenen Erfolge zu beleuchten.

Zusammen mit dem **Archiv der Hansestadt Wismar** und dem **Landesarchiv Mecklenburg-Vorpommern in Greifswald** erprobte Dirk Alvermann († 17.10.2023), Leiter des **Universitätsarchivs Greifswald**, im Rahmen eines DFG geförderten Digitalisierungsprojekts zwischen 2019 und 2021 die Möglichkeiten zum Einsatz der HTR für große gleichförmige Bestände. In dem Kooperationsprojekt wurden Spruchakten der Juristischen Fakultät, Urteilsbegründungen des Wismarer Ratsgerichts und Urteilsbegründungen der Assessoren am Wismarer Tribunal (ca. 250.000 insgesamt Seiten) ausgewählt.[32]

Die Akten wurden zunächst durch einen Stadtarchivar (Wismar) geordnet. Die Paginierung und Digitalisierung führten extra hierfür angestellte Projektmitarbeitenden durch. Die Bearbeitung der Digitalisate erfolgte nach dem Goobi-Worflow (Intranda).[33] Für das Projekt wurden drei verschiedenen Plattformen parallel verwendet (Ariadne,

[31] Hodel, 2023, S. 164-166.
[32] Vgl. Heigl, Jörn, 2023, S. 52.
[33] Intranda GmbH, 2020.

Goobi-Workflow und Transkribus).[34]

Die fertigen Scans wurden mit Metadaten und Strukturdaten angereichert. Anschließend erfolgte der Upload auf die Transkribus-Server (aktenweise) durch eine Projektkraft in Greifswald. Die eigentliche Bearbeitung in Transkribus erfolgt wiederum in Wismar, wo zuerst die Layoutanalyse händisch durch Studentische Hilfskräfte durchgeführt wurde.

Die eigentliche automatische Texterkennung erfolgte zunächst bei nur zehn Seiten mit einem generalisierten, öffentlichen Modell[35]. Anschließend wurden die Fehler der HTR manuell korrigiert.[36] Auf Grundlage der so erzeugten GT wurden spezialisierte Modelle trainiert. Nach dem Erreichen einer CER von 7% bei Nachfolgemodellen (mit einer GT von insgesamt 260.000 Wörter) wurde die Schrifterkennung auf die restlichen Texte angewendet. Das Ergebnis wurde durch Streichungen im Text und Papierschäden bei ansonsten besserer abschneidender CER getrübt.[37]

Im Greifswalder Universitätsarchiv wurden die Spruchakten der Juristischen Fakultät mit der HTR-Software bearbeitet. Das Vorhaben wurde durch eine Projektkoordinatorin überwacht, die u.a. beim Datentransfer zwischen Goobi und Transkribus den Überblick behalten musste. Hier ordnete der Archivar mit einer Hilfskraft den Bestand zunächst. Das Einspielen der Erschließungsdaten von Ariadne in die Digitalisierungssoftware Goobi wurden durch einen FaMI durchgeführt. Nach dem Einspielen der Digitalisate in Transkribus führten Studentische Hilfskräfte die Layout-Analyse durch und korrigierten die automatisiert gesetzten Baselines und Textregionen. Allerdings konnte dieser Prozess für die Spruchakten des 17. Jahrhunderts erst nach Einsatz eines speziellen Transkribus-Tools abgeschlossen werden (Page-to-Page-Layout-Analysis – P2PaLA), welches z.B. für die Segmentierung von Marginalien, Kopf- und Fußnoten oder Konzeptabschnitten sorgte.[38]

In Greifswald erstellten zwei wissenschaftliche Kräfte die GT, indem sie zunächst händisch transkribierten, da kein Grundmodell für frühneuzeitliche Kurrentschrift vorhanden war. Sie transkribierten pro Stunde im Durchschnitt zwei Seiten.

[34] Vgl. Heigl, Jörn, 2023, S. 57.
[35] READ-COOP SCE, 2023, Öffentliche AI-Modelle in Transkribus.
[36] Vgl. Heigl, Jörn, 2023, S. 52.
[37] Vgl. Heigl, Jörn, 2023, S. 53.
[38] Vgl. Heigl, Jörn, 2023, S. 58.

Nach der Anwendung des ersten auf dem eigenen Material basierenden Modells, konnte viel schneller GT hergestellt werden, da die Texte nur noch korrigiert werden mussten. Ab dann konnten pro Stunde ca. 8 Seiten GT hergestellt werden [39] Letztendlich entstanden im Ergebnis Spezialmodelle zu jeweiligen zeitlichen Abschnitten, ohne dass ein zuverlässiges Modell für alle drei Jahrhunderte kreiert werden konnte. Die Kontrolle der HTR-Modelle erfolgte mit Testsets für regelmäßige statistische Auswertungen. Dafür wurden 3-12 exemplarischen Seiten pro Trainingsband ausgewählt, die nicht im Trainings- oder Validierungsset einbezogen wurden, um die Belastbarkeit des CER-Ergebnisses zu kontrollieren. Dieser hohe Kontrollaufwand lohnt sich den Projektbeteiligten Elisabeth Heigl und Nils Jörn zufolge nur bei heterogenen Massenbeständen.[40]

Als Hauptarbeitsergebnis für beide Projekte kann die Übernahme der Digitalisate, mitsamt der Transkriptionen in die Digitale Bibliothek Mecklenburg-Vorpommern und die weitere Verknüpfung in die Deutsche Digitale Bibliothek (DDB) angesehen werden.

Über Verzeichnungssoftware „Ariadne“ wurde ein Permalink zu den Digitalisaten gesetzt, sodass man dort direkt vom elektronischen Findbuch zum Digitalisat springen kann.[41] Das Projekte sorgte für neue Impulse für die Öffentlichkeitsarbeit. Unter anderem gab es eine Nachnutzung der digitalisierten Daten durch das Geschichtsportal des Archivs der Hansestadt Wismar „zeitreise-wismar.de“. Auch wurden online nun regelmäßig Archivalien des Monats präsentiert.[42]

Alle drei Zentralbestände sind in einer von Transkribus gehosteten Read-and-Search-Seite unter dem Namen „Rechtsprechung im Ostseeraum (1580-1871)“ gleichzeitig durchsuchbar.[43] Diese kann mit „Smart Search“ genutzt werden, welche die automatisch generierten Volltexte nach dem Prinzip des Keyword-Spotting durchsucht.[44]

[39] Vgl. Heigl, Jörn, 2023, S. 58-59.
[40] Heigl, Jörn, 2023, S. 59.
[41] Anm: Bsp. für Datenbankeintrag in Ariadne (Universitätsarchiv Greifswald /2.5./I Juristische Fakultät) St. 467-1): https://ariadne-portal.uni-greifswald.de/?arc=1&type=obj&id=5442083 [Zugriff am 24.01.2024].
[42] Hansestadt Wismar, 2024, Zeitreise Wismar.
[43] Universität Greifswald, o. D., Rechtsprechung im Ostseeraum.
[44] Heigl, Jörn, 2023, S. 61.

Das **Hochschularchiv der ETH Zürich** verwendet Transkribus für die Aufbereitung verschiedener Bestände aus dem Verwaltungsarchiv und den Privatarchivbeständen (Nachlässen). Das Hauptaugenmerk in ihrem Projekt lag auf den Schulratsprotokollen der Polytechnik (später in ETH umbenannt). Bei den von 1854 bis 1902 reichenden handschriftlichen Dokumenten handelt es sich um eine der wichtigsten Quellen der Hochschule. Daneben gab es das Ziel automatisierte Transkriptionen für die zahlreichen Tagebücher von Joseph Wolfgang von Deschwanden (erster Direktor der ETH) und Arnold Heim (Geologe) zu generieren. Dies wurde durch die Erstellung eigener HTR-Modelle mittels weniger händisch transkribierter Tagebücher ermöglicht. Die Ergebnisse waren für die Projektverantwortlichen, wenn auch fehlerbehaftet, mit einer CER zwischen 3.8 und 4,7% ausreichend für eine für die nächsten Jahre geplante Veröffentlichung über das Portal „e-manuscripta“. [45] Der Aufwand für die Erstellung eines Modells lohnt sich nach Einschätzung der im Hochschularchiv Beschäftigten proportional zur Menge des von der Person verfassten Materials, welches in der Kultureinrichtung vorhanden ist. Die Aufgabe der Qualitätskontrolle bzw. Kritik an der unter HTR-Unterstützung entstandenen Editionen liegt lt. Johannes Wahl und Nadine Fischer von der ETH Zürich bei den Nutzenden und nicht beim Archiv selbst.[46]

Der **Archivverbund Bautzen** setzte zwischen 2021 und 2023 Transkribus ein, um 300 Bücher mit Ratsprotokollen aus dem 17. bis 19. Jahrhundert zu transkribieren und online zugänglich zu machen. Mit etwa 55.000 Seiten dokumentieren diese Bücher die Sitzungen des Rats aus wohlhabenden Patrizierfamilien, aus denen unter anderem die Bürgermeister gewählt wurden. Das Projekt wurde von der Initiative „WissensWandel“ der Beauftragten der Bundesregierung für Kultur und Medien unterstützt.
Die digitalisierten Dokumente wurden an Transkribus übergeben und nicht selbst bearbeitet. Der Bestand wurde mithilfe des generalisierten Modells „Transkribus Early Kurrent“ transkribiert. Durch Training eines spezialisierten Modells, welches auf „Transkribus Early Kurrent“ basiert, verbesserten sich die Ergebnisse erheblich.
Die gesamte Dokumentensammlung wurde 2023 überarbeitet und die präziseren und korrigierten Ergebnisse sind nun ebenfalls über „read and search“ zugänglich.

[45] Wahl, Fischer 2022, [Zugriff am 24.01.2024].
[46] Ebd.

Transkribus ermöglicht die Einrichtung von Permalinks mit Verknüpfungen zu Archivportal-D und Findbuch.net, was die Kombination von Findmittelrecherche und direkter Nutzung ermöglicht. [47]
Die Nutzerzahlen für die „read and search“-Veröffentlichung erscheint mit 20 Nutzenden im Monat etwas gering.[48]Als Öffentlichkeitswirksam kann die Kooperation bezeichnet werden, da die Regionalpresse berichtete und Berichte per Twitter (jetzt „X“) auf größere Beachtung stießen. [49]

Günther Mühlberger gab 2019 einen umfangreichen Überblick über weitere damals laufende Projekte.[50]

1.4 Aufwand und Nutzen als Entscheidungsfaktoren

Die Entscheidung über den Einsatz der Software Transkribus ergibt sich aus der Abwägung des damit verbundenen Aufwands und des Nutzens, den ihr Einsatz für Einrichtung und Nutzende hat.
Ein entscheidender Faktor, um den Aufwand und die eigenständige Machbarkeit eines HTR-Projekts für ein Archiv zu bemessen, ist die zur Verfügung stehende Anzahl an Mitarbeitenden im Archiv und ihre Qualifikation. In diesem Kontext sind paläographische Fähigkeiten, eine allgemeine Affinität im Umgang mit neuer (Datenbank-)Software und auch Erfahrung bei der Erstellung von Editionen von Vorteil.
Ein weiterer Faktor ist die Eignung der eigenen Bestände für eine Anwendung mit HTR. Stellt eine Handschrift die Software vor große Herausforderungen oder ist die Heterogenität so groß, dass ein Training auf ein Bestandsegment kompliziert ist, führt das zu einer längeren Bearbeitungszeit oder halbgaren Ergebnissen.
Der Aufwand ergibt sich auch über den Faktor Zeit. Dieser hängt von der Geschwindigkeit beim Erlernen von Transkribus und der Anwendung der Software auf einen Bestand ab. Genauso spielt die Performance der Plattform selbst eine Rolle.

[47] READ-COOP SCE, o.D., 200 Jahre Bautzener Stadtgeschichte.
[48] Vgl. Richter-Laugwitz, 2023, S.5.
[49] READ-COOP SCE, o.D., 200 Jahre Bautzener Stadtgeschichte.
[50] Mühlberger et al., 2019, 964-966.

Ein ebenfalls zu berücksichtigender Faktor ist bei Transkribus der finanzielle, der bei der Auswahl und dem Erwerb von Lizenzen mehrere Optionen zulässt.

Der Nutzen von Transkribus liegt, so kann aus den Beispielen der angeführten Pionierarchiven geschlussfolgert werden, in aller erster Linie in der Durchsuchbarkeit und Lesbarkeit der Quellen und damit besseren Recherchemöglichkeiten, was wiederum barrierearme Nutzung bedeuten kann. Daraus ergibt sich mehr Aufmerksamkeit durch Nutzende, die das Angebot der Einrichtung des Archivs als Ganzes auch stärker wahrnehmen könnten. Der Nutzen ergibt sich auch über die inhaltliche und formelle Eignung der gewählten Bestände und ihrer Strahlkraft. Selbstverständliche Vorteile ergeben sich aus einer erheblichen Verringerung der Arbeitszeit an Transkriptionen und damit der Einrichtung von digitalen Editionen, wenn dies ein Ziel der Einrichtung ist. Zum Nutzungsgrad der bislang durch Transkribus in den Projektarchiven entstandenen digitalen Angebote gibt es z.Zt. noch keine belastbaren Zahlen.

Wirtschaftlichkeit ist aus dem Abwägen der beiden Entscheidungsfaktoren dann gegeben, wenn der Nutzen gegenüber dem Aufwand überwiegt und die zeitlichen und finanziellen Investitionen beim Handling von Transkribus die Vorteile nicht ausgleichen oder gar übertreffen.

Im Folgenden wird für das Archiv der MPG untersucht, wie die Rahmenbedingungen für Aufwand und Nutzen zu beurteilen sind.

2. Rahmenbedingungen für das Training und die Anwendung von HTR-Modellen im Archiv der Max-Planck-Gesellschaft

Die Rahmenbedingungen umfassen die personelle Aufstellung des Archivs, die Struktur und Umfang der Bestände, die Digitalisierungsinfrastruktur und ebenfalls den Stand der Digitalisierung.

Die folgenden Untersuchungen fanden von Oktober 2023 bis zum Januar 2024 statt und stützten sich insbesondere auf die im Literaturverzeichnis aufgeführten Werke und Websites und eigenen Erfahrungen bei der Nutzung der Web- und Expert-Version von Transkribus. Für Letzteres wurde dem Autor ein Account mit „Scholarship-Zugang“, welcher mit 4.000 Credits aufgeladen war, zur Verfügung gestellt.

Für die Einschätzung der Zuverlässigkeit der verwendeten HTR-Modelle wurde die Character Error Rate (CER) in Stichproben für einzelne Seiten errechnet bzw. angegeben. Um den Aufwand des Projekts annähernd zu bestimmen, wurden verschiedene Workflowmodelle gegeneinander abgewogen und die Arbeitszeit für die einzelnen Schritte aufgezeichnet.

2.1 Personelle Anforderungen und vorhandene Kompetenzen

Eine gewisse Technikaffinität und Geübtheit mit Java-Anwendungen bei der Benutzung von Transkribus ist von Vorteil. Das Programm (hier: Expert Client) ist, optisch wenig übersichtlich und ein Überblick über die zahlreichen Werkzeuge kann nicht durch jeden gleich schnell erlangt werden.[51] Mitarbeitende mit guten Text- und Sprachkenntnissen in Hinsicht auf das Material könnten zur Verbesserung von automatisiert generierten Transkripten beitragen und so wiederum später die Trainingsdaten verbessern.[52]
Wie zuvor durch Beispiele anderer Archive festgestellt, gehören nicht ausschließlich IT-Beauftragte und Paläographen (des Archivs) zu den Beteiligten eines HTR-Projekts.

[51] Graf, 2020.
[52] Vgl.: Milioni, 2020, S. 39.

Die Revision eines Bestandes, die Koordination von Digitalisierung und Transkription, die Editierung des Layouts in der Software und die Qualitätsüberprüfung sind weitere notwendige Aufgaben, welche weniger Anforderungen an die ausführenden richten.

Das Archiv der Max-Planck-Gesellschaft (MPG) verfügt über folgender Personalstärke:[53]

-11 Mitarbeitenden insgesamt:

-2x Archivare BA

-3x Archivare MA (davon 1x Archivleitung, 1x Aufbau Digitale Langzeitarchivierung)

- 1x Bibliothekarin (Teilzeit)

-2 FaMIs

-1x Projektkraft MA-Geschichte

-1x Informatikerin -Zuständig für Digitale Langzeitarchivierung, Archivsoftware (Teilzeit)

-1x Bürokauffrau

- 1x Technischer Bearbeiter

Zweifelsohne ist die Personalstärke in Universitätsarchiven und Archiven anderer Wissenschaftsfördergesellschaften, gar Spezialarchive anderer Nischen, oft kleiner. Auch deshalb wird in diesem Kapitel das Handling eines Transkriptionsprojekts für eine kleinere Personenzahl beschrieben.[54]
Die wichtigsten Akteure und deren hier relevante Kompetenzen werden im Folgenden beschrieben.

Die Archivarinnen und Archivare (B.A./M.A.) des Archivs der MPG sind jeweils für eigene Bestandsegmente zuständig und haben mindestens zehn Jahre Erschließungs- und Bewertungserfahrung sowie zumindest teilweise langjährige Expertise im Lesen von Kurrenthandschriften. Dies bezieht sich hier auf die Teile der Bestände der Kaiser-Wilhelm-Gesellschaft, der Vorgängerorganisation der MPG (bis 1948) und ihrer wissenschaftlichen Mitglieder.

[53] Anm. Aktuelle Aufführung des Personals im Archiv der MPG: https://www.archiv-berlin.mpg.de/ansprechpartner [Zugriff am: 20.01.2024].
[54] Vgl. z.B. Aktuelle Personalstärke des Universitätsarchivs Greifswald (4 Personen) - https://www.uni-greifswald.de/universitaet/einrichtungen/archiv/ueber-uns/mitarbeitende/ [Zugriff am 23.01.2024], Personalstärke des Universitätsarchivs der Freien Universität Berlin (10 Personen) - https://www.fu-berlin.de/sites/uniarchiv/mitarbeiter_innen/index.html [Zugriff am 23.01.2024].

Einer der Fachangestellten für Medien- und Informationsdienste (FaMI), der Autor dieser Arbeit, ist, abgesehen von der oberflächlichen Beschäftigung mit digitalen Editionen im Rahmen der Fernweiterbildung an der FH Potsdam, nicht näher in Kontakt mit Editionen oder Transkriptionen gekommen. Weitere Erfahrungen wurden erst beim Verfassen der Bachelorarbeit und der Beschäftigung mit Transkribus gemacht. Des Weiteren sind rudimentäre Kenntnisse in der Paläografie aber kaum Kenntnisse im Fachvokabular der Physik, also keine genaue Abschätzung der Plausibilität zum Vorkommen bestimmter Fachbegriffe in den zu bearbeitenden Archivalien möglich. Es besteht nur wenig Vorwissen in Bezug auf OCR-Verfahren.

Die beiden zuständigen Mitarbeiterinnen für digitale Langzeitarchivierung (DLZA) arbeiten mit diversen archivtypischen Dateiformaten (Mets, TEI, ALTO, XML) in Verbindung mit der im Archiv eingesetzten Archivsoftware und haben so ein fundiertes Verständnis für den Umgang mit Austauschformaten, die bei der Arbeit mit Transkribus verwendet werden. Die Mitarbeiterinnen sind im Austausch mit den Digitalisierungsdienstleistern des Archivs und verwalten die Ablage aller Digitalisate auf dem Archivserver. Davon abgesehen ist ihr Wissen bzgl. des Backend von Datenbanken im Allgemeinen nützlich für die Mitwirkung an einem HTR-Projekt.

Die Archivleiterin ist im Anstoß neuer Projekte geübt und hat Erfahrung darin, in koordinierender Position Verantwortung zu übernehmen.
Ähnliche Zusammensetzungen von Kompetenzen finden sich in vielen kleineren Archiven.

2.2 Auswahlmöglichkeiten in den Beständen und Stand der Digitalisierung

Das 1975 gegründete Archiv der Max-Planck-Gesellschaft legt seinen Schwerpunkt der Bestandsüberlieferung auf die Vor- und Nachlässe von hervorragenden Persönlichkeiten aus der Wissenschaft und damit insbesondere auf die Wissenschaftlichen Mitgliedern der Kaiser-Wilhelm- und Max-Planck-Gesellschaft (MPG). Nicht wenige aus diesem Kreis sind Nobelpreisträger. Der zweitgrößte Anteil in der Überlieferung liegt bei den Institutsbeständen und Unterlagen der

Generalverwaltung der jeweiligen Gesellschaft. In Summe kommt das Archiv auf 4.000 lfm. Meter Archivgut.[55]

Digitalisiert wurden durch verschiedene Dienstleister bis Ende des Jahres 2023 - 70.227 Verzeichnungseinheiten mit 9.043.461 TIF (Images) und damit umgerechnet rund 1.130 lfm. Damit ist etwas mehr als ein Viertel des Gesamtbestandes des Archivs digitalisiert und für die Nutzenden im archivrechtlichen Rahmen leicht zugänglich.[56] Hierbei handelt es sich um die wirklichen Kernbestände des Archivs, wozu die Institutsbetreuung (II. Abt., Rep. 66), diverse Gremien der MPG (II. Abt.), der gesamte Bestand der Kaiser-Wilhelm-Gesellschaft (KWG – I. Abt.) und zahlreiche der meistgefragten Nachlässe zählen. Zuzüglich zum digitalisierten Schriftgut liegen viele Foto- und Audio-Digitalisate vor.[57]
Bei einem Großteil des Schriftguts handelt es sich um maschinenschriftliche Dokumente, da die meisten Unterlagen nach der Gründung der KWG im Jahr 1911 entstanden sind und, wie in den meisten Archiven, hauptsächlich Verwaltungshandeln dokumentieren. Das Archiv nimmt bzgl. des prozentualen Anteils der Digitalisate am Gesamtbestand im positiven Sinne eine Sonderstellung bei den Wissenschaftsarchiven ein.
Anhand der Fachliteratur und durch die exemplarisch beschriebenen Projekte lässt sich ableiten, dass Massen von Unterlagen mit gleichförmigem Schriftbild, welche doch schwer zu lesen sind, am ehesten für erfolgreiche HTR-Projekte geeignet sind. Günstig ist es, wenn bereits digitalisierte Bestände mit häufig genutzten Unterlagen von hohem historischem Wert für solche Vorhaben nominiert werden.

Im Archiv der MPG kommen aus formattechnischen Überlegungen, zunächst unabhängig vom Inhalt die Unterlagenarten Brief, Postkarte, Notizbuch, Laborbuch, Register, Tagebuch und Taschenkalender infrage, um die Transkribus-Layoutanalyse vor lösbare Herausforderungen zu stellen.

[55] Starkloff, 2024, Starseite Archiv der Max-Planck-Gesellschaft.
[56] Die Zahlen zur Menge an Digitalisaten stammen vom internen Laufwerk des Archivs der MPG.
[57] Starkloff, 2023, Nutzung von digitalisiertem Archivgut.

Die Überlieferung lässt keine Anwendung auf Massenakten zu. Stattdessen sollte ein punktueller Einsatz in verschiedenen Nachlässen und Unterlagen der Generalverwaltung der KWG in Erwägung gezogen werden.

Folgende Bestände kamen nach eigener Untersuchung nach diesen Gesichtspunkten in der Überlieferung infrage:

- Nachlass des Biochemikers und Nobelpreisträgers Otto Warburg (hier: Laborbücher zur Photosynthese, biografische Aufzeichnungen) – III. Abt., Rep.1
- Nachlass des Kernphysikers Otto Hahn (hier: Taschenkalender oder Korrespondenz mit Lise Meitner (Edition vorhanden und Abschriften am Bestand) – III. Abt., Rep. 14
- Sammlung zum Chemiker Fritz Haber – (hier: Briefe mit Bezug zur Ammoniaksythese) – Va. Abt., Rep. 5
- Nachlass Carl-Friedrich von Weizsäcker (hier: Briefe von Verwandtschaft in Politik und Forschung) – III. Abt., Rep. 111
- Nachlass des Kunsthistorikers Ernst Steinmann (Tagebücher) – III. Abt. Rep. 63, Nr. 31-56[58]

Der Nachlass von Otto Hahn liegt als einziger vollständig digitalisiert vor. Für diesen Bestand und den Nachlass Ernst Steinmann existieren Editionen und damit Erleichterungen zum Einspielen von Groundtruth (GT) in Transkribus.

2.3 Wahl der Briefe Lise Meitners für Transkribus

Folgende inhaltliche und praktische Gründe führten zur Auswahl der Briefe der österreichischen Kernphysikerin Lise Meitner an ihren Kollegen und Nobelpreisträger Otto Hahn aus dem Kaiser Wilhelm-Institut für Physik mit der Laufzeit aus den Jahren 1912-1966 (AMPG, III. Abt., Rep.14, Nr. 4869-4959, Nr. 6890-6898) für das Projekt. [*siehe Findbuchauszug Nachlass Otto Hahn in ActaPro – Anlage 1*]

Die Briefe Lise Meitners bieten der Wissenschaftsgeschichte einen beispiellosen Grad

[58] Anm. Bestandssignaturen nachgewiesen unter der Archivdatenbank ActaPro: https://ursamajor.archiv-berlin.mpg.de/actaproweb/archive.xhtml [Zugriff am 25.01.2024].

an Transparenz bzgl. der Entwicklungen im Gebiet der praktischen Physik zur Kernspaltung für den genannten Zeitraum. Lise Meitner leistete 1938 einen wichtigen Beitrag zur Deutung der im selben Jahr durch den Kollegen Otto Hahn durchgeführten erstdokumentierten Kernspaltung. Zugleich sind die Briefe von hohem sozialgeschichtlichem Wert für die Erforschung der Lebensumstände einer bedeutenden Wissenschaftlerin zu Beginn des 20. Jahrhunderts. Die anhaltend hohe Zahl von Anfragen zu diesem Bestandssegment mit dem Ziel biografischer Forschungen lässt auf ein anhaltendes öffentliches Interesse schließen. Eine digitale Edition würde daher eine gewisse Aufmerksamkeit auf sich ziehen.

Für die Jahre 1912 bis 1924 veröffentlichte Sabine Ernst bereits Transkriptionen, die damals in Kooperation mit dem Archiv entstanden.[59] In Anbetracht der zeitlichen Begrenzung für die Bearbeitung dieser Bachelorarbeit lag es nahe, die 1992 erschienene Briefedition mit Transkriptionen der Korrespondenz Meitners als Grundlage für ein HTR-Modell zu verwenden. Die Einspeisung von OCR-Daten aus der Edition, ohne die Notwendigkeit Transkriptionen aufwändig selbst erstellen zu müssen, war mitausschlaggebend für die Entscheidung.

