Schwerpunkt Geschichtswissenschaften digital

Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft

Anwendung, Einordnung und Methodenkritik

von Tobias Hodel

I. Einleitung

Die Erkennung von Handschriften ist kein Kernthema der Geschichtswissenschaft, sondern eher eine wenig diskutierte Kompetenz von Historikerinnen und Historikern, die oft nicht nur gedruckte, sondern auch handschriftliche Dokumente auswerten. Wenn sie, in Archiv oder Bibliothek, nach Schriftstücken suchen, die sie für ihre Fragestellungen heranziehen können, und wenn sie diese Schriftstücke dann transkribieren, und sei es nur auszugsweise, investieren sie sehr viel Zeit. Von dieser Erfahrung ausgehend, scheint die maschinelle Erkennung von Text im Allgemeinen und Handschriften im Besonderen vor allem als zeitsparendes Hilfsmittel von Interesse für den Wissenschaftsbetrieb. Wechselt man die Perspektive und fragt, welche Bedeutung handschriftliches Material als empirische Grundlage historischer Forschung hat, kommt man zu einer anderen Einschätzung. Insbesondere wenn es um die Zeit vor 1900 geht, macht die handschriftliche Überlieferung einen großen Anteil der historischen Dokumente aus, die zu Quellen historischer Forschung werden können. Handschriftliche Dokumente computergestützt lesbar zu

source: https://doi.org/10.48350/173343 | downloaded: 3.5.2024

DOI 10.1515/hzhz-2023-0006

machen, würde also eine fundamentale Erweiterung der historischen Quellenbasis eröffnen. Einschränkend muss vorab konstatiert werden, dass der Fokus auf diesen Seiten auf Alphabetschriften liegt, für Sprachen mit umfangreicheren Zeichensätzen, insbesondere Chinesisch, ist die Technologie noch weniger ausgereift.

Die neuesten Entwicklungen in den Computerwissenschaften, der Mathematik und den digitalen Geisteswissenschaften diskutieren unterdessen Zugriffe auf historische Dokumente, bei denen es nicht nur um maschinelles Lesen (und Transkribieren) geht. Die neuen Techniken erlauben auch, die maschinell aufbereiteten Texte anders zu erforschen als bisher. Das lineare Lesen muss dann nicht mehr der primäre Zugang sein. Die damit verbundenen Chancen, Herausforderungen und auch die möglichen Gefahren, die im Folgenden diskutiert werden sollen, basieren auf der Anwendung von maschinellen Lernverfahren, insbesondere dem sogenannten *deep learning*. Weil diese Technologie nicht nur in den Geisteswissenschaften zu Anwendung kommt, ist die Auseinandersetzung damit zugleich eine Möglichkeit, den Einsatz von künstlicher Intelligenz auch in grundsätzlicher Hinsicht zu problematisieren. Immerhin bescheinigt man dem Einsatz dieser Technologie das Potential, die gesamte Gesellschaft in ihren sozialen Strukturen und politischen Vorgängen fundamental zu verändern. In jüngster Zeit wird deshalb nicht nur auf die Chancen, sondern ebenso auf die Schattenseiten des *deep learning* hingewiesen. Die Technologie kann etwa zu Marginalisierung von Minderheiten und zur Verstärkung rassistischer Stereotype führen.[1] Umso mehr scheint geboten, sich mit Verfahren des maschinellen Lernens zu befassen. Der vorliegende Aufriss verfolgt damit ein doppeltes Ziel: Die Texterkennung und dessen Potential für die Geschichtswissenschaft zu ergründen auf der einen Seite, Konsequenzen des Einsatzes von *machine learning* Verfahren offenzulegen auf der anderen.

II. Die Digitalisierung historischer Dokumente – ein Abriss

Ein zentrales Versprechen der Digitalisierung für die Geschichtswissenschaften lag von Beginn an im vereinfachten Zugriff auf historische Dokumente. Seit den Anfängen der breiten Digitalisierungswellen in den Geisteswissenschaften in den frü-

1 Siehe dazu die zentrale Studie zu Suchmaschinen von *Safiya Umoja Noble*, Algorithms of Oppression. How Search Engines Reinforce Racism. New York 2018.

hen 1990er Jahren wurden umfangreiche Bestände von historischen Dokumenten und Forschungsliteratur in digitale Bilder verwandelt und über das gleichzeitig stark ausgebaute und (inzwischen) niederschwellig zugängliche Internet (insbesondere das world wide web) in dieser Form weit über Archive, Museen und Bibliotheken hinaus in digitalen Kopien breit verfügbar.

1. Do-it-yourself-Erweiterung der Quellenbasis

Dank einiger Entwicklungsschritte im Bereich der analogen und digitalen Fotografie können seit bald fünf Jahren gar qualitativ gute Aufnahmen mit Smartphones gemacht werden, die weit verbreitet sind und ständig mitgeführt werden können. Für die digitale Bereitstellung von Digitalisaten muss nicht mehr auf Digitalisierungsoffensiven von Privaten und der öffentlichen Hand gewartet werden.

