Altbestandserschließung mit KI-Anwendungen

Ein Werkstattbericht der Universitätsbibliothek Tübingen

Dorothee Huff und Kristina Stöbener

Abstract

An der Universitätsbibliothek Tübingen wird seit 2019 maschinelle Texterkennung (HTR) eingesetzt, um handschriftliche Dokumente zu transkribieren und durchsuchbar zu machen. Da die automatischen Transkriptionsergebnisse nicht fehlerfrei sind, wurde getestet, diese mithilfe von Large Language Models (LLMs) wie ChatGPT zu verbessern und aufzuwerten. Anschließend wurde überlegt, inwieweit die erzeugten Volltexte für eine (semi-) automatische Erschließung nachgenutzt werden können. Anwendungen wie ChatGPT zeigen hier Potential, auch nicht fehlerfreie automatisch generierte Volltexte für die Erzeugung von Inhaltsangaben und Schlagworten zu nutzen. Anhand von Beispielen soll gezeigt werden, wie ein derartiges Vorgehen aussehen kann und welche Ergebnisse zu erwarten sind.

Automatic text recognition (HTR) has been used at Tübingen University Library since 2019 to transcribe handwritten documents and make them searchable. As the automatic transcription results are not error-free, tests were carried out to improve and enhance them using Large Language Models (LLMs) such as ChatGPT. Further on, the extent to which the generated full texts can be reused for (semi-)automatic indexing was considered. Applications such as ChatGPT show potential for using even non-error-free automatically generated full texts to generate content information and keywords. Based on examples we will show how such a procedure can look like and what results can be expected.

1. Einführung

In den letzten Jahrzehnten haben Bibliotheken große Anstrengungen unternommen, ihre historischen Bestände durch Digitalisierung der Originaldokumente zeit- und ortsunabhängig online zugänglich zu machen. Die erzeugten Bilder verlagern jedoch zunächst nur die Betrachtung des Originals auf den Bildschirm, ohne dass eine Umwandlung zu Daten stattfindet, die ein Computer nicht nur widerspiegeln, sondern auch weiterverarbeiten kann. Um im Kontext der zunehmend datengetriebenen wissenschaftlichen und gesellschaftlichen Entwicklungen relevant zu bleiben, müssen Bibliotheken an dieser Stelle tätig werden und Teil der digitalen Innovation sein.[1] Nach der Bereitstellung der Bilddateien ist der nächste logische Schritt die Anreicherung der Digitalisate im Präsentationssystem einer Bibliothek mit einem durchsuchbaren und prozessierbaren Volltext, um den steigenden Ansprüchen an die Präsentation und Nachnutzbarkeit der Daten zu entsprechen.[2] Die Universitätsbibliothek Tübingen beschäftigt sich mittlerweile seit 2019 mit dem Thema der automatischen Texterkennung von Handschriften (HTR) mit dem Ziel, historische Bestände aus der Universitätsbibliothek und dem Universitätsarchiv mithilfe von Volltexten zusätzlich zur digitalen Präsentation für die Nutzerinnen und Nutzer leichter zugänglich zu machen und neue Forschungsfragen sowie die Bearbeitung von großen Textmengen zu ermöglichen.[3]

Historische Bestände sind oft schwach strukturiert und wenig inhaltlich erschlossen. Auch fehlen ihnen nicht selten erschließende Parameter wie Register, die einen Rückschluss auf die Textinhalte erlauben würden. Eine intellektuelle Verschlagwortung ist aber ein aufwendiges Unterfangen. Es steht daher zu klären und beispielhaft zu testen, ob die erzeugten Transkriptionsdaten von handschriftlichen Materialien nicht nur den Nutzerinnen und Nutzern für die wissenschaftliche Weiterverwertung bereitgestellt, sondern auch von der Bibliothek selbst mit weiteren Anwendungen der sogenannten Künstlichen Intelligenz (KI) aus dem Bereich des Machine Learning (ML) nachgenutzt werden können. Zunächst soll geprüft werden, ob sich automatische Transkriptionsdaten auf diese Weise aufwerten lassen. Zudem soll überlegt werden, ob und inwieweit die erstellten Volltexte auch zu einer Tiefenerschließung der Dokumente herangezogen werden können. Anwendungen wie ChatGPT zeigen hier Potential, auch nicht fehlerfreie automatisch generierte Volltexte für die Erzeugung von Inhaltsangaben sowie die Extraktion von Schlagworten, Personen- und Ortsangaben zu nutzen. Anhand von Beispielen soll gezeigt werden, wie ein derartiges Vorgehen aussehen kann und welche Hindernisse und Ergebnisse zu erwarten sind.

1 Vgl. Candela, Gustavo/ Sáez, María Dolores/ Escobar Esteban, MPilar/ Marco-Such, Manuel: Reusing digital collections from GLAM institutions, in: Journal of Information Science 48,2 (2022) S. 251-267, Online: https://doi.org/10.1177/0165551520950246, S. 251.

2 Vgl. Neudecker, Clemens/ Zaczynska, Karolina/ Baierer, Konstantin/ Rehm, Georg/ Gerber, Mike/ Schneider, Julián Moreno: Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten, in: Franke-Maier, Michael/ Kasprzik, Anna/ Ledl, Andreas/ Schürmann, Hans (Hrsg.): Qualität in der Inhaltserschließung (Bibliotheks- und Informationspraxis 70) Berlin 2021, S. 137-166, Online: https://doi.org/10.1515/9783110691597-009, S. 137.

3 Vgl. Huff, Dorothee/ Stöbener, Kristina: Projekt OCR-BW. Automatische Texterkennung von Handschriften, in: O-Bib. Das Offene Bibliotheksjournal 9,4 (2022) S. 1-19, Online: https://doi.org/10.5282/o-bib/5885.

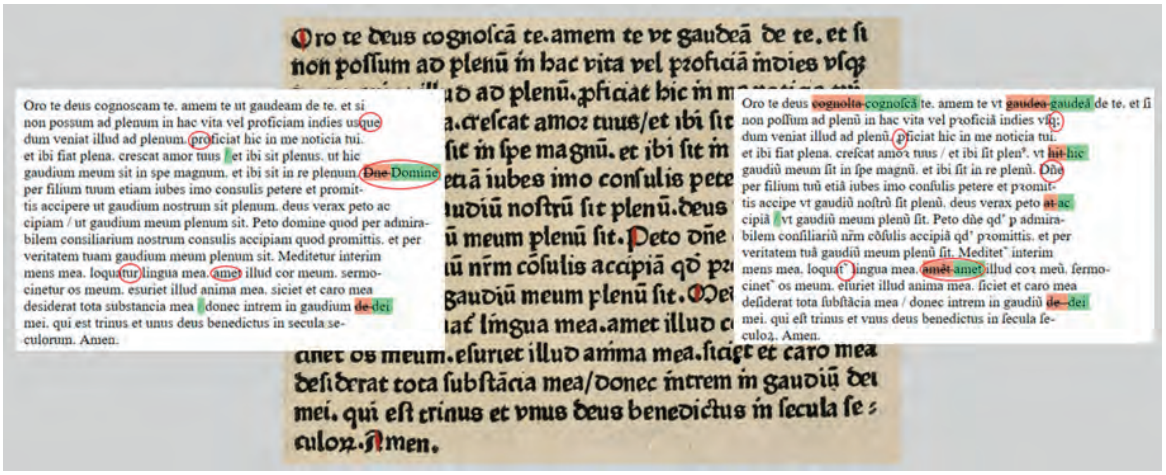


Abb. 1: Auf der linken Seite das Ergebnis der Texterkennung mit dem normalisierten Modell, rechts das des diplomatischen Modells (Tübingen, UB, Gb 321.2, Teil 2).

2. Weiterverarbeitung automatischer Transkriptionsdaten mit KI

2.1 Aufwertung von automatischen Transkriptionsdaten

Die UB Tübingen hat für verschiedene Textkorpora Ground-Truth-Daten erstellt, anhand dieser Texterkennungsmodelle erzeugt und mit den Modellen weitere Seiten und Dokumente erkannt. Diese automatischen Transkriptionsdaten sind zum einen in der Regel selbst bei noch so gutem Ergebnis nicht fehlerfrei und liegen zum anderen gemäß der im Projekt OCR-BW verfolgten Transkriptionsrichtlinien in der Regel zeichengetreu unter Beibehaltung von Sonderzeichen und Abkürzungen vor.[4] Beide Aspekte erschweren eine Volltextsuche im Dokument. Im Idealfall würde die UB Tübingen den Nutzerinnen und Nutzern sowohl eine diplomatische wie auch eine normalisierte Version zur Verfügung stellen, die hinsichtlich ihrer Fehler korrigiert worden sind, was bisher nicht zu leisten war. Hier bieten Large Language Models (LLMs) wie ChatGPT[5] potentiell neue Möglichkeiten für die Aufwertung von automatischen Transkriptionsdaten. Im Folgenden soll beispielhaft gezeigt werden, inwieweit Künstliche Intelligenz zur Korrektur fehlerhafter Transkriptionen sowie zum Wechsel des Transkriptionslevels eingesetzt werden kann. LLMs scheinen für diese Aufgabe grundsätzlich prädestiniert zu sein, da ihre Stärke im Erzeugen von Text anhand der Berechnung wahrscheinlicher Wortfolgen besteht:[6] „HTR – itself a product of ML – has potential for further integration with AI tools and systems."[7] Von daher sollte es im Rahmen des Möglichen liegen, dass auf diese Weise ein Text erzeugt werden kann, der richtiger ist als eine fehlerbehaftete Transkription mit Sonderzeichen. Inwieweit sich dies realisieren lässt und ob Halluzinationen auftreten, wird zu prüfen sein.

Als erstes Beispiel dient ein Inkunabelkorpus, für welches als Vergleichsgrundlage je ein Set normalisierter und diplomatischer Transkriptionsdaten erstellt wurde, auf deren Grundlage entsprechende Texterkennungsmodelle trainiert worden sind. Mit dem diplomatischen Modell erzeugte Transkriptionsdaten wurden mit dem Ziel in ChatGPT eingegeben, diese zu korrigieren und zu normalisieren. Das Prompting wurde in mehreren Schritten verfeinert, bis das Ergebnis den Erwartungen entsprochen hat. So musste neben der grundlegenden Bitte um eine Normalisierung der Sonderzeichen und Abkürzungen sowie um eine Textkorrektur spezifiziert werden, dass die zeilengenaue Formatierung[8] beibehalten werden sollte, dass der lateinische Text nicht ins Deutsche übersetzt werden sollte sowie dass die vorliegende Orthographie und Groß- und Kleinschreibung beibehalten werden sollten.

Auf der ersten der beiden Vergleichsseiten des Inkunabeldatensatzes (Tübingen, UB, Gb 321.2, Teil 2, Bl. 83r) liegt die Zeichenfehlerrate (CER) bei der automatischen Texterkennung mit dem normalisierten Modell bei 0,92 % und mit dem diplomatischen Modell bei 1,17 %. Eine Berechnung der CER für die von ChatGPT korrigierte Transkription ergab zunächst eine Fehlerrate von 2,17 % und somit keine Verbesserung, sondern im Gegenteil eine Verschlechterung. Eine genauere Prüfung der Daten zeigte jedoch, dass die im Abgleich mit den normalisierten Ground-Truth-Daten gefundenen Fehler nicht unbedingt falsch waren. Wie schon bei der Darlegung des Prompting-Prozesses angedeutet, korrigiert ChatGPT mitunter mehr, als eigentlich gewünscht ist, was sich auch durch

4 Vgl. Huff/ Stöbener, Projekt OCR-BW, 2022, S. 4-5.

5 Im Rahmen der vorgestellten Tests wurde die Version ChatGPT 4o eingesetzt.

6 ChatGPT und vergleichbare Anwendungen sind Large Language Models, die darauf trainiert worden sind, Text zu erzeugen. ChatGPT wurde für die dargelegten Tests verwendet, weil es zu der Zeit die Anwendung war, mit welcher sich importierte Dateien in verschiedenen Formaten am besten verarbeiten ließen.