Allerdings muss klargestellt werden, dass die Nachnutzungsmöglichkeiten der Briefe zum Stand der Fertigstellung dieser Arbeit rein hypothetisch bleiben. Da im Nachlassbestand bereits unveröffentlichte Transkriptionen für die gesamte Laufzeit vorhanden sind, die lediglich technischer Aufbereitung bedürften, handelt es sich hier um einen Versuchsaufbau, der tatsächliche, günstigere Umstände ignoriert, um einen klassischen Workflow durchzuspielen, bei dem Transkriptionen nicht in dem hohen Maße vorhanden sind.

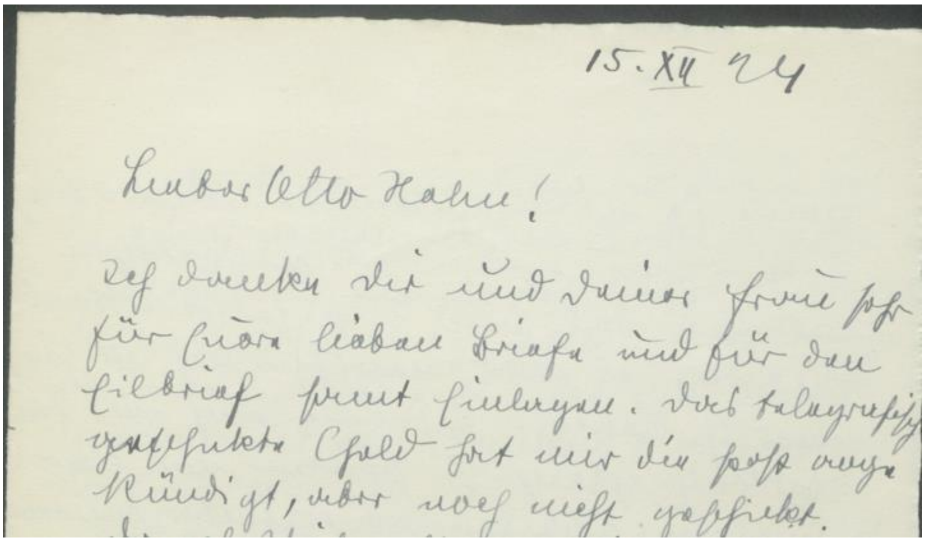
Des Weiteren ist eine Veröffentlichung aufgrund des noch gültigen Urheberrechts, also 70 Jahre nach dem Tod (Lise Meitner † 1968) bislang nicht ohne weiteres möglich, zumal sich weitere urheberrechtliche Problematiken aus der Erwerbungsgeschichte ergeben, die hier aus Platzgründen nicht näher ausgeführt werden können.[60]

[59] Vgl. Ernst, 1992.
[60] Anm. Der Enkel Otto Hahns, (Dietrich Hahn) beanspruchte Teile des Copyrights für sich. Für detailliertere Informationen zum Erwerb der Briefe Lise Meitners durch das Archiv der MPG, siehe: Ernst, 1992, S. II (Geleitwort).

Es sei erwähnt, dass hochgeladene Dateien bei Transkribus nicht automatisch mit der Community geteilt werden, sondern privat bleiben. Noch urheberrechtlich geschütztes Material kann so auch für Testzwecke genutzt werden.[61]

Die 1214 Seiten der Briefe von Lise Meitner in 99 Archivsignaturen ergeben zusammen ca. 0,3 lfd Meter Schriftgut.[62] Es handelt sich größtenteils um Kurrentschrift mit einzelnen Ausläufern des Sütterlin-Stils.[63]

Für die Jahre 1912 bis 1937 sind Unterschiede im Schreibmaterial auffälliger als auffällige Stilveränderungen in der Handschrift. Lise Meitner beschrieb Postkarten, kleine und große Briefbögen, vergilbtes Papier und verwendete häufig Bleistift statt Tinte. Dies stellt die Software manchmal vor die Herausforderung, gute Kontrastwerte herzustellen. Die Lesbarkeit ist auch für Anfänger in der Paläographie eher als leicht denn als mittelschwer einzustufen.

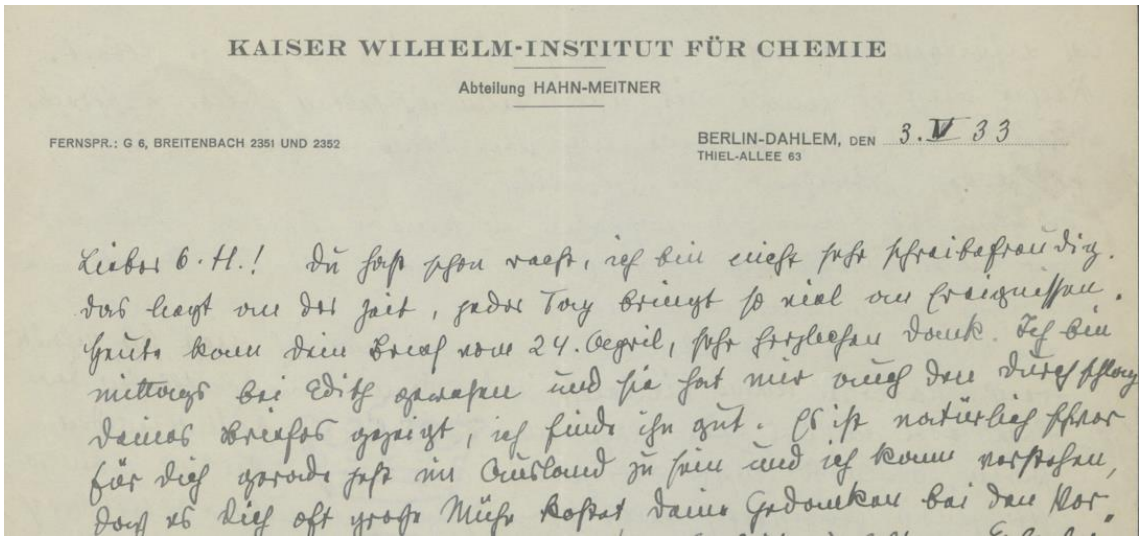

15. XII 24

Lieber Otto Hahn!

Ich danke dir und deiner Frau sehr für Euren lieben Briefe und für den Eilbrief samt Einlagen. Das telegraphisch geschickte Geld hat mir die Post angekündigt, aber noch nicht geschickt.

Abbildung 1: Handschrift von Lise Meitner (Ausschnitt eines Briefs vom 15.12.1924), In: III. Abt., Rep. 14, Nr. 6893

[61] Vgl. Milioni, 2020, S. 17.
[62] AMPG, III. Abt., Rep.14, Nr. 4869-4959, Nr. 6890-6898.
[63] Vgl. z.B.: Münter, o.D., Deutsche Sprache / deutsche Schrift.


KAISER WILHELM-INSTITUT FÜR CHEMIE

Abteilung HAHN-MEITNER

FERNSPR.: G 6, BREITENBACH 2351 UND 2352

BERLIN-DAHLEM, DEN 3.V 33
THIEL-ALLEE 63

Liebes O.H.! Du hast schon recht, ich bin nicht fest schreibfaul. Das liegt nun seit, jeder Tag bringt so viel an Eindrücken. Heute kam Dein Brief vom 24. April, fast Lichtgeschwindigkeit. Ich bin mittags bei Edith gewesen und sie hat mir auch den Brief gezeigt. Deinen Brief zeigst, ich finde ihn gut. Es ist natürlich schwer für dich gerade jetzt im Ausland zu sein und ich kann verstehen, daß es dich oft große Mühe kostet deine Gedanken bei der Ar-

Abbildung 2: Handschrift von Lise Meitner (Ausschnitt Brief vom 03.05.1933), In: III. Abt., Rep. 14, Nr. 6895

Später sollten Proben aus verschiedenen Jahrzehnten und verschiedenen Beschreibstoffen ausgewählt werden, um eine möglichst repräsentative Auswahl für den Trainings-Korpus des Modells zu erhalten. Dies war jedoch aus zeitpraktischen Gründen nicht mehr möglich.[64]

Im Allgemeinen bleibt das Schriftbild bis in das Jahr 1938 sehr gleichförmig. Ab dem 20.11.1938 übernimmt Meitner in einer radikalen Weise Elemente der modernen deutschen Schreibschrift und verwendet nur noch für wenige Buchstaben (großes „N“, „A“, „U“, kleines „e“, „r“) die alte Kurrent-Schreibformen. Ab den späten 1940er Jahren übernimmt Meitner die moderne Deutsche Schreibschrift gänzlich.

Die Erstellung eines präzisen spezialisierten Modells für die Handschrift Meitners erscheint damit als schwer umzusetzen, da sehr radikale Unterschiede von einem Modell mit einem kleinen Trainingsset schwer zu interpretieren sind. Zum Austesten eines eigenen Modells an den Schriften Meitners 1924-1938 eignet sich der Bestand dennoch. Die Frage, ob generische, auf viele Texttypen trainierte Modelle hierfür besser eingesetzt wären drängt sich auf.

[*Wandel im Schriftbild von Lise Meitner 1938-1939 - Anlage 2]*

[64] Anm. Die für die HTR-Modelle verwendete Edition (Ernst, 1992) umfasst nur die Jahre 1912-1924.

2.4 Technische Ausgangslage und qualitative Voraussetzungen für ein „Training“

Die Digitalisate liegen in einer Qualität von 300dpi als Master-TIF mit einer Farbtiefe von 24Bit (RGB) vor. Zusätzlich wurden JPG- und PDF/A-Dateien in einer 150dpi-Qualität als Nutzungsderivate in die Fileablage (keine Workflow-Software wie z.B. Goobi) des Servers gespeichert. Eine OCR-Texterkennung durch die Digitalisierungsfirma Ossenberg mithilfe der Software FineReader von Abbyy gehörte ebenfalls für alle Scans zum Standardlieferumfang. Diese Texterkennung für Maschinenschrift kommt bei Handschriften allerdings an ihre Grenzen. Die Derivate weisen kein Bildrauschen auf.

Nach den DFG-Praxisregeln „Digitalisierung“ sollten auf den digitalen Repräsentationen der Archivalien auch kleinste relevante Details noch zu erkennen sein, wenn die Datei optisch auf ein Viertel der Ausgangsgröße verkleinert wird. Weitere Vorgaben an die Scans die Mindestqualität einer Auflösung von 300dpi, eine Farbtiefe von 16 Bit in RGB ausreichend und die Speicherung als TIF für die Master-Dateien. Bildrauschen und andere optischen Störungen sollen vermieden werden. [65] Die Praxisregeln der DFG wurden hier, wenn auch erst im zweiten Trainingsdurchlauf in diesen Punkten eingehalten.

Die Briefe Meitners im Nachlass Otto Hahn sind vollständig verzeichnet und nach Laufzeit geordnet. Die Verzeichnungsdaten sind auch über Kalliope und das Archivportal-D nachgewiesen.[66] Zu einer tieferen inhaltlichen Erschließung, abgesehen von der Einzeldatenerfassung der Briefe, kam es nicht.

Die in den DFG-Praxisregeln „Digitalisierung“ beschriebenen Vorgaben für Digitalisate sind auch der Maßstab für das Ausgangsmaterial eines HTR-Projekts. Welche Qualitätsstandards an Textoutput durch automatisierte HTR-Transkription für welche Zwecke erforderlich sind und welche Kosten für das jeweilige Ziel angemessen sind, kann nach den Praxisregeln nur durch die Zielsetzung des jeweiligen Projekts beantwortet und nicht pauschal beschrieben werden.

---

[65] Altenhöner, Berger, Bracht, Klimpel, Meyer, Neuburger, Stäcker, Stein, 2023, S. 7.
[66] *Link:* Archivportal-D: Permalink auf Datenbankeintrag AMPG, III. Abt., Rep. 14, Nr. 4905: http://www.archivportal-d.de/item/ANNCRDJCH4FAU6YXEWS6JLIOHIJG3ND2 [Zugriff am 27.01.2024].

Allgemein wird eine Textgenauigkeit unter 95% als nicht erstrebenswert (CER unter 5%) bewertet, bzw. sollte diese nach der DFG nicht mit HTR/OCR-Dienstleistern vereinbart werden. Intrinsische Phänomene wie Verschmutzung, Widerdruckschatten, manuelle Unterstreichungen oder Annotationen könnten sich problematisch auf die HTR/OCR-Bearbeitung auswirken und sollten beachtet werden. Weiterhin empfiehlt die DFG den XML-Standard ALTO für die Nachnutzung der Daten aus einem HTR-Projekt. [67]

3. Softwareanwendung an Briefen von Lise Meitner

3.1 Wahl der Tools und der Methodik

Erste Schritte im Umgang mit der Software wurden ausschließlich mit der Web-Applikation Transkribus Lite gegangen.[68] READ-Coop weist auf der Website darauf hin, dass der Expert-Client nicht länger mit Updates versorgt wird und der Support dementsprechend „bald“ eingestellt würde.[69] Die Bedienung der Software per Browser gestaltete sich von Anfang an als intuitiv und recht einfach. Die Nutzung des Expert-Clients wurde vom Autor deutlich zu spät forciert, da die Aussage auf der Website von Read Coop dazu verleitete, zu glauben, dass keine Notwendigkeit für die Verwendung der Software bestünde. Wegen der Konzentration auf die Web-Anwendung ergaben sich im Laufe des Versuchs folgende Vor- aber vor allem auch Nachteile.

Die Arbeitsoberfläche des Expert-Clients ist deutlich komplexer aufgebaut als die der Web-Applikation. Die Werkzeuge sind standardmäßig über wenige Klicks direkt über die obere Bedienungsleiste erreichbar. Das inkludiert die Tools für das Anlegen von neuen Sammlungen (Collections), die Bearbeitung des Layouts (Baselines, Regionen, Linien) des ausgewählten Images, die Texteditiermaske für Transkription, die Auswahl der Modelle für HTR und die Trainingssektion für eigene Modelle. Ohne eigene Anpassungen (Anpassung über „Menu View“) kann der Überblick schnell verloren gehen. Die Webapplikation richtet sich mit dem schlichten, aufgeräumten Aufbau und einer stufenweisen Anordnung weniger Werkzeuge eher an Gelegenheitsnutzer. Die

[67] Altenhöner, Berger, Bracht, Klimpel, Meyer, Neuburger, Stäcker, Stein, 2023, S. 31-33.
[68] *Link:* Transkribus-Lite: https://app.transkribus.eu/ [Zugriff am 25.01.2024].
[69] READ-Coop, o. D., Desktop-Client herunterladen.

Funktionen sind dadurch aber auch deutlich begrenzt. Sowohl bei der Layout-Editierung, der Transkription, dem Einpflegen von GT aus externen Quellen als auch bei dem Training von Modellen ist weniger Feinjustierung möglich.

Die Versionsgeschichten pro Image sind im Expert-Client, anders als in der Browserversion, nahtlos nachvollziehbar. Ein weiteres Feature des Clients im Layout-Editor, welches online in Transkribus fehlt, ist die Möglichkeit Wortebenen und Linien einzeln händisch zu editieren. Das Editieren der Basislinien reicht normalerweise zusammen mit dem Bestimmen der Textregionen für die Vorbereitung der Transkription aus. Bei besonders anspruchsvollen Seitenlayouts ist eine Justierung von Worthöhe und Form jedoch von Vorteil.
Am stärksten verlangsamten sich die Testdurchläufe mit Transkribus durch den Verzicht auf den Expert-Client bei der Einpflege der bestehenden OCR-Daten der Edition von Sabine Ernst „Lise Meitner an Otto Hahn“ und dem Upload der Digitalisate.

Nach der Fertigstellung der ersten zwei von drei Modellen wurde der bereits installierte Expert-Client eingesetzt. Aufbauend auf der Arbeit vieler Stunden wurde anschließend hauptsächlich mit dem Expert-Client gearbeitet, um die Qualität der bereits erstellten Modelle zu analysieren und zu evaluieren. Im nächsten Kapitel wird die Arbeit mit dem Web-Interface beschrieben, wobei der Wechsel zum Client abgrenzend dargestellt wird.

Während des Uploads von Digitalisaten, des Einstellens von Ground-Truth und des Modelltrainings sowie der Evaluation der Ergebnisse wurde der Zeitaufwand pro Arbeitsschritt in Stunden erfasst. Er wird im folgenden Kapitel nach jedem Abschnitt angegeben.

3.2 Arbeitsprozess zur Erstellung eigener Modelle aus der Handschrift

OCR-Prozess zur Nutzung bereits vorhandener Groundtruth (GT)

Der erste Schritt für die Erstellung eines auf die Handschrift von Lise Meitner trainierten HTR-Modells war die Erstellung von OCR-Daten aus der Edition „Lise Meitner an Otto Hahn“ von Sabine Ernst. Dies sollte als Grundlage zur Generierung von Groundtruth (GT) dienen. Da dieses 1992 veröffentlichte Edition bislang nie digital

erschien, musste sie zunächst über den archiveigenen Nutzerscanner Zeutschel Zeta (2014) mit einer Auflösung von 300dpi als PDF-Multipage eingescannt werden.[70] Die PDF-Datei wurde in Adobe Acrobat Pro geöffnet, um die mit den Transkripten beschriebenen Seiten 13 bis 125 in Unicode lesbar zu machen.[71]

Die Edition von Sabine Ernst gibt die Briefe Meitners nicht zeilengetreu wieder, sondern erlaubt sich, für einen besseren Lesefluss eigene Zeilenumbrüche zu setzen und Rechtschreibfehler zu korrigieren. Dadurch war die Bearbeitung der Texte aufwändiger, als wenn die Schriftzeichen originalgetreu notiert worden wären. Die Übertragung der OCR-Daten erfolgte manuell. Dies erforderte einige Formatierungsarbeiten im Texteditor von Transkribus.

Das Scannen, die Texterstellung mittels OCR nahm etwa ***eine Stunde*** in Anspruch.

Das Tool Text2Image stellt in Transkribus ein Werkzeug dar, dass dazu verwendet werden kann, vorhandene Transkriptionen auf Seitenebene mit einer Zeilensegmentierung abzugleichen.[72] Dieses Werkzeug ist nicht im zuerst favorisiert genutzten Web-Interface vorhanden und zum Zeitpunkt der Einspeisung wusste der Autor noch nichts von dessen potenzieller Nützlichkeit. Die Möglichkeit dieses Werkzeug in der Transkribus GUI zu verwenden, bestand allerdings auch nicht, da die verwendete Version unter „Text2Image“ nicht als Werkzeug listete und auch extern zu dem Zeitpunkt kein Download als Add-On möglich war.[73]
Das Update auf die neueste Version konnte wegen Schwierigkeiten mit der Java-Kompatibilität bis zum Abgabetermin nicht mehr umgesetzt werden. Günther Mühlberger und Nikolina Milioni beschrieben in ihren Abhandlungen den Einsatz von Text2Image als Best-Practice-Lösung für die Erstellung von Trainingsmaterial (GT) für Anwender, die bereits über Transkriptionen in digitaler Form verfügen.[74] Lediglich die Funktion „Sync local transcriptions with doc“ konnte bei den eigenen Tests genutzt werden, um eine txt.-Datei mit den Transkriptionen mit einem passenden Image zu

[70] *Link:* Vgl. Technische Daten des Scanners: https://www.zeutschel.de/produkte/zeta-scanner/. [Zugriff am 20.01.2024].
[71] *Link:* Acrobat Reader DC November 2023 (23.00620380): https://helpx.adobe.com/de/acrobat/release-note/release-notes-acrobat-reader.html [Zugriff am 18.01.2024].
[72] READ-Coop, o. D., Text2Image.
[73] Anm. Transkribus Expert-Client in der Version 1.26.0 wurde verwendet.
[74] Mühlberger et al. 2019, S. 960 und Milioni, 2020, S. 19-20.

verknüpfen. Sie wurde in Vorbereitung auf das eigentliche Training nicht weiter genutzt. Es fehlte die entscheidende Funktion, die Zeichenketten passend mit den Zeilen zu matchen. Dieser Arbeitsschritt musste händisch erfolgen.

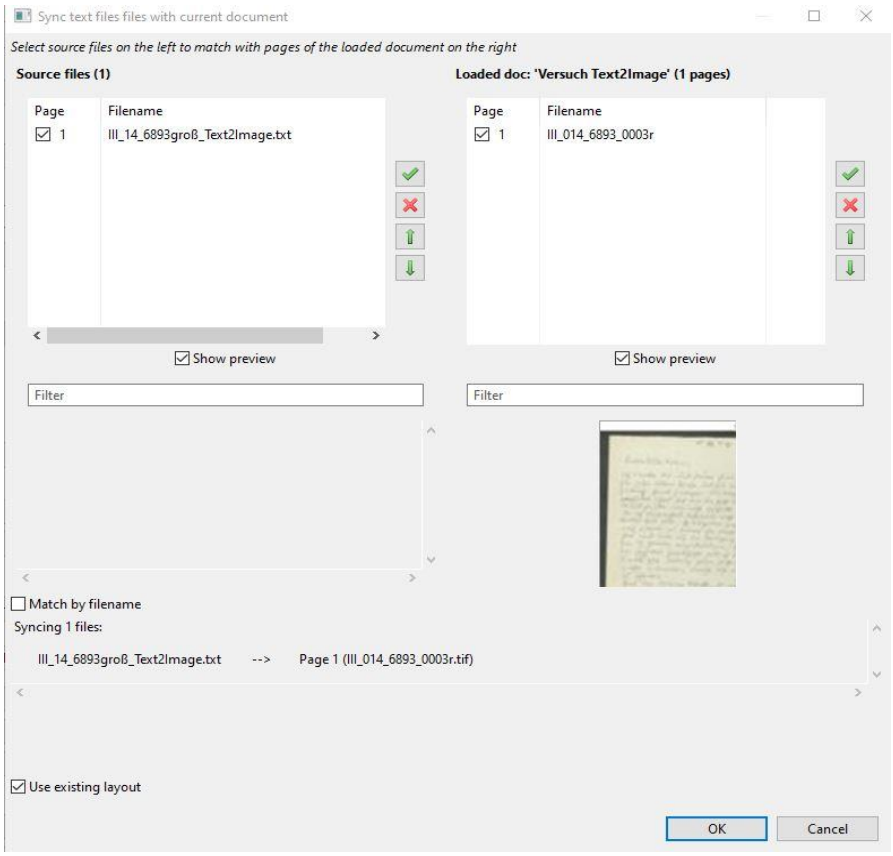


Abbildung 3: Funktion „Sync text files with current document“ im Expert-Client

Einspielen der Dokumente

Als Grundlage für die Transkription wurden PDF-Dateien der folgenden Signaturen aus dem Nachlass Otto Hahn als Multipage-PDF mit einer Qualität von 150dpi (basierend auf JPG-Images) in eine dafür erstellten „Collection“ mit dem Web-Interface eingespeist:

- III. Abt., Rep. 14, Nr. 6890
- III. Abt., Rep. 14, Nr. 6892
- III. Abt., Rep. 14, Nr. 6893

Die Collection „Lise Meitner an Otto Hahn (Training)“ im eigenen Transkribus-Profil sollte am Ende der Trainingsvorbereitung alle Dateien mit den Transkriptionen und Layoutinformationen enthalten, die als Groundtruth für das erste Modell des Versuchs

benötigt werden. Es handelte sich insgesamt um 108 Seiten Groundtruth. Der Upload in der Browser-Anwendung konnten am Stück nur in sehr kleinen Dateien erfolgen. Lediglich JPGs bis 8 MB und PDFs bis 10 MB PDFs wurden von Transkribus-Lite verarbeitet.

Im weiteren Verlauf der Modellvorbereitung wurde wegen komfortableren Steuermöglichkeiten und kaum Upload-Beschränkungen der Expert-Client zum Hochladen von Dokumenten benutzt. Dies geschah zunächst über das Extrahieren und Hochladen von Bildern aus den PDF-Dateien. Später wurde der Import von größeren Images (TIF) via FTP durchgeführt.

Das Einspeisen der Daten verbrauchte zusammen mit der Organisation der Dateien in einer Fileablage ***zwei Stunden***.

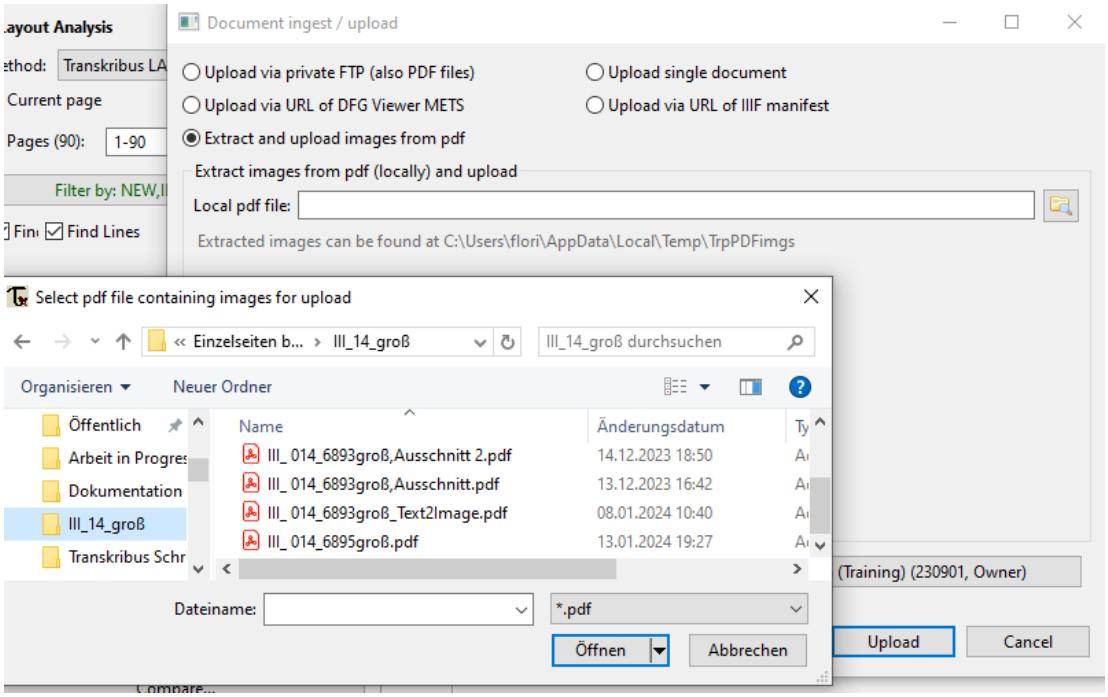


Abbildung 4: Upload-Optionen im Expert-Client

Layoutanalyse/Segmentierung/Einfügen von OCR-Daten

Als nächstes erfolgte eine halbautomatische Layoutanalyse der eingestellten Seiten. Hierfür wurde das Standard Analyse-Modell „Universal Lines“ gewählt. Das Modell vom Transkribus-Team ist standardmäßig voreingestellt und bedient bei einer CER von

8,9%CER viele Seitenlayouts mit verschiedensten Anforderungen.[75] Die Ergebnisse waren pro Seite bei der annähernd konsistent geradlinigen Handschrift von Lise Meitner lediglich mit ca. 8% Anpassungsbedarf direkt verwendbar, da das Layoutmodell die Grundlinien und die Textregionen insgesamt zuverlässig erkannte. Die häufigsten Probleme bereiteten dem Layoutmodell Korrekturvermerke oberhalb des Haupttextes, kleiner werdende Schrift nahe des Blattrandes und bei Postkarten die Änderung der Leserichtung (von unten nach oben). Diese Besonderheiten wurden nicht oder falsch zugeordnet mit Grundlinien versehen. Zur Korrektur wurden händisch im Web-Interface abschnittweise fälschlich unverbundene Linien einer Zeile zusammengeführt. Teilweise mussten Linien neu gezogen, gelöscht oder dem Text entsprechend korrekt angeordnet werden. Die Seitenzahl am oberen Bildschirmrand wurde nicht in die GT einbezogen und somit auch nicht in der Layoutanalyse berücksichtigt.