Damit wird die niederschwellige und rasche Auswertung digital vorhandener Dokumente weiter befördert und es erlaubt, die Quellenbasis der gesamten Disziplin zu erweitern. Dank der wertigen Linsen und hochauflösenden Fotozellen, die in Mobiltelefonen verbaut werden, spazieren wir mit einer Ausrüstung in Archive und Bibliotheken, die vor wenigen Jahren noch professioneller Fotoapparate bedurft hätte. Mit Unterstützung von kleinen Zelten (etwa dem ScanTent)[2] oder professionellen Scanner, die fix eingerichtet sind, können innerhalb von kurzer Zeit eine Vielzahl von Digitalisaten erstellt und mittels Texterkennung bearbeitet werden. Vorausgesetzt natürlich, dass die besuchten Institutionen die selbstgemachte Digitalisierung ihrer Bestände erlauben. Diesbezüglich ist aktuell im deutschsprachigen Raum ein Trend zu einer liberaleren Haltung spürbar.

Die Digitalisate erweitern konstruktiv, was von Seiten der Erinnerungsinstitutionen ohnehin selbst zur Verfügung gestellt wird und untergraben dadurch die Fokussierung auf zentrale Stücke oder Serien, die die Selbstdarstellung einer Institution verstärken. Obwohl noch nicht abschließend gelöst ist, wie die Dokumente idealerweise geteilt und gemeinsam bearbeitet werden können, gehen Digitalisate den ersten Schritt zu einer breiteren Quellenaufbereitung und -auswertung.[3] Vielleicht ist es denn auch das Versprechen der Texterkennung, die als Rückkoppelung die

2 Siehe ‹https://scantent.eu›.

3 Ein Ansatz, der sich vorwiegend auf Transkriptionen spezialisiert, ist das Basler Projekt „Transcriptiones“. Siehe ‹http://transcriptiones.ch/›.

Digitalisierung als Forschungsgrundlage befeuert und somit mental den Wandel zu einer Geschichtswissenschaft unterstützt, die auf großen Quellenkorpora besteht.

2. Die Entwicklung der OCR als erster Schritt, um Dokumente zugänglich zu machen

Bereits in den 1990er Jahren folgte die Entwicklung der OCR-Technologie (*optical character recognition*), die es ermöglichte, einen Großteil dieser digitalen Kopien maschinell zu durchsuchen. Digitalisierte Texte für Forschung und Öffentlichkeit leichter zugänglich zu machen, gehörte – unter anderem – auch zu den Zielen großangelegter Scan-Aktionen, beim Google Books-Projekt ebenso wie bei den europäischen und nationalen Digitalisierungsbemühungen, die das gedruckte Erbe in nichtkommerziellen Händen sichern wollten.[4] Beides trug stark zum Aufschwung der *digital humanities* bei, die anfangs noch *humanities computing* genannt wurden.[5]

Die Erkennung von gedruckten Texten mittels OCR basierte darauf, Bildausschnitte zu segmentieren und mit Formenbibliotheken abzugleichen. Relativ rasch wurde dabei ein Plateau der Erkennqualität erreicht, die zwar für moderne Drucke ausreichte, komplexere und weniger normalisierte Printerzeugnisse jedoch nicht in sinnvoller Qualität erkannte. Bis 2010 konnte etwa Fraktur, die im 19. Jahrhundert dominante Schrifttype des Deutschen, kaum erkannt werden, sodass eine Vielzahl früherer Scan- und OCR-Projekte zahlreiche Dokumente aus dem 19. Jahrhundert weitgehend ignorierten oder in minderwertiger Qualität aufbereiteten.

Die starke Verbesserung bei der Erkennung von Fraktur (größtenteils im Rahmen des EU-Projekts „Improving Access to Text", kurz IMPACT, entwickelt) fand parallel mit einer allgemeinen Verbesserung der Texterkennung durch die Verwendung von Wörterbüchern statt.[6] Dank dem Einsatz von – aus linguistischer Perspektive simplen – Sprachmodellen und der Sammlung und Publikation großer Wörterbücher wurde es ab 2010 möglich, die Texterkennung für standardisierte Sprachen, etwa

4 Insbesondere das französische Portal Gallica wurde gezielt als Gegenentwurf zu Google Books geplant und umgesetzt (s. *Bruno Racine*, Google et le nouveau monde. Paris 2010; *Alain Jacquesson*, Google Livres et le futur des bibliothèques numériques. Paris 2010; *Jean Noël Jeanneney*, Google and the Myth of Universal Knowledge. A View from Europe. Chicago 2007). Weitere Unternehmen mit dem Ziel, Scans und Volltext zur Verfügung zu stellen, sind *archive.org*, die *digital public library of america* (dpla) und das Meta-Portal europeana, das digitale und digitalisierte Kulturgüter aus dem gesamten europäischen Raum versammelt.