7 Siehe Nockels, Joseph/ Gooding, Paul/ Terras, Melissa: The implications of handwritten text recognition for accessing the past at scale, in: Journal of Documentation 80,7 (2024) S. 148-167, Online: https://doi.org/10.1108/JD-09-2023-0183, S. 156.

8 Dies ist wichtig, damit die Daten ohne Aufwand wieder in die Transkriptionsplattform beziehungsweise das Präsentationssystem eingefügt werden können.

Universität Marburg

TU München

Universitätsbibliothek Salzburg

Berlin-Brandenburg International School

zambelli

EINFACH MACHEN. AUS METALL.

Zambelli Bibliotheken

Lernen und Wohlfühlen

Die Zambelli Bibliothekseinrichtungen begleiten wissenschaftliche und öffentliche Bibliotheken, die sich mit neuen Gegebenheiten auseinandersetzen und sich weiterentwickeln wollen. Wir helfen Ihnen Ihre Bibliothek so auszustatten, dass attraktive und funktionale Lernräume entstehen. Dabei können Sie sich auf in Sicherheit und Funktion bewährte Einrichtungs lösungen verlassen.
Gemeinsam schaffen wir gestalterisch-kreative Raumkonzepte.

Nehmen Sie mit uns Kontakt auf!
regalsysteme@zambelli.com

www.zambelli.com/regalsysteme

das entsprechende Prompting nicht ganz vermeiden ließ. So sind viele Fehler auf Unterschiede in der Groß- und Kleinschreibung zurückzuführen. Darüber hinaus hat ChatGPT die gewünschte Normalisierung und Korrektur mitunter noch einen Schritt weitergeführt und zusätzliche Eingriffe in den Text vorgenommen. So wurden beispielsweise mittelalterliche orthographische Besonderheiten der lateinischen Sprache wie die Ersetzung von „t" durch „c" oder von „ae" durch „e" wieder ins klassische Latein umgewandelt. Wenn das Ergebnis hinsichtlich der-

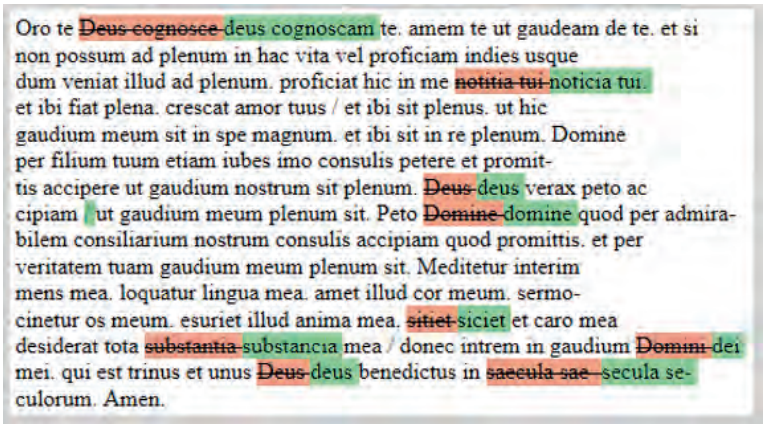

Oro te ~~Deus cognosce~~ deus cognoscam te. amem te ut gaudeam de te. et si
non possum ad plenum in hac vita vel proficiam indies usque
dum veniat illud ad plenum. proficiat hic in me ~~notitia tui~~ noticia tui.
et ibi fiat plena. crescat amor tuus / et ibi sit plenus. ut hic
gaudium meum sit in spe magnum. et ibi sit in re plenum. Domine
per filium tuum etiam iubes imo consulis petere et promit-
tis accipere ut gaudium nostrum sit plenum. ~~Deus~~ deus verax peto ac
cipiam / ut gaudium meum plenum sit. Peto ~~Domine~~ domine quod per admira-
bilem consiliarium nostrum consulis accipiam quod promittis. et per
veritatem tuam gaudium meum plenum sit. Meditetur interim
mens mea. loquatur lingua mea. amet illud cor meum. sermo-
cinetur os meum. esuriet illud anima mea. ~~sitiet~~ siciet et caro mea
desiderat tota ~~substantia~~ substancia mea / donec intrem in gaudium ~~Domini~~ dei
mei. qui est trinus et unus ~~Deus~~ deus benedictus in ~~saecula saec~~ secula se-
culorum. Amen.

Abb. 2: Das von ChatGPT erzeugte unbereinigte Ergebnis im Vergleich zu den normalisierten Ground-Truth-Daten (Tübingen, UB, Gb 321.2, Teil 2).

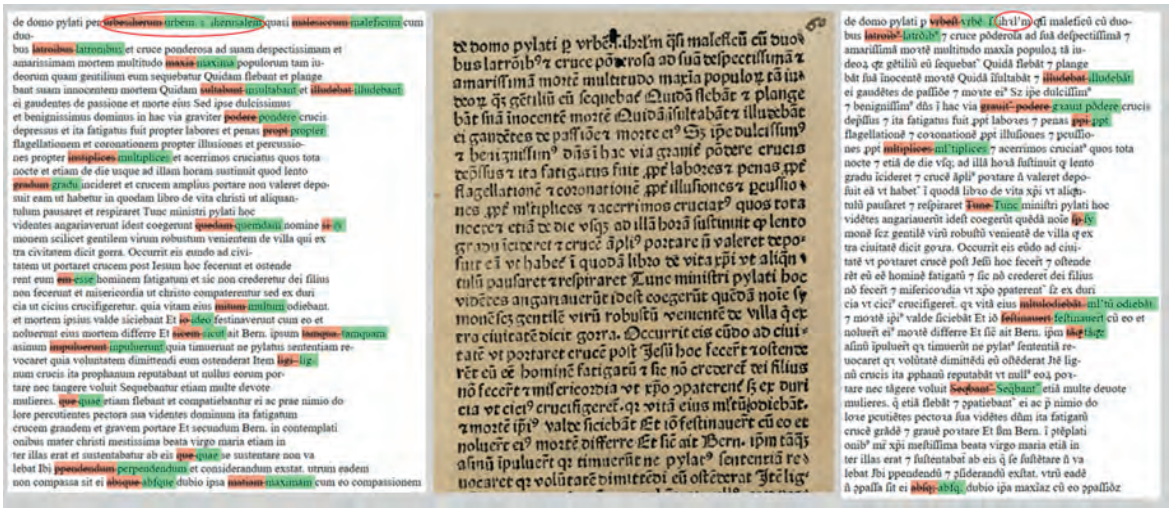


Abb. 3: Auf der linken Seite das Ergebnis der Texterkennung mit dem normalisierten Modell, rechts das des diplomatischen Modells (Tübingen, UB, Gb 461, Teil 1).

artiger zusätzlicher Eingriffe bereinigt wird, die zwar unterschiedlich aber eigentlich nicht falsch sind, stellt sich das Ergebnis mit einer CER von 0,92 % deutlich besser dar. Somit wäre man in diesem Fall ohne den Mehraufwand der Erstellung eines weiteren Ground-Truth-Datensatzes von der diplomatischen Transkription zu einer normalisierten Version gelangt, die in Bezug auf die Fehlerrate genauso gut war wie das entsprechend trainierte normalisierte Modell. Neben den bereits erwähnten Eingriffen in den Text werden zudem mitunter Wörter durch Synonyme ausgetauscht wie hier „dei" durch „Domini", was weiterhin als Fehler gewertet wurde, da dies nicht nur einen ortho-

graphischen, sondern einen inhaltlichen Eingriff in den Text darstellt, auch wenn das Textverständnis davon nicht beeinträchtigt wird.

Auf der zweiten Beispielseite des diplomatischen Modells (Tübingen, UB, Gb 461, Teil 1, Bl. 60r) zeigen sich die eben angeführten Tendenzen noch deutlicher. Hier liegt die CER der ChatGPT-Version im Vergleich zu den normalisierten Ground-Truth-Daten bei 6,62 %. Nach der Bereinigung um Fehler, die eine Volltextsuche nicht einschränken würden, weil sie nur die Groß- und Kleinschreibung, Zeichensetzung und Anpassung an das klassische Latein betreffen, sinkt der Wert auf 3,88 %. Damit ist die CER in diesem Fall weiterhin signifikant höher als jene, die das normalisierte Modell erreicht hat (2,11 %), während die Zeichenfehlerrate des diplomatischen Modells mit 1,09 % fast um die Hälfte niedriger ist.

Interessant ist, dass der Text (Abb. 4) so gut wie keine Rechtschreibfehler im Sinne von nichtexistenten Wortformen enthält. Fast alle falsch erkannten Wörter sind nicht per se orthographisch falsch, sondern geben eine im Vergleich zum Original falsch eingesetzte Wortform wieder. Dies betrifft oftmals Verbformen und Fallendungen von Substantiven und Adjektiven. Darüber hinaus wurden aber auch wieder Wörter komplett ausgetauscht. Völlig falsch ist hier zum Beispiel die Ersetzung von „mater" mit „mirabili". Letzteres scheint nur auf den ersten Blick ein passendes Adjektiv zu „Christi" zu sein, erweist sich jedoch aufgrund der unterschiedlichen Deklination als nicht richtig. Zum vorhergehenden Wort „contemplationibus", was ChatGPT durch das Synonym „meditationibus" ausgetauscht hat, passt zwar der Kasus, aber nicht der Numerus. Mitunter ist die Überkorrektur durch ChatGPT jedoch auch vorteilhaft. So wurde sogar ein Druckfehler gefunden und „abfque" richtig zu „absque" umgewandelt. Auch für die Suche nach Named Entities zeigen sich Vorteile. Im Gegensatz zum Ergebnis der automatischen Texterkennung mit dem normalisierten und dem diplomatischen Mo-

dell hätten Nutzerinnen und Nutzer wahrscheinlich eine höhere Chance, „Jerusalem“ mit der ausgegebenen Form „Ierusalem“ in Einklang zu bringen.[9] Auch das Ausschreiben des Namens „Bernardus“ ist womöglich suchfreundlicher als die Abkürzung „Bern.“. Überhaupt erleichtert die nicht dem Prompt entsprechende Großschreibung von Eigennamen das Lesen. Auch die Umwandlung in klassisches Latein könnte ein Vorteil bei der Suche nach Begriffen sein, da die klassische Schreibweise in der Regel die in Wörterbüchern angelegte Form ist.

Genauso setzt sich bei der Korrektur automatischer Transkriptionsdaten deutscher Texte der Eindruck fort, dass ChatGPT es zwar schafft, einen fehlerfreien Text zu produzieren, dieser jedoch inhaltlich vom Original abweichen kann. Ein Extrembeispiel ist eine mit ChatGPT korrigierte automatische Transkription eines Expeditionstagebuchs von Edwin Hennig (UAT 407/80). Hier sah sich ChatGPT anscheinend berufen, aus der Alltagsbeschreibung eines Paläontologen auf Expedition eine Art Fantasyroman zu kreieren (immerhin ohne Rechtschreibfehler).