[*Darstellung einer manuellen Layoutanalyse in Transkribus – Anlage 3*]

Am Ende mussten die Buchstaben der Zeile auf der Grundlinie liegen und ihre Unterlängen darunter verlaufen, sodass später eine Übereinstimmung zwischen den Zeilen im Bild und den transkribierten Texten entstehen konnte.

Die zeitintensivste Tätigkeit bei der Modellvorbereitung war das Kopieren der Transkriptionsdaten in die entsprechenden Zeilen der Dokumente im Texteditior von Transkribus. Der Textinhalt einer vollständigen Briefseite wurde hierfür vom OCR-Text aus Adobe Acrobat in die erste Zeile einer Seite in den Editor kopiert. Danach wurden die einzelnen Textabschnitte Zeile für Zeile so ausgeschnitten und wieder eingefügt, dass sie am Ende korrekt auf der ganzen Seite angeordnet waren. Manchmal musste hierbei die Zeilenreihenfolge korrigiert und neu angeordnet werden.

Bestimmte Elemente im Text wurden hervorgehoben, damit sie im Modell berücksichtigt werden konnten. Dabei handelt es sich im Trainingsset um Durchstreichungen, hochgestellte Zahlen und Unterstreichungen. Sonderzeichen, in diesem Fall griechische Buchstaben (α, β, Ω), die Meitner gelegentlich zur Beschreibung von Strahlungsarten benutzte, konnten per Virtual Keyboard im Interface aufgerufen

[75] *Link:* Vgl. Advanced Layout Configuration Settings: https://help.transkribus.org/advanced-layout-configuration-settings [Zugriff am 21.01.2024].

werden. Transkribus erlaubt so die Eingabe aller Unicode-Buchstaben.[76]
Transkribus bietet im Editor ebenfalls die Möglichkeit bestimmte Wörter mit Tags zu versehen, um diese als Datierungen, Personen oder Orte zu kennzeichnen. Das Tagging-Tool dient momentan noch am ehesten der Kennzeichnung besonderer Entitäten, die bei der Veröffentlichung als digitale Edition hervorgehoben werden sollen. Z. Zt. gibt es noch keine Vorteile, das Tool für die Erstellung von Groundtruth zu verwenden. Ein einstellbarer Parameter ist das Ignorieren bestimmter getaggter Wörter, die während des Trainingsprozesses nicht ausgelesen werden sollen.[77]

Der Aufwand pro Seite betrug für die Korrektur der Layoutsegmentierung und das Einspielen der OCR-Daten durchschnittlich 15 Minuten. Dabei ist zu beachten, dass bei einigen Seiten (vor allem Postkarten) große Korrekturen mit dem Layout-Editor nötig waren, während andere Seiten nur einen Bruchteil der Bearbeitungszeit in Anspruch nahmen.

Die 108 Seiten waren somit *in **27 Stunden*** fertig für das Training bearbeitet.

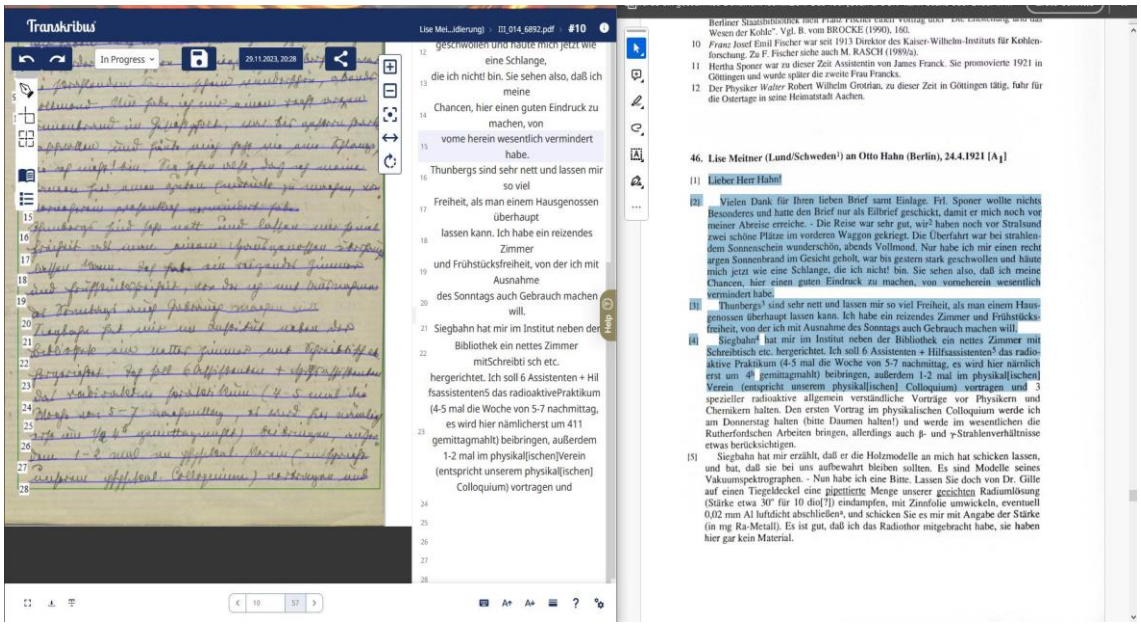


Abbildung 5: Einarbeitung der Edition Lise Meitner an Otto Hahn in das Transkribus Web-Interface (Transkribus Lite)

[76] Vgl. Mühlberger 2018, S. 150.
[77] Vgl. Mühlberger,et al., 2019, S. 960.

Modelltraining – Erster Versuch

Die Grundlage für das erste Modell speiste sich aus allen 108 eingestellten Seiten. Davon wurden 11 einem Validierungsset zugeordnet (10,19%). Damit verblieben 97 Seiten für das Trainingsset. PyLaia. das HTR-Modul von Transkribus wurde verwendet, um das Training durchzuführen.
In den Advanced Settings der Trainingsmaske wurden im ersten Durchlauf nur marginale Änderungen an den Standartwerten vorgenommen. Die maximalen Epochs (Trainingsdurchläufe) wurden mit 100 eher niedrig gewählt. Der „Early Stopping"-Wert wurde bei 20 belassen. Dies bedeutet, dass das Training abgebrochen wird, wenn die CER des Validierungssets nach 20 Durchläufen nicht mehr weiter sinkt.[78]
Da das Validierungsset allerdings nur 11 Seiten enthielt und damit nicht allzu viel Variationen der Dokumentenarten aufweist, ist die Wahrscheinlichkeit nicht klein, dass ein fehlerhafter Durchlauf das Training zu früh beendet. So wurde das Training auch hier schon im 20. Durchlauf gestoppt.
[*Abbildung von Trainingseinstellung für Modell „Lise Meitner 1912-1924" – Anlage 4]*

Aufgrund des recht kleinen Volumens an GT wurde dem Training ein Basismodell hinzugefügt. Hierbei handelt es sich um das durch das Transkribus Team kuratierte Modell „The German Giant I", das einen großen Pool von handschriftlichen Kurrent und Sütterlin-Quellen und auch einige Druckschriften als Trainingsdaten verwendete. Es basiert auf dem Modell Transkribus German Handwriting M1. Das Modell kommt bei einem sehr vielfältigen Set mit einer Größe von über 15 Mio. Wörtern auf eine beachtliche CER von 8,3%.[79]

Das Modelltraining dauerte hierbei für den Anwender eine Dreiviertelstunde, wobei die Zeit in der Warteschlange auf dem Transkribus-Server zur Bearbeitung nicht mit einberechnet ist. Aufgrund von ungenügenden Ergebnissen (siehe Kapitel 3.3) und der kleinen Anzahl von nur 20 Trainingszirkeln musste die Modellerstellung mit erheblichen Änderungen wiederholt werden.

[78] Vgl.READ-Coop, o. D., How to Train and Apply Handwritten Text Recognition Models in Transkribus eXpert.
[79] Vgl. READ-Coop, o. D, German Giant I.

Modelltraining – Zweiter Versuch

Für das Nachfolgemodell wurde die gleiche Groundtruth verwendet.
Während des Einspielens der Daten aus der Edition zuvor fiel bereits auf, dass der Detailgrad der Schrift in der Ansichtsmaske von Transkribus im Vergleich zur Betrachtung im Acrobat Reader bei den verwendeten PDFs, die aus einer Zusammenstellung von JPG-Derivaten mit 150dpi bestanden, zu gering ist, um bei der Modellerstellung zufriedenstellende Ergebnisse erzielen zu können. Dieser Verdacht bestätigte sich bei späteren Versuchen (siehe Auswertung Kapitel 3.3).
Andere Nutzer der Software berichten davon, dass die HTR-Modelle, unabhängig von den DFG-Praxisregeln ab einer Auflösung von 100dpi gleich zuverlässig arbeiten würden, was der Autor überprüfen wollte.[80] Nun wurden die Master-TIF-Dateien genutzt. Sie wurden in PDFs konvertiert und dabei nur leicht komprimiert, um den Informationsverlust für das Modell möglichst gering zu halten.[81]

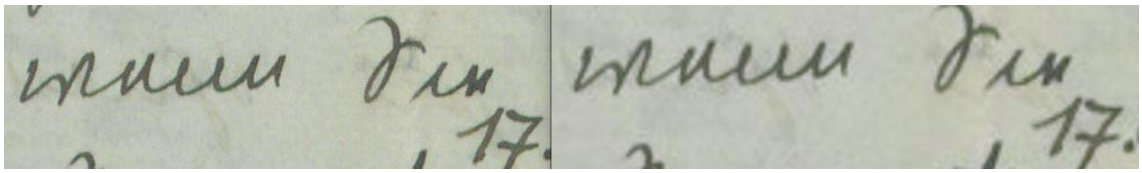


Abbildung 6: Ausschnitt aus III. Abt., Rep. 14, Nr. 6890 mit 120% Zoom - links: leicht komprimiertes JPG (258 dpi), rechts: JPG-Derivat (150 dpi)

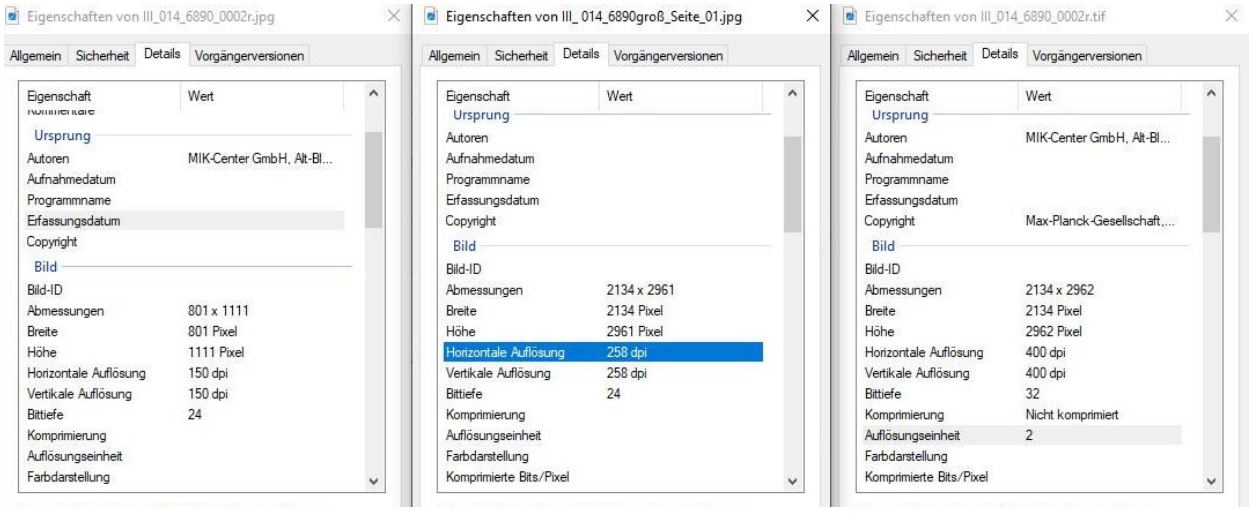


Abbildung 7: v.l.n.r. JPG-Derivat (150 dpi), verwendete nur leicht komprimiertes JPG (258 dpi), Master TIF (400 dpi)

[80] Vgl. Alvermann, 2022, S. 191.
[81] Anm.: Dieser Arbeitsschritt erwiesen sich später als unnötig, da der Upload der TIFs direkt über den Expert-Client deutlich schneller von statten geht.

Der Upload der größeren Dateien konnte via FTP nur mit dem Expert-Client durchgeführt werden.

Die JPG-Dateien mit 258dpi mussten anschließend wieder mit den OCR-Daten der Edition gemappt werden. Bei dem Vorgang musste das Copy-Paste-Verfahren mangels „Text2Image"-Tool nochmals angewandt werden, diesmal allerdings mit zeilengenauem Text, was den Aufwand verringerte.

Auf halber Strecke des Kopiervorgangs wurde die Arbeit mit dem Expert-Client fortgesetzt. Ab diesem Zeitpunkt (12. Dezember 2023) wurde auch für andere Arbeitsschritte weitgehend die potentere Version von Transkribus verwendet. Das eigentliche Training der hochauflösenden Files mit der Engine „PyLaia" erfolgte diesmal mit 150 Epochs, dem Grundmodell „Transkribus German Kurrent M2" und ausgeschalteter „Early-Stoppings"-Funktion.[82] Aus dem Training- und Validationsset wurden wenige Dokumente zur Bereinigung entfernt, da Sie wegen vertikaler Schrift oder sehr klein geschriebenen Marginalien nicht verarbeitet werden konnten. Der Trainingsdurchlauf dauerte knappe ***zwei Stunden*** und führte zu einer annehmbaren CER beim Modell „Lise Meitner 2.0".

Finales Modell

Im Anschluss wurde das Modell nochmals in einem leicht veränderten Verhältnis zwischen Training- und Validierungsset mit 180 Epochs und sonst gleichen Parametern trainiert. Als Grundmodell diente das vorherige Modell „Lise Meitner 2.0". Dabei erzielte es mit einer CER von 6,5% das beste Trainingsergebnis im Validationsset. Auch weitere Experimente mit dem gleichen Umfang an GT konnten das Ergebnis nicht mehr verbessern.

Die wiederholte Einstellung von OCR-Daten in die JPGs mit geringer Kompression benötigte weitere fünf Stunden, was bei der Berechnung des Gesamtaufwands allerdings nicht näher berücksichtigt wird. Zum einen ist der Aufwand vermeidbar, da bereits im ersten Trainingsset Scans mit höherer Auflösung hätten ausgewählt werden können. Zum anderen wäre der Aufwand für die Anpassung des Editionstextes an die richtigen Zeilen mit dem Expert-Client deutlich einfacher gewesen.

[82] Vgl. READ-Coop, o. D., German Kurrent M2.

Der nächste Schritt auf Projektebene im Archiv wäre es, die erstellten Modelle zur Generierung automatischer Transkripte der Briefe von Lise Meitner außerhalb des Zeitrahmens der vorliegenden Edition zu nutzen. Diese müssten anschließend korrigiert werden, um daraus wiederum einen größeren Pool aus GT zu bilden, der für

> Für das Training des Modells und die Feineinstellungen der PyLaia-Engine benötigte der Autor etwa ***vier Stunden***.

zuverlässigere Nachfolgemodelle genutzt werden könnten.[83]

3.3 Test und Vergleich der Genauigkeit der eigenen Modelle

Die HTR-Transkriptionsqualität, die aus 100 Seiten GT hervorging, ist auf den ersten Blick enttäuschend, da sie hinter der Genauigkeit anderer Modelle mit vergleichbarer Größe zurückblieb (siehe z.B: Milionis - Model B mit einer CER von 7.48% bei 4934 Wörtern).[84]

Die drei erstellten Modelle verbesserten sich in Reihenfolge ihrer Entstehung, wie zu erwarten war, zusammen mit der Lernkurve des Anwenders. Da für jedes in Transkribus erstellte Modell eine Lernkurve angezeigt wird, die den Lernverlauf in CER und Word Error Rate (WER), die Parameter, die Anzahl der Wörter für Trainings- und Validationsset, abbildet, konnte die Modellqualität recht schnell eingeschätzt und visualisiert werden.

Das erste Modell *„Lise Meitner 1912-1924"* krankte an unüberlegter Parameterwahl und einem zu kurzen Training.

Das Modell enthält 14.831 Wörter und 2024 Linien (Zeilen). Der Graph unterhalb des Textabschnitts zeigt die Genauigkeit des Modells in den verschiedenen Trainingsdurchläufen (Epochs). Fortschritt des Trainingssets und die rote Linie den Fortschritt des Evaluierungssets an. Je paralleler die Linien zur x-Achse und y-Achse verlaufen, desto besser ist die CER.[85]

---

[83] Vgl. Vorgehensweise bei der Modellerstellung für die Urteilsbegründungen des Wismarer Ratsgerichts durch das Stadtarchiv Wismar: Heigl, Jörn, 2023, S. 53
[84] Milioni, 2020, S. 25ff.
[85] Anm.: Die Darstellung der Modellergebnisse von Nikolina Milioni dienten diesem Kapitel als Vorbild: Milioni, 2020, S. 24-27.

Das Modell kam über eine CER von 19% im Validierungsset und 32% im Trainingsset nicht hinaus. Die Word Error Rate (WER) lag bei knapp 60%.

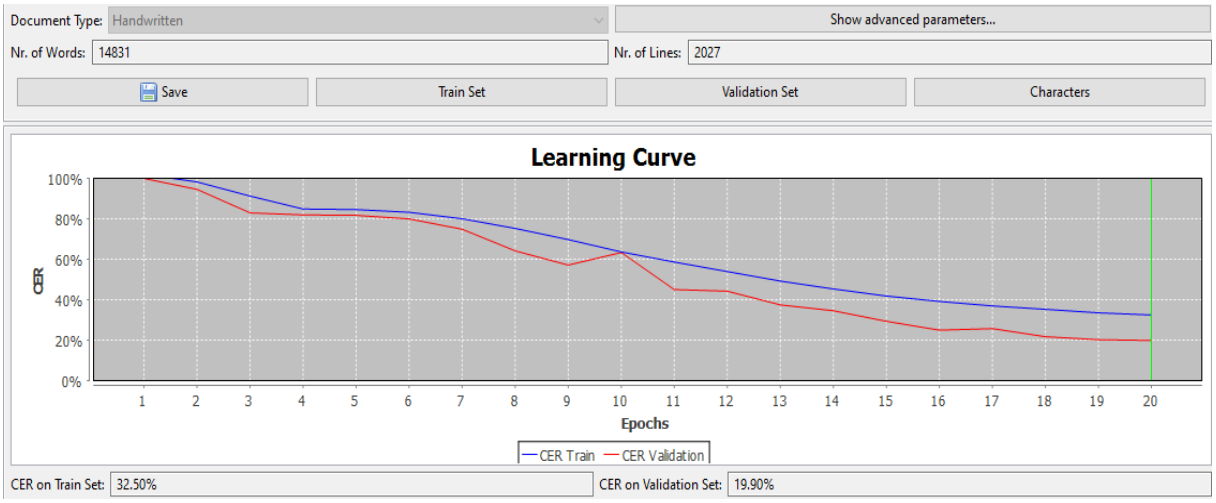


Abbildung 8: Lernkurve für das 1. Modell „Lise Meitner 1912 - 1924" mit 20 Epochs

Das Nachfolgemodell *„Lise Meitner 2.0“* mit einer höheren Zahl an Epochs und JPGs in einer höheren Auflösung erzielte eine Gesamt-CER von 11,3%. Das Modell enthält 12.905 Wörter auf 1954 Zeilen.

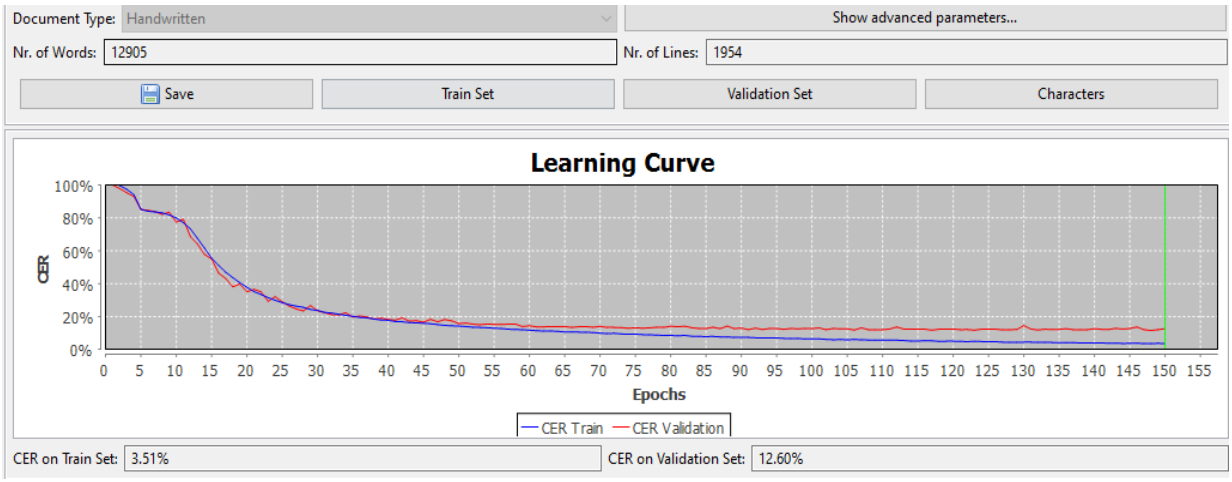


Abbildung 9: Lernkurve für das 2. Modell „Lise Meitner 2.0" mit 150 Epochs

Das finale Modell *„Lise Meitner pur“* enthält 12854 Wörter und 1955 Zeilen und nutzte „Lise Meitner 2.0“ als Basismodell. Diesmal sank die Gesamt-CER auf 6,1%, wobei die CER am Trainingsset sogar auf 1,4% sank.

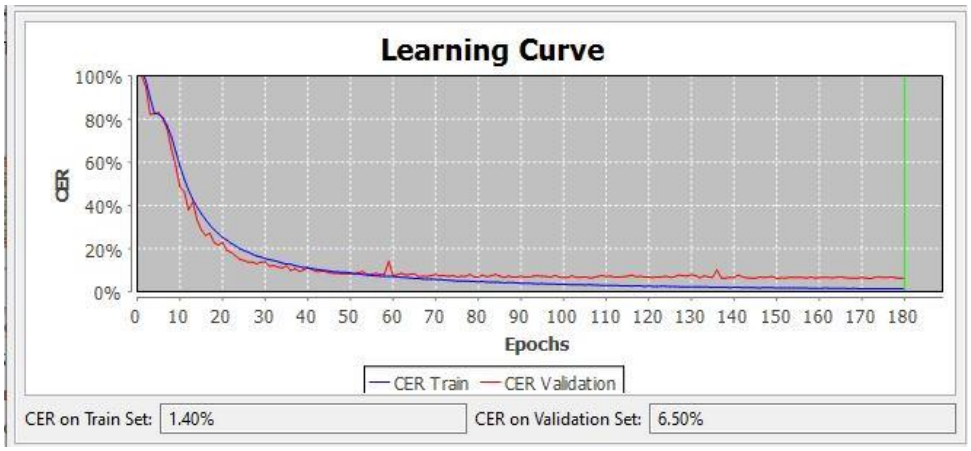


Abbildung 10: Lernkurve für das Modell „Lise Meitner pur" mit 180 Epochs

Übersicht zu den erstellten Modellen

Modell-Name	Wörter	CER (Mittelwert)	Epochs	Seiten Train-Set	Seiten Valid-Set
Lise Meitner 1912-1924	14831	19,90%	20	96	11
Lise Meitner 2.0	12905	12,60%	150	90	10
Lise Meitner pur	12854	6,50%	180	90	10

Überprüfung an einem Testset

Zur Feststellung der Qualität der Modelle an Testseiten gibt Transkribus dem Nutzer ein Fehlermesstool an die Hand, mit dem zwei Versionen einer Seite miteinander verglichen und die Abweichungen wiederum als CER und WER ausgegeben werden. Die verwendeten Seiten waren weder im Trainings- noch im Validationsset enthalten und sollten die Anwendbarkeit der Modelle für die Handschrift Lise Meitners praktisch testen. Dafür wurden weitere vier Seiten händisch transkribiert. Auch für diesen Zweck gab es im Archiv schon Transkripte als Vorlage, die allerdings noch nicht veröffentlicht wurden und nur in Schreibmaschinenschrift vorlagen. Sie mussten tlw. noch überarbeitet werden.

Die HTR-Ergebnisse im Testset wichen im negativen Sinne hinsichtlich der CER und WER tlw. stark gegenüber den Validierungssets ab. Dies kann teilweise mit der geringen Samplerate erklärt werden. Die Differenz der CER-Ergebnisse zwischen Modell und Testseiten ist allerdings beim finalen Modell „Lise Meitner pur“ wesentlich größer.