5 *Patrik Svensson*, Humanities Computing as Digital Humanities, in: Digital Humanities Quarterly 3/3, 2009; *Patrick Sahle*, Digital Humanities? Gibt's doch gar nicht!, in: Zeitschrift für digitale Geschichtswissenschaften, Sonderband 1, 2015. Online: ‹https://doi.org/10.17175/sb001_004›.

6 Zu IMPACT siehe: ‹http://www.impact-project.eu/›.

das Deutsche des 20. Jahrhunderts, nochmals wesentlich zu verbessern. Beides, die Erkennung von Fraktur und der Einbezug von Wörterbüchern, half jedoch insgesamt wenig, um Drucke vor 1830 besser zu erkennen. Die zu wenig normalisierte Schriftsprache stellte sich als zu hohe Hürde heraus. Der Einbezug von Wörterbüchern führte zur Hyperkorrektur, bei der ähnliche Wörter der Vorlage zu im Wörterbuch enthaltenen Schreibungen normalisiert wurden: „Marchen" wurden zu „Märchen", obwohl die Formerkennung das Trema, das aus einem „a" ein „ä" macht, nicht erkannt haben konnte. Offenbar schätzte das System die Wahrscheinlichkeit höher ein, dass die Formerkennung versagt hatte, als dass ein Wort außerhalb des Wörterbuchs an der fraglichen Stelle vorkam. Und auch die Fraktur blieb ein Problem, da die Buchstabenschnitte zu stark variierten und einander ähnelnde Buchstabenformen wie „Stab-s" und der Kleinbuchstabe „f" nicht unterschieden werden konnten.[7]

Zu den Texten, die von OCR nach wie vor nicht gut erkannt werden, gehören Editionen vormoderner Quellen. Teils sind die Ergebnisse schlicht unbrauchbar. Weil für vormoderne Sprachen jedoch keine Sprachmodelle vorliegen und wegen der variierenden Schreibung nur sehr aufwändig zu erstellen sind, kann der erkannte Text nur bedingt weiterverarbeitet werden. Auch für Suchzugriffe auf Editionen ist OCR nur mit Umsicht zu gebrauchen, weil die gewonnenen Resultate mehrheitlich nicht belastbar sind. Das gilt auch für Editionen, die lateinische Quellen aufbereiten: Dabei liegt das Problem im Sprachwechsel zwischen Kommentar und Quellen.

Gerade für die historische Forschung, die seit Jahrzehnten die Auswertung von einzelnen Texten, aber auch von großen Textkonvoluten betreibt, ist der leichtere Zugriff und die Möglichkeit, umfangreiche Texte und ganze Textbestände zu durchsuchen, zu einem wichtigen Treiber geworden, auf neue Weise Zusammenhänge zu identifizieren.[8] Einschränkend ist aber zu sagen, dass auf empirischer Ebene bislang

7 Im Rahmen von OCR-D wird die Problematik erstmals auch historisch untersucht: *Nikolas Weichselbaumer, Mathias Seuret, Saskia Limbach* et al., The Rapid Rise of Fraktur, in: Christof Schöch (Hrsg.), DHd 2020. Spielräume Digital Humanities zwischen Modellierung und Interpretation. Paderborn 2020, 229–232. Online: ‹https://doi.org/10.5281/zenodo.3666689›.

8 Siehe als Beispiel eine Auswertung mit Googles nGram Viewer: *Joanna Guldi*, The History of Walking and the Digital Turn: Stride and Lounge in London: 1808–1851, in: JModH 84/1, 2012, 116–144. Online: ‹http://www.jstor.org/stable/10.1086/663350›. Als Versuch der Zusammenstellung, welches Potential digitale Auswertungen mitbringen, siehe: *Lisa Spiro*, Access, Explore, Converse: The Impact (and Potential Impact) of the Digital Humanities on Scholarship, in: Juliette Hueber / Antonio Mendes da Silva (Eds.), Keys

vorwiegend die Forschungen zum 19. und 20. Jahrhundert von der Texterkennung mit OCR profitiert haben, da deren Quellengrundlage massiv verbreitert und um Dokumente ergänzt wurde, die zuvor nicht im Fokus der Forschung gestanden hatten. Für die Vormoderne, insbesondere die Frühe Neuzeit und das Mittelalter, blieben die Gewinne im Wesentlichen auf den vereinfachten Zugriff auf die ältere Literatur und – mit den beschriebenen Problemen – auf gedruckte Quelleneditionen beschränkt. Weitere Ungleichheiten ergeben sich mit Blick darauf, dass nicht alle Regionen der Welt über die notwendige Infrastruktur verfügen. Bestehende Sammlungen sind im globalen Süden viel seltener digitalisiert oder digital zugänglich, so dass die unterschiedlichen Möglichkeiten maschinell gestützter Forschung globale Ungleichheiten reproduzieren oder gar verschärfen können.[9] Das betrifft gerade die neuesten Entwicklungen im Bereich der Texterkennung, die (noch) recht ressourcenintensiv und kostspielig sind.