Ein weiterer Durchlauf mit einem Prompt, der auf eine reine Korrektur der Rechtschreibfehler ohne inhaltliche Änderungen abzielte, führte schließlich zum Erfolg. Auch hier ließen sich zwar trotz entsprechendem Prompting weitere Eingriffe durch ChatGPT wie eine Anpassung an die heutigen Rechtschreibregeln nicht verhindern, was zum Beispiel die Verwendung von „ss“ und „ß“ betrifft. Auch durch einen Punkt gekürzte Begriffe wie „d.“ für „die“ wurden teilweise selbstständig vervollständigt.[10] Um diese Aspekte für die Vergleichbarkeit bereinigt, ist das Ergebnis durchaus ansehnlich. Ziel war in diesem Fall die Prüfung, wie gut es funktioniert, eine mit einem generischen Texterkennungsmodell ohne eigenen Trainingsaufwand erzeugte automatische Transkription durch ChatGPT korrigieren zu lassen. Das eigens für dieses Korpus trainierte Modell (UAT_M6, ID 28578) erzielte für diesen Text eine CER von 3,52 %. Das eingesetzte generische Modell („German Giant“, ID 50870) war mit einer Fehlerrate von 4,30 % nicht viel schlechter. Der von ChatGPT korrigierte Output lag zunächst mit einer CER von 5,38 % höher als diese Werte. Nach einer Anpassung der Transkription hinsichtlich der nicht als Fehler gewerteten Unterschiede für einen fairen

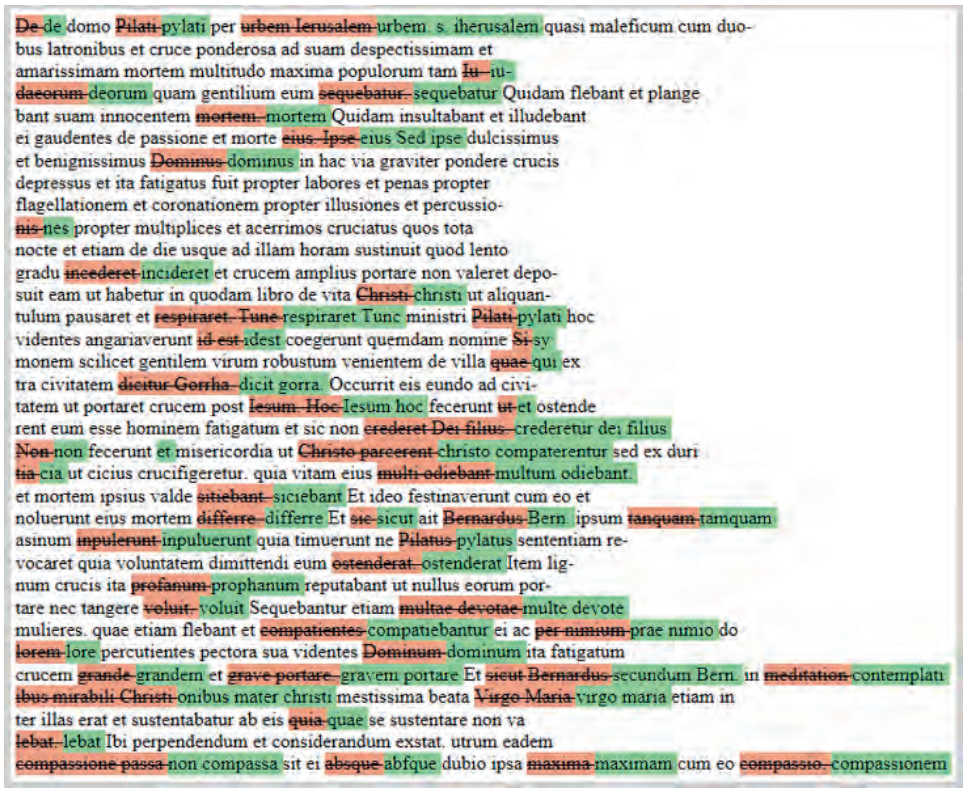
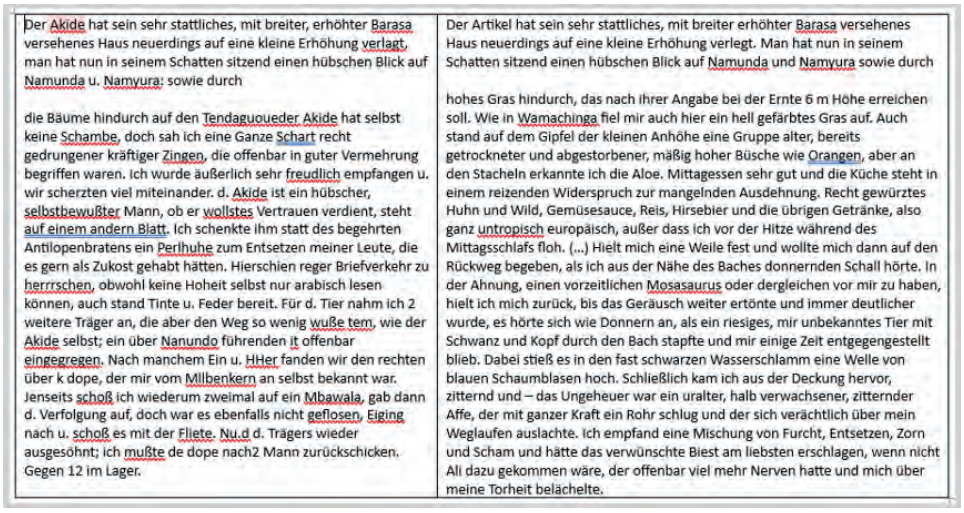

De de domo Pilati pylati per urbem Ierusalem urbem. s. iherusalem quasi maleficum cum duobus latronibus et cruce ponderosa ad suam despectissimam et amarissimam mortem multitudo maxima populorum tam Iu- iudaeorum deorum quam gentilium eum sequebatur sequebatur Quidam flebant et plange bant suam innocentem mortem. mortem Quidam insultabant et illudebant ei gaudentes de passione et morte eius. Ipse eius Sed ipse dulcissimus et benignissimus Dominus dominus in hac via graviter pondere crucis depressus et ita fatigatus fuit propter labores et penas propter flagellationem et coronationem propter illusiones et percussionis nes propter multiplices et acerrimos cruciatus quos tota nocte et etiam de die usque ad illam horam sustinuit quod lento gradu incederet incideret et crucem amplius portare non valeret deposuit eam ut habetur in quodam libro de vita Christi christi ut aliquantulum pausaret et respiraret. Tune respiraret Tunc ministri Pilati pylati hoc videntes angariaverunt id est idest coegerunt quemdam nomine Si sy monem scilicet gentilem virum robustum venientem de villa quae qui ex tra civitatem dicitur Gorrha dicit gorra. Occurrit eis eundo ad civitatem ut portaret crucem post Iesum. Hoc Iesum hoc fecerunt ut et ostende rent eum esse hominem fatigatum et sic non crederet Dei filius. crederetur dei filius Non non fecerunt et misericordia ut Christo parcerent christo compateretur sed ex duritia cia ut cicius crucifigeretur. quia vitam eius multi odiebant multum odiebant. et mortem ipsius valde sitiebant siciebant Et ideo festinaverunt cum eo et noluerunt eius mortem differre differre Et sic sicut ait Bernardus Bern ipsum tanquam tamquam asinum inpulerunt inpuluerunt quia timuerunt ne Pilatus pylatus sententiam revocaret quia voluntatem dimittendi eum ostenderat ostenderat Item lignum crucis ita profanum prophanum reputabant ut nullus eorum portare nec tangere voluit voluit Sequebantur etiam multae devotae multe devote mulieres. quae etiam flebant et compatientes compatiebantur ei ac per nimium prae nimio dolorem lore percutientes pectora sua videntes Dominum dominum ita fatigatum crucem grande grandem et grave portare gravem portare Et sicut Bernardus secundum Bern. in meditation contemplatibus mirabili Christi onibus mater christi mestissima beata Virgo Maria virgo maria etiam in ter illas erat et sustentabatur ab eis quia quae se sustentare non valebat lebat Ibi perpendendum et considerandum exstat. utrum eadem compassione passa non compassa sit ei absque abfque dubio ipsa maxima maximam cum eo compassio compassionem

Abb. 4: Das von ChatGPT erzeugte unbereinigte Ergebnis im Vergleich zu den normalisierten Ground-Truth-Daten (Tübingen, UB, Gb 461, Teil 1).


Der Akide hat sein sehr stattliches, mit breiter, erhöhter Barasa versehenes Haus neuerdings auf eine kleine Erhöhung verlagt, man hat nun in seinem Schatten sitzend einen hübschen Blick auf Namunda u. Namyura: sowie durch

die Bäume hindurch auf den Tendaguoueder Akide hat selbst keine Schambe, doch sah ich eine Ganze Schart recht gedrungener kräftiger Zingen, die offenbar in guter Vermehrung begriffen waren. Ich wurde äußerlich sehr freudlich empfangen u. wir scherzten viel miteinander. d. Akide ist ein hübscher, selbstbewußter Mann, ob er wollstes Vertrauen verdient, steht auf einem andern Blatt. Ich schenkte ihm statt des begehrten Antilopenbratens ein Perlhuhe zum Entsetzen meiner Leute, die es gern als Zukost gehabt hätten. Hierschien reger Briefverkehr zu herrrschen, obwohl keine Hoheit selbst nur arabisch lesen können, auch stand Tinte u. Feder bereit. Für d. Tier nahm ich 2 weitere Träger an, die aber den Weg so wenig wuße tem, wie der Akide selbst; ein über Nanundo führenden it offenbar eingegregen. Nach manchem Ein u. HHer fanden wir den rechten über k dope, der mir vom Millbenkern an selbst bekannt war. Jenseits schoß ich wiederum zweimal auf ein Mbawala, gab dann d. Verfolgung auf, doch war es ebenfalls nicht geflosen, Eiging nach u. schoß es mit der Fliete. Nu.d d. Trägers wieder ausgesöhnt; ich mußte de dope nach2 Mann zurückschicken. Gegen 12 im Lager.