Das Modell „Lise Meitner 2.0“ schloss gegenüber den Testseiten durchschnittlich um 4,3% (CER) schlechter ab, als bei den Daten seines Validierungssets.

Das Modell „Lise Meitner pur“ schloss dagegen bei den Testseiten durchschnittlich um 6,015% (CER) schlechter ab im Verhältnis zu den Daten seines Validierungssets.

Damit verfehlte das favorisierte eigene Modell mit einer CER von 12,55% die angestrebte Höchstmarke von 10%.[86] „Meitner 2.0“ ist mit einer CER von 16,86% nicht für die Weiterführung automatischer Texterkennung im ausgewählten Bestandssegment verwendbar. Die Belastbarkeit der Ergebnisse aus dem Training muss wegen der dargelegten hohen Diskrepanz in diesem Fall infrage gestellt und überprüft werden.

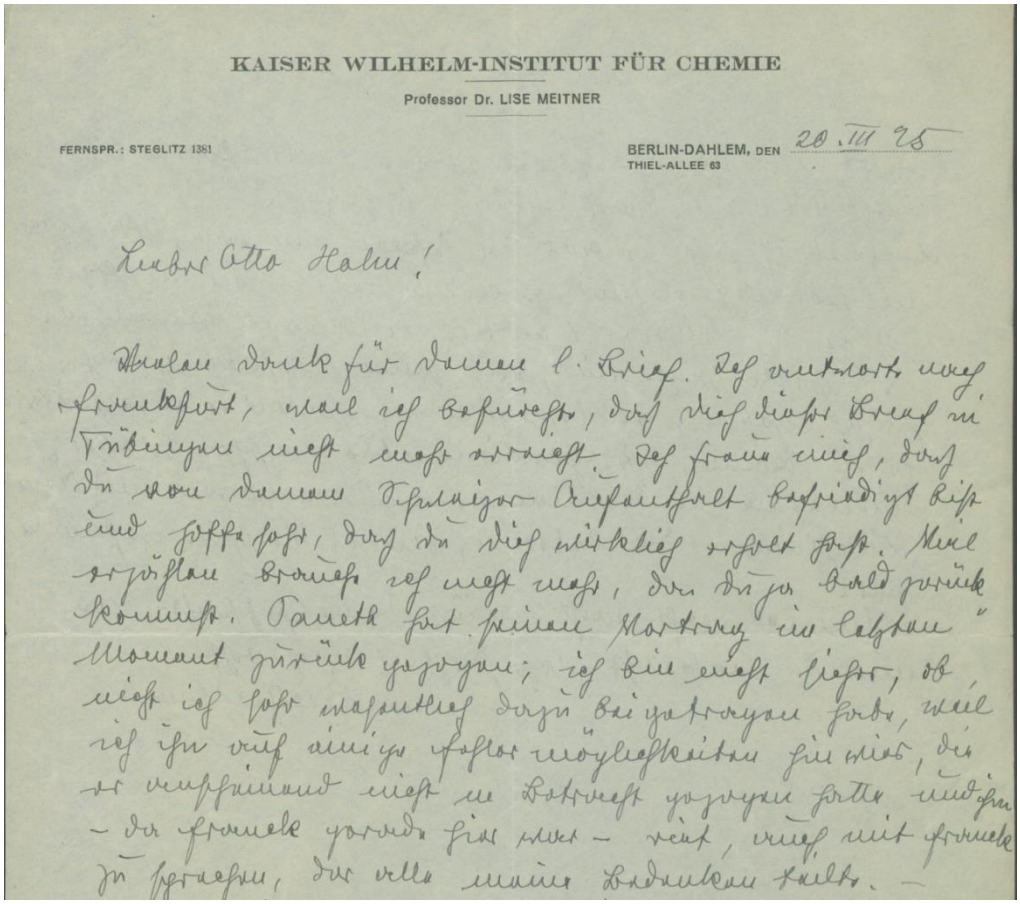
Als Faustregel führte einer der Initiatoren von Transkribus Günther Mühlberger die Regel an, dass die Ergebnisse sowohl repräsentativ als auch robust sind, wenn die Werte, die gegen das Trainings- als auch das Testset gemessen werden, nicht zu weit voneinander abweichen. Die CER in Bezug auf das Trainingsset neigt zur Überanpassung, da das Netz die Daten auswendig lernt und dadurch besonders gute Ergebnisse bei bereits verarbeiteten Zeichen erzielt. Dagegen sind die Ergebnisse bei der Anwendung auf neue Daten oftmals wesentlich schlechter.[87] Die Ergebnisse zwischen Trainingsset und Validierungsset der beiden Modelle „Lise Meitner 2.0/Lise Meitner pur“ wichen jeweils zwischen 5 und 7% (CER) voneinander ab.

Da das finale Modell „Lise Meitner pur“ den Vorgänger „Lise Meitner 2.0“ als Basis-Modell für das Training verwendete, ist es wahrscheinlich, dass die Ergebnisse des Validationssets weniger aussagekräftig sind. Zum Teil wanderten Seiten des Trainingssets des Basis-Modells ins Validationsset von „Lise Meitner pur“. Damit wurde das Modell schon indirekt auf die Seiten trainiert, die eigentlich für unabhängige

[86] Anm. Siehe zum Zielwert auch 1.2 -Modelltraining.
[87] Vgl. Mühlberger, 2018, S. 151

Verifikationen dienen sollten. Da dieses Modell trotz dieser Feststellung bei den nachfolgend beschriebenen Testseiten am besten abschnitt (CER und WER), wurde es innerhalb dieser Arbeit weiter als finales Modell „Lise Meitner pur“ berücksichtigt.

*Das **Beispiel A*** ist ein Brief aus dem Jahr 1925, in dem sich die Handschrift von Lise Meitner noch nicht wesentlich verändert hatte. Beachtenswert ist, dass die CER aus den Validationsergebnissen vom Modell „Lise Meitner pur“ in diesem Beispiel nur um 0,45% besser ausfällt als beim Vorgängermodell.



KAISER WILHELM-INSTITUT FÜR CHEMIE

Professor Dr. LISE MEITNER

FERNSPR.: STEGLITZ 1361

BERLIN-DAHLEM, DEN THIEL-ALLEE 63

20. III 25

Liebe Otto Hahn!

Abbildung 11: Ausschnitt des Digitalisats III. Abt., Rep. 14, Nr. 6893, S. 11

~~Lisht MedtrITTT PR ADis~~ KAISER WILHELM-INSTITUT FÜR CHEMIE
~~Twem 2. AWII~~ Professor LISE MEITNER
~~2. 29~~ 20. III. 25
~~liber Ite etn~~ Lieber Otto Hahn!
~~lielen~~ Vielen Dank für ~~demen lien~~ Deinen l. Brief. Ich ~~antwerte~~ antworte nach
Frankfurt, weil ich befürchte, daß Dich dieser Brief in
Tübingen nicht ~~nehr~~ mehr erreicht. Ich freue mich, daß
Du von Deinem ~~Schweger~~ Schweizer Aufenthalt ~~befriedgtbist,~~ befriedigt bist
und ~~joffesehr, daß De~~ hoffe sehr ‚dass Du Dich wirklich erholt ~~haszt. Val~~ hast. Viel
erzählen brauche ich nicht ~~nicht, dae Di~~ mehr, da du ja ~~balt zurnk~~ bald zurück
~~konnst. Panett~~ kommst. Paneth hat seinen ~~Vertrag~~ Vortrag im letzten
Moment ~~zurückgezogen,~~ zurück gezogen; ich bin nicht sicher, ~~o~~ ob
nicht ich sehr wesentlich dazu beigetragen habe, ~~wul~~ weil
ich ihn auf ~~eineze Fehlermöglichketen hin wies,~~ einige Fehler möglichkeiten hinwies, die
er anscheinend nicht in Betracht ~~gezozen~~ gezogen hatte und ~~hi~~ ihm
- ~~Da~~ da Franck gerade hier ~~war~~ war - riet, auch mit ~~Frank~~ Franck
zu sprechen, der alle meine Bedenken teilte.
~~Gangtgrund~~ Hauptgrund zu meinem ~~heuteizen~~ heutigen Brief ~~st Holgender,~~ ist folgender:
Herr v. ~~Leue~~ Laue ist von einem ~~Prof. Feies (iede~~ Prot. Spiess (Medi
~~Zieier)~~ ziner aus Frankfurt gebeten ~~werden,~~ worden, ihm einen
radioaktiven ~~Fachmane~~ Fachmann zu ~~neinen,~~ nennen, der ~~hm~~ ihm raten
könnte, wie man radioaktive ~~Substanen~~ Substanzen an
~~gwissen~~ gewissen Geweben ~~adsorbinren~~ adsorbieren könnte u. er ~~bitter~~ bittet
Dich nun, da Du gerade in ~~FFrankfurt~~ Frankfurt bist, Dich

*Abbildung 12: Abgleich aus III. Abt., Rep. 14, Nr. 6893, S. 11; Modell „**Lise Meitner 2.0**" mit der Ground Truth (grün unterlegte Korrekturen) – CER: 13,31%, WER: 34,62%*

~~Aaak ZeeeELLII 2A Aaccete~~ KAISER WILHELM-INSTITUT FÜR CHEMIE
~~L W. 2 WüAEE~~ Professor LISE MEITNER
20. ~~I. 26.~~ III. 25
Lieber ~~Ao~~ Otto Hahn!
~~lielen~~ Vielen Dank für Deinen l. Brief. Ich ~~antwerte nch~~ antworte nach
Frankfurt, weil ich ~~befüochte,~~ befürchte, daß ~~Dieh~~ Dich dieser Brief in
~~Tübngen~~ Tübingen nicht mehr erreicht. Ich freue mich, daß
Du ~~ven Deinen Schweer Anfenthalt~~ von Deinem Schweizer Aufenthalt befriedigt ~~bist,~~ bist
und hoffe ~~sehr, daß~~ sehr ‚dass Du Dich wirklich erholt hast. Viel
erzählen brauche ich nicht mehr, da ~~Duja bals~~ du ja bald zurück
~~kömnst.~~ kommst. Paneth hat seinen Vortrag im letzten
Moment zurück ~~gezogen,~~ gezogen; ich bin nicht sicher, ~~o~~ ob
~~neht~~ nicht ich sehr wesentlich dazu ~~bei getragen~~ beigetragen habe, ~~senn~~ weil
ich ihn auf ~~minige~~ einige Fehler möglichkeiten hinwies, die
er ~~auscheinend~~ anscheinend nicht in Betracht ~~gezagen~~ gezogen hatte und ihm
~~a~~ - da Franck gerade ~~Eie~~ hier war ~~iet,~~ - riet, auch mit ~~Frack~~ Franck
zu sprechen, ~~dur~~ der alle meine Bedenken ~~teilbe.~~ teilte.
~~spaupt grund~~ Hauptgrund zu ~~meinen~~ meinem heutigen Brief ist ~~folgendes.~~ folgender:
Herr v. Laue ist von einem ~~P rof Lpiess (Miede.~~ Prot. Spiess (Medi
~~ener)~~ ziner aus ~~Frankturt~~ Frankfurt gebeten worden, ihm ~~emen.~~ einen
radioaktiven ~~Fachmainn~~ Fachmann zu ~~meinnen,~~ nennen, der ~~ihin~~ ihm raten
~~Nönnte, nie~~ könnte, wie man radioaktive ~~Substangen~~ Substanzen an
~~genissen~~ gewissen Geweben ~~adsorturen kömnte~~ adsorbieren könnte u. er ~~bllet,~~ bittet
Dich ~~min, de~~ nun, da Du gerade in ~~Frünkfurt~~ Frankfurt bist, ~~mch~~ Dich

*Abbildung 13: Abgleich aus III. Abt., Rep. 14, Nr. 6893, S. 11; Modell „**Lise Meitner pur**" mit der Ground Truth (grün unterlegte Korrekturen) – CER 12,86%, WER 37,02%*

Das ***Beispiel B*** aus dem 1927 zeigt einen besonders gut leserlichen Brief von Meitner, der wiederum mit den beiden eigenen Modellen abgeglichen wurde. Die CER des finalen Modells ist dem des Vorgängers um 3,31% überlegen.

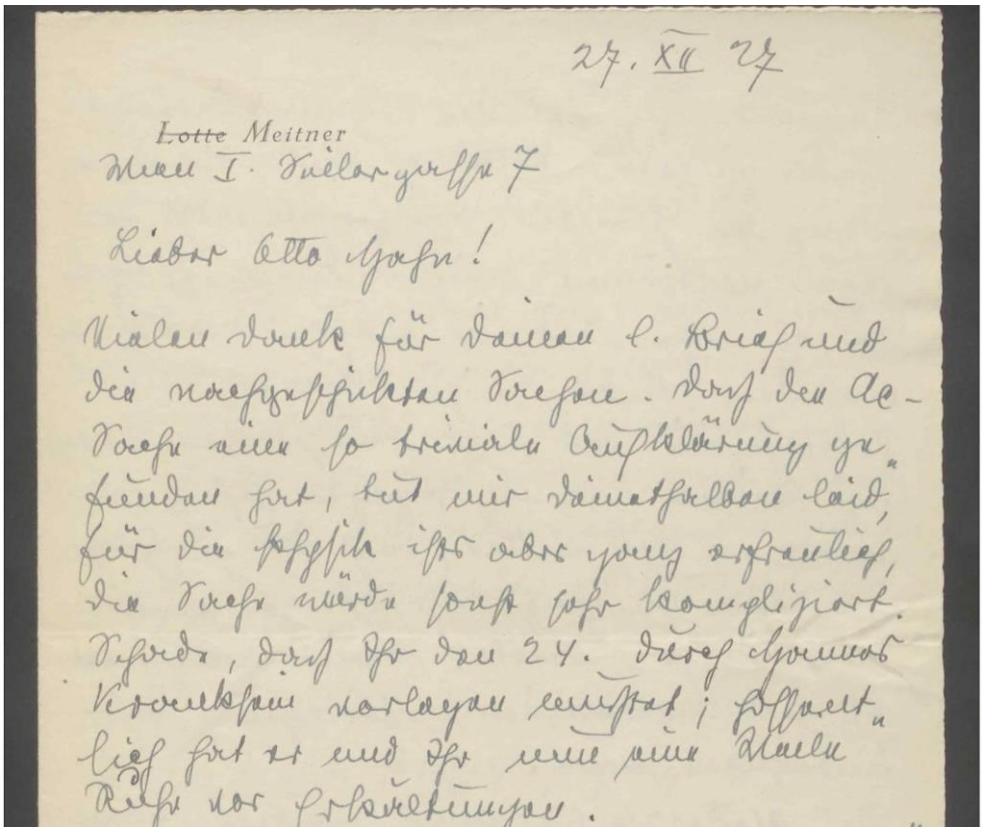


Abbildung 14: Ausschnitt Digitalisat III. Abt., Rep. 14, Nr. 6894, S.

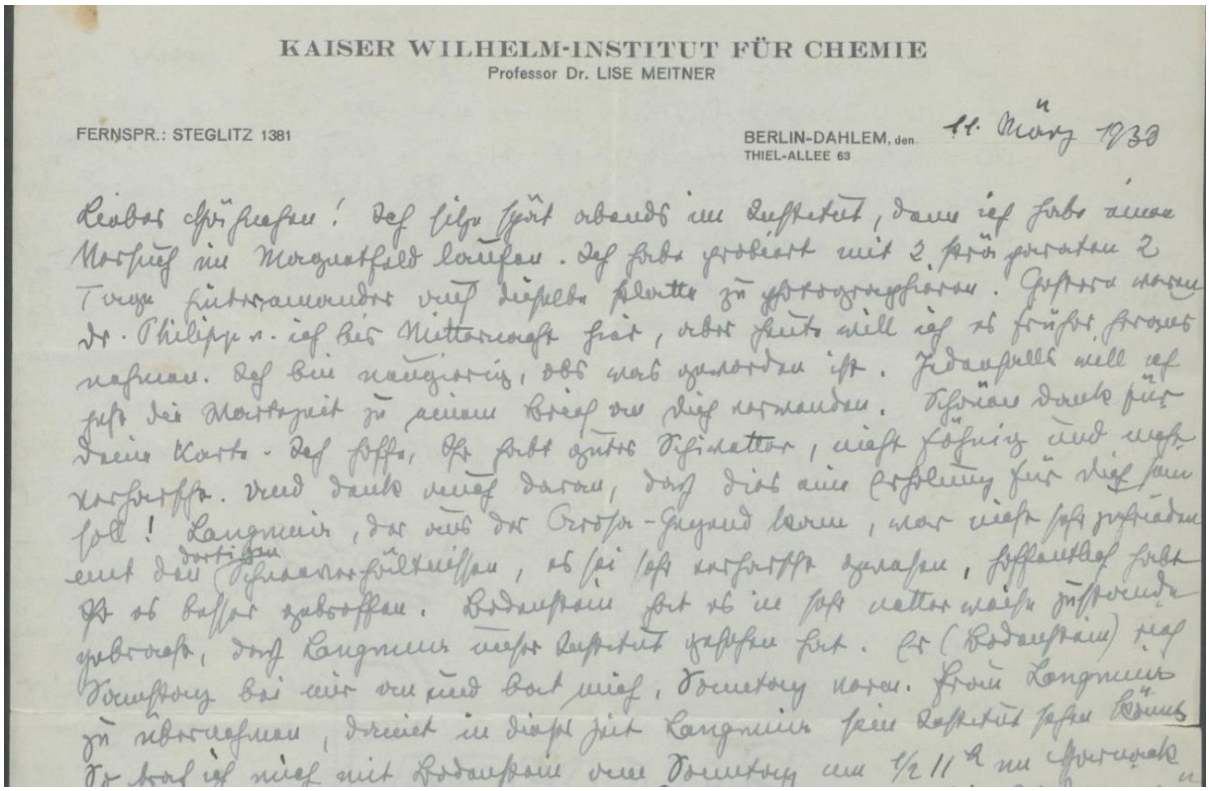
AI5 0006
27. 2I 27.XII 27
füree 1119121 Lotte Meitner
Wien E Sellergasse I. Seilergasse 7
Liber Ott Hehn! Lieber Otto Hahn!
Vielen vielen Dank für demen len Briefund deinen l. Brief und
die nachgeschickten Sachen. Daß die 2e. Al-
Sache eine so tewale Aufklarung ge triviale Aufklärung ge-
funden hat, but tut mir Deinethalben leid leid,
für die Pehysine Physik ists aber ganz erfreulich,
die Sache wrde wurde sonst sehr könzliziert kompliziert.
Schade, daß Ihr den 27. Durch games 24. durch Hannos
Kranksen vorlegen müßtet, hoffent., Krankseim verlegen mustet; hoffent-
lich hat er und Ihr nun eine Meile Weile
Riche ver Erkaltungen Ruhe vor Erkältungen.
Mein Vortrag in Kiel ist am 17 haue 19. Jänner.
(Du ermierst Dich erinnerst dich doch, daß er erst am
18. 12. sein sollte u u. ich ihe, ihn, um das erste
Kolleg am 13. nicht ausfallen zu
lassen, auf den 9h. voriegt habe) 19. verlegt habe).
Da ich gerabe ich, gerade weil ich als einzgge einzige nicht
in Frankfurtwar, Frankfurt war, einen Wert drauf

*Abbildung 15: III. Abt., Rep. 14, Nr. 6894, S. 6; Modell „**Lise Meitner 2.0**" mit der Ground Truth (grün unterlegte Korrekturen) – CER: 13,86%, WER: 41,56%*

AE 0006
27. E 27. 27.XII 27
tu 1 122 Lotte Meitner
Wien Z Seller gasse F I. Seilergasse 7
Lieber Otto Hahn!
Vielen vielen Dank für deimnen lben deinen l. Brief und
die nachgeschickten Sachen. Daß die 2a. Al-
Sache eine so trwiale Aufklarung ge triviale Aufklärung ge-
funden hat, tut mir Demethalben Deinethalben leid,
Für für die Physik ists aber ganz erfreulich erfreulich,
de die Sache würde wurde sonst sehr kompliziert kompliziert.
Schade, daß Ihr den 24. Durch gannes durch Hannos
Kranksen Krankseim verlegen mußtet, hoffent. mustet; hoffent-
lich hat er und Ihr nun eine Weile
Ruht Ruhe vor Erkaltuungen Erkältungen.
Mein Vorbrag Vortrag in Kiel ist aml am 19. Jaue Jänner.
(Du erinerst Dich erinnerst dich doch, daß er erst am
13. 12. sein sollte u. ich ihn, um das erste
Kolleg am 63. 13. nicht ausfallen zu
lassen, auf den 9. verliegt habe) 19. verlegt habe).
Da ich, gerade well weil ich als einzige nicht
in grankfurtwar, Frankfurt war, einen Wert drauf

*Abbildung 16: Abgleich aus III. Abt., Rep. 14, Nr. 6894, S. 6; Modell „**Lise Meitner pur**" mit der Ground Truth (grün unterlegte Korrekturen) – CER: 10,5%, WER 32,47%*

Beispiel C von 1933 zeigt eine Seite mit komplexeren Fachbegriffen und vielen Eigennamen (z.B. „Klumpenbildung“ und Hydralisierbarkeit“). Die CER im finalen Modell ist um 4,9% geringer als die des Vorgängermodells „Lise Meitner 2.0“.



KAISER WILHELM-INSTITUT FÜR CHEMIE
Professor Dr. LISE MEITNER

FERNSPR.: STEGLITZ 1381

BERLIN-DAHLEM, den. 11. März 1933
THIEL-ALLEE 63

Abbildung 17: Ausschnitt Digitalisat III. Abt., Rep. 14, Nr. 6895, S. 2

behtt LnalISTTTST F0 AuDr5 KAISER WILHELM INSTITUT FÜR CHEMIE
W 00. 11E WTTU Professor Dr. LISE MEITNER
LRRRR LalIST 1 FERNSPR.: STEGLITZ 1381
ohvlernavm 1 Miarz seh BERLIN-DAHLEM, den 11. März 1930
WAI 1 THIEL-ALLEE 63
Lieber Hä flichen! Liebes Hähnchen ! Ich lise sit sitze spät abends in Sistitut dene im Institut, denn ich habe eine einen
Versuch en im Magnetfeld laufen. Ich habe prohiert probiert mit 2 Pträparaten2 Präparaten 2
Tag Einteremander Tage hintereinander auf tühelte dieselbe Platte zuphotogrephieren. Gestere ver zu photographieren. Gester waren
Dr. Mtilz. Philipp u. ich bes Mitteracht bis Mitternacht hier, aber heute will ich es früher heranus heraus
nehnen nehmen. Ich bin nengierng, neugierug, obs was geworden ist. Jideehalls vill Jedenfalls will ich
jetzt die Wartezeit zu einem Brief an Dich dich verwenden. Schonen danck Schönen Dank für
Denne Deine Karte. Ich soffe, hoffe, Ihr habt geter Schwetter, gutes Schiwetter, nicht fohnig föhnig und nicht
verharscht and Denk verharscht. Und denke auch daran, daß dies eine frfolung Erholung für Dich san dich sein
soll? Leipmier soll! Langmeier, der ans aus der Arrsa jegend kamn, wer Arosa-Gegend kam, war nicht schrzehreten sehr zufrieden
mit diusshieeversoltuissee, den Schneeverhältnissen, es sti schr eerhertht gwwesen, stfeikch habte sei sehr verharscht gewesen, hoffentlich habt
dertihhen dortigen
Ihr es besser gebroffen. Brdenstem hit getroffen. Bodenstein hat es in sehr nitter wche ustände netter Weise zustande
gebracht, daß Leizmi niser Iestitut Langmeier unser Institut gesehen hat. Er (Bedeistein e (Bodenstein) rief
Samstag bei nur mir an und bat mich, Sonntag vorm. Frau Lovpiie Langmuir
zu ebernehmen, damt übernehmen, damit in dieser Zeit Eevomie sein Eestitutsehen könn Langmuir kein Institut sehen könne.
So traf ich mich mit Bedenstei Bodenstein am Soüntag im 1e 1R im Harnack Sonntag um 1/2 II h am Harnack¬
haus, wo Lengmmar Langmuirs wohnten. Vor fühten wämmen Wir fuhren zusammen in die Stadt, ich
gnig ging mit Frau5 Frau L. in die Steinbrandtausstellung, Rembrandtausstellung, die Herren is ins
hhyikelisch chemische Iestitut. Geizen Eh physikalisch chemische Institut. Gegen 1 h holte in Bedenstein ab¬ uns Bodenstein ab
wis führen Alle wir fuhren alle nach kannsen Wannsee zu ihm, wo aüßerden ausserdem noch Reucke Plancks,
(olmere Volmers u. Schleicks Schlencks waren. Etwas vor 1t 1/2 5 führen wii fuhren wir wieder
ins Harnackhauszurück Harnackhaus zurück zu dem Ffiziellen offiziellen Thee. Auf der Fahrt
dahin (orher (vorher hatte ichkaum ich kaum mit Leupeme gesprohen) Bagte Langmuir gesprochen) fragte er nich mich,
wann er zu uns ins Institut kommen könnte. Wer verabreden Wir verabreden,
daß ich ihn Montag früh um 1 X 10 h im Harnackhaus abholen würde würde,
was ich dann auch tat. Ich erzählte ihm zu erst zuerst im allgemeinen allgemeinen,
was für gheüische chemische und physikalische Peroblebei Probleme bei uns bearbeitet
würden und führte ihn dann herum. Von der Cheene Chemie zeigte ihe ihm
Werner die Versuche eber die Kluipenbeldung i Abhanzigkeit über Klumpenbildung in Abhängigkeit
ou Sauregehalt von Säuregehalt und vor Hdrolisierbarkeit (Unterschiedlieches Ver von Hydrolysierbarkeit (Unterschiedliches Ver¬
dalben halten von B Bi und Ra slzen (ict Ba salzen) etc. Er hätte hatte es sehr eett geusammen, nett zusammen

*Abbildung 18: Abgleich aus III. Abt., Rep. 14, Nr. 6895, S. 1; Modell „**Lise Meitner 2.0**" mit der Ground Truth (grün unterlegte Korrekturen) – CER: 19,37%, WER 46,88%*