Einen Schub erfuhr die Handschriftenerkennung vor etwa vier Jahren mit dem Einbezug von Technologien des maschinellen Lernens: Neben Handschriften lassen sich so Schriftschnitte der Zeit vor 1830 zurück bis zu den Anfängen des Druckes in so guter Qualität erkennen, dass die Suchfunktion und die Textextraktion auch für die Forschung zur Vormoderne interessant werden. Die Druckbuchstaben werden dabei wie sehr regelmäßige Handschriften behandelt. Ohnehin wurden bei der Einführung neuer Lösungen zur besseren Texterkennung Druck- und Handschriften jeweils gleichermaßen bedacht. Die ersten Versuche zur Erkennung historischer und zeitgenössischer Handschriften in den 1990er Jahren führten jedoch nicht zu befriedigenden Lösungen. Und auch die Einführung von Sprachmodellen in den Erkennprozess, bei dem Buchstabenfolgen mit Methoden der Wahrscheinlichkeitsrechnung eruiert werden sollten (sog. Hidden Markov Models), brachten nicht den erhofften Erfolg. Um 2010 wurde ein damals völlig neuer Ansatz gewählt, bei dem neuronale Netze, die dem Bereich des *deep learning* zugeschlagen werden, für die Erkennung von Handschriften erprobt wurden. Bereits ein neuronales Netz mit simpler Architektur war fähig, die bislang erreichte Qualität der Texterkennung stark zu übertreffen. In der Folge konnten auf der Basis von *deep learning* Algorithmen erstellt

for Architectural History Research in the Digital Era. Handbook. Paris 2017 (Actes de colloques). Online: ‹http://books.openedition.org/inha/4925›.

9 *Lara Putnam*, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, in: AHR 121/2, 2016, 377–402. Online: ‹https://doi.org/10.1093/ahr/121.2.377›.

werden, die die Handschriftenerkennung für eine Vielzahl an Handschriften ermöglicht, und es ist inzwischen absehbar, dass maschinelles Lernen auch die Grundlage für zukünftige Verbesserungen und Erweiterungen im Bereich von Schrifterkennung sein wird. Um eine kritische Auseinandersetzung zu ermöglichen, sind Grundkenntnisse darüber erforderlich, wie *deep learning* in der Handschriftenerkennung technisch funktioniert.

III. Wie funktioniert Handschriftenerkennung?

Bei der maschinellen Handschriftenerkennung – *Handwritten Text Recognition* oder kurz HTR – strukturieren oder annotieren neuronale Netze, die in einem Trainingsprozess ‚selbständig' optimiert wurden, die eingespeisten Bilddaten. Diese Art der Datenverarbeitung beginnt auf vielen Feldern die bislang dominierende Analyse mithilfe von herkömmlichen Algorithmen zu ersetzen, die klaren Handlungsabläufen folgen. Bei einem herkömmlichen Algorithmus handelt es sich, so die Definition in der Wikipedia (Stand 2021), um eine „eindeutige Handlungsvorschrift zur Lösung eines Problems oder einer Klasse von Problemen. Algorithmen bestehen aus endlich vielen, wohldefinierten Einzelschritten".[10] Das Training und vor allem die Anwendung neuronaler Netze folgt nun genau nicht mehr einer solchen eindeutigen Handlungsvorschrift. Vielmehr werden Handlungen erst zugeordnet (trainiert) und in einem zweiten Schritt angewandt. Wie genau die maschinellen Handlungsergebnisse entstehen, wird dabei nicht vom Menschen kontrolliert und kann auch nicht wirklich nachvollzogen werden.[11] Wenn im Zusammenhang mit neuronalen Netzen von Algorithmen gesprochen wird, ist also etwas anderes gemeint, als die Definition von Wikipedia beschreibt. Ein Algorithmus – oft wird auch kurz von einem Modell gesprochen – bedeutet dabei eine durch Training optimierte Modellanwendung, die einen Input zu einem Output umformt.

Neuronale Netze sind analog zu geläufigen Vorstellungen über das Funktionie-

10 Persistente URL: https://de.wikipedia.org/w/index.php?title=Algorithmus&oldid=201404175.

11 Siehe dazu die Blogposts *Alexander Mordvintsev/Christopher Olah/Mike Tyka*, Inceptionism: Going Deeper into Neural Networks, Google AI Blog, 17.06.2015. Online: ‹https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html›, *Alexander Mordvintsev/Christopher Olah/Mike Tyka*, DeepDream – A Code Example for Visualizing Neural Networks, Google AI Blog, 01.07.2015. Online: ‹http://ai.googleblog.com/2015/07/deepdream-code-example-for-visualizing.html›.