Der Artikel hat sein sehr stattliches, mit breiter erhöhter Barasa versehenes Haus neuerdings auf eine kleine Erhöhung verlegt. Man hat nun in seinem Schatten sitzend einen hübschen Blick auf Namunda und Namyura sowie durch

hohes Gras hindurch, das nach ihrer Angabe bei der Ernte 6 m Höhe erreichen soll. Wie in Wamachinga fiel mir auch hier ein hell gefärbtes Gras auf. Auch stand auf dem Gipfel der kleinen Anhöhe eine Gruppe alter, bereits getrockneter und abgestorbener, mäßig hoher Büsche wie Orangen, aber an den Stacheln erkannte ich die Aloe. Mittagessen sehr gut und die Küche steht in einem reizenden Widerspruch zur mangelnden Ausdehnung. Recht gewürztes Huhn und Wild, Gemüsesauce, Reis, Hirsebier und die übrigen Getränke, also ganz untropisch europäisch, außer dass ich vor der Hitze während des Mittagsschlafs floh. (...) Hielt mich eine Weile fest und wollte mich dann auf den Rückweg begeben, als ich aus der Nähe des Baches donnernden Schall hörte. In der Ahnung, einen vorzeitlichen Mosasaurus oder dergleichen vor mir zu haben, hielt ich mich zurück, bis das Geräusch weiter ertönte und immer deutlicher wurde, es hörte sich wie Donnern an, als ein riesiges, mir unbekanntes Tier mit Schwanz und Kopf durch den Bach stapfte und mir einige Zeit entgegengestellt blieb. Dabei stieß es in den fast schwarzen Wasserschlamm eine Welle von blauen Schaumblasen hoch. Schließlich kam ich aus der Deckung hervor, zitternd und – das Ungeheuer war ein uralter, halb verwachsener, zitternder Affe, der mit ganzer Kraft ein Rohr schlug und der sich verächtlich über mein Weglaufen auslachte. Ich empfand eine Mischung von Furcht, Entsetzen, Zorn und Scham und hätte das verwünschte Biest am liebsten erschlagen, wenn nicht Ali dazu gekommen wäre, der offenbar viel mehr Nerven hatte und mich über meine Torheit belächelte.

Abb. 5: Auszug aus einem Expeditionstagebuch von Edwin Hennig (Tübingen, UA, 407/80) an der Stelle, wo der Text gegenüber der automatischen Transkription, die verbessert werden sollte, eine inhaltlich neue Wendung erfährt.

9 Das diplomatische Modell hat die vorliegende Zeichenfolge zwar richtig ausgegeben, aber man muss wissen, dass man nach dieser suchen muss, um einen Treffer für „Jerusalem“ zu erzielen. Das normalisierte Modell hat es hingegen nicht geschafft, die Abkürzung korrekt aufzulösen.

10 Hingegen wurde „u.“ für „und“ in der gekürzten Form beibehalten.

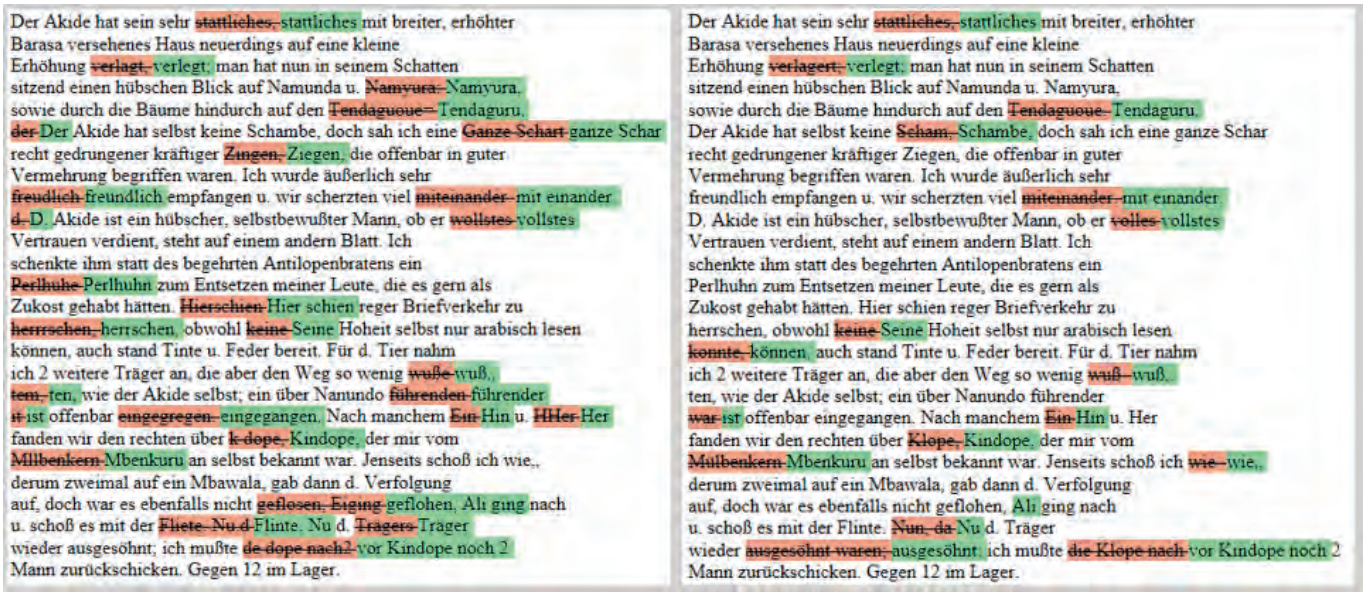


Abb. 6: Auf der linken Seite die Transkription mit dem generischen Modell „German Giant", auf der rechten Seite die auf dieser Grundlage korrigierte und an die Transkriptionsrichtlinien angepasste Version.

Vergleich sank der Wert auf 3,30 % und hat somit sogar knapp das auf diesen Schreiber spezialisierte Modell geschlagen. Noch besser wird das Ergebnis jedoch, wenn man die mit dem spezialisierten Modell erzeugte Transkription durch ChatGPT korrigieren lässt. Auf diese Weise ließ sich die Fehlerrate auf 2,73 % senken.

Ein Vergleich des ChatGPT-Ergebnisses mit der automatischen Transkription durch das generische Modell zeigt, dass ChatGPT sehr gut kleinere Rechtschreibfehler korrigiert, die auch für einen menschlichen Leser die Lesbarkeit des gemeinten Wortes nur wenig einschränken. Schwierig wird es für Mensch und Maschine an den Stellen, wo die ausgegebenen Buchstaben keine Identifizierung des originären Wortes ermöglichen. Dies ist im vorliegenden Dokument vornehmlich bei Begriffen in Suaheli der Fall, wo dem generischen Modell wahrscheinlich die entsprechenden Trainingsdaten gefehlt haben. Hier hat das auf den Schreiber spezialisierte Modell einen Vorteil, welches anhand der Ground-Truth-Daten gelernt hat, diese Wörter zu lesen. Umgekehrt korrigiert ChatGPT den vom generischen Modell richtig erkannten Begriff „Schambe", der eine Plantage bezeichnet, in den alltagssprachlich im Deutschen gebräuchlicheren Begriff „Scham" und zeigt so, dass der erfolgreiche Einsatz von LLMs davon abhängig ist, welche Trainingsdaten diesen zugrunde liegen, und dass sich so ein Bias ergeben kann. Im damaligen Deutsch-Ostafrika gebräuchliche Ausdrücke gehören anscheinend nicht dazu: „Explizit Minderheiten und sozial schwächere Gruppen sind in diesen Trainingsmaterialien oftmals unter- oder falsch repräsentiert, und dementsprechend beeinflusst fallen die Modellresultate aus."[11] Neben orthographischen Fehlern hat ChatGPT auch in die Grammatik eingegriffen, wie sich vor allem bei überkorrigierten Verbformen zeigt, die dem Sprachmodell anscheinend richtiger erschienen. Während die Auflösung von Abkürzungen eine Leseerleichterung sein kann, stellt die eigenmächtige Anpassung von Verbformen einen nicht gewollten Eingriff in den Originaltext dar.

Für die Weiterverwendung solcher durch ChatGPT korrigierter Transkriptionsdaten ist außerdem zu beachten, dass ChatGPT bei den im Rahmen dieser Arbeit vorgenommenen Versuchen bessere Ergebnisse auf reinen Textdaten erzielt hat als auf XML-Daten. Dies liegt wahrscheinlich in dem Umstand begründet, dass der Text in einem XML-Format nicht zusammenhängend vorliegt, sondern die Zeilen voneinander abgetrennt in ein Schema mit weiteren, den Text strukturierenden Elementen eingebunden sind. Für eine Weiterverarbeitung sind jedoch oftmals die ebenfalls im XML-Schema eingebundenen Koordinaten interessant, die die Textzeilen der Transkriptionen mit denen des Digitalisats verknüpfen, um die Daten wieder in eine Transkriptionsplattform oder ein Präsentationssystem einzuspielen. Für die hier vorgenommenen Tests wurden zeilengenaue TXT-Daten verwendet, die eine semiautomatische Weiterverarbeitung der Daten ermöglichen. Für eine Einbindung einer derartigen Prozessierung mit ChatGPT oder einem anderen Large Language Model in einen Workflow müsste jedoch eine andere Lösung gefunden werden.

Insgesamt entsteht der Eindruck, dass die Nachbearbeitung von automatischen Transkriptionsdaten mit Large Language Models interessante Möglichkeiten bietet,

11 Siehe Hodel, Tobias: Large Language Models, oder weshalb wir künstliche Intelligenz im Archiv finden sollten, in: Fähle, Daniel/ Müller, Peter (Hrsg.): Smart und intelligent – Digitale Unterstützung für die Arbeit im Archiv. Vorträge des 82. Südwestdeutschen Archivtags am 22. und 23. Juni 2023 (Werkhefte des Landesarchivs Baden-Württemberg 31) Ostfildern 2024, S. 77-84, Online: https://doi.org/10.48350/197467, S. 80.

aber noch Grenzen hat. Je nach Text können bessere oder schlechtere Ergebnisse erzielt werden. Je weniger Fehler die automatische Transkription beinhaltet und je weniger komplex der Text hinsichtlich Sonderzeichen, Abkürzungen und Wortschatz ist, desto besser kommt ChatGPT damit zurecht. Wenn die Transkription jedoch bereits an sich schon viele Fehler beinhaltet, welche einzelne Wörter komplett unkenntlich machen, wenn sie stark gekürzt ist, Eigennamen und/oder Begriffe aus einer in den Trainingsdaten des LLM wahrscheinlich unterrepräsentierten Sprache enthält, die schon der zugrunde liegenden Texterkennung in der Regel mehr Schwierigkeiten bereiten als Wörter des alltäglichen Sprachgebrauchs, so erschwert dies die automatische Datennachkorrektur. Somit ist dieses Vorgehen für manche Dokumenttypen besser geeignet als für andere, beispielsweise für Transkriptionsdaten von Texten in deutscher Kurrentschrift, für die es bereits sehr gute generische Modelle gibt. Hier bietet sich eine zusätzliche Option, die Daten nach einer Transkription auf Knopfdruck noch einmal ohne viel Ressourcenaufwand maschinell zu verbessern und die Zeichenfehlerrate zu senken, bevor sie den Nutzerinnen und Nutzern zur Verfügung gestellt werden.

Aber selbst wenn das Ergebnis zufriedenstellend ausfällt, gilt es, sich einiger Fallstricke bewusst zu sein, die die Integrität der Daten beeinflussen können. Wie dargelegt wurde, ist zu beachten, dass nicht wie gewünscht nur Fehler korrigiert und Sonderzeichen aufgelöst werden, sondern dass ChatGPT die Aufforderung zur Korrektur in einem weiteren Sinne auffassen kann und mitunter zusätzliche Eingriffe in den Text vornimmt. Je nachdem kann der derart korrigierte Text zwar orthographisch fehlerfrei sein, aber inhaltlich mehr oder weniger signifikant vom Original abweichen. Dies war bei allen getesteten Dokumenten trotz Gegensteuerung durch Prompting in unterschiedlicher Ausprägung der Fall und hat gezeigt, dass grundsätzlich Vorsicht geboten ist, da ein auf den ersten Blick stimmiger Text erzeugt wird und durch ChatGPT vorgenommene Änderungen nicht gleich auffallen.