Aaa LIS22 1 Aa ALe KAISER WILHELM INSTITUT FÜR CHEMIE
2 . AAs Wel Professor Dr. LISE MEITNER
EEMIEER. PREIIL 19 FERNSPR.: STEGLITZ 1381
davvenvm 1. Marz 1950 BERLIN-DAHLEM, den 11. März 1930
A e 2. 2. THIEL-ALLEE 63
Lieber Hähichen! Liebes Hähnchen ! Ich sise sitze spät abends im Institut, denn ich habe einen
Versuch im Magnetfeld laufen. Ich habe probiert mit3 mit 2 Präparaten 2
Tage Einteremander hintereinander auf dieselbe Plätte zuphotographieren. Gestern wer Platte zu photographieren. Gester waren
Dr. Milissee Philipp u. ich bis Mitternacht hiet, hier, aber heute will ich es früher heraus
nehmen. Ich bin neugierig, neugierug, obs was geworden ist. Jedegfalls vill Jedenfalls will ich
jetzt die Wartezeit zu einem Brief an Dich dich verwenden. Schönen Dank für
Deie Deine Karte. Ich hoffe, Ihr habt gutes Schwetter, Schiwetter, nicht föhnig und nicht
verharsche and denk verharscht. Und denke auch daran, daß dies eine Erholung für Dich dich sein
soll! Leugmme, Langmeier, der aus der Arosa Gegend kan, Arosa-Gegend kam, war nicht sehr zuhrieden, zufrieden
mit de Schieeverhältnissen, den Schneeverhältnissen, es sei sehr verharscht gwesen, gewesen, hoffentlich habt, habt
derhehen. dortigen
Ihr es besser gebroffen. Brdensten bit getroffen. Bodenstein hat es in sehr netter weise Weise zustande
gebracht, daß Lengmmns nieser Langmeier unser Institut gesehen hat. Er (Bödeistein fie (Bodenstein) rief
Sanstag Samstag bei mir an und batmich, bat mich, Sonntag verm. vorm. Frau Lanpmmn Langmuir
zu übernehmen, damit in dieses dieser Zeit Leupmins sein Iestitut Langmuir kein Institut sehen könnn, könne.
So braf traf ich mich mit Bödenstein Bodenstein am Sonntag um 12 12 im Harnack 1/2 II h am Harnack¬
hans, haus, wo Lenpmuas Langmuirs wohnten. Wir führen fuhren zusammen in die Städt, Stadt, ich
ging mit Frau F. L. in die Sembrandt ausstellung Rembrandtausstellung, die Herren ins
ysikalisch physikalisch chemische Instiobtut. Geigen 1 helte Institut. Gegen 1 h holte uns Bedenstein Bodenstein ab
mer führen Alle wir fuhren alle nach kannsee Wannsee zu ihm, wo außerdem ausserdem noch Pancke Plancks,
Pomees und Volmers u. Schlencks waren. twas wer 1 Etwas vor 1/2 5 fehren fuhren wir wieder
eins ins Harnackhaus zutück zurück zu den Offiziellen Thie. dem offiziellen Thee. Auf der Fahrte Fahrt
dahin Vorher (vorher hatte ich kaum mit Leupmine Langmuir gesprochen) hagte es mich fragte er mich,
gann wann er zu eins uns ins Insetut Institut kommen könnte. Wir verabreden. verabreden,
daß ich ihn Montag früh mu 1h um 10 h im Harnackhaus abholen würden würde,
was ich dann auch tat. Ich erzählte ihin ihm zuerst in allgemeinen im allgemeinen,
was für chemische und physikalesche Problebei physikalische Probleme bei uns bearbeitet. bearbeitet
würden und führte ihn dann herum herum. Von der Geee Chemie zeigte hin. ihm
Verner Werner die Versuche über die Klumgenbildung Klumpenbildung in Abhangegkeit Abhängigkeit
on Sauregehalt von Säuregehalt und von Hdrolihierbarkeit (Untertiedliches Ver Hydrolysierbarkeit (Unterschiedliches Ver¬
aalten halten von B Bi und Ba salzen) et. Es hätte salzen) etc. Er hatte es sehr nett Insammen zusammen

*Abbildung 19: Abgleich aus III. Abt., Rep. 14, Nr. 6895, S. 1; Modell „**Lise Meitner pur**" mit der Ground Truth (grün unterlegte Korrekturen) – CER: 14,44%, WER: 40,29%*

Auf der zweiten Seite des Schreibens ist der gleiche Trend zu erkennen, nur dass die Differenz der CERs zwischen den beiden Modellen mit über 8% noch größer ist.

Previous Advanced Compare Results

	Sco...	Type	Results
decoding 2.7.0 - Model: 57630, Lise Meitner 2.0, LM: none	Do...	HTR	CER/WER: 20,93%/57,46%
decoding 2.7.0 - Model: 57639, Lise Meitner pur, LM: none	Do...	HTR	CER/WER: 12,44%/37,06%

Abbildung 20: Abgleich im „Comparing Tool" - III. Abt., Rep. 14, Nr. 6895, S. 2 gemessen gegen die Ground Truth

Der Arbeitsaufwand für die Durchführung einer Analyse an den vier Testseiten zur Überprüfung der Genauigkeit der eigenen Modelle im Bestand betrug ***vier Stunden.***

Die am häufigsten auftretende Fehler in den HTR-Texten beider Modelle sind nicht einfach auf bestimmte Phänomene herunterzubrechen. Die Druckbuchstaben mit lateinischen Lettern der Briefköpfe und handschriftliche Abkürzungen in Druckstil wurden quasi nie erkannt, da die zugrundeliegende Groundtruth diese Schrifttypen auch nicht beinhaltete. Ansonsten kann man fast selbstverständlich feststellen, dass die Modelle bei unordentlichen oder kleineren Schriftpassagen auch deutlich mehr Mühe hatten, ähnlich wie das mnschliche Auge. Weitere typische Fehler bestanden aus der Verwechslung der Kleinbuchstaben „t" und „b", sowie der Schwierigkeit beim Differenzieren zwischen „i" und „c". Punkte der Buchstaben „i", „j" und der Umlaute wurden auch häufig nicht erkannt.

Die Arbeitszeit erhöhte sich durch den Kenntnismangel im Umgang mit den Vergleichstools vom Expert-Client und könnte bei späteren Projekten noch effizienter genutzt werden.

3.4 Test und Vergleich der Genauigkeit freier, generischer Modelle

Die GT wurde im nächsten Schritt mit den HTR-Ergebnissen der hier relevantesten öffentlichen Modelle abgeglichen. Hierfür wurden die zwei Algorithmen verwendet, die schon bei der Erstellung der Lise Meitner-Modelle als Grundmodell dienten.

Das erste Modell „Transkribus German Kurrent M1" umfasst knapp 580.000 Wörter einer großen Zahl deutscher Texte aus Kurrent, Sütterlin und Fraktur aus dem 17.-20.

Jahrhundert und gibt eine solide CER von 6% an. Es erzielte auch gute Ergebnisse (CER von 7%) bei einzelnen Seiten aus dem Trainingsset Lise Meitner.[88]

Das zweite generische, vielversprechende Modell „German Giant I“ wird vielfach von Genealogen verwendet und baute auf das Modell „German Kurrent M2“ auf.

Für den Vergleich wurden wieder die Seiten genutzt, die bereits für das Testen der eigenen Modelle verwendet wurden. Alle Ergebnisse deuten darauf hin, dass die Verbesserung der Transkriptionsqualität durch das eigene Modelltraining gegenüber den öffentlichen Algorithmen bei der gewählten Methodik und dem Bestand marginal ausfällt.

In folgendem Ausschnitt des Vergleich-Tools des Expert-Clients konnte für ***Beispiel A*** durch das Modell „Transkribus German Handwriting M1“ ein 2,3% günstigeres Ergebnis erzielt werden.

Previous Advanced Compare Results

Created	Status	Queries	Duration	Scope	Type	Results
21.12.23 12:54:22	Completed	Page(s) : 3 | Option : Quick Compare | Ref: ...	0.33 sec.	Document 170...	HTR	CER/WER: 10,56%/32,69%
21.12.23 12:43:04	Completed	Page(s) : 3 | Option : Quick Compare | Ref: ...	0.50 sec.	Document 170...	HTR	CER/WER: 12,86%/37,02%

Abbildung 21: Rot markiert - Ergebnisse German Giant in III. Abt., Rep. 14, Nr. 6893, S.11

Die GT des Samples ***Beispiel B*** konnte mit dem Modell „German Giant“ mit einer CER-Ungenauigkeit von nur 5,66% nachgebildet werden.
[*detaillierter Vergleich zum Abschneiden vom Modell „German Giant“ – Anlage 5*]

	Ground Truth	Charakters	German Giant-Modell	Charakter Errors: German Giant	Lise Meitner pur Modell	Charakter Errors: Lise Meitner pur-Modell
Summe Fehler	0	0		41		61
Zeichenanzahl gesamt:	679		672		662	
CER gesamt:	0		5,66%		10,50%	
WER gesamt:	0 (insgesamt 120 Wörter)		17,5% (21 Wörter)		32,47% (35 Wörter)	

Abbildung 22: Ergebnisse German Giant im Vergleich zu „Lise Meitner pur“ in: III. Abt., Rep. 14, Nr. 6894, S. 6

[88] Beschreibung des Vorgängermodells Transkribus German Kurrent M2: READ-Coop, o. D., German Kurrent M2. Anm. Eine Beschreibung des aktuellen Modells M1 ist online leider nicht abrufbar.

Beispiel C konnte durch kein generisches Modell besser „gelöst“ werden. Die eigens trainierten Modelle waren überlegen, allerdings blieben die Ergebnisse weiter vergleichbar. Die CER-Differenz zwischen dem Modell „German Giant I.“ und „Lise Meitner pur“ betrug 2,36%. Bei dem Test setzte sich „German Giant“ gegen das generische Modell „German Handwriting M2“ durch.

	Duration	Scope	Type	Results
del: 50870, The German Giant I, LM: none	0.76 sec.	Document 174...	HTR	CER/WER: 16,90%/40,38%
del: 57639, Lise Meitner pur, LM: none	0.47 sec.	Document 174...	HTR	CER/WER: 14,44%/40,29%

Abbildung 23: Ergebnisse German Giant und Lise Meitner pur in III. Abt., Rep. 14, Nr. 6895, S.1

Die zweite Seite des gleichen Briefs, *Beispiel D,* wurde erneut durch „Lise Meitner pur“ im Vergleich zu den generischen Modellen am zielgenauesten mit einer CER von 12,44% transkribiert. Die Differenz zum nächstbesten Modell „German Handwriting M2“ betrug 2,2%.

Previous Advanced Compare Results

	Duration	Scope	Type	Results
decoding 2.7.0 - Model: 35909, Transkribus German handwritin...	0.65 sec.	Document 174...	HTR	CER/WER: 14,62%/39,05%
decoding 2.7.0 - Model: 57630, Lise Meitner 2.0, LM: none	0.77 sec.	Document 174...	HTR	CER/WER: 20,93%/57,46%
decoding 2.7.0 - Model: 57639, Lise Meitner pur, LM: none	0.56 sec.	Document 174...	HTR	CER/WER: 12,44%/37,06%

Abbildung 24: Ergebnisse German Handwrting M1, Lise Meitner 2.0 und Lise Meitner pur in III. Abt., Rep. 14, Nr. 6895, S.2

Durchschnittlich schnitten die generischen Modelle bei den Testseiten mit einer um **2,8 Prozentpunkte** kleineren (günstigeren) CER ab.

<u>Überprüfung des Trainingsets</u>

Zur besseren Einordnung der unabhängigen Stichproben wurde alle Groundtruth noch mit HTR-Transkriptversionen der beiden untersuchten generischen Modellen abgeglichen. Dafür mussten diese Versionen generiert und durch das „Advanced Comparing Tool“ des Expert-Clients wiederum mit allen 101 Seiten GT verglichen werden.

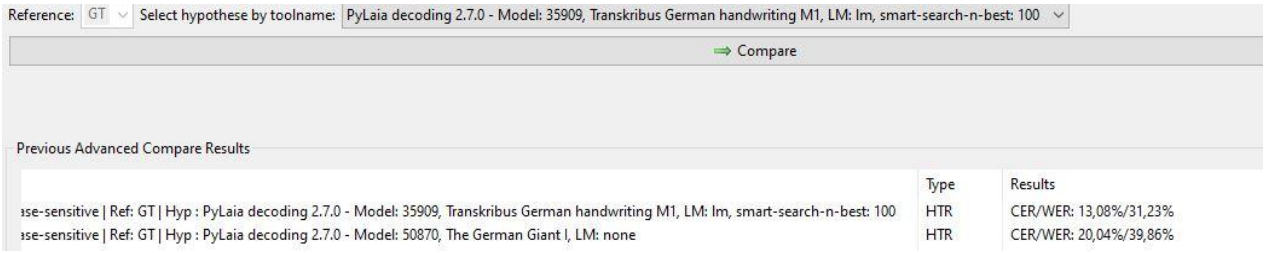


Abbildung 25: CER/WER-Wert in der Signatur II. Abt., Rep. 14, Nr. 6890 für die Modelle German Giant I. und Transkribus German Handwriting M2 (58 Seiten)

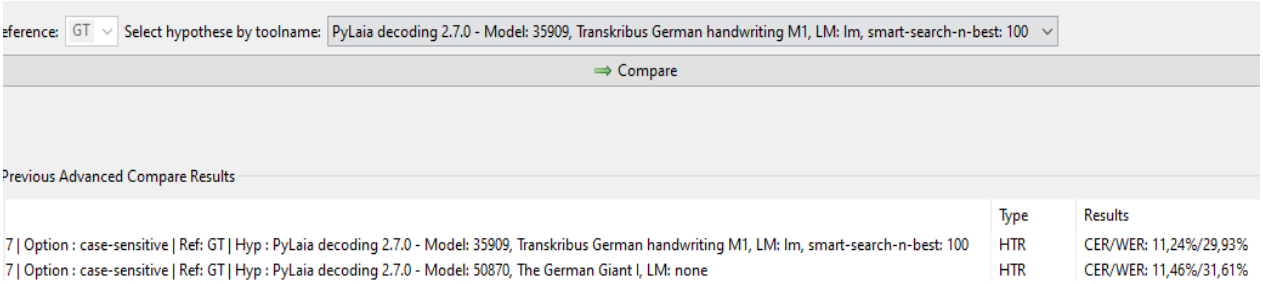


Abbildung 26: CER/WER-Wert in der Signatur II. Abt., Rep. 14, Nr. 6893 für die Modelle German Giant I. und Transkribus German Handwriting M2 (37 Seiten)

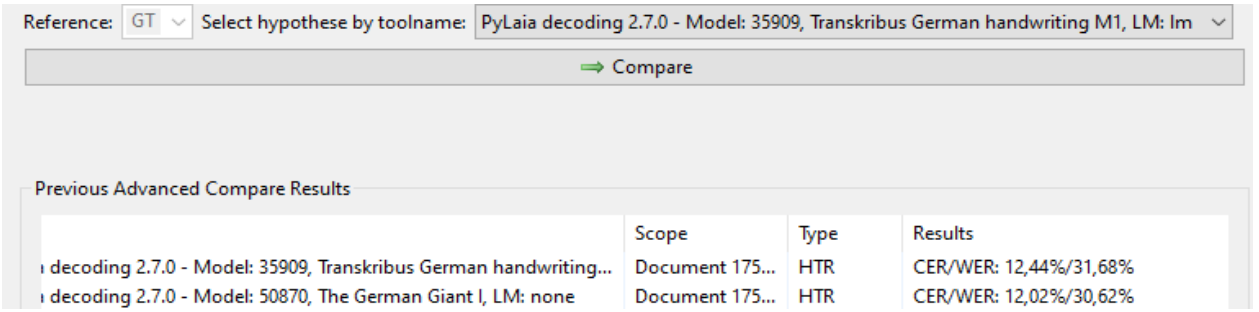


Abbildung 27: CER/WER-Wert in der Signatur II. Abt., Rep. 14, Nr. 6895 für die Modelle German Giant I. und Transkribus German Handwriting M2 (6 Seiten)

Die Ergebnisse beider Modelle rangieren im Trainingsset im Bereich von 12-15% CER und können somit grundsätzlich zum Anstoß eines Transkriptionsprojekts verwendet werden. Von beiden getesteten Modellen schloss das generische Modell „Transkribus German Handwriting M1“ mit einer CER von 12,5% am besten ab.

Die Überprüfung des Trainingssets durch die beiden generischen Modelle dauerte insgesamt **eineinhalb Stunden**.

Die Testergebnisse mit neuem Schriftmaterial waren vergleichbar oder besser als die Leistung der eigens trainierten Modelle bei der Anwendung auf neue Seiten aus dem Bestand. Hieraus leitete der Autor die Hypothese ab, dass die Verwendung eines

generischen Modells aufwandsärmer ähnliche Ergebnisse erzielen kann, wenn das bearbeitete Material wenig stetige Eigenschaften besitzt und das Trainingsset (ca. 100 Seiten) recht klein ist. Zumindest kann dies für den vorliegenden Bestand festgehalten werden.

Für die Arbeitsschritte des Experiments in Transkribus wurde durch den Autor, zusammenfassend dargestellt folgender Zeitaufwand betrieben:

Aufgaben	***Zeitaufwand in Min.***
Vorbereitung von vorhandenen Transkriptionen/ Eigenständige Transkription	*60*
Einspeisen von Digitalisaten in Transkribus	*120*
Layoutanalyse und Korrektur der Groundtruth (GT)	*1620*
Modelltraining	*240*
Evaluation	*240*
Gesamt:	*2010 = 33h + 30min*

Im nachfolgenden Kapitel werden Aufwand, Nutzen und Chancen für das Archiv dargestellt, die sich aus konsequentem Lernen aus den gemachten Erfahrungen mit Transkribus ergeben.

4. Aufwand-/Nutzenanalyse am Beispiel

4.1 Schlüsse aus den Softwaretests für die Planung einer Anwendung durch das Archiv

Der eigenständige Aufbau eines HTR-Modells für ein Teilsegment des Nachlassbestands Otto Hahn, spezieller für die Handschrift Lise Meitners, diente als Test, um den ungefähren Aufwand eines solchen Vorgehens zu ermessen, um diesen auf die Infrastruktur des Archivs der Max-Planck-Gesellschaft übertragen zu können. Dabei wurde beobachtet, dass die grundsätzlich in der Fachliteratur vermittelten Vorteile der

Verwendung eigens erstellter Modelle, die mittels vorhandener Groundtruth (GT) erzeugt werden, für kleinere Bestände, zumindest im vorliegenden Versuchsaufbau, so nicht bestehen müssen.
Das generische Basismodell „Transkribus German Handwriting“ konnte auf den Seiten des Testsets um durchschnittlich 2,8% besser abschneiden als das endgültige, individualisierte Modell. Vergleicht man die CER-Ergebnisse des Validierungssets „Lise Meitner pur“ trotz der zweifelhaften Repräsentativität für das Konvolut, so ist festzuhalten, dass hier das individualisierte Modell um ca. 6% besser abschnitt als die generischen Modelle. Aufgrund der unter 3.2 dargestellten mangelnden Belastbarkeit der CER-Ergebnisse des Modells „Lise Meitner pur“ entscheidet sich der Autor dafür, die Ergebnisse aus der Analyse der Testsets stärker zu gewichten (Kapitel 3.3, S. 40).

Der Aufwand rechtfertigte im Fallbeispiel nicht das erzielte Ergebnis von einem finalen Modell mit einer CER über 10% (im Testset). Eine gute Umsatzplanung für Transkriptionsvorhaben von kleinen Bestandssegmenten muss dementsprechend das Potenzial generischer HTR-Modelle in einem hohen Maße mit einbeziehen, wenn man sich dazu entschließt von der Technologie Gebrauch zu machen und keine Ressourcen verschwenden möchte. [89]

Die Organisation innerhalb des Archivs muss sich, je nach dafür nominiertem Bestandsegment, daran orientieren, wie Testversuche mit den generischen Modellen qualitativ (CER/WER) an vorhandener GT verlaufen. Unter diesen Vorzeichen muss der Planungsaufwand für Projekte ähnlich dem des Beispiels „Briefedition Lise Meitner“ eingeschätzt werden.
Es stellen sich für dieses Kapitel folgende Fragen:
Als wie groß ist der Planungsaufwand demnach bei der ausschließlichen Nutzung von vorhandenen im Vergleich zu eigenen HTR-Modellen zu messen? In welchem Maß müssen sich die sonstigen, etablierten Arbeitsabläufe im Archiv an ein solches Projekt anpassen? Wie zeitaufwendig ist die Durchführung der einzelnen Arbeitsschritte für das Personal?

[89] Vgl. Hodel, 2023, S. 163

4.2 Planungsaufwand und Organisation innerhalb des Archivs

Im folgenden Abschnitt wird der organisatorische Aufwand besprochen, den das Archiv leisten muss, um ein Transkriptionsprojekt wie das zuvor beschriebene professioneller durchzuführen, ohne auf zusätzliche Projektstellen angewiesen zu sein oder weite Teile des Workflows outzusourcen. Die Ausführungen beziehen sich auf die Anwendungsmöglichkeiten für die in Kapitel 2.2 angeführten Bestandssegmente.

Die aus den Erfahrungen der eigenen Versuche inspirierten, durch den Autor empfohlenen Arbeitsschritte, die einer gründlichen Planung bedürfen, bestehen aus:

- Festlegung von Zweck und Anwendungsart von Transkriptionen
- Bestimmung des Bestandsegments
- Austausch mit dem Transkribus-Team u. a. über geeignete Subskriptionsmodelle
- Ggf. Bearbeitung/Preprocessing der Digitalisate
- Einspielen von Testdokumenten in Transkribus
- Vorbereitung bereits vorhandener Groundtruth (GT) aus Editionen oder eigenständige Transkription
- Segmentierung, Layoutanalyse
- Einspielen der GT in die Digitalisate
- Testen von generischen, freien Modellen an wenigen Seiten GT

Je nach Ergebnis ergibt sich anschließend chronologisch ein Workflow auf einem der zwei möglichen Wege:

Möglichkeit 1: Eigenes Modell	***Möglichkeit 2: Generisches Modell***
Modelltraining	Nutzung von generischem Modell auf ganzes Segment
Kontrolle der HTR-Ergebnisse mit Testsets	Kontrolle der HTR-Ergebnisse mit Testsets
Korrektur Modell	Partielle Korrektur der Transkripte (aus generischem Modell)
Anwendung des eigenen Modells auf Bestand	Einleitung nächster Schritte zur Veröffentlichung – z.B. mittels „Read-and-Search“
Nachkorrektur, Einleitung nächster Schritte zur Veröffentlichung z.B. mittels „Read-and-search“	

Um für das hier betrachtete Archiv mit realistischen Zahlen zu arbeiten, bleibt der Autor bei dem Beispiel von rund 1200 Seiten von einer oder wenigen Händen aus dem frühen 20. Jahrhundert aus dem eigenen Versuchsaufbau, die für das hypothetische Projekt transkribiert werden sollen.

4.2.1 Zielsetzung und Bestandswahl

Die Bestandswahl und die Zweckbestimmung sind grundsätzliche Entscheidungen für die Leitungsebene. Im Gespräch mit dem bestandsbildenden Archivar müssen die Bestandswahl und der Zweck der Arbeit mit Transkribus erörtert werden.

Die organisatorische Planung beginnt damit eine klare Vorstellung dafür zu entwickeln, welchen Platz das doch in jedem Fall aufwendige Transkribieren von Schriftgut innerhalb der Aufgabenerfüllung des Archivs einnehmen soll.

Es gilt klarzustellen, ob man Transkripte eher für die Öffentlichkeitsarbeit, als Erschließungshilfe noch unverzeichnete Bestände oder grundsätzlich zur Erweiterung der Auswertungsmöglichkeiten bereits erschlossener Bestände als weiteres Angebot an seine Nutzenden einsetzen möchte. Auch wenn die vorliegende Arbeit sich auf die Umsetzung der letztere Option beschränkt, sollen im Folgenden der Vollständigkeit halber auch die anderen Möglichkeiten besprochen werden.
Die einzelnen Anwendungsgebiete können sich dabei auch überlappen und ergänzen. Nichtsdestotrotz führt die Entscheidung, sich auf eines der Einsatzziele zu konzentrieren, automatisch zu einer Modifikation des zuvor skizzierten Arbeitsablaufs. Sie beeinflusst selbstverständlich auch die Bestandswahl.

Ein Beispiel für ein Projekt der Öffentlichkeitsarbeit wäre z.B. im Fall des Archivs der MPG die Hervorhebung von einzelnen historisch signifikanten Autografen namhafter Nobelpreisträger (z. B.: Fritz Haber, Otto Hahn, Otto Warburg) der Max-Planck-Gesellschaft oder ihrer Vorgängerorganisation aus ihrem Bestand. Diese könnten, verknüpft mit den Bildern ihrer Digitalisate, für die Präsentation auf einer von Transkribus gehosteten „Read & Search“-Website als Eye-Catcher fungieren.[90] Sie

[90] Vgl. z. B: Datenbank zu den Regierungsratsprotokollen des Kantons St. Gallen 1803-1931 als Teil einer Online-Präsentation der Transkriptionsergebnisse mithilfe vom Angebot Transkribus/Sites: Staatsarchiv St. Gallen, o.D.

könnten anschließend zusätzlich mit dem Jubiläumsprojekt „Pioniere des Wissens“ verknüpft werden, um noch mehr Reichweite zu generieren.[91] Zeichengenaue, fehlerfreie Transkripte wären bei diesem Vorhaben das gewünschte Endergebnis, welches auf Schauwerte ausgelegt ist und neugierig machen soll.

Automatisierte Transkripte könnten als Hilfestellung zur Tiefenerschließung der bislang unerschlossenen Briefe aus dem Nachlass Carl Friedrich von Weizsäcker (AMPG, III. Abt., Rep. 111) dienen, welcher ohnehin für die Digitalisierung nominiert ist. Entscheidend ist für diesen Anwendungsfall weniger exakter Genauigkeit der HTR-Endprodukte (CER/WER), sondern die Herstellung einer annähernd guten Lesbarkeit, die dem Erschließenden Zeit dabei abnimmt, sich in den Briefen zurechtzufinden.