ren des menschlichen Gehirns aufgebaut. Demnach fließen Informationen dabei in Form von elektrischen bzw. digitalen Impulsen durch vernetzte Nervenzellen, die Knoten in einem System mehr oder minder oft benutzter Verbindungen sind. Je nach Funktion der Nervenzellen werden die eingehenden Impulse vor der Weiterleitung an den nächsten Knoten verstärkt, gedämpft oder umgewandelt. Das Gehirn erzeugt auf einen Impuls hin eine Reaktion wie bspw. ein Schmerzempfinden; das künstliche neuronalen Netz gibt stattdessen einen Wert aus bzw. bewertet den Input anhand eines vordefinierten Rasters. Im Fall der Handschriftenerkennung gehören zu diesem Raster alle theoretisch vorkommenden Schriftzeichen – Groß- und Kleinbuchstaben, Leerzeichen, Satzzeichen und Ähnliches. Eine Eingabe, etwa ein Bild(-ausschnitt) eines Schriftzugs führt zu einer Ausgabe, etwa einem Wort(-teil). Die Frage, wie diese Resultate zustandekommen, wie die Algorithmen denn eigentlich funktionieren, wird nur selten gestellt. Was genau zwischen Eingabe und Ausgabe innerhalb und zwischen den künstlichen ‚Nervenzellen' geschieht (den sogenannten Schichten eines maschinellen neuronalen Netzes), ist im System der informatischen Architektur schwer nachzuvollziehen, auch wenn sich Zwischenergebnisse der Programme bzw. einzelner Programmschichten teilweise ableiten und interpretieren lassen. Zurzeit wird die Nachvollziehbarkeit des Weges zwischen Input und Output als Problem untergeordneter Bedeutung angesehen, weil die Resultate für sich sprechen: Künstliche neuronale Netze nehmen mit erstaunlichem Erfolg Bewertungen oder Zuordnungen vor, die menschlichen Leistungen nahekommen.

Bei der Erkennung von Handschriften werden digitale Bilder von handschriftlichen Dokumenten vorbereitend in Bilder von Zeilen zerlegt. Nur sehr kleine Ausschnitte dieser Bildzeilen – wenige Pixel große Teile – dienen dann den neuronalen Netzen als Input. Um das neuronale Netz zu trainieren, muss die gewünschte Ausgabe – die Transkription – bekannt sein und dem neuronalen Netz in Form einer Zeichenkette übermittelt werden. Alle in einem Trainingsset übermittelten Zeichenketten bilden zusammen ein geschlossenes Set aller möglicherweise zu erkennenden Zeichen. Dieses Set ist – je nach Vorlage – typischerweise auf den (historischen) Zeichenbestand einer oder mehrerer Sprachen begrenzt. Der maschinell errechnete Output wird mit der vorgelegten Transkription verglichen, also mit dem Ergebnis, das ein Mensch erstellt hat. Darauf folgt eine Phase der Optimierung, in der die einzelnen ‚Neuronen' – einzelne kleine Programmteile – maschinell korrigiert werden, um eine Ausgabe zu erreichen, die näher an der vorliegenden Transkription liegt.

Dieser Prozess wird mit möglichst viel Trainingsmaterial, also Bildschnipseln von Wörtern oder ganzen Zeilen mehrfach wiederholt, um das Netz so fein wie möglich anzupassen. Wie bei einem Kind, das lesen lernen soll, besteht dabei die Gefahr, dass einzelne Sequenzen, etwa ganze Zeilen, auswendig gelernt werden. In einem solchen Fall, dem sogenannten *overfitting*, kann ein trainiertes Netz zwar die bekannten (trainierten) Eingaben fast perfekt identifizieren, also die korrekte Transkription für die trainierten Bilder von Handschriften reproduzieren, es ist jedoch nicht fähig, ähnliche Inputs korrekt wiederzugeben. Selbst der Zeichenbestand des auswendig gelernten Abschnitts in veränderter Reihenfolge überfordert den maschinellen Algorithmus. Um diesem Effekt vorzubeugen, wird im Trainingsprozess sogenannter *noise* erzeugt. Dabei werden minimale Fehler in den Input eingeschleust, Bildausschnitte von wenigen Pixeln, damit der Input für das neuronale Netz jedes Mal leicht verändert wird.[12] Und auch zu einem späteren Zeitpunkt, wenn im neuronalen Netz nur noch digitale Signale von einer zur nächsten Schicht weitergegeben werden, werden zufällige Störungen eingebaut, um *overfitting* zu vermeiden.