Die von ChatGPT selbstständig vorgenommenen Überkorrekturen[12] sind allerdings nicht nur nachteilig. Wenn ganze Wörter ausgetauscht, ausgelassen und ergänzt werden, wird natürlich der Originaltext in unerwünschter Weise kompromittiert, selbst wenn im rein sprachlichen Sinne womöglich eine Verbesserung erzielt wird. Daher ist es beim Prompten notwendig, möglichst genaue und detaillierte Anweisungen zu geben und diese je nach Ergebnis weiter zu verfeinern. Kleinere Eingriffe normalisierender oder zum Teil sogar schon editorischer Art können jedoch die Lesbarkeit des Textes erhöhen. Im Rahmen der hier durchgeführten Tests waren derartige Korrekturen, wie die Großschreibung von Satzanfängen, Eigennamen und Nomina Sacra, eine selbstständige Zeichensetzung, eine Anpassung an die wörterbuchrelevante Orthographie – sei es nun gemäß dem klassischen Latein oder der heutigen deutschen Rechtschreibung – sowie das unaufgeforderte Ausschreiben abgekürzter Begriffe, ein ungeplanter Effekt, welcher jedoch für die Zielsetzung, mit Hilfe von ChatGPT aufbereitete Transkriptionsdaten mit der höchstmöglichen Lesbarkeit zur Verfügung zu stellen, bewusst weiter verfolgt werden könnte.

Zum anderen bietet KI auch gezielt die Möglichkeit, das Transkriptionslevel zu wechseln, weil es nicht immer möglich ist, von einer zeichengetreuen Transkription mithilfe von Skripten automatisch zu einer normalisierten Transkription zu gelangen, da zum Beispiel Abkürzungszeichen in verschiedenen Varianten aufgelöst werden können oder Wörter so stark kontrahiert sind, dass es mehrere Optionen der Auflösung gibt beziehungsweise ein und dasselbe Wort auf unterschiedliche Weisen abgekürzt sein kann. Dieses Vorgehen ermöglicht es, zwei Transkriptionsoptionen anzubieten, ohne den Aufwand eines zweiten Modelltrainings mit einem zweiten Ground-Truth-Datensatz zu betreiben. Die Normalisierung per ChatGPT hat sogar den Vorteil, dass nicht alle Abkürzungsoptionen in den Trainingsdaten vorliegen müssen. Während es kein Problem ist, einem Texterkennungsmodell die Normalisierung von regelmäßig vorkommenden Sonderzeichen mit immer gleicher Bedeutung beizubringen, ist dies bei komplexeren und/oder selteneren Fällen schwieriger. Die Trainingsdaten müssten alle Eventualitäten in ausreichender Häufigkeit abdecken, da ein Texterkennungsmodell nur erkennen kann, was es kennt. Dies würde bei stark gekürzten Texten einen großen Datenumfang bedeuten, der selbst dann wahrscheinlich nicht jeden Einzelfall berücksichtigen kann. Hier könnte die Datenaufwertung mit ChatGPT den Vorteil haben, dass derartige LLMs auf einer riesigen Datenbasis arbeiten und von daher womöglich weniger Schwierigkeiten mit komplizierteren Abkürzungen haben.

2.2 Erschließung anhand automatischer Transkriptionsdaten

Ein durchsuchbarer Volltext, ob nun in Form einer automatischen oder manuellen Transkription, ist jedoch vor allem dann vorteilhaft, wenn man bereits ein spezifi-

12 Hier zeigt sich, dass automatische Texterkennung und Large Language Models technisch auf die gleiche Funktionsweise zurückgehen. Auch beim Einsatz von Texterkennungssoftware kann es mitunter zu Überkorrekturen kommen, wenn die Gewichtung der Konfidenzen dem neuronalen Netz ein anderes Wort als das tatsächlich vorliegende wahrscheinlicher erscheinen lässt. Auch hier findet also nicht nur eine Ausgabe des Zeichenbestands statt, so wie er gelesen wird, sondern der gelesene Zeichenbestand wird im Kontext interpretiert, was oftmals funktioniert und das Ergebnis verbessert, aber eben auch zu Fehllesungen führen kann.

sches Anliegen an einen Text hat beziehungsweise weiß, wonach man suchen möchte. Historische Bestände wie Protokolle, Tagebücher und Briefnachlässe bieten mit ihren Titeln wie „Tagebuch, Teil 1", „Protokolle des Akademischen Senats, Band 73" oder „Briefe von Verfasser X an Empfänger Y" allerdings kaum Hinweise auf ihren Inhalt. Im Gegensatz zu moderneren Druckwerken gibt es in vielen Fällen weder Inhaltsverzeichnisse, Register oder gar Klappentexte, die inhaltliche Rückschlüsse zulassen und Interesse für die Materialien generieren. Selbst eine detailliertere Strukturdatenerschließung der Digitalisate gliedert derartige Dokumente meist nur chronologisch. In solchen Fällen schwach strukturierter Materialien kann die Zugänglichkeit bisher nur durch eine zeitintensive Verschlagwortung oder Regestierung erhöht werden.[13] Volltexte erleichtern zwar den Zugang zu historischen Dokumenten und machen sie durchsuchbar, aber man muss bereits wissen, wonach man sucht beziehungsweise auf gut Glück anhand der vorhandenen Eckdaten wie Titel, Entstehungszeit und Ort – insofern diese vorliegen – die Texte durchsuchen.

Können nun Volltexte mit weiteren KI-Methoden für eine (semi-)automatische Tiefenerschließung von Altbeständen mit geringem Ressourcenaufwand nachgenutzt werden? Anwendungen wie ChatGPT zeigen hier Potential, selbst fehlerbehaftete automatisch generierte Transkriptionsdaten für die Ermittlung von Erschließungsdaten, die Erzeugung von Inhaltsangaben sowie die Extraktion von Schlagworten, Personen- und Ortsangaben heranzuziehen. Dabei werden auch hier wieder Ground-Truth-Daten mit automatisch generierten Transkriptionen verglichen. Es soll beispielhaft gezeigt werden, wie ein derartiges Vorgehen aussehen kann, für welche Art von Erschließung solche Methoden geeignet sind und wo ihre Grenzen liegen. Ziel ist dabei eine auswertungsoffene Erschließung, die zwar inhaltliche Informationen vermittelt, aber nicht tendenziös ist.

Als Testbeispiel wurde ein Band der Expeditionstagebücher von Edwin Hennig (UAT 407/80) genutzt, der von einer paläontologischen Expedition an den Tendaguru im heutigen Tansania (damals Deutsch-Ostafrika) berichtet. Der Band hat 171 Seiten, die mit 35.334 Wörtern beschrieben sind, welche sich wiederum aus 233.795 Zeichen zusammensetzen. Für diese berechnet der OpenAI-Tokenizer einen Wert von 74.713 Tokens. Als Verarbeitungsobergrenze von ChatGPT wurden zum Zeitpunkt der Tests 8.192 Tokens angegeben.[14]

Für den kompletten Band wurden Ground-Truth-Daten als Vergleichsgrundlage erzeugt. Diesen sollen automatische Transkriptionen gegenübergestellt werden, die unterschiedliche CER-Stufen repräsentieren. Auf diese Weise wird eine Einschätzung möglich, inwieweit automatische Transkriptionen für die Weiterverarbeitung mit ChatGPT genutzt werden können und von welcher Qualität diese sein müssen. Wie auch schon beim Training von Texterkennungsmodellen ist neben der Ergebnisqualität der Ressourcen-Aufwand ein wichtiges Kriterium, um den Nutzen eines solchen Vorgehens für die Arbeit mit historischen Beständen einzuschätzen.

Zunächst wurde ChatGPT hinsichtlich der Erstellung einer präzisen und sachlichen Zusammenfassung des Inhalts gepromptet. Als Eckdaten für die Bewertung der Qualität der Ergebnisse wurden folgende Aspekte herangezogen: Nennung und Einordnung von Personen, Zeit- und Ortsangaben sowie ein grundlegendes inhaltliches Verständnis des Textes. Für diesen Anwendungsfall wurde die Token-Grenze ignoriert und jeweils die gesamten Daten des Bandes an ChatGPT zur Weiterverarbeitung übergeben. Der auf den Ground-Truth-Daten beruhenden Inhaltsangabe gelang es problemlos, diese Daten aus dem Text zu extrahieren und mit 189 Wörtern zusammenzufassen (Abb. 7).[15] Dabei war es ChatGPT möglich, zu verstehen, dass die Expedition von Dr. Edwin Hennig und Dr. Janensch geleitet wurde, auch wenn sich Edwin Hennig nur durch einen Besitzeintrag identifiziert und einmalig namentlich nennt. Genauso wurde der zeitliche Rahmen der Expedition wie auch deren Ziel richtig angegeben. Neben diesen harten Fakten stimmt auch der erzählende inhaltliche Teil, was nach Erstellung der Ground-Truth-Daten und somit guter Textkenntnis subjektiv eingeschätzt wurde.

Bei der Heranziehung automatischer Transkriptionsdaten für die Erstellung von Inhaltsangaben zeigen sich schnell die Grenzen dieses Vorgehens. Wenn nur einmalig genannte Fakten wie Personennamen oder Zeitangaben falsch transkribiert worden sind, dann setzen sich diese Fehler in der Bearbeitung mit ChatGPT fort. Die mit dem Texterkennungsmodell „German Giant" (Abb. 8) erstellten Transkriptionsdaten (4,30 % CER) ermöglichen es ChatGPT, die Person des Edwin Hennig als relevant zu identifizieren, auch wenn der Name falsch geschrieben

13 Vgl. Huff, Dorothee/ Keyler, Regina: Das Projekt OCR-BW: Automatische Texterkennung auch für Archive, in: Fähle, Daniel/ Müller, Peter (Hrsg.): Smart und intelligent – Digitale Unterstützung für die Arbeit im Archiv. Vorträge des 82. Südwestdeutschen Archivtags am 22. und 23. Juni 2023 (Werkhefte des Landesarchivs Baden-Württemberg 31) Ostfildern 2024, S. 25-33, Online: https://doi.org/10.53458/books.238, S. 27.

14 Diese Grenze gilt für GPT-3.5 und GPT-4. Für ChatGPT 4o war zum damaligen Zeitpunkt noch keine Obergrenze definiert, sondern nur der ungefähre Wert angegeben, dass 1.2x weniger Tokens bearbeitet würden. Um zu gewährleisten, dass die Token-Obergrenze eingehalten würde, wurde der Wert von 8.192 auch für ChatGPT 4o als Grenze angenommen.

15 Die Inhaltsangabe wurde noch einmal von ChatGPT auf 124 Wörter reduziert und manuell schließlich auf 106 Wörter gekürzt sowie entsprechend der bevorzugten GND-Ansetzung der erwähnten Namen und Orte modifiziert: https://opendigi.ub.uni-tuebingen.de/opendigi/UAT_407_080 [abgerufen am 03.11.2024].