Als Beispiel einer praktischen Anwendungsmöglichkeit von Transkripten für die Herstellung eines barrierearmen Zugangs zu einem gefragten Teilsegment eines Bestandes für Nutzende sollen hier die Tagebücher des Kunsthistorikers Ernst Steinmann in dessen Nachlass (III. Abt., Rep. 63A, Nr. 31-56) dienen. Hierzu existieren bereits tlw. Transkriptionen eines Historikers und es ist geplant, dass der Bestand demnächst vollständig digitalisiert wird. Die wichtigsten Anforderungen an ein diesbezügliches Transkriptionsendprodukt sind hierbei Durchsuchbarkeit und die Erreichung einer CER mit einem Mindeststandard unter 10%.[92]
Für die Schaffung des vereinfachten Nutzungszugangs zu den Tagebüchern von Ernst Steinmann müssen die unter Kapitel 4.3 beschriebenen Arbeitsschritte erfolgen. Eine manuelle Arbeit in der Transkribus GUI für die Vorbereitung von Groundtruth mithilfe von Layoutanalyse- und Korrektur, sowie der Texteditierung mit den Editionsdaten ist dabei unerlässlich.

Jedes der Beispiele stellt andere Anforderungen an die Planungsarbeit des Archivs und die Qualität der Transkripte, weshalb die Aufstellung eines einheitlichen Workflows nicht zielführend wäre.

[91] Vgl.: z.B. Online-Ausstellungsprojekt „Pioniere der Wissenschaft“ anlässlich des 75. Jubiläums der Max-Planck-Gesellschaft: https://www.nobel.mpg.de/de/vermessung-des-kosmos [aufgerufen am 18.01.2024].
[92] Orientierung an den Zielen des Universitätsarchivs Greifswald für die Spruchakten der Juristenfakultät: Vgl.: Alvermann 2022, S. 188; Einschätzung von Günther Mühlberger zum Wert von Transkripten als Forschungsquelle bei einer CER über 10%: Muehlberger, 2019, S. 962.

Da alle Anwendungsfälle zwar zum Tragen kommen können aber die wahrscheinlichste Variante die Schaffung von erweiterten Auswertungsmöglichkeiten von Beständen für Nutzende ist und Ausführungen zu anderen Anwendungsfällen hier den Rahmen sprengen würden, konzentrieren sich alle weiteren Erläuterungen in dem Kapitel auf die für den Zweck organisatorischen Maßnahmen und den damit verbundenen Aufwand.

4.2.2 Zusammenarbeit mit dem Team von Transkribus und Archiven

Bevor das Archiv für die Abstimmung der Zusammenarbeit an Transkribus herantritt, gilt es die am Projekt interessierten Mitarbeitenden mit dem Einführungsmaterial zur Bedienung der Software vertraut zu machen. READ-Coop ermöglicht Ungeschulten detaillierte Anleitungen und Webinare in unterschiedlichem Detailgrad. [93]
Der Kontakt mit Transkribus erfolgt über das Service-Center. Für die Transkription von bis zu 4.000 Seiten im Jahr empfahl der Mitarbeiter Martin Zahl den „Scholar-Plan“[94].
Mit diesem Subskriptionspaket hat das Archiv neben den bereits vorgestellten Features, Zugriff weitere Exportfunktionen (METS, TEI, ALTO) der fertigen Transkripte und auf ein Veröffentlichungsportal (Transkribus Sites), mit dem eine oder mehrere Online-Datenbanken aus den eingestellten Collections der Öffentlichkeit vorgestellt werden können.[95]
Das Preismodell von Transkribus ist auf den Erwerb von „Credits“ durch die Nutzenden ausgerichtet, wobei ein Credit jeweils eine HTR-Transkription (eine Seite) wert ist. Geht man davon aus, dass die Preiskategorie „Scholar“ gewählt wird, welche 6.000 Credits (also 6.000 Seiten) für 856,90€ anbietet, beläuft sich der Preis pro automatisch transkribierter Seite auf 0,14€.[96]

Ein Austausch über speziell zugeschnittene Leistungen mit dem Team von Transkribus macht dann besonders Sinn, wenn sich die Zuständigen im Archiv für den Anschluss der Transkriptionsdaten an die eigene Erschließungssoftware entscheiden. Ein API-Zugang

[93] Anm. Anwender finden Tutorials im Help-Center von Transkribus: https://help.transkribus.org/ [Zugriff am 20.01.2024]., Übersicht zu den ersten Anwendungsschritten mit dem Interface: https://help.transkribus.org/first-steps-in-transkribus [Zugriff am 20.01.2024].
[94] READ-Coop, 2023, Plans.
[95] READ-Coop, 2023, Sites.
[96] READ-Coop, 2023, Plans.

(Programmierschnittstelle - Application Programming Interface) wird i. d. R. nur für Organisationen eingerichtet, die größere Mengen an Daten verarbeiten.

Zu Beginn des Projekts ist es zweckdienlich, mit anderen transkribierenden Archiven in Kontakt zu treten und über die Nutzung von HTR-Modellen in den Austausch zu gehen, wenn diese Handschriften ähnlichen Charakters bearbeitet haben. Das gilt für das Archiv der MPG, wie auch weitere kleine Spezialarchive, da internes Knowhow zum technischen Vorgang noch die Ausnahme ist. Für das Archiv der MPG kommen also auch die in Kapitel 1.3 vorgestellten Einrichtungen infrage.

4.2.3 Arbeitsgruppe im Archiv und Aufgabenverteilung

Um die notwendigen Arbeitsschritte und deren personelle Verteilung zu koordinieren, ist es sinnvoll, eine Arbeitsgruppe „Transkribus“ einzuberufen. Dies erfolgt nach Bekanntgabe der Bestandswahl und der Projektziele. Wer in das Projekt eingebunden wird und welche Mitarbeitende von Anfang an interessiert und geeignet sind, sollte sich aus einer Dienstberatung ergeben, die im Falle des MPG-Archivs monatlich stattfindet.[97]

Es erfolgt eine Zuordnung entsprechend der Kompetenzen. Eine Mindestanzahl von vier Personen ist bei der Vielfalt an Aufgagemben und der Größe des Archivs realistisch. Durch die Trennung von Entscheidungsebene und operativer Ebene ist ein zügiger Arbeitsablauf begünstigt. Mehr als zwei Bearbeiter im Text- und Layouteditor bei einer relativ kleinen „Collection“ würden eher zu einer gegenseitigen Behinderung als zu einer effizienteren Arbeitsweise führen.

Es ergibt sich für das Beispiel ein Stab aus technisch Zuständigen (FaMI, Informatikerin) und ein inhaltlich paläographisch arbeitender Stab (Archivar, Archivleitung). Die Leitung der Arbeitsgruppe sollte beim bestandsbildenden Archivar liegen, der die beste Kenntnis über die zu bearbeitenden Archivalien hat.

Die Bearbeitung des Schriftguts in der Transkribus-GUI kann kooperativ gleichzeitig an einer Sammlung (Collection) erfolgen, da das Programm es zulässt angemeldete

[97] Anm.: Für eine genaue Darstellung der Verteilung von Kompetenzen in dem Bsp.-Projekt siehe Kapitel 2. 1.

Anwender zur Mitarbeit über die Funktion „User Manager“ hinzuzufügen. Für die gleichzeitige Arbeit sind Regeln festzulegen und Seiten eindeutig ihren Bearbeitern zuzuordnen, um aufwendige Redundanzen zu vermeiden.[98]

In der Arbeitsgruppe muss außerdem eine Evaluationsstrategie entwickelt werden. Die Vorbereitung eines repräsentativen Testsets, welches die Heterogenität des Bestandssegments z.B. mit der Wahl der Laufzeitverteilung berücksichtigt, ist ein wichtiger Teil davon.[99]

Unabdingbar ist das Aussetzen von bestimmten Tagesgeschäften zu vereinbaren, um für beteiligte Mitarbeitende die Fokussierung auf das Projekt sicherzustellen und Überlastung entgegenzuwirken. Für einen gewissen Zeitraum muss die Erschließung bei mindestens einem Archivar bspw. zurückgefahren werden, um den Projekt-Bestand zu sichten und ggf. Besonderheiten für die Einspeisung der Digitalisate festzuhalten.

Wenn ein Digitalisierungsprojekt direkt mit dem Transkriptionsvorhaben verknüpft ist, muss das Ausschließen des Bestands, wie auch sonst bei den Vorbereitungen einer Bestandsdigitalisierung, für die Nutzung im Lesesaal kommuniziert werden.

Mit dem entsprechenden Dienstleister der Digitalisierung, hier Ossenberg Digitalisierung und Software GmbH, müssen die Anforderung an die Derivate der Digitalisats-Master übermittelt werden.

Der organisatorische Aufwand ist umfangreich. Die Auswirkungen auf andere Arbeitsabläufe müssten spürbar sein. Wie viel Zeitaufwand in den einzelnen Arbeitsschritten für das Personal steckt, zeigt der nächste Abschnitt.

4.3 Personalaufwand und Arbeitsschritte

Die einzelnen Arbeitsschritte für das Archivteam am Beispiel der Briefedition Lise Meitner und deren zeitlicher Aufwand werden in diesem Abschnitt chronologisch in einem plausiblen, aber fiktiven Szenario durchgespielt, mit dem Ziel, das Workflowkonzept auf andere ähnlich große Projekte übertragen zu können (siehe

[98] Werbedarstellung von Transkribus: READ-Coop, 2023, Collaboration.

[99] Vgl.: Mühlberger, 2018, S. 151.

Kapitel 2.2).Es handelt sich um Hochrechnungen der Arbeitszeit anhand der Arbeitsschritte aus dem eigenen Versuch, verknüpft mit den Erfahrungsbeschreibungen aus anderen kleinen Archiven, die keine exakte Genauigkeit bieten können. Die reduzierte Bearbeitungszeit, die IT-affinere Nutzer vermutlich im Gegensatz zum Autor haben, wurde in der Rechnung noch nicht berücksichtigt. Die technischen Vorgänge in der Software Transkribus werden vom Autor nur dann detailliert ausgeführt, wenn sie sich wesentlich von den in Kapitel 3 (Softwareanwendung) beschriebenen Schritten unterscheiden. Die Schritte weichen zum Teil davon ab, weil die Transkribus-Einstellungen während des Experiments nicht optimal gewählt wurden und da einige Aufgaben im Team abgewandelt durchgeführt werden sollten.

[*Übersicht über die Aufgabenverteilung – Anlage 6*]

4.3.1 Von der Initialisierung bis zur Evaluation generischer Modelle

Für die Vertragsverhandlungen bis zum Abschluss eines Scholar Plan-Abonnements mit READ Co-Op können ein paar Tage eingeplant werden. Der geeignete Plan umfasst jährlich 6.000 Credits mit dem Potenzial 6.000 Seiten zu transkribieren und 1.000 davon online über Transkribus-Sites zu veröffentlichen. Dazu muss die Informatikerin zusammen mit der Bürokauffrau Kontakt mit dem Einkaufsreferat der Generalverwaltung der Max-Planck-Gesellschaft aufnehmen und die Angebote von Transkribus nochmals prüfen.[100] Es muss mit einem Arbeitsaufwand von ***insgesamt 4-5 Stunden*** gerechnet werden. Der Arbeitsaufwand übersteigt nicht den Aufwand von sonstigen routinierten Vertragsverhandlungen, sondern ist bspw. gegenüber den deutlich kostspieligeren Digitalisierungsprojekten als recht gering zu messen, da diese immer eine über eine Ausschreibung und ein Vergabeverfahren erfolgen muss.

Anschließend ist die Installation des Programms für alle drei Nutzer im Archiv notwendig. Diese Aufgabe übernimmt die Informatikerin für sich sowie für FaMI und den bestandsverantwortlichen Archivar. Die Aufgabe kann auch von anderen Beteiligten erfüllt werden, jedoch ist das Funktionieren von Transkribus an Java gebunden. Die Installationsdatei muss über das Terminal ausgeführt werden, was eher

[100] Preismodelle von Transkribus (Angebot: „Scholar“ mit 6.000 Credits): READ-Coop, 2023, Plans.

in der Kompetenz eines Mitarbeitenden mit großer IT-Affinität liegt. Die Installation dauert in der Regel nur einige Minuten und das Programm kann für das Team entsprechend eigener Erfahrungswerte in etwas weniger als ***einer Stunde*** einsatzbereit sein.

Pro Mitarbeitenden brauch es einen eigenen Account, allerdings muss nur ein Account mit der Transkribus-eigenen Währung (Credits) aufgeladen werden. Es besteht die einfache Möglichkeit die Credits innerhalb der Arbeitsgruppe zu teilen, um dann mit HTR-Transkriptionen zu beginnen.[101] Die Durchmischung der Verwendung des Web-Interfaces mit dem Expert-Client innerhalb der Arbeitsgruppe wäre ungünstig bei Absprachen, da sich die Versionen stark, wie in Kapitel 3.1 beschrieben, unterscheiden.

Sobald das Programm einsatzbereit ist, kann eine Schulungsveranstaltung mit allen vier Projektmitgliedern durchgeführt werden. Es genügt, wenn sich dafür eine/einer der Beteiligten, vorzugsweise die Informatikerin, so intensiv mit dem Thema über Webinare, den Anleitungen auf der Website und dem Programm vertraut macht, dass sie den Kollegen in der Arbeitsgruppe, die für das Projekt entscheidenden Funktionen verständlich vermitteln kann. Alternativ kann ein Meeting direkt mit Transkribus-Mitarbeitenden gebucht werden, was dem Autor beim E-Mailkontakt mit Transkribus angeboten wurde. Die Beschäftigung mit Webinaren, Anleitungen nahmen im Fall des Autors einige Zeit in Anspruch, bis das Programm in den Grundzügen bedient werden konnte. Eine Arbeitszeit von ***drei Stunden pro Anwende*****r** (FaMI und Archivar) und ***fünf Stunden für die Informatikerin*** scheint für die Einarbeitung realistisch zu sein.

Parallel zu dem Einstiegsprozess in die Software, führt der Archivar in seiner Funktion als Projektleiter eine Bestandsrevision durch und stellt über Stichproben fest, ob das gesamte vorausgewählte Segment für die Einspeisung geeignet ist. Wesentliche Auffälligkeiten in den Veränderungen des Schriftbilds und Layout-Besonderheiten werden notiert. Das können z.B. eingefügte Tabellen oder andere grafische Strukturen sein, mit denen Transkribus bzw. der HTR-Algorithmus „Pylaia“, in der Layoutanalyse noch nicht zurechtkommt.[102] Es wird ein repräsentativer Mix der Daten für das erste Datenset zusammengestellt, welches später zu einem Trainingsset werden könnte.

[101] READ-Coop, o. D., Managing Credits.
[102] Vgl.: Mühlberger et al., 2019, S. 967.

Dieser Prozess benötigt bei den ca. 1200 Seiten ***etwa 6 Stunden***, wenn Stichproben bei jeder zehnten Seite genommen und drei Minuten pro Seite veranschlagt werden.

Das Einspielen der Images sollte wie beschrieben in hoher Auflösung erfolgen, damit auch kleinste Details durch das HTR-Modul verarbeitet werden können. Eine Orientierung an der Best-Practice des Projekts des Universitätsarchivs Greifswald „Rechtsprechung im Ostseeraum“ mit 300dpi liegt nahe.[103]
Für das Einspeisen von TIF-Dateien dieser Menge und das erstmalige Anlegen von Collections in Transkribus kann anhand des Selbstversuchs in Kapitel 3.2 von einer Dauer von ***zwei Stunden*** ausgegangen werden.

Die automatische Layoutanalyse wird isoliert als erster Bearbeitungsprozess in der Software nur gewählt, wenn vorab keine Transkripte als GT vorhanden sind, die über das Text2Image-Tool recht komfortabel auf die entsprechenden Seite Zeile für Zeile gemappt werden können.[104] Für diesen Arbeitsschritt, der im Eigenversuch nicht nachgestellt werden konnte, muss pro Seite eines Digitalisats eine entsprechende txt-Datei aus den OCR-Daten der Edition erzeugt werden. Eine Funktion des Tools ist es zwar Textstellen, die im Image nicht auftauchen zu ignorieren, diese wurde bislang aber als so fehleranfällig rezipiert, dass das manuelle Entfernen der Metadaten zeitlich sparsamer sein dürfte.[105] Für die Erzeugung dieser Dateien durch den FaMI oder den IT-kundigen Mitarbeitenden kann ***pro Datei etwa mit einem Aufwand von fünf Minuten*** gerechnet werden. Die Layoutanalyse wird durch Text2Image gleich automatisch mit ausgeführt. Es kann davon ausgegangen werden, dass das nützliche Werkzeug auch in späteren Softwareversionen als Feature mit angeboten wird.[106]
Für die Briefe von Lise Meitner kommen beide Anwendungen zum Tragen. Die GT der Briefe aus 108 Seiten mit Laufzeiten bis 1924 können über die eingescannte Edition mit dem Tool Text2Web übertragen werden. .[107]
Aus den Lehren bzgl. der Variabilität im Schriftbild Lise Meitners (siehe Kapitel 2.3) ist zu ziehen, dass weitere Seiten aus Signaturen späterer Laufzeiten des Bestandes vor

[103] Heigl, 2020.
[104] Vgl. Empfehlung von Milioni für den Worflow bei bereits vorhandenen Transkripten: Milioni, 2020, S. 19.
[105] Milioni, 2020, S. 19-20.
[106] Vgl. Erfolge durch das Züricher Staatsarchiv beim Einspielen von fast 200.00 Seiten aus dem 19. Jhd. mit „Text2Image“: Mühlberger et al., 2019, S. 964.
[107] Anm. Gemeint sind innerhalb der Edition: Ernst, 1992, S. 13-125.

einer manuellen/halbautomatischen Transkription mit einer Layoutanalyse bearbeitet werden sollten. Bei einem hier vorhandenen Grundstock von 100 Seiten, in der Datierung zusammenhängender GT, geht der Autor von mindestens notwendigen weiteren 40 Seiten aus, um für genug Varianz zu sorgen. Die Reduktion der älteren Laufzeiten ist im Verhältnis auch möglich. Für die Segmentierung wird das Standard Layout-Tool „Universal Lines“ verwendet.

Bei der Nachbearbeitung der Layoutformatierung aus den Ergebnissen des Text2Image-Prozesses kann **pro Seite** mit einem Arbeitsaufwand von weiteren **3 Minuten** gerechnet werden. Die manuelle Korrektur der Layoutanalyse für die anderen Seiten kann erfahrungsgemäß ähnlich schnell, in **4-5 Minuten** erfolgen.
Für diesen Vorgang ist die Arbeitsteilung zwischen FaMI und Informatikerin geeignet. Die Informatikerin kann die Bearbeitung der txt-Dateien und die Datenübertragung via „Text2Image“ erledigen, wogegen der FaMI die Korrekturen der Layoutanalyse bei beiden Verfahren übernimmt.

Daraus ergibt sich zusammengetragen folgende Arbeitszeit:

Verarbeitung der Trainings- und Validierungsdaten (GT):
→ 108 Seiten x 8 Minuten = ***864min (davon 324 min FaMI/ 540 min Informatikerin)***
Layoutanalyse der Seiten (ohne GT)
→ 40 Seiten x 5 Minuten = ***200 Minuten (durch FaMI))***
Die gesamte Bearbeitungszeit für die Layoutanalyse und das GT-Einspielen beträgt im Beispiel: ***1064 Minuten ≈ 17-18 Stunden***

Man beachte, dass hierbei ein Set angedacht ist, dass sowohl für die Überprüfung der Leistungsfähigkeit der generischen Modelle als auch potenziell als Trainingsset für ein eigenes Modell weiter verwendbar ist.

Im folgenden Schritt sind die Mitarbeitenden mit großen paläografischen Kenntnissen gefragt auch für die verbliebenen 40 Seiten GT zu transkribieren. Das kann bereits unter Zuhilfenahme von generischen Modellen erfolgen, deren Output noch korrigiert wird.[108] Es ist darauf zu achten bei der Erzeugung von automatischen Transkriptionen die Engine „PyLaia“ die schon vorhandenen Grundlinien, und Textregionen verwenden

[108] Vgl. Heigl, Jörn 2023, S. 58f.

zu lassen. Die Geschwindigkeit des Transkribierens ist sehr vielen Faktoren abhängig, die ohne großangelegte Studie nicht standardisiert in Zahlen zu bemessen sind. Dirk Alvermann ging davon aus, dass erfahrene „Transcriber“ unter Zuhilfenahme von generischen Modellen zu Beginn mit einer Korrekturzeit von zwei Seiten pro Stunde rechnen müssen. Seine Angaben beruhten auf den Kurrent-Akten aus dem 17-19. Jahrhundert des Wismarer Tribunals, die im Stadtarchiv Wismar transkribiert wurden.[109] Daraus ergibt sich für die beiden Archivare (Bestandsführender Archivar und Archivleitung) im Projekt eine ungefähre Arbeitszeit von ***20 Stunden***. Die geradlinige Handschrift von Lise Meitner macht eine etwas schnellere Arbeitsweise in dem Fall wahrscheinlicher.

Anschließend kann der FaMI das Testen von geläufigen, in vorigen Kapiteln vorgestellten, generischen Modellen mit einer niedrigen CER an der gesamten GT (148 Seiten) übernehmen, indem die Images in einer neuen Version automatisch transkribiert werden (siehe Kapitel 3.4). Es wird das „Comparing-Tool“ zur Berechnung der CER/WER für das zusammengestellte Set verwendet. Dabei sollte die Informatikerin bei der Interpretation des Analyse-Tools Hilfestellungen leisten. Für verschiedene Laufzeiten der digitalisierten Briefe sind ggf. auch unterschiedliche Modelle geeignet, weshalb sich der bestandsbildendende Archivar mit dem FaMI über sinnvolle Grenzsetzungen für den Wechsel von einem Modell zu einem anderen austauschen sollte, wenn Tests die Verwendung verschiedener Modelle nahelegen (z.B. eine Verwendung von „Transkribus German Handwriting M1“ für Briefe von 1924-1940 und „Modern German Handwriting (20th Century)“ für spätere Handschriften.[110]
Fallen die CER-Ergebnisse unter 10% kann entsprechend der Zielvorgaben von dem Anstoß eines unter Umständen komplexen Trainingsprozesses prinzipiell abgesehen werden.

Daneben sollte dokumentiert werden, wie die Ergebnisse einer „Fuzzy-Search“ (basierend auf Keyword-Spotting) ausfallen. [111]

[109] Alvermann 2022 ,S. 196f.
[110] Vgl. Vorgehensweise im Universitätsarchiv Greifswald bei der Erstellung von Phasenmodellen: Heigl, Jörn, 2023. S. 59.
[111] Vgl. Mühlberger, 2018, S. 153 f.

Dabei sollte nach speziellen Wörtern gesucht werden, die durch die Modelle falsch transkribiert wurden. Werden diese zielgenau gefunden, ist das entsprechend für das Modell positiv zu dokumentieren.

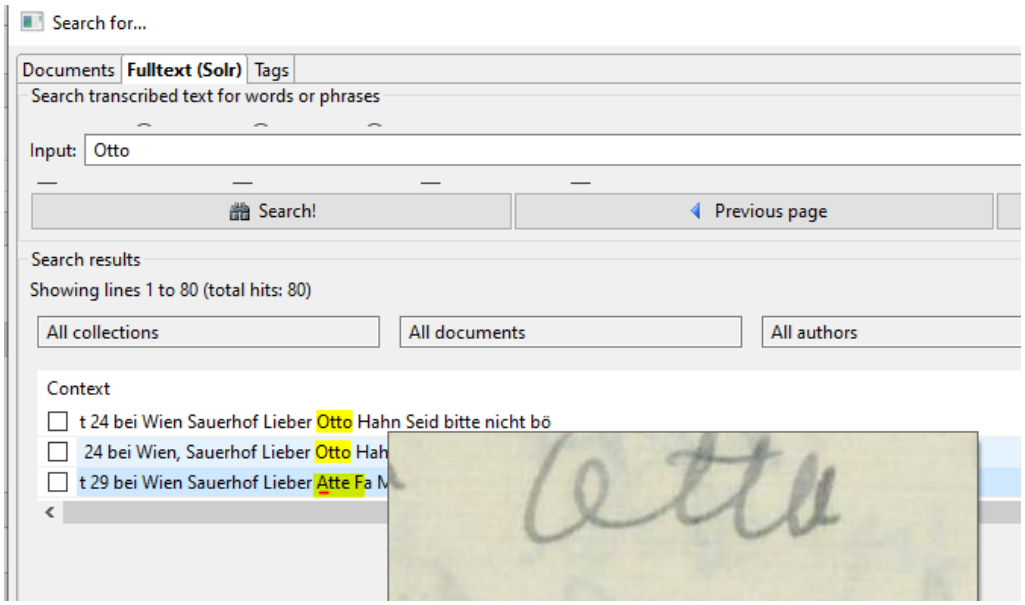


Abbildung 28: Bsp.: Fuzzy-Search im Korpus des Trainingsets „Lise Meitner 2.0" mit dem Ergebnis "Atte" für „Otto“

Die Ermittlung der CER/WER von 2 bis 3 generischen Modellen und das stichprobenartige Testen der „Fuzzy-Search“ kann innerhalb von ***3-4 Stunden*** geleistet werden.

4.3.2 Eignung des generischen Modells – Szenario A

Wenn die Ergebnisse im Bereich einer CER von 6-8% rangieren, kann die Nutzung von den am besten abscheidenden generischen Modellen auf das ganze Segment der ausgewählten Digitalisate angewendet werden. Dieser Vorgang wird abgesichert durch Kontrollen von einzelnen Testseiten mit weit auseinandergehenden Laufzeiten. Diese müssen händisch korrigiert werden. Schneiden diese ähnlich gut wie die Seiten der GT ab, kann das Transkriptionsvorhaben ohne weitere Arbeitsschritte in Transkribus als gelungen angesehen werden.

Der FaMI führt mit Unterstützung der Informatikerin die HTR des geeigneten Modells auf alle restlichen Seiten aus. Der bestandsführende Archivar testet das Potenzial der automatischen Transkription, in dem er ein Sample der zuvor generierten Transkripte von 10 Seiten korrigiert und erneut einen Fehlerquotienten berechnet.

Die HTR-Anwendung eines oder mehrerer generischer Modelle auf den ganzen Bestand nimmt ***eineinhalb Stunden Arbeitszeit*** in Anspruch.

Die Korrektur von Samples (10 S.) der generierten Seiten kann durch den Archivar innerhalb vom ***3 Stunden*** erledigt werden. Die Evaluation der Testergebnisse benötigt etwa ***zwei Stunden***.

Als nächstes können in Szenario A weitere Schritte zur Veröffentlichung der digitalen Edition eingeleitet werden, die keinen Anspruch auf hundertprozentige textliche Korrektheit erhebt.