Erkannter Text mit Sprachmodell	Erkannter Text ohne Sprachmodell	Vom Autor transkribierter Text
schulthes und ràtt unnd darzů von dem grossen rȧtt|sechzehn̄ die dann von den vier vennēn uf der hochenn|mitwucheñ darzů geordnet werden in dem ratthusz sich ver¬|anmen unnd zů anfang so gebütt min herr der schul¬|ches allen denen so dasind by iren eyden geswornen eyden	schul**c**hes unnd ràtt unnd darzů von dem grossen rȧtt |s**a**chzechn̄ die **b**ann von den vier vennēn uf**f** der hochenn¬|**n**itwuchn̄ darzů geordnet werden in dem ratthus sieh ver¬|anmer unnd zů anfang so gebütt min herr der schul¬|ches allen denen so dasind by iren eyde**m** geswornēn eyden	schulthes und ràtt unnd darzů von dem grossen rȧtt|sechzehn̄ die dann von den vier vennr̄n uf der hochenn|mitwucheñ darzů geordnet werden in dem ratthusz sich ver¬|samnen unnd zů anfang so gebütt min herr der schul¬|thes allen denen so *da sind* by iren eyden geswornen eyden

Kommentar:
Fett markiert sind Fehllesungen, die durch das Sprachmodell korrigiert werden. Die unterstrichen markierten Fehllesungen bleiben in beiden Lesungen falsch, weil Wörter vom Sprachmodell nicht über Zeilenumbrüche hinweg erkannt werden können. Aufgrund eines Fehlers in der Layouterkennung wird versammen (versammeln, unterstrichen) ebenfalls in beiden Erkennvorgängen falsch gelesen, da das „s" nicht als Teil der Zeile angesehen wird. *Kursiv* markiert ist eine Änderung der Getrennt- und Zusammenschreibung gemäß den modernen Transkriptionsrichtlinien.

12 Die Metapher des Sehens ist nur bis zu einem gewissen Punkt zutreffend. Zwar werden als Input Bildausschnitte eingespeist, aber auch zu einem späteren Zeitpunkt, wenn nur noch digitale Signale von einer zur nächsten Schicht weitergegeben werden, werden zur Vermeidung des *overfitting* zufällig Störungen des Netzes eingebaut.

Das Resultat eines solchen Trainingsvorgangs ist ein angepasstes neuronales Netz, das man als Modell bezeichnet. Dieses Modell ist in der Lage, dieselbe oder ähnliche Handschriften zu transkribieren, wobei der Ausdruck „transkribieren" einen wichtigen Schritt vor der Ausgabe einer Zeichenkette verschleiert: Die ausgegebene Zeichenkette (*string*) basiert auf einer primär generierten, in der Regel ausgeblendeten Variantentabelle (*confidence matrix*), die für alle möglicherweise vorkommenden Zeichen einen Wahrscheinlichkeitswert für deren Vorkommen in einem bestimmten Ausschnitt errechnet. Erst danach werden die Zeichen mit den höchsten Wahrscheinlichkeitswerten für die Ausgabe einer Zeichenkette zusammengestellt.

Bei der Errechnung der Ausgabewerte stehen der Maschine weder paläographisches Wissen noch Sprachkenntnisse zur Verfügung. Nach dem ersten Bewertungsvorgang kann das Ergebnis aber durch den Vergleich mit zwei sprachbasierten Komponenten verbessert werden. Erstens lassen sich die Ergebnisse mit einem Wörterbuch abgleichen, genauer gesagt, mit einer Liste von in einer Sprache überhaupt vorkommenden Zeichenketten (*strings*). So wäre die Zeichenkette „srpen" im Tschechischen vertreten – sie bedeutet „August" –, aber im Deutschen nicht. Zweitens kann eine Tabelle beigezogen werden, die angibt, wie häufig in einer Sprache bestimmte Zeichenfolgen vorkommen, nämlich Zeichenketten von fünf Zeichen, sogenannte Character 5-grams. Der Abgleich mit einem guten Wörterbuch verbessert zwar die Erkennung prozentual, erzeugt aber zugleich die Gefahr der Hyperkorrektur, also der Normalisierung eines Wortes hin zu einer im Wörterbuch angebotenen Schreibung, die an der zu erkennenden Stelle vielleicht gerade nicht vorkommt.[13] 5-gram Sprachmodelle sind etwas anders ausgerichtet. Sie basieren darauf, dass in geschriebenen Sprachen einige Zeichenfolgen mehr oder weniger ‚sinnvoll' bzw. wahrscheinlich sind. Aus einer Vielzahl von bereits erfassten Texten werden dabei alle möglichen Folgen von jeweils fünf Zeichen, inklusive Leer- und Satzzeichen, in einer Tabelle zusammengestellt, die auch die Häufigkeit des Vorkommens jeder Zeichenfolge als Wert enthält. Durch den Abgleich mit dieser Tabelle kann eine maschinell ermittelte Zeichenfolge als mehr oder weniger wahrscheinlich bewertet werden. Die Zeichenkette „srpen" dürfte in einem deutschen Sprachmodell mit einer Wahrscheinlichkeit von Null bewertet werden, doch ein Korrekturvorschlag

13 Siehe dazu auch *Tobias Hodel*, Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities, in: Lise Jaillant (Ed.), Archives, Access and AI. Working with Born-Digital and Digitized Archival Collections. Bielefeld 2022, 157–177, 171 f.

und damit die Gefahr einer Hyperkorrektur würde ausbleiben, während ein Wörterbuch womöglich eine Korrektur zu ‚setzen' vorschlagen würde.