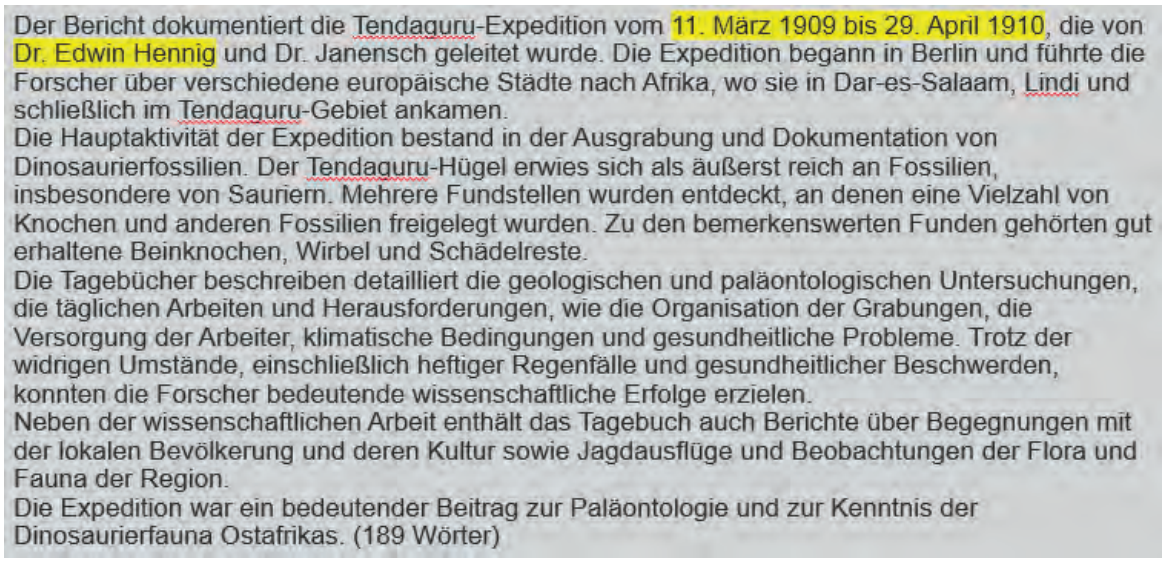

Der Bericht dokumentiert die Tendaguru-Expedition vom 11. März 1909 bis 29. April 1910, die von Dr. Edwin Hennig und Dr. Janensch geleitet wurde. Die Expedition begann in Berlin und führte die Forscher über verschiedene europäische Städte nach Afrika, wo sie in Dar-es-Salaam, Lindi und schließlich im Tendaguru-Gebiet ankamen.
Die Hauptaktivität der Expedition bestand in der Ausgrabung und Dokumentation von Dinosaurierfossilien. Der Tendaguru-Hügel erwies sich als äußerst reich an Fossilien, insbesondere von Sauriern. Mehrere Fundstellen wurden entdeckt, an denen eine Vielzahl von Knochen und anderen Fossilien freigelegt wurden. Zu den bemerkenswerten Funden gehörten gut erhaltene Beinknochen, Wirbel und Schädelreste.
Die Tagebücher beschreiben detailliert die geologischen und paläontologischen Untersuchungen, die täglichen Arbeiten und Herausforderungen, wie die Organisation der Grabungen, die Versorgung der Arbeiter, klimatische Bedingungen und gesundheitliche Probleme. Trotz der widrigen Umstände, einschließlich heftiger Regenfälle und gesundheitlicher Beschwerden, konnten die Forscher bedeutende wissenschaftliche Erfolge erzielen.
Neben der wissenschaftlichen Arbeit enthält das Tagebuch auch Berichte über Begegnungen mit der lokalen Bevölkerung und deren Kultur sowie Jagdausflüge und Beobachtungen der Flora und Fauna der Region.
Die Expedition war ein bedeutender Beitrag zur Paläontologie und zur Kenntnis der Dinosaurierfauna Ostafrikas. (189 Wörter)

Abb. 7: Inhaltsangabe auf Basis der Ground-Truth-Daten.

wurde. Werner Janensch als erster Expeditionsleiter und am häufigsten im Tagebuch erwähnte Person wird nicht genannt. Bei der Angabe des Expeditionszeitraums/Verfassungszeitraums stimmen die Jahreszahlen, aber Tag und Monat des Enddatums sind falsch. Dagegen stimmen die Ortsangaben, und auch die allgemeine Inhaltsangabe gibt die relevanten Informationen gut wieder.

Die Verwendung einer mit dem Modell „German_Kurrent_XIX_pylaia“ erzeugten automatischen Transkription, deren Zeichenfehlerrate mit 14,07 % deutlich höher liegt, bekräftigt diesen Eindruck noch einmal (Abb. 9). Die Datenqualität hat es ChatGPT offenbar nicht gestattet, die Person Edwin Hennig zu identifizieren. Dafür wird der Expeditionsleiter Werner Janensch als Verfasser angegeben. Die Expedition wird nun in einem falschen Jahr verortet und genauere Angaben zum Tag und Monat fallen weg. Ortsangaben werden überhaupt nicht gemacht. Hinsichtlich der inhaltlichen Aspekte ist festzustellen, dass diese weniger detailliert ausfallen als in den beiden oben besprochenen Versionen, aber grundsätzlich zutreffen. ChatGPT lässt an dieser Stelle die eigene Unsicherheit erkennen, indem es von einem „vermutlich kolonialen Kontext“ spricht.

Es zeigt sich also, dass die Erstellung von inhaltlichen Zusammenfassungen ein aussichtsreiches Anwendungsfeld für die Weiterverarbeitung von Transkriptionsdaten mit Large Language Models ist. Die relevanten Informationen wurden wohl aufgrund ihrer Mehrfachnennung im Dokument, die einzelne Fehllesungen ausgeglichen haben dürfte, auch aus den automatisch erzeugten Volltexten zuverlässig herausgefiltert – bei einer niedrigen Fehlerrate jedoch weit detaillierter als bei einer höheren. Harte Fakten sollten jedoch noch einmal manuell überprüft werden. Zumindest in diesem Fall würde sich das Ergebnis von ChatGPT gut mit den grundlegenden Erschließungsdaten ergänzen, die für den Menschen auf einen Blick erkennbar sind, nämlich die auf dem Umschlag beziehungsweise im Einband vermerkten Informationen zu Verfasser und Zeitraum.

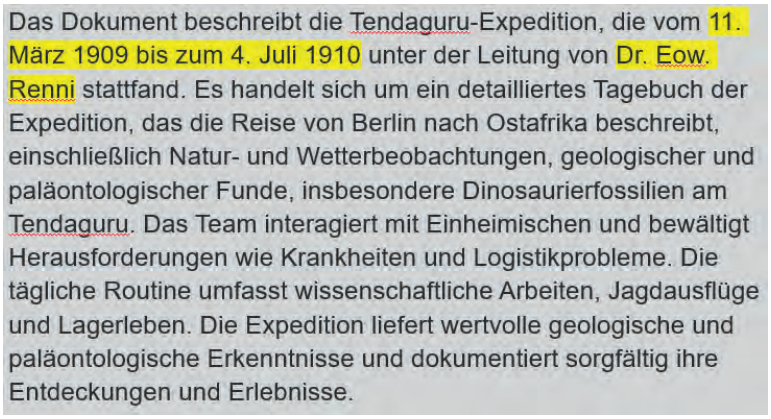

Das Dokument beschreibt die Tendaguru-Expedition, die vom 11. März 1909 bis zum 4. Juli 1910 unter der Leitung von Dr. Eow. Renni stattfand. Es handelt sich um ein detailliertes Tagebuch der Expedition, das die Reise von Berlin nach Ostafrika beschreibt, einschließlich Natur- und Wetterbeobachtungen, geologischer und paläontologischer Funde, insbesondere Dinosaurierfossilien am Tendaguru. Das Team interagiert mit Einheimischen und bewältigt Herausforderungen wie Krankheiten und Logistikprobleme. Die tägliche Routine umfasst wissenschaftliche Arbeiten, Jagdausflüge und Lagerleben. Die Expedition liefert wertvolle geologische und paläontologische Erkenntnisse und dokumentiert sorgfältig ihre Entdeckungen und Erlebnisse.

Abb. 8: Inhaltsangabe auf Basis des Modells „German Giant“.

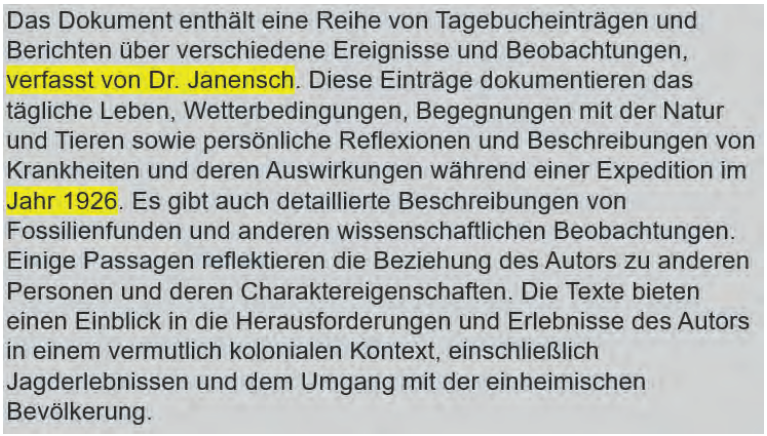

Das Dokument enthält eine Reihe von Tagebucheinträgen und Berichten über verschiedene Ereignisse und Beobachtungen, verfasst von Dr. Janensch. Diese Einträge dokumentieren das tägliche Leben, Wetterbedingungen, Begegnungen mit der Natur und Tieren sowie persönliche Reflexionen und Beschreibungen von Krankheiten und deren Auswirkungen während einer Expedition im Jahr 1926. Es gibt auch detaillierte Beschreibungen von Fossilienfunden und anderen wissenschaftlichen Beobachtungen. Einige Passagen reflektieren die Beziehung des Autors zu anderen Personen und deren Charaktereigenschaften. Die Texte bieten einen Einblick in die Herausforderungen und Erlebnisse des Autors in einem vermutlich kolonialen Kontext, einschließlich Jagderlebnissen und dem Umgang mit der einheimischen Bevölkerung.

Abb. 9: Inhaltsangabe auf Basis des Modells „German_Kurrent_XIX_pylaia“ (14,07 % CER).

In einem weiteren Schritt wurde getestet, ob die Transkriptionsdaten für eine Sacherschließung herangezogen werden können. Dies wurde am Beispiel von Personennamen erprobt. In einem ersten Schritt sollten die fünf am häufigsten genannten Personen ermittelt werden. Dabei dient die Zuverlässigkeit der Findung der Namen als Qua-

litätsmarker. Vergleichsdaten wurden in diesem Fall durch manuelles Tagging und eine quantitative Auswertung dieser Tags produziert. Gegen diese Daten wurden wiederum sowohl die Ground-Truth-Daten wie auch die von ChatGPT anhand der automatischen Transkriptionsdaten in verschiedenen CER-Stufen ermittelten Personennamen auf ihre Zuverlässigkeit getestet. Anschließend wurde ChatGPT auch zur Identifizierung der Personennamen herangezogen. Dabei sollten die Namen vervollständigt werden, insoweit sie unvollständig genannt worden sind, und um Lebensdaten sowie GND-Nummern ergänzt werden.

Das Ergebnis des manuellen Taggings lautet wie folgt in absteigender Reihenfolge der Anzahl der Nennungen. Für die Personen „Ali" und „C. Besser" konnte kein vollständiger Name ermittelt werden. Die im Dokument vorliegenden Namensansetzungen werden in eckigen Klammern angegeben.