4.3.3 Training eines eigenen Modells – Szenario B

Sollte keines der generischen Modelle ein zufriedenstellendes Ergebnis unter der CER von 10% liefern, kann ein eigenes Modell mit der bereits zahlreich vorhandenen GT trainiert werden. Die gesammelte GT kann dabei anteilig zu 90% für ein Trainingsset genutzt werden. Die 10-prozentige Auswahl für das Validierungsset wird vom Archivar hierbei so vorgenommen, dass die Diversität des Schriftbilds und die Verwendung verschiedener Schreibstoffe darin abgebildet sind, um einen genügenden Grad an Repräsentativität für das Gesamtsegment zu erreichen.

Die Informatikerin kann unter Verwendung des zuvor am bestperformenden Basismodells das Training durchführen und dabei zunächst die Standardeinstellungen von der Engine „PyLaia“ beibehalten. Die Genauigkeit des Modells wird, wie unter Kapitel 3.3 beschrieben, von Transkribus für Trainings- und Validierungssets direkt errechnet. Anhand der erzielten Ergebnisse kann erörtert werden, ob direkt ein Feintuning an den Parametern der Trainingseinstellungen erfolgen soll oder, ob die Ergebnisse so vielversprechend sind, dass man sie direkt an einem Testset mit diesem ersten eigenen Modell überprüfen soll. Dafür könnte man die Höchstgrenze einer CER von 9% wählen. Der Archivar kann zum Testen des Modells weitere 10 Seiten

Transkriptionen mithilfe des eigenen Modells erstellen und anschließend korrigieren. Danach wird wiederum die Fehlerrate dieses Testsets ermittelt.

Sollten die Ergebnisse noch in einem Bereich über 9% liegen, ist eine Korrektur bzw. Weiterentwicklung des Modells, bspw. mit modifizierter Wahl der „Epochs“ (Trainingslerndurchläufe) oder Einstellung im Pre-Processing-Bereich der nächste Schritt.

Für die Evaluation ist hier im Vergleich zu *Szenario A* ein größerer Aufwand zu betreiben, denn die verschiedenen eigenen Modelle sind nicht nur gegenübereinander, sondern auch im Vergleich zu den generischen Modellen auszuwerten.

Das Modelltraining und die Feineinstellung beansprucht die Informatikerin für etwa ***dreieineinhalb Stunden***. Die Wartezeit auf die HTR-Engine „PyLaia“ bis zum Durchlauf aller Epochs variiert stark je nach Auslastung der Transkribus-Plattform.

Die Korrektur von Samples (10 Seiten) kann vom Archivar innerhalb vom ***3 Stunden*** erledigt werden. Die Evaluation an den Testseiten und den Modellen untereinander werden durch den FaMI und der Informatikerin ausgeführt, was bezogen auf die eigene Erfahrung (Kapitel 3.2) etwa ***4 Stunden*** in Anspruch nimmt.

An dieser Stelle sind die Optionen für den weiteren Verlauf des Trainings so verzweigt, dass eine Hochrechnung der Arbeitszeit nicht weiter möglich ist.

Je nachdem, ob die Ergebnisse im zweiten Trainingsdurchlauf noch nicht zufriedenstellend sind, können weitere Modelle mit wachsender Groundtruth verwendet werden. Die Erstellung neuer GT mit den vorliegenden Digitalisaten wird unter Zuhilfenahme fortgeschrittenerer eigener Modelle ab einer CER von 10% immer einfacher und damit pro Seite schneller umsetzbar, da immer weniger Wörter korrigiert werden müssen und die Zeichenfehler aus dem Kontext der Seite heraus erkennbar werden.[112] Das bedeutet, dass sich auch der FaMI, der nur rudimentäre Kenntnisse in Paläographie besitzt, ab diesem Punkt selbst bei der Korrektur automatisch erzeugter Transkripte beteiligen kann.

Die Option einer Aufteilung der GT zur Erstellung von Phasenmodellen, wie auch in *Szenario A* beschrieben, kann in Erwägung gezogen werden, ist aber für die meisten

[112] Vgl. Hodel 2023, S. 164.

kleinen Projekte zu aufwendig, da sowohl die Trainingsdaten als auch die Zieldigitalisate genau aufgespalten werden müssten.[113]

Bei einem kleinen Projekt (1200-3000 Seiten), dass die Aufmerksamkeit von vier Mitarbeitenden bindet, ist es jedoch sinnvoll, einen klaren Schlussstrich ab einer erreichten CER zwischen 7 und 8% zu ziehen, um den oben beschriebenen Arbeitsaufwand nicht stark zu erhöhen. Eine Genauigkeit von unter 5% CER ist unter den genannten Umständen eines recht heterogenen Teilbestands unrealistisch.

Ist die Projektleitung mit dem eigenen Modell zufrieden, erfolgt durch die Informatikerin die Anwendung auf den gesamten Bestand. Eine letzte Kontrolle von Samples wird zur Qualitätssicherung durch den bestandsführenden Archivar durchgeführt. Hierbei ist wichtig, dass sowohl die Layoutausgabe als auch die Textausgabe kontrolliert wird.

Während der gesamten Testdauer ist zu beachten, dass pro automatisch generiertes Transkript ein Credit vom Transkribus-Account verbraucht wird, unabhängig davon, ob diese mit eigenen oder öffentlichen Modellen erzeugt werden. Somit sind nicht unendlich Experimentierdurchläufe möglich und eine HTR gut zu durchdenken, wenn nicht weitere Subskriptionsleistungen für den Projektzeitraum hinzugebucht werden sollen.

Es kann davon ausgegangen werden, dass bei der Modelloptimierung ***weitere 2-5 Arbeitsstunden*** auf die Informatikerin zukommen. Die Anwendung auf den ganzen Bestand fordert weitere ***eineinhalb Stunden***.

Die Evaluation der Samples durch den Archivar fordert ***etwa 3 Arbeitsstunden***.

[113] Vgl. Ausführungen zu Phasenmodellen In: Heigl, Jörn, 2023, S. 59.

4.3.4 Zusammenfassung Personeller Aufwand und Schlussfolgerungen

Nachfolgend wird der gesamte Arbeitsumfang (in h) für eine Projektlaufzeit von 6-8 Wochen für das Archiv der MPG aufgeführt:

Arbeitsaufwand der Arbeitsgruppe insgesamt → *82 ½h (Generisches Modell)*
93 ½h (Eigenes Modell)

anteilig:

- Archivleitung: Gesamt: 18h
- Archivar (Projektleitung): Gesamt: 32h / Gesamt: 33h
- Informatikerin/Zuständige für Digitalisierung: Gesamt: 17 ½h / Gesamt: 26 ½h
- FaMI: Gesamt: 15h / Gesamt: 16h

[gesamter Übersichtsplan zur Arbeitsaufteilung im Bsp. - Anlage 6]

Der geschätzte Arbeitsaufwand für die Erstellung einer kleineren Menge von 1.100 automatisch generierten Transkriptionen liegt, ungeachtet spezifischer Herausforderungen, die sich erst nach einer genaueren Durchsicht des Teilbestandes ergeben können, etwa zwischen 82 und 93 Stunden für das angestrebte Ergebnis einer CER von unter 10%.[114]

Das Beispiel aus dem Archiv der MPG kann lediglich auf Archive übertragen werden, die ähnlich strukturiert sind, einen vergleichbaren Personalaufbau haben und ebenfalls Briefe aus dem frühen 20. Jhd. transkribieren wollen.

Die Zeitersparnis gegenüber der manuellen Transkription ist offensichtlich. Die Anzahl der Arbeitsstunden könnte so verteilt werden, dass das skizzierte Projekt innerhalb von sechs Wochen abgeschlossen werden könnte. Dazu müsste im Bsp. der am stärksten beanspruchte Projektmitarbeiter (Projektleitung) etwa fünf Stunden je Arbeitswoche für das Projekt aufwenden.

[114] Anm.: Da für die das Einspielen von GT bereits 100 Seiten verwendet wurden (1.200 S. – 100 S.= 1.100S).

Auffällig ist, dass eine große Zeitersparnis, die über die alleinige Verwendung von generischen Modellen erreicht werden sollte, im vorliegenden Gedankenexperiment nicht völlig aufgegangen ist. Eine Zeitersparnis von *11 Stunden* rechtfertigt im hier gezeichneten Beispiel nicht die möglichen Qualitätseinbußen und die vergebene Chance, die GT des Bestandes zu nutzen.
Es ist jedoch zu beachten, dass der Aufwand für die hier beschriebenen Präzisionstests auch reduziert werden kann. Dazu würde entsprechend weniger GT eingespielt, weniger eigenständig transkribiert werden (usw.). Die Ergebnisse der Qualitätstests sind hiernach entsprechend weniger repräsentativ, wobei hier von Projekt zu Projekt genau über die Verhältnismäßigkeit des Aufwandes nachgedacht werden muss.

Die personellen Hürden, über die man zum erfolgreichen Einsatz der Software Transkribus kommen kann, sind gering genug, um das Potenzial der HTR auch als kleines Archiv ausschöpfen zu können. Der Grad der Beschäftigungstiefe mit Transkribus kann an die eigenen Anforderungen und Kapazitäten so angepasst werden, dass auch ohne große Projektfördermittel kleinere Projekte eigenständig bewältigbar bleiben. Der Grad des Nutzens ist eng an anschließende Anwendungs- bzw. Veröffentlichungsvorhaben geknüpft.

5. Ausblick und Fazit

5.1 Öffentliche Bereitstellung von Transkriptionen / Weitere Anwendungsarten

Die vorausgehende exemplarische Untersuchung der Machbarkeit eines Transkriptionsprojekts berücksichtigten den Weg hin zur Verfügungstellung für Nutzende noch nicht im Detail. Für die Nachnutzung der Briefe Lise Meitner bestehen noch nicht die rechtlichen und technischen Voraussetzungen für eine öffentliche Nachnutzung. Urheberrechtliche Fragen sind noch genauso in Klärung, wie das Potenzial zur Weiterentwicklung der erstellten Modelle.
Für die Tagebücher des einstigen Leiters der Bibliotheca Hertziana Ernst Steinmann (III. Abt., Rep. 63, Nr. 31-56), die im Archiv der MPG ebenfalls viel genutzt wurden, ist ein

praktischer Einsatz von Transkribus wahrscheinlicher. Es handelt sich bei dem Nachlass um einen der am meisten benutzten Bestände, das Urheberrecht ist seit langem abgelaufen, die Kurrentschrift ist schwer lesbar, aber einheitlich im Stil und eine durch einen Archivnutzer und Biografen in Aussicht gestellte Edition zu einem Teil der Tagebücher könnte den Anstoß zur Generierung von Groundtruth geben.[115]

Eine häufig genutzte Methode zur direkten Veröffentlichung der Digitalisate mit eingebetteten Transkriptionen ist der durch Transkribus angebotene kostenpflichtige Service „Transkribus Sites“.[116]

Diese Art der „Spotlightveröffentlichung“ wurde oft neben der Verknüpfung der bereits existierenden Erschließungsdaten mit den Digitalisaten in Verbundportalen eingerichtet. Im Fall der Spruchakten aus dem Universitätsarchiv Greifswald wurden die Verzeichnungsdaten des Online-Findmittels Ariadne mit den Digitalisaten und Volltexten in der Digitalen Bibliothek MV verlinkt, von wo die Images gehostet wurden.[117]

Zur Umsetzung einer möglichen Veröffentlichung von Digitalisaten mit entsprechenden Transkriptionen sollte im Archiv der MPG auf die Ressourcen und Erfahrungen von Institutionen innerhalb der Max-Planck-Gesellschaft gebaut werden. Hierbei kann bspw. mit der Bibliotheca Hertziana eine Kooperation angestrebt werden. Diese kunst- und kulturhistorische Einrichtung der MPG hat bereits ein Projekt „HumanitiesConnect“, für eigene ausgewählte (antike) Buchbestände und weiterer Bestände anderer Max-Planck-Institute mit „Transkribus Sites“ umgesetzt.[118]

Die Max-Planck-Digital Library unterstützt die Mitarbeitenden der Max-Planck-Gesellschaft mit Publikationsdienstleistungen und auf dem Gebiet des Forschungsdatenmanagements. Mit dem dortigen Know-How könnten neue Auswertungsansätze nicht nur für den HTR-Output von Transkribus, sondern auch für

[115] Anm. Joseph Imorde verfasste bereits viele biografische Einträge zu Ernst Steinmann und nutze dessen Nachlass, um Abschriften der Tagebücher zu erstellen: Vgl. z.B.: Imorde, Joseph, "Steinmann, Ernst" in: Neue Deutsche Biographie 25 (2013), S. 217-218 [Online-Version]; URL: https://www.deutsche-biographie.de/pnd117259012.html#ndbcontent [Zugriff am 27.01.2024].
[116] READ-Coop, 2023, Sites.
[117] Vgl. Universität Greifswald. o. D., Spruchakten.;
[118] Vgl. Bastianello, o.D.

die bereits existierenden OCR-Daten der zahlreichen Digitalisate (hauptsächlich Gremienbestände der Gesellschaft) ausgelotet werden.

Überlegenswert für kleine Archive im Allgemeinen ist auch Crowdsourcing zur Korrektur automatisierter Transkriptionen, wie es z.B. das Hochschularchiv (ETH-Zürich) auf der Plattform „e-maunscripta“ für einzelne Bestände vorsieht. Hierbei ist es allerdings schwierig praktikable Strategien der Qualitätssicherung zu finden, wobei aus gutem Grund auch schon längst viele HTR-generierte Texte öffentlich sind, die eben auch keine hundertprozentige Genauigkeit aufweisen.

Die Möglichkeit der Nutzung von Transkribus an Terminals in Archivlesesälen via Lizenz war ein Einsatzszenario, dass den Autor der Arbeit, im Glauben neue innovative Ansätze zu bedenken, erst näher über den Gesamtthemenkomplex nachdenken ließ. Tatsächlich ist sogar die Verwendung eigener Kameras/Smartphones im Staatsarchiv des Kanton Zürich zu diesem Zweck bereits üblich. Dort steht für optimale Lichtverhältnisse ein Scanzelt der READ-Coop zur Verfügung. „In Verbindung mit der DocScan App können Bilder direkt in die Plattform […] von READ-Coop geladen und dort mit automatischer Handschriftenerkennung (HTR-Technologie) weiter bearbeitet werden.“[119] Die rechtlichen Voraussetzung für solche Einsätze ist die Gemeinfreiheit der herausgegebenen Unterlagen, was sich z.B. wiederum schlecht für heterogene Briefbestände aus dem 20. Jhd. eignet.

HTR kann auch in der Erschließungsarbeit Potentiale entfalten, wie ein Beispiel zeigt. Das Amsterdamer Stadtarchiv führte bereits 2017 ein Projekt durch, in dem sogenannte „Vreemdelingskaarten“ (Karteikarten) der Polizeidirektion Amsterdam HTR-erfasst wurden, die Auskunft über 200.000 Flüchtlinge und Migranten aus der Zeit des dritten Reiches geben. Die Namen und Daten sind nun mithilfe von Transkribus automatisiert erschlossen worden und Teil der Verzeichnungsdaten. [120]

Noch ist die automatisierte Verzeichnung wie in diesem Beispiel nur für standardisierte Vorlagen, wie Personenstandskarteikarten oder die handschriftlichen operativen Karteikarten der Staatssicherheit der DDR denkbar und kaum für heterogene,

[119] Staatsarchiv Zürich, 2024.
[120] Vgl. Alvermann, 2022, S. 197.

handschriftliche Bestände geeignet. Jedoch ist in den nächsten Jahren wie in allen Bereichen der KI mit rasanten Weiterentwicklungen zu rechnen, die nicht nur die Nutzung innerhalb des Archivalltags prägen werden.

5.2 Fazit

Die Frage, ob sich Transkribus als Werkzeug und damit eine Kooperation mit READ-Coop für wissenschaftliche Spezialarchive oder andere kleine Archive in einer Aufwand-Nutzen-Abwägung eignet, kann auch nach der vorliegenden technischen und organisatorischen Untersuchung anhand des eigenen Workflow-Beispiels nicht einheitlich beantwortet werden. Zu unterschiedlich sind die Anforderungen und Ausgangsbedingungen sowohl im Bereich der Bestandszusammensetzung als auch in personeller und finanzieller Hinsicht von kleinen Spezialarchiven, als dass eine allgemeingültige Handlungsempfehlung gegeben werden könnte.

Die Hürden das Werkzeug Transkribus an die eigenen Bedürfnisse anzupassen, sind denkbar gering. Das Tool bietet mit seinen verschiedenen Versionsarten (Webanwendung und Expert-Client) viel Flexibilität, um sowohl für HTR-Experten als auch für Neulinge nützlich zu sein.

Für kleine Archive ist das selbstständige Training eines spezialisierten Modells an wenig umfangreichen heterogenen Beständen, die im Format sehr unterschiedlich aufgebaut sind, nur mühselig umsetzbar und kann in Summe ineffizient sein, wie im Versuchsaufbau (Kapitel 3) herausgearbeitet wurde. Die Anwendung von potenten generischen, öffentlichen Modellen kann dem eigenen Training aus Aufwandsgründen in diesen Fällen vorgezogen oder vorausgestellt werden. Das selbstständige Modelltraining via Transkribus lohnt, wenn das gewählte Bestandsegment groß genug ist, um genug Material für Modelltraining und Nachnutzung zu bieten, häufig genutzt wird und das Schriftbild homogene Eigenschaften besitzt. Der Aufwand wird minimiert, wenn bereits Editionen bzw. Transkripte für den Anstoß eines Trainings vorhanden sind. In den wenigsten Fällen kann ein HTR-Projekt als Einzelleistung neben dem regulären Archivalltag sorgfältig durchgeführt werden. Es bedarf einer genauen Zielsetzung, einer kompetenzbasierten Arbeitsteilung in technische und inhaltliche Teilbereiche und

grundsätzlich einer gewissen Einarbeitungszeit sowie persönlichen Einsatz, bis Transkribus beherrscht wird.

Speziell im Archiv der Max-Planck-Gesellschaft ist ein Einsatz der Software nur punktuell lohnend, da wenig zusammenhängende Bestandssegmente mit Handschriften existieren. Spezifisch bei den Briefen Lise Meitners ist eine Nachnutzung der eigenen Modelle, die in der Arbeit getestet wurden, nicht nur wegen der mangelhaften Fehlerquote (CER) kritisch zu betrachten. Aufgrund des unstetigen Stils der Handschrift über die Jahrzehnte und eines Modells, welches nur einen kleineren Ausschnitt dieser Schriftbildvariationen abdeckt, kann es als nicht zuverlässig genug betrachtet werden, um den restlichen Bestand der Korrespondenz ausreichend präzise zu transkribieren.

Das erstellte Workflowbeispiel kann aber auf eines der im Kapitel 2.2 genannten Autographensegmente nichtsdestotrotz Anwendung finden, sofern eine Bestandsrevision strukturell günstige Eigenschaften zutage fördert.

Literaturverzeichnis

ALTENHÖNER, Reinhard, Andreas BERGER, Christian BRACHT, Paul KLIMPEL, Sebastian MEYER, Andreas NEUBURGER, Thomas STÄCKER und Regine STEIN, 2023. DFG-Praxisregeln „Digitalisierung“. Aktualisierte Fassung 2022. [online]. 16.02.2023. [Zugriff am: 09.01.2024]. DOI 10.5281/ZENODO.7435724. Verfügbar unter: https://zenodo.org/record/7435724

ALVERMANN, Dirk, 2022. Handwritten Text Recognition als Schlüsseltechnologie für integrierte Digitalisierungs- und Erschließungsprozesse. In: Nutzung 3.0--zwischen Hermeneutik und Technologie? Beiträge zum 25. Archivwissenschaftlichen Kolloquium der Archivschule Marburg. Marburg. In: Veröffentlichungen der Archivschule Marburg, Hochschule für Archivwissenschaft, Nr. 69 Archivschule Marburg, S. 183-198. ISBN 978-3-923833-87-0

BASTIANELLO, Elisa, [kein Datum], Transkribus in der Bibliotheca Hertziana – Max-Planck-Institut für Kunstgeschichte [online]. Erfolgsgeschichten. Innsbruck: READ-COOP SCE [Zugriff am 12.01.2024]. Verfügbar unter: https://readcoop.eu/de/erfolgsgeschichten/bibliotheca-hertziana/

ERNST, Sabine, 1992. Lise Meitner an Otto Hahn, Briefe aus den Jahren 1912 bis 1924: Edition und Kommentierung / Sabine Ernst. Mit einem Geleitwort von Fritz Krafft. Stuttgart: Wiss. Verl.-Ges. ISBN 3804712541

GOOGLE IRELAND LIMITED, 2024, Cloud Vision API [online]. Handschriften in Bildern erkennen. Dublin (Irland): Google Ireland Limited [Zugriff am 25.01.2024]. Verfügbar unter: https://cloud.google.com/vision/docs/handwriting?hl=de

GRAF, Klaus, 2020. Marc Rothballer: Transkribus – Erfahrungsbericht zu maschinellem Lernen und Handwritten Text Recognition in der Heimat- und Familienforschung. Archivalia [online]. 26.06.2020. [Zugriff am: 27.07.2023]. Verfügbar unter: https://archivalia.hypotheses.org/124394

HANSESTADT WISMAR, 2024, Zeitreise Wismar, [online]. Geschichtsportal aus dem Archiv der Hansestadt Wismar. Hansestadt Wismar: Hansestadt Wismar Der Bürgermeister Thomas Beyer [Zugriff am 25.01.2024]. Verfügbar unter: https://zeitreise-wismar.de/

HEIGL, Elisabeth, 2020. Auflösung [online]. Rechtsprechung im Ostseeraum. Greifswald: Universität Greifswald, 03.03.2020 [Zugriff am 21.01.2024]. Verfügbar unter: https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/de/graphical-resolution/

HEIGL, Elisabeth und Nils JÖRN, 2023. Handschriftenerkennung als Teil der Digitalisierungs- und Erschließungsprozesse im Archiv. In: Rainer HERING (Hrsg.), Der 8. Norddeutsche Archivtag in Stralsund. Nordhausen: Traugott Bautz. S. 47–62. Bibliothemata, Band 32. ISBN 978-3-95948-604-0

HODEL, Tobias, 2023. Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft: Anwendung, Einordnung und Methodenkritik. In: Historische Zeitschrift. 01.02.2023. Bd. 316, Nr. 1, S. 151–180. ISSN 2196-680X, 0018-2613. Verfügbar unter: DOI 10.1515/hzhz-2023-0006

HOITINK, Yvette, 2023. Using AI for Transcriptions: VerledenTekst. Dutch Genealogy [online]. 10.06.2023 [Zugriff am: 21.12.2023]. Verfügbar unter: https://www.dutchgenealogy.nl/using-ai-for-transcriptions-verledentekst/World

INTRANDA GMBH, 2020, Goobi für den Workflow [online]. Göttingen: intranda GmbH [Zugriff am 25.01.2024]. Verfügbar unter: https://www.intranda.com/digiverso/goobi/workflow/

MICROSOFT CORPORATION, 2024. Azure AI/Vision Studio (Portal) [online]. Extract text from images. Redmon WA (USA): Microsoft Corporation [Zugriff am 24.01.2024]. Verfügbar unter: https://portal.vision.cognitive.azure.com/demo/extract-text-from-images

MILIONI, Nikolina, 2020. Automatic Transcription of Historical Documents. Transkribus as a Tool for Libraries, Archives and Scholars [Master Degree Project]. Uppsala. Verfügbar unter: https://www.diva-portal.org/smash/get/diva2:1437985/FULLTEXT01.pdf (Download)

MÜHLBERGER, Guenter, Louise SEAWARD, Melissa TERRAS, Sofia ARES OLIVEIRA, Vicente BOSCH, Maximilian BRYAN, Sebastian COLUTTO, Hervé DÉJEAN, Markus DIEM, Stefan FIEL, Basilis GATOS, Albert GREINOECKER, Tobias GRÜNING, Guenter HACKL, Vili HAUKKOVAARA, Gerhard HEYER, Lauri HIRVONEN, Tobias HODEL, Matti JOKINEN, Philip KAHLE, Mario KALLIO, Frederic KAPLAN, Florian KLEBER, Roger LABAHN, Eva Maria LANG, Sören LAUBE, Gundram LEIFERT, Georgios LOULOUDIS, Rory MCNICHOLL, Jean-Luc MEUNIER, Johannes MICHAEL, Elena MÜHLBAUER, Nathanael PHILIPP, Ioannis PRATIKAKIS, Joan PUIGCERVER PÉREZ, George RETSINAS, Verónica ROMERO, Robert SABLATNIG, Joan Andreu SÁNCHEZ, Philip SCHOFIELD, Giorgos SFIKAS, Christian SIEBER, Nikolaos STAMATOPOULOS, Tobias STRAUSS, Tamara TERBUL, Berthold ULREICH und Mauricio VILLEGAS, 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In: Journal of Documentation. 09.11.2019. Bd. 75, Nr. 5, S. 954–976. [Zugriff am 11.05.2023]. ISSN 0022-0418. Verfügbar unter: DOI 10.1108/JD-07-2018-0114

MÜHLBERGER, Günter, 2018. Archiv 4.0 oder warum die automatisierte Texterkennung alles verändern wird. In: Tagungsdokumentationen zum Deutschen Archivtag Wolfsburg, 2017. 2018. Bd. 22, S. 145–156

MÜNTER, Ursula [kein Datum]. Deutsche Sprache / deutsche Schrift [online]. Berlin: Münter, Ursula [Zugriff am: 20.12.2023]. Verfügbar unter: http://www.kurrentschrift.net/index.php?s=schrift

MYSCRIPT, 2023. AI, neural networks and handwriting recognition (2024). Nantes (Frankreich): MyScript SAS [Zugriff am 23.01.2024]. Verfügbar unter: https://www.myscript.com/ai/

READ-COOP SCE, [kein Datum]. 200 Jahre Bautzen'Stadtgeschichte zugänglich machen [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am: 02.01.2024]. Verfügbar unter: https://readcoop.eu/de/erfolgsgeschichten/making-200-years-of-bautzens-city-history-accessible/

READ-COOP SCE, [kein Datum]. Desktop-Client herunterladen [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am 31.12.2023]. Verfügbar unter: https://readcoop.eu/de/transkribus/download/

READ-COOP SCE, [kein Datum]. German Giant I [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am: 02.01.2024]. Verfügbar unter: https://readcoop.eu/de/modelle/the-german-giant-i/.