Beim Abgleich der vom neuronalen Netz erstellten Tabelle (*Confidence Matrix*) erfolgt im Zusammenspiel mit dem Sprachmodell eine erneute Bewertung, die zur Ausgabe der nunmehr wahrscheinlichsten Zeichenkette führt. Insgesamt ergibt sich eine beträchtliche Verbesserung der Resultate. Bei brauchbaren Modellen der Handschriftenerkennung, deren Zeichenfehlerquote bei 5–10 Prozent liegt, reduziert sich die Fehlerquote nach der Verwendung von Sprachmodellen um 0,7–1 Prozent. Bei guten Modellen mit Fehlerquote von unter 5 Prozent reduzieren sich die Fehler noch um 0,3–0,7 Prozent. Doch nicht nur die Quantität der Fehler sinkt. Der Abgleich mit Sprachmodellen hat außerdem den Effekt, dass gerade gänzlich sinnlose Fehllesungen bzw. völlig ungebräuchliche Zeichenfolgen stark reduziert werden.

Unabhängig davon, ob Wörterbücher oder n-gram-Listen beigezogen werden, beginnt jedes Training ohne Kontextwissen, sofern man nicht auf ein schon bei anderen Gelegenheiten trainiertes neuronales Netz zurückgreift, also auf ein bestehendes Modell. Auch Trainingsvorgänge, aus denen das System früher gegebenenfalls Modelle für das Erkennen von Handschriften entwickelt hat, bleiben ohne Einfluss – will heißen, dass das neuronale Netz anfänglich zufällig ausgerichtet ist und erst im Laufe des Trainings eingegebene Formen einem der vordefinierten Zeichenwerte zuordnet. Die neuronalen Netze speichern dabei auch keinerlei Bilder, um diese etwa mit dem aktuellen Input zu vergleichen. Es handelt sich vielmehr um völlig vom Bild abgekoppelte Systeme, die zwar Bild(-teile) interpretieren können, selbst aber nur elektrische Signale weitergeben, die keine Ähnlichkeit mit einer Handschrift mehr aufweisen. Die Technologie ist also nicht vergleichbar mit der OCR-Technologie, die auf die optische Erkennung von einzelnen Zeichen abstellt und auf Bibliotheken basiert, die mögliche Schreib- und Druckvarianten für existierende Buchstaben, Zahlen, Satz- und Sonderzeichen enthalten, mit denen der Input abgeglichen wird. Die Art der Fehler der beiden Ansätze unterscheiden sich denn auch. Bei HTR werden etwa Zeichen häufiger verdoppelt, da es für den Algorithmus nur beschränkt eine Rolle spielt, ob ein Zeichen vollständig oder nur teilweise geschrieben wurde, während OCR unvollständig ausgeführte Zeichen eher ‚vergisst', also gar nicht ausgibt, und unvollständige Zeichen tendenziell einem vorgängigen oder nachfolgenden Zeichen zurechnet, also bei der Segmentierung von Bildern in einzelne Zeichen versagt.

Doch auch die Erkennung mit HTR zieht eine Vielzahl kleinerer und größerer Probleme nach sich: Ein Modell kann etwa nur solche Zeichen dekodieren, die im Rahmen des Trainings vorgegeben wurden. Wer (vormoderne) Rechnungen transkribieren lassen möchte, wird vermutlich scheitern, wenn das neuronale Netz anhand von Chroniken trainiert wurde, weil die Ligatur für „Pfund" dort kaum vorkommt, während sie in Rechnungen sehr gebräuchlich ist (℔). Problematisch ist auch, dass viele zurzeit verbreitete Transkriptionsrichtlinien vorschreiben, bestimmte Schreibungen bei der Transkription modernisiert wiederzugeben, insbesondere bei den Buchstaben „u" und „v": Während ein Schriftstück die Schreibung „vnd" enthält, erscheint gemäß modernen Transkriptionsrichtlinie also womöglich „und" in der menschlichen Transkription. Wird eine solche Transkription als Trainingsmaterial verwendet – also als erwünschter Output bereitgestellt – folgt das neuronale Netz dem menschlichen Vorbild. Es unterscheidet dann etwa zwischen den Buchstaben „v" und „u" aufgrund des konsonantischen bzw. vokalischen Einsatzes in einem Text. Das erzeugt die Tendenz, die Konvention unabhängig davon zu übernehmen, wie der Buchstabe als grafisches Zeichen tatsächlich geschrieben wurde. Dasselbe gilt für die Getrennt- und Zusammenschreibung, die in vormodernen Dokumenten oft vom heutigen Standard abweicht, für die Erkennung von Handschriften ist das ein zu vernachlässigendes Problem. Die Technologie wird jedoch in einer Vielzahl von Szenarien eingesetzt, die gebieten, diesem und ähnlichen Phänomen weiterhin Aufmerksamkeit zu schenken.