1. Werner Janensch [J, J., Jan., Janensch, Dr. J, Dr. J., Dr. Janensch]
   Expeditionsleiter, Paläontologe ⟶ 170 Nennungen
2. Bernhard Sattler [Sattler, Hr./Herr Sattler]
   Ingenieur, Vertreter der DOAG ⟶ 36 Nennungen
3. Ali [Ali, Boy Ali]
   Boy (persönlicher Diener) ⟶ 34 Nennungen
4. C. Besser [Hr./Herr Besser]
   Vertreter der DOAG ⟶ 16 Nennungen
5. Boheti bin Amrani [Boheti]
   Aufseher und Präparator ⟶ 13 Nennungen

Zunächst wurde geprüft, wie gut ChatGPT diese Namen in den Ground-Truth-Daten findet, in welchen die Namen richtig transkribiert vorliegen. Die Personen sollten in absteigender Reihenfolge der Häufigkeit der Nennungen sortiert und in Listenform ausgegeben werden. Zudem sollte angegeben werden, wie häufig die jeweilige Person im Dokument genannt wird, wobei alle Instanzen eines Namens ganz gleich welcher Ansetzung zusammengezählt werden sollten. Eine Schwierigkeit vor allem bei der Person des Expeditionsleiters Werner Janensch ist das Vorliegen des Namens in verschiedenen Formen und Abkürzungsvarianten. In diesem Fall wie auch bei den nachfolgenden CER-Stufen wurde zunächst der komplette Datensatz ohne Beachtung der Token-Obergrenze in ChatGPT eingegeben.

Ergebnis Ground-Truth-Daten: Abgesehen von der Person „C. Besser" wurden alle Namen von der Originalliste gefunden. Anstelle von „C. Besser" wird „Bornhardt" genannt, der tatsächlich im Dokument fünfmal Erwähnung findet, jedoch nicht als Person, sondern als Literaturangabe gewertet wurde. Anzumerken ist noch, dass die Nennungen von Werner Janensch zwar deutlich unter dem tatsächlichen Wert liegen, aber wenn man allein die Namensform „Dr. Janensch" zählt, kommt ChatGPT den im Originaldokument vorhandenen 73 Nennungen dieser Namensform sehr nahe.

1. Dr. Janensch ⟶ 74 Nennungen (-96)
2. Ali ⟶ 35 Nennungen (+1)
3. Boheti ⟶ 8 Nennungen (-5)
4. Bornhardt ⟶ 6 Nennungen (+1)
5. Sattler ⟶ 4 Nennungen (-32)

- C. Besser wird nicht gefunden

Ergebnis „M6" (2,43 % CER): Der Vergleich mit den auf Grundlage eines eigens für die Handschrift Edwin Hennigs trainierten Texterkennungsmodells erzeugten Daten ergibt eine Liste, die zwar nicht hinsichtlich der Anzahl der Nennungen und der Reihenfolge, so aber doch von den gefundenen Namen her fast den anvisierten Vergleichsdaten entspricht.

1. Dr. Janensch ⟶ 30 Nennungen (-140)
2. Herr Sattler ⟶ 13 Nennungen (-23)
3. Herr Besser ⟶ 8 Nennungen (-8)
4. Boheti ⟶ 6 Nennungen (-7)
5. Salim ⟶ 5 Nennungen (+3)
6. Ali ⟶ 5 Nennungen (-29)

Ergebnis „German Giant" (4,30 % CER): Hier fallen deutliche Unterschiede im Gegensatz zu den manuell erzeugten Tagging-Daten ins Auge. Von der Originalliste der fünf am häufigsten genannten Personen wurden nur zwei gefunden und auch hier im Fall von Werner Janensch wieder mit einer deutlich geringeren Anzahl der Nennungen. Dafür findet sich „Herr Schwarz" auf der Liste, der tatsächlich genauso oft wie von ChatGPT angegeben im Dokument erwähnt wird. „Dr. Eow. Renni" ist offensichtlich eine Fehllesung von „Dr. Edwin Hennig", wie schon oben hinsichtlich der Inhaltsangaben erwähnt. „Herr Jakler" könnte man entsprechend als Verlesung von „Sattler" interpretieren. Dafür werden die beiden nichteuropäischen Expeditionsteilnehmer nicht in die Liste inkludiert.

1. Herr Besser ⟶ 10 Nennungen (-6)
2. Dr. Janensch ⟶ 8 Nennungen (-162)
3. Herr Schwarz ⟶ 5 Nennungen (+/- 0)
4. Dr. Eow. Renni ⟶ 4 Nennungen (+3)
5. Herr Jakler ⟶ 3 Nennungen (gemeint Sattler? -33)

- Ali und Boheti bin Amrani werden nicht gefunden

Nach den gezeigten Testbeispielen, in welchen der Ressourcenaufwand so niedrig wie möglich gehalten wurde, indem ChatGPT die unkorrigierten Transkriptionsdaten des gesamten Dokuments komplett zur Verfügung gestellt wurden, wurde geprüft, ob ein höherer Aufwand das Ergebnis entsprechend verbessern könnte. Dazu wurde ChatGPT zum einen eine durch das LLM selbst korrigierte Version der Transkriptionsdaten zur Verfügung gestellt,

zum anderen aber auch getestet, ob das Einhalten der Token-Grenze einen Unterschied macht.

Ergebnis „German Giant" korrigiert: Hier fällt auf, dass sich das Ergebnis erheblich von der Namenserhebung auf Grundlage der unkorrigierten Transkriptionsdaten mit dem German-Giant-Texterkennungsmodell unterscheidet. Wenn man die Ergebnisse rein in ihrer ausgegebenen Form vergleicht, ohne dass ausgegebene Zeichenketten als Verschreibungen interpretiert werden, so ist der Unterschied zur unkorrigierten Transkription größer als zu den Vergleichsdaten. Interessanterweise wird der Name von Werner Janensch in diesem Fall häufiger gezählt als in den manuell erstellten Tags. Möglicherweise wurden bei der angewiesenen Verbesserung der automatisch erzeugten Transkriptionsdaten aufgrund der hohen Häufigkeit des Namens auch weitere Zeichenketten entsprechend überkorrigiert, so dass mehr Instanzen gefunden werden konnten. Auch wird hier ein Herr Kaiser erheblich öfter gezählt, als sein Name im Originaldokument fällt. Allerdings wird der Begriff „Kaiser" in anderen Kontexten noch zweimal erwähnt, was wohl als weitere Namensnennung fehlinterpretiert wurde.

1. Dr. Janensch ⟶ 203 Nennungen (+33)
2. Herr Sattler ⟶ 29 Nennungen (-7)
3. Boheti ⟶ 12 Nennungen (-1)
4. Herr Kaiser ⟶ 7 Nennungen (+6)
5. Dr. Renni ⟶ 1 Nennung (+/-0 für Hennig)

Ergebnis „German Giant" tokenisiert: Für einen abschließenden Vergleich wurden die automatisch mit dem Texterkennungsmodell German Giant erzeugten Transkriptionsdaten entsprechend der Bearbeitungsobergrenze von 8.192 Tokens in zehn Abschnitte aufgeteilt, die ChatGPT jeweils mit demselben Prompt wie bisher auch zur Verfügung gestellt worden sind. Anschließend wurden die Ergebnisse der einzelnen Textabschnitte zusammengerechnet. Generell zeigt sich, dass sich mit diesem Vorgehen das beste Ergebnis erzielen lässt. Die ausgegebene Liste entspricht exakt dem manuellen Tagging, was die Personen und ihre Reihenfolge anbelangt. Dafür weichen die Nennungen jedoch vom Original ab und Boheti wird mit einer Nennung nur ganz knapp aufgeführt. Wenn es aber darum geht, dass ChatGPT die am häufigsten genannten Personen finden soll, damit diese verschlagwortet werden können, so wurde die Aufgabe von ChatGPT mit dieser Vorgehensweise wie gewünscht gelöst.

Davon abgesehen zeigt die Verarbeitung im Detail jedoch einige Schwierigkeiten bei der Datenverarbeitung mit ChatGPT auf. Wie bereits erwähnt, wurde immer derselbe Prompt verwendet und die Dokumente wurden jeweils im gleichen Format zur Verfügung gestellt. Es wurde jedoch für jeden Durchlauf ein neuer Chat eröffnet. Dieses Vorgehen hat gezeigt, dass derselbe Prompt nicht mit reproduzierbaren Ergebnissen gleichzusetzen ist. Teilweise wurden die gewünschten Namenslisten ausgegeben, zum Teil jedoch auch die Substantive des Textes oder in einem Fall anscheinend sogar alle Wörter aufgelistet und gezählt.[16] Warum die Ergebnisse jeweils so ausgefallen sind, war nicht nachvollziehbar. Wenn Daten jedoch nacheinander in denselben Chat mit immer demselben Prompt eingegeben werden, dann wendet ChatGPT auch denselben Lösungsweg an.

1. Dr. Janensch ⟶ 124 Nennungen (-46)
2. Herr Sattler ⟶ 24 Nennungen (-12)
3. Ali ⟶ 22 Nennungen (-12)
4. Herr Besser ⟶ 11 Nennungen (-5)
5. Boheti ⟶ 1 Nennungen (-12)

Auch für Erwähnungen von Orten und Organisationen wurde derselbe Test erfolgreich durchgeführt. Etwas schwammiger war hingegen die Suche nach relevanten inhaltlichen Schlagworten, die eine gewisse Interpretationsleistung voraussetzen. Es geht anders als bei beispielsweise Personennamen in diesem Fall nicht nur darum, die Instanzen einer spezifischen Kategorie im Gesamttext quantitativ zu identifizieren und aufzulisten – wobei auch dabei anhand von Mustern festgestellt werden muss, welche Entitäten dazugehören –, sondern es ist qualitativ zu bewerten, was bedeutsam ist und was nicht. Außerdem hat sich gezeigt, dass es einen Unterschied macht, ob Begriffe aus dem deutschen Wortschatz zugeordnet werden sollen oder Begriffe aus der suahelischen Sprache. Hier war die inhaltliche Zuordnung für ChatGPT deutlich schwieriger. So ergab die Frage nach in dem Expeditionstagebuch erwähnten einheimischen Bevölkerungsgruppen eine Auflistung von Wörtern in Suaheli. Eine Überprüfung dieser hat jedoch ergeben, dass sie nur teilweise die gewünschten Bezeichnungen enthielt. Daneben wurden auch Tier- und Landschaftsnamen aufgelistet. Auch in diesem Fall wird die wohl durch die Trainingsdaten von ChatGPT bedingte Schwierigkeit der Bearbeitung von Sprachen außereuropäischer Herkunft erkennbar.[17]

Insgesamt lässt sich hinsichtlich einer Schlagwortsuche auf Grundlage von automatischen Transkriptionsdaten feststellen, dass die relevanten Personen und Begriffe deutscher Sprachherkunft gefunden werden, insofern

16 In diesen Fällen wurden die ausgegebenen Listen nach Personennamen durchsucht und die Ergebnisse entsprechend miteinbezogen.