READ-COOP SCE, [kein Datum]. German Kurrent M2 [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am: 06.01.2024]. Verfügbar unter: https://readcoop.eu/de/modelle/german-kurrent-and-sutterlin-17th-20th-century/

READ-COOP SCE, [kein Datum]. How to Train and Apply Handwritten Text Recognition Models in Transkribus eXpert [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am 05.01.2024] Vormals verfügbar unter: https://readcoop.eu/de/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/ [Anm. verwendete Version zu finden unter: Anlagen/ how-to-train-a-handwritten-text-recognition-model-in-transkribus.html]

READ-COOP SCE, [kein Datum]. Managing Credits [online] Help Center. Innsbruck (Österreich): READ-COOP SCE [Zugriff am: 06.01.2024]. Verfügbar unter: https://help.transkribus.org/managing-credits

READ-COOP SCE, [kein Datum]. Text2Image [online]. Dokumentation. Innsbruck (Österreich): READ-COOP SCE [Zugriff am 31.12.2023]. Verfügbar unter: https://readcoop.eu/de/transkribus/docu/text2image/

READ-COOP SCE, 2023, Collaboration [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am 25.01.2024]. Verfügbar unter: https://www.transkribus.org/collaboration

READ-COOP SCE, 2023, Öffentliche AI-Modelle in Transkribus [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am 25.01.2024]. Verfügbar unter: https://readcoop.eu/de/transkribus/oeffentliche-modelle/

READ-COOP SCE, 2023, Plans [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am 25.01.2024]. Verfügbar unter: https://www.transkribus.org/plans

READ-COOP SCE, 2023, Sites [online]. Innsbruck (Österreich): READ-COOP SCE [Zugriff am 25.01.2024]. Verfügbar unter: https://www.transkribus.org/plans

RICHTER-LAUGWITZ, Grit, 2023. Jahresbericht 2022 Archivverbund Stadtarchiv/Staatsfilialarchiv Bautzen [online]. Bautzen [Zugriff am 10.01.2024]. Verfügbar unter: https://www.archivverbund-bautzen.de/fileadmin/media/archivverbund/Jahresbericht_2022.pdf (download)

RUFENACHT, Mattia, 2020. Die Evolution der Dokumentenerfassung. *Parashift* [online]. 19 Mai 2020. [Zugriff am: 23 Januar 2024]. Verfügbar unter: https://parashift.io/de/die-evolution-der-dokumentenerfassung/

STAATSARCHIV ST. GALLEN [kein Datum]. Regierungsratsprotokolle des Kantons St. Gallen 1803-1931 [online]. St. Gallen (Schweiz): Staatsarchiv St. Gallen [Zugriff am 18.01.2024]. Verfügbar unter: https://app.transkribus.eu/sites/sg

STAATSARCHIV ZÜRICH, 2024. Recherche im Staatsarchiv [online]. Archivalien fotografieren. Zürich (Schweiz): Staatsarchiv Kanton Zürich [Zugriff am 28.01.2024]. Verfügbar unter: https://www.zh.ch/de/politik-staat/recherche-im-staatsarchiv.html#-1406638641

STARKLOFF, Kristina, 2023. Nutzung von Digitalisiertem Archivgut. Berlin: Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. [Zugriff am 25.01.2024]. Verfügbar unter: https://www.archiv-berlin.mpg.de/140305/digitalisierte_archivgut

STARKLOFF, Kristina, 2024. Startseite Archiv der Max-Planck-Gesellschaft. Berlin: Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. [Zugriff am 25.01.2024]. Verfügbar unter: https://www.archiv-berlin.mpg.de/

TRANSKRIBUS, 2023. We've reached 150,000 #Transkribus users! Thank you to each and every one of you for being a part of our incredible community dedicated to preserving and unlocking our written past. In: twitter.com [online]. 04.10.2023 [Zugriff am 23.01.2024]. Verfügbar unter: https://twitter.com/Transkribus/status/1709478460391911482

TRANSKRIPTORIUM, (o.D.), Providing True Open Access to Digitalized Content [online]. Valencia (Spanien): tranSkriptorium [Zugriff am 28.01.2024]. Verfügbar unter: http://www.transkriptorium.com/news/news

UNIVERSITÄT GREIFSWALD, o. D., Rechtsprechung im Ostseeraum (1580-1871) [online]: Greifswald: Universität Greifswald [Zugriff am 24.01.2024]. Verfügbar unter: https://transkribus.eu/r/jurisdiction/

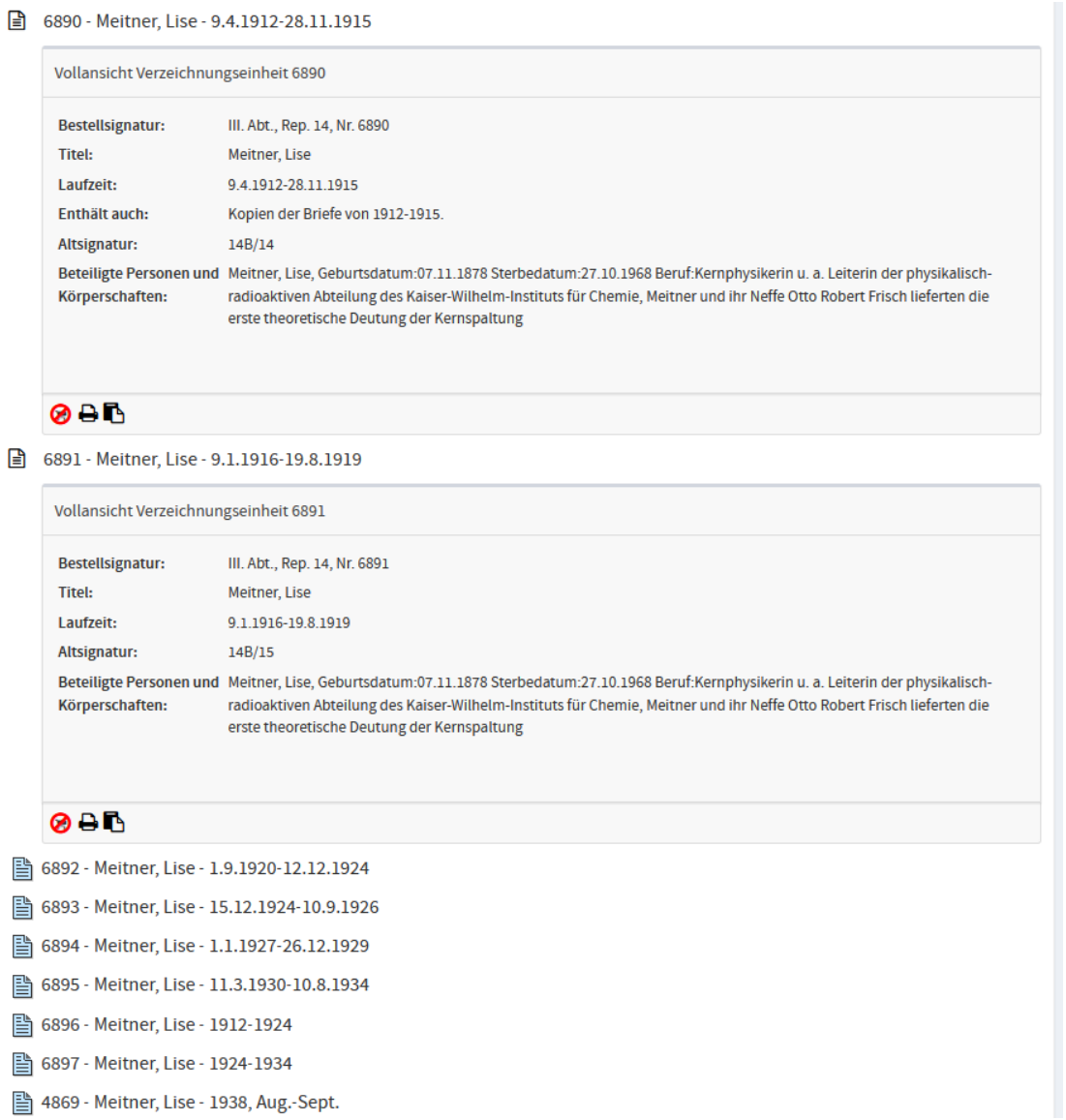
UNIVERSITÄT GREIFSWALD, [kein Datum]., Spruchakten [online]. Digitale Bibliothek MV. Greifswald: Universität Greifswald [Zugriff am 20.01.2024]. Verfügbar unter: https://www.digitale-bibliothek-mv.de/viewer/toc/PPNUAG_SprAkt/

Universität Würzburg, 2022. Historische Schriften digital erkennen [online]. Würzburg. Julius-Maximilians-Universität Würzburg [Zugriff am: 22.12.2023]. Verfügbar unter: https://www.uni-wuerzburg.de/aktuelles/einblick/single/news/historische-schriften-digital-erkennen/

WAHL, Johannes und Nadine FISCHER, 2022. Künstliche Intelligenz und neuronale Netze als Tor zur Vergangenheit – Die Verwendung von Transkribus im Hochschularchiv. ETHeritage [online]. 16.12.2022. [Zugriff am: 21 Dezember 2023]. Verfügbar unter: https://etheritage.ethz.ch/2022/12/16/kuenstliche-intelligenz-und-neuronale-netze-als-tor-zur-vergangenheit-die-verwendung-von-transkribus-im-hochschularchiv/

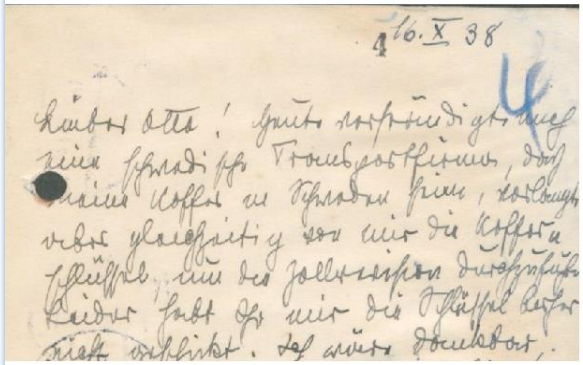
Anlagen

Anlage 1

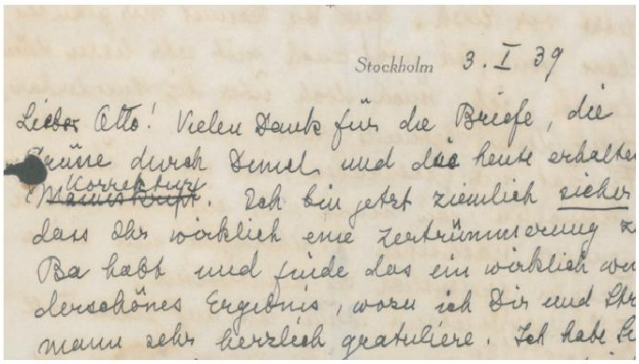


Ausschnitt des Online-Findbuchs (ActaPro-Benutzung) zum Bestand AMPG III. Abt., Rep. 14

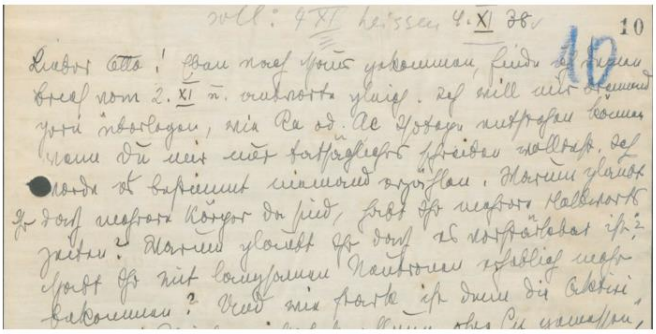
Anlage 2



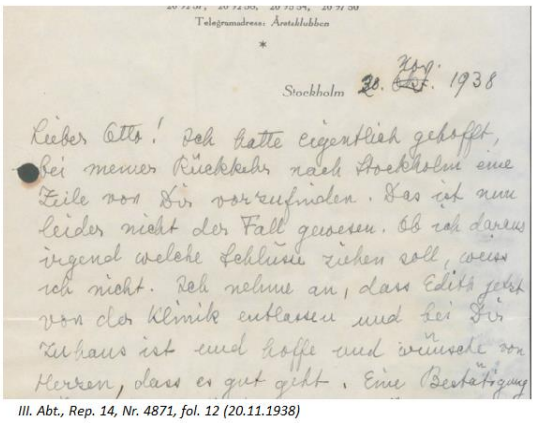
III. Abt., Rep. 14, Nr. 4871, fol. 5 (16.10.1938)



III. Abt., Rep. 14, Nr. 4875, fol. 4 (03.01.1939)



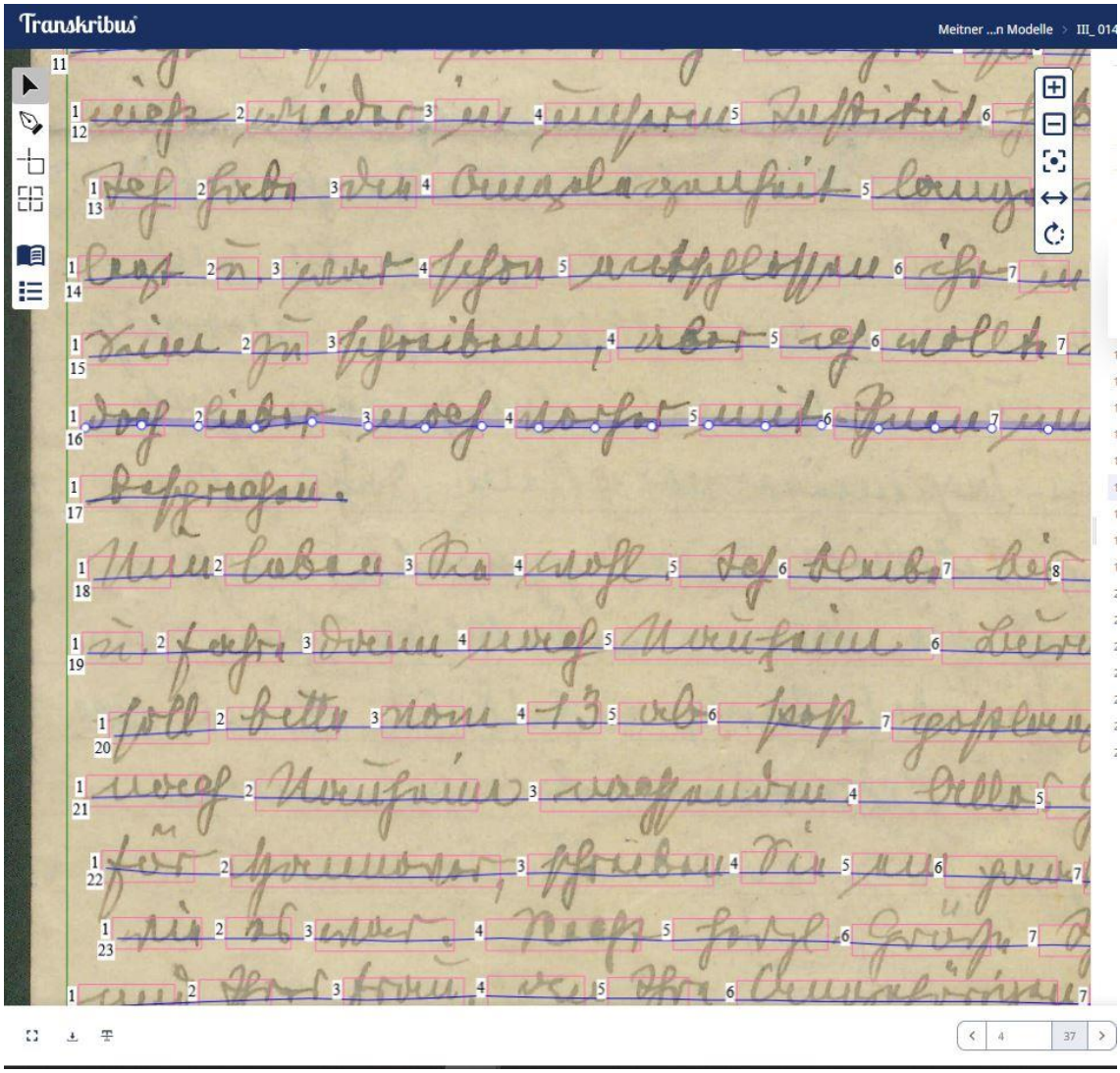
III. Abt., Rep. 14, Nr. 4871 fol. 11 (04.11.1938)



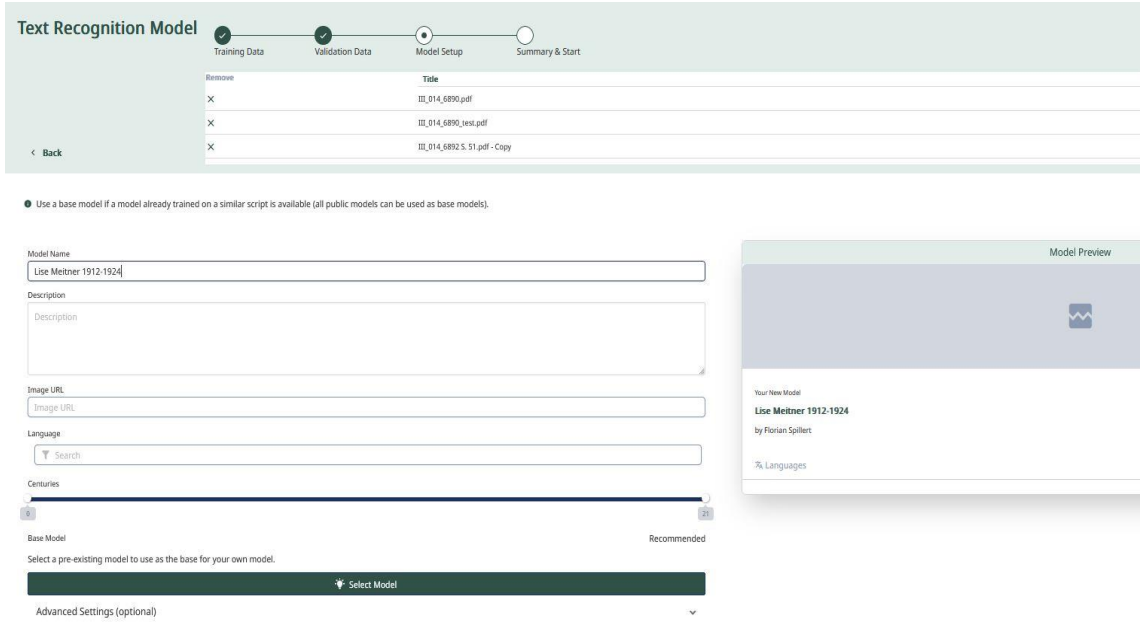
III. Abt., Rep. 14, Nr. 4871, fol. 12 (20.11.1938)

Wandel im Schriftbild von Lise Meitner (16.10.1938 – 03.01.1939), In: III. Abt., Rep. 14, Nr. 4871

Anlage 3



Darstellung von manueller Layoutanalyse in Transkribus-Lite



Trainingseinstellungen für das Modell „Lise Meitner 1914-1924“

	Ground Truth	Charakters	German Giant-Modell	Charakter Errors: German Giant	Lise Meitner-Modell	Charakter Errors: Meitner-Modell	Lise Meitner pur Modell	Charakter Errors: Lise Meitner pur-Modell
	0006	4	0006	0	A	4	AE	4
	27.XII 27	9	27.XI 27	1	2 22 Ic	6	27. E 27.	3
	Lotte Meitner	13	Lotte Meitner	0	birben 10222121	13	tu 1 122	9
	Wien I. Seilergasse 7	21	Mien I. Sellergasse 7	2	wien 2 Sellergaße F	5	Wien Z Seller gasse F	5
	Lieber Otto Hahn!	17	Lieber Atto Hahn!	1	Lieber btte Hahn	4	Lieber Otto Hahn!	0
	vielen Dank für deinen I. Brief und	35	vielen Dank für deinen I. Brief und	0	Kielen dank für denen 2 Briefunnd	5	Vielen Dank für deimnen Iben Brief und	3
	die nachgeschickten Sachen. Daß die Al-	39	die nachgeschickten Sachen. Doch die Al¬	0	die nachgeschikten Sachen Daß die 22	6	die nachgeschickten Sachen. Daß die 2a	2
	Sache eine so triviale Aufklärung ge-	37	Sache eine so triviale Auhklärung ge-	1	Sache eine so trmale Auhkklarunng ge	5	Sache eine so trwiale Aufklarung ge	3
	funden hat, tut mir Deinethalben leid,	38	funden hat, tut mir dennethalben leid	1	fuunden hat, but mir Denethallen lenc	5	funden hat, tut mir Demethalben leid,	2
	für die Physik ists aber ganz erfreulich,	38	für die sehpsite ihts aber ganz erfreulich	9	zur die sehysihe ists aber ganz erhreul	11	Für die Physik ists aber ganz erfreulich	2
	die Sache wurde sonst sehr kompliziert.	39	die Sache wurde sonst sehr kompliziort	1	die Sache mirde sonht sehr Konpliziert	4	de Sache würde sonst sehr kompliziert	3
	Schade, daß Ihr den 24. durch Hannos	36	Schade, doch Ihr den 24. durch Hannos	2	Schade, daß Ihr den 21. durch gammes	4	Schade, daß Ihr den 24. Durch gannes	4
	Krankseim verlegen mustet; hoffent-	35	Kranksem vorlegen müstet; hoffent	3	Kranksen verlegen Inßtet, hofsenit	9	Kranksen verlegen mußtet, hoffent.	2
	lich hat er und Ihr nun eine Weile	34	lich hat er und Ihr nun eine Weile	0	lich hat er und Ihr unn eine Weile	1	lich hat er und Ihr nun eine Weile	0
	Ruhe vor Erkältungen.	21	Ruhr vor Erkaltungen	2	Siche ver Frkealtunsen	17	Ruht vor Erkaltuungen	3
	Mein Vortrag in Kiel ist am 19. Jänner.	39	Mein Vortrag in Kiel ist am 19. Jänne	1	Wen Vorbrag in kiel ist am Sc ihane	21	Mein Vorbrag in Kiel ist aml 19. Jaue	6
	(Du erinnerst dich doch, daß er erst am	40	Du ernnerst doch doch, doch er erst am	4	duermmerht Dich doch, daß er erst a	6	Du erinerst Dich doch, daß er erst am	3
	12. sein sollte u. ich ihn, um das erste	40	12. sein sollte u. ich ihn, um das erste	0	12 sein sollte u ich ihn, un das erhte	4	13. sein sollte u. ich ihn, um das erste	1
	Kolleg am 13. nicht ausfallen zu	40	Kolley am 13. niht ausfallen zu	2	kolleg a 12 niht anrhallen zu	6	Kolleg am 63. nicht ausfallen zu	1
	lassen, auf den 19. verlegt habe).	34	lassen, auch den 19. vorlegt habe)	2	lasen, aich den ih vernegt habe	7	lassen, auf den 9. verliegt habe)	2
	Da ich, gerade weil ich als einzige nicht	41	da ich, gerade weil ich als einzge mehe,	5	da ich, gerade weil ich al enzige nicht	2	Da ich, gerade well ich als einzige nicht	1
	in Frankfurt war, einen Wert drauf	34	de franbefört war, einen Wort drauch	4	in grankhurt war, einen Vert drauf	3	in grankfurtwar, einen Wert drauf	2
Summe Fehler	0	0		41		148		61
Zeichenanzahl gesamt:	679		672		648		662	
CER gesamt:	0		5,66%		21,80%		10,50%	
WER gesamt:	0 (insgesamt 120 Wörter)		17,5% (21 Wörter)		45,8% (55 Wörter)		32,47% (35 Wörter)	

Vergleich zum Abschneiden von generischen Modellen mit eigenen Modellen in: III. Abt., Rep. 14, Nr. 6894, S.6

Anlage 6 (1 von 2)

Übersichtsplan der Aufgabenverteilung (Ziel: 1200 Transkripte)

Alternative Workflow-Wege → Generisches Modell (grün)
Eigenes Modell (rot)

Arbeitsaufwand der Arbeitsgruppe insgesamt *82 ½h (Generisches Modell)*
93 ½h (Eigenes Modell)

Mitarbeitende	***Aufgaben***	***Arbeitsaufwand in Stunden (h)***
Archivleitung	- Entscheidung über Ziele - Archivalienauswahl - Koordination - Manuelles Transkribieren (20 Seiten)	Ca. 3h 2h ca. 3h 10h ➢ Gesamt: 18h
Archivar (bestandsverantwortlich)	- Einarbeitung Software - Leitung der Arbeitsgruppe - Bestandsrevision - Auswahl der Archivalien - Manuelles Transkribieren (20 Seiten) - Korrektur von automatischen Transkriptionen - Evaluation der Tests (generisches Modell) - Anlegen von Validierungsset - Korrektur von Samples (Produkte des eigenen Modells)	3h Ca. 5h 6h 2h 10h 3h 3h 1h 6h ➢ Gesamt: 32h ➢ Gesamt: 33h

Anlage 6 (2von2)

Mitarbeitende	***Aufgaben***	***Arbeitsaufwand in Stunden (h)***
Informatikerin/Zuständige für Digitales Langzeitarchiv	- Einarbeitung Software - Installation vom Transkribus (Expert Client) - Einspielen der Digitalisate - Mapping von GT mit Text2Image/Txt-Erstelltung und Ordnerstruktur - Anwendung: generische Modelle auf Bestand - Training eines Modells - Evaluation: Testset und Modelle untereinander - Modelloptimierung - Anwendung eigenes Modell auf Bestand - Einrichtung von Transkribus-Sites	5h 1h 2h 9h ½ h 3 ½ h 2h 2 ½ h 1 ½ h (nicht untersucht) ➢ Gesamt: 17 ½h ➢ Gesamt: 26 ½h
FaMI	- Einarbeitung (Software) - Layoutanalyse und Korrektur - Testen von generischen Modellen - Anwendung von generischem Modell auf Bestand - Evaluation: Testsets und Modelle untereinander	3h Ca. 9h Ca. 2h 1h 2h ➢ Gesamt: 15h ➢ Gesamt: 16h
Bürokauffrau/Sekretariat (nicht Teil der Arbeitsgruppe)	*- Verhandlungen und Vertragsabschluss mit Transkribus* *Kontakt zum Support*	5h (gesamt)

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit eigenständig und ohne fremde Hilfe angefertigt habe. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Berlin, 28.01.2024

Florian Spillert