IV. Leistungsfähigkeit der Handschriftenerkennung und der Dokumentenanalyse

Der Blick in die Wettbewerbe der großen informatischen Konferenzen zur Dokumentenanalyse zeigt gleichzeitig, dass die Erkennung von Alphabetschriften, insbesondere lateinischer und Kurrenthandschriften, aus computerwissenschaftlicher Sicht – für den Moment zumindest – als gelöst gilt.[14] Die maschinelle Erkennung von Handschriften ist mit dem Einsatz neuronaler Netze in eine Phase eingetreten, in der Aufwand und Ertrag in einem akzeptablen Verhältnis stehen. Wenn mehrere

14 Weder an der ICDAR 2019 noch an der ICFHR 2020 fanden dazu Wettbewerbe statt.

Hundert oder gar Tausende von Seiten in ähnlichen Handschriften transkribiert werden sollen, lohnt sich der Aufwand, Trainingsdaten zu erstellen. In den letzten Jahren ist die Qualität der Handschriftenerkennung sprunghaft angestiegen; und zugleich braucht es dafür weniger Trainingsmaterial. Ende 2021 lassen sich Modelle trainieren, die mit bereits 50 Seiten (bzw. ca. 1000 Zeilen oder ca. 7000 Wörtern) Fehlerquoten von unter 5 Prozent der transkribierten Zeichen erzielen. Allerdings kann es in Einzelfällen schwierig sein, den Umfang des notwendigen Trainingsmaterials im Voraus abzuschätzen, weil die Resultate von der Komplexität und Regelmäßigkeit einer Handschrift abhängig sind. Als Faustregel bewährt sich: Für Menschen schwer zu lesende Schriften sind dies auch für die Maschine.[15]

Inzwischen muss aber nicht für jedes zu transkribierende Dokument immer ein von Grund auf neu optimiertes Modell erarbeitet werden. In vielen Fällen können vortrainierte Modelle beigezogen werden, die sich mit nur zwei Dutzend Seiten neuen Trainingsmaterials auf eine individuelle Handschrift hin optimieren lassen. Dabei ist es auch möglich, nach Ort oder Zeit zusammengefasste Dokumentenkorpora als Gruppe mit vortrainierten Modellen zu erkennen. Erste Tests zeigen, dass sich mit HTR verschiedene Drucktypen in einem generischen, d.h. einem allgemeinen Modell versammeln lassen, mit dem eine gute Erkennung möglich ist.[16]

Für einige Schriften, etwa Kurrent des 19. Jahrhunderts, sind inzwischen schon sogenannte generische Modelle entwickelt worden.[17] Diese Modelle vermögen eine Vielzahl einzelner Hände zu bündeln, sodass nicht nur eine individuelle Handschrift, sondern ganze Handschriftentypen erkannt werden. Gerade für deutsche Kurrent des 16. bis 19. Jahrhunderts oder für gotische Kursiven aus dem Mittelalter ist die Entwicklung solcher Modelle ein entscheidender Schritt, weil viele schreibende Hände dieser Zeit jeweils nur wenig umfangreich überliefert sind. In zahlreichen Fällen reicht die vorhandene Überlieferung also quantitativ nicht dafür aus, in-

15 Zu *best-practices* der Erkennung siehe *Tobias Hodel*, Best-practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale, in: Schöch (Hrsg.), DHd 2020 (wie Anm. 8), 84–87. Online: ‹https://doi.org/10.5281/zenodo.3666689›.

16 Dazu eignet sich etwa das freiverfügbare Druckmodell „Transkribus – print 0.3“. Online: ‹https://readcoop.eu/model/print-multi-language-danish-dutch-german-finnish-french-latin-swedish/›.

17 Zur Generalisierbarkeit von Handschriftenmodellen siehe hier und im folgenden *Tobias Hodel/David Schoch/Christa Schneider* et al., General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example, in: Journal of Open Humanities Data 7, 2021. Online: ‹https://doi.org/10.5334/johd.46›.

dividuell zugeschnittene Modelle zu trainieren. Generische Modelle sind aber auch deshalb wünschenswert, weil der Aufwand für Trainingsprozesse noch immer groß ist, was die Einsatzchancen der Texterkennungstechnologie reduziert. Solche Modelle schaffen hier Abhilfe, da sie einerseits Text bereits erfolgreich in einer ersten Version zugänglich machen und andererseits als Grundlagenmodell (sog. *base model*) für spezialisierte Handschriftenmodelle dienen können. Solche *base models* sind neuronale Netze, die bereits auf gewisse Schrifttypen trainiert sind und sich mit ‚wenigen' Zeilen Trainingsmaterials (300–400 Zeilen) auf eine Handschrift spezialisieren lassen.

Erste Erfahrungen belegen, dass sich mit bereits bestehender Technik solche generischen Modelle produzieren lassen. In Tests mit Kurrent und gotischen Buchschriften konnten damit Handschriften erkannt werden, die zuvor nicht trainiert worden waren. Dabei dringt man zu einer Qualität von weniger als