17 Siehe hierzu Bender, Emily M./ Gebru, Timnit/ McMillan-Major, Angelina/ Shmitchell, Shmargaret: On the Dangers of Stochastic Parrots. Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021) S. 610-623, Online: https://doi.org/10.1145/3442188.3445922, S. 617: "As people in positions of privilege with respect to a society's racism, misogyny, ableism, etc., tend to be overrepresented in training data for LMs […], this training data thus includes encoded biases, many already recognized as harmful."

das Ergebnis der automatischen Texterkennung nicht zu schlecht ist. Der Text sollte zumindest insoweit korrekt transkribiert sein, dass der Inhalt erkennbar ist und dass von sich wiederholenden relevanten Begriffen einige richtig erkannte Instanzen vorhanden sind. Die Identifikation der Instanzen erfolgt jedoch selbst anhand von schlechteren Daten zuverlässiger, wenn die Token-Obergrenze eingehalten und das Dokument ChatGPT entsprechend aufgeteilt zur Verfügung gestellt wird. In die Beurteilung muss miteinbezogen werden, dass die Ergebnisse nur mit Vorsicht evaluiert werden können und exemplarischen Charakter haben, da sie oftmals nicht genauso reproduzierbar sind. Hier könnte man mehrere Durchläufe starten und die jeweiligen Ergebnisse vergleichen oder die Anfragen der Reihe nach in immer demselben Chat stellen, der Antworten in der gewünschten Weise hervorbringt. Eine Lösung zur Absicherung der Ergebnisse könnte auch in diesem Fall darin bestehen, dass man die gefundenen Entitäten durch eine Volltextsuche bestätigen lässt. Jedoch muss es vielleicht auch nicht darum gehen, exakte Werte zu erhalten. Es ist zu überlegen, ob Näherungsdaten nicht auch genügen, um die Zugänglichkeit zum Dokument zu erhöhen. Dem ist allerdings entgegenzusetzen, dass so die Bewertungsgrundlage für die Nutzerinnen und Nutzer verzerrt sein kann, auf welcher sie das Dokument inhaltlich einschätzen. Large Language Models sind also eine Black Box, wenn man ihre genaue Architektur und Trainingsdaten nicht kennt.[18]

Nachdem nun Schlagworte vorliegen, stellt sich die Frage, ob ChatGPT auch den nächsten Schritt gehen und zum Beispiel die GND- oder Wikidata-Identifikationsnummern zu den Namen und Orten angeben kann. Hier waren die Ergebnisse sehr durchwachsen. Während zum Teil die richtigen Identifikatoren ausgegeben wurden, konnten sie in anderen Fällen auch völlig frei erfunden sein. Auf den ersten Blick korrekt anmutende Listen mit Nummern im richtigen Format hielten einer Überprüfung nicht stand. Zum Teil wurden die Nummern sogar mit Links hinterlegt, die beim Anwählen ins Leere führten. Dies zeigt eindringlich, dass Large Language Models sich zunächst auf die Produktion von wahrscheinlichem Text verstehen und keine Datenbank sind.[19] Dieser Text muss jedoch nicht inhaltlich richtig sein. Die Identifizierung von GND-Nummern durch ChatGPT allein ist also als unzuverlässig zu bewerten und kann Halluzinationen erzeugen. Eine Verbesserung des Ergebnisses lässt sich dadurch herbeiführen, dass der ChatGPT-Prompt einen Verweis auf eine einschlägige Website beinhaltet, die die gewünschten Informationen enthält. Im Fall von Personennamen hat sich beispielsweise eine Einbindung der Deutschen Biographie[20] als effektiv erwiesen. Die GND an sich ließ sich nicht einbinden, wie auch ein aktueller Zugriff auf das Internet nicht möglich ist.[21] Somit erscheint auch bei diesem Einsatzgebiet von LLMs eine halbautomatische Lösung als der sicherere Weg: Man lässt sich durch ChatGPT die Personennamen heraussuchen, verifiziert diese durch eine Volltextsuche in den Transkriptionsdaten und recherchiert die GND-Nummern anschließend selbst.[22] Der größte Aufwand, nämlich das Durchsuchen des Textes nach relevanten inhaltlichen Aspekten, kann so dennoch an die KI ausgelagert werden.

3. Fazit

Insgesamt hat sich gezeigt, dass Anwendungen der künstlichen Intelligenz im weiteren Sinne beziehungsweise des Machine Learnings im engeren Sinne ein Mittel der Wahl sein können, um auch handschriftliche Bibliotheksbestände über eine reine Digitalisierung hinaus zu erschließen. Nicht nur für Archivalien, sondern auch für Bibliotheksgut lässt sich feststellen, dass „neue Technologien auf jeden Fall als Chance verstanden werden [sollten], um die Massen an Archivalien neu und anders deuten zu lassen."[23] Mit dem technischen Fortschritt, der über die letzten Jahre zu beobachten war, den aktuell absehbaren und zukünftig sicher noch kommenden Entwicklungen ergibt sich für Bibliotheken die Möglichkeit, den Zugang zu handschriftlichen Dokumenten zu erleichtern und neue Nutzergruppen zu erschließen. Bei immer besseren Ergebnissen sinkt der Arbeitsaufwand und aktuell noch notwendige Arbeitsschritte werden in Zukunft womöglich obsolet. Durch Umwandlung von „handwritten texts towards collections as data"[24] können große Textmengen maschinell für verschiedene Fragestellungen bearbeitet und Dokumente außerhalb des traditionellen Kanons beziehungsweise für nicht traditionell textverarbeitende Disziplinen zugänglich gemacht werden.

Die Weiterverarbeitung von automatisch erzeugten Transkriptionsdaten mit LLMs wie ChatGPT zeigt Potential, Lösungen zu finden, die mehr und neue Nutzungsszena-

18 Vgl. Hodel, Tobias: Supervised and unsupervised. Approaches to machine learning for textual entities, in: Jaillant, Lise (Hrsg.): Archives, Access and Artificial Intelligence (Digital Humanities Research 2) Bielefeld 2022, S. 157-178, Online: https://doi.org/10.1515/9783839455845-007, S. 159.

19 Vgl. Spina, Salvatore: Artificial Intelligence in archival and historical scholarship workflow. HTS and ChatGPT, in: Umanistica Digitale 16 (2023) S. 125-140, Online: https://doi.org/10.6092/issn.2532-8816/17205, S. 130-131.

20 https://www.deutsche-biographie.de/home [abgerufen am 17.08.2024].

21 Vgl. Spina, Artificial Intelligence in archival and historical scholarship workflow, 2023, S. 130-131.

22 Eine Lösung könnte eine Kombination von LLMs und zum Beispiel Textdaten, Datenbanken oder Knowledge Graphen sein. In Form eines RAG können die Texte eingebunden werden, so dass das LLM nicht von anderen Daten fremdbeeinflusst wird.

23 Siehe Hodel, Large Language Models, 2024, S. 84.

24 Siehe Nockels/ Gooding/ Terras, The implications of handwritten text recognition, 2024, S. 155.

rien inkludieren, die Datenqualität erhöhen und den Ressourcenaufwand verringern. Damit die so geschaffenen Ergebnisse aber verlässlich und reproduzierbar sind, werden noch weitere Entwicklungen nötig sein, als sie die aktuellen Out-of-the-box-Lösungen derzeit bieten. Mit Einbindung von LLMs sind noch weitere Features vorstellbar, die jedoch erst einmal entwickelt und integriert werden müssen. In diesem Bereich müssen sich Bibliotheken in Zukunft den sich so ergebenden Chancen und Herausforderungen bei der Veröffentlichung digitaler Sammlungen stellen.

Literaturverzeichnis

Bender, Emily M./ Gebru, Timnit/ McMillan-Major, Angelina/ Shmitchell, Shmargaret: On the Dangers of Stochastic Parrots. Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021) S. 610-623, Online: https://doi.org/10.1145/3442188.3445922.

Candela, Gustavo/ Sáez, María Dolores/ Escobar Esteban, MPilar/ Marco-Such, Manuel: Reusing digital collections from GLAM institutions, in: Journal of Information Science 48,2 (2022) S. 251-267, Online: https://doi.org/10.1177/0165551520950246.

Hodel, Tobias: Large Language Models, oder weshalb wir künstliche Intelligenz im Archiv finden sollten, in: Fähle, Daniel/ Müller, Peter (Hrsg.): Smart und intelligent – Digitale Unterstützung für die Arbeit im Archiv. Vorträge des 82. Südwestdeutschen Archivtags am 22. und 23. Juni 2023 (Werkhefte des Landesarchivs Baden-Württemberg 31) Ostfildern 2024, S. 77-84, Online: https://doi.org/10.48350/197467.

Hodel, Tobias: Supervised and unsupervised. Approaches to machine learning for textual entities, in: Jaillant, Lise (Hrsg.): Archives, Access and Artificial Intelligence (Digital Humanities Research 2) Bielefeld 2022, S. 157-178, Online: https://doi.org/10.1515/9783839455845-007.

Huff, Dorothee/ Keyler, Regina: Das Projekt OCR-BW: Automatische Texterkennung auch für Archive, in: Fähle, Daniel/ Müller, Peter (Hrsg.): Smart und intelligent – Digitale Unterstützung für die Arbeit im Archiv. Vorträge des 82. Südwestdeutschen Archivtags am 22. und 23. Juni 2023 (Werkhefte des Landesarchivs Baden-Württemberg 31) Ostfildern 2024, S. 25-33, Online: https://doi.org/10.53458/books.238.

Huff, Dorothee/ Stöbener, Kristina: Projekt OCR-BW. Automatische Texterkennung von Handschriften, in: O-Bib. Das Offene Bibliotheksjournal 9,4 (2022) S. 1-19, Online: https://doi.org/10.5282/o-bib/5885.

Neudecker, Clemens/ Zaczynska, Karolina/ Baierer, Konstantin/ Rehm, Georg/ Gerber, Mike/ Schneider, Julián Moreno: Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten, in: Franke-Maier, Michael/ Kasprzik, Anna/ Ledl, Andreas/ Schürmann, Hans (Hrsg.): Qualität in der Inhaltserschließung (Bibliotheks- und Informationspraxis 70) Berlin 2021, S. 137-166, Online: https://doi.org/10.1515/9783110691597-009.

Nockels, Joseph/ Gooding, Paul/ Terras, Melissa: The implications of handwritten text recognition for accessing the past at scale, in: Journal of Documentation 80,7 (2024) S. 148-167, Online: https://doi.org/10.1108/JD-09-2023-0183.

Spina, Salvatore: Artificial Intelligence in archival and historical scholarship workflow. HTS and ChatGPT, in: Umanistica Digitale 16 (2023) S. 125-140, Online: https://doi.org/10.6092/issn.2532-8816/17205.

© Valentin Marquardt / Uni.Tübingen

Dorothee Huff

Studium der Mittleren und Neueren Geschichte, Lateinischen Philologie des Mittelalters und der Neuzeit, Kulturantrophologie/ Europäischen Ethnologie sowie Bibliotheks- und Informationswissenschaft. Seit Oktober 2024 Leiterin der Abteilung Handschriften und Historische Drucke mit Restaurierungswerkstatt und Digitalisierungszentrum, Referentin für Bestandserhaltung, Fachreferentin für Geschichte, osteuropäische Geschichte und Buchwesen an der Universitätsbibliothek Tübingen.
dorothee.huff@uni-tuebingen.de

© Wichern GmbH, Berlin

Kristina Stöbener

Studium der Geschichte und Germanistik, Handschriftenbearbeiterin (Erschließung mittelalterlicher Handschriften) in der Staatsbibliothek zu Berlin, Bibliotheksreferendariat an der Herzog August Bibliothek Wolfenbüttel, 2019–2024 Leiterin der Abteilung Handschriften und Historische Drucke an der Universitätsbibliothek Tübingen, seit Oktober 2024 stellv. Leiterin des Referats Nachlässe und Autographen der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz.
Kristina.Stoebener@sbb.spk-berlin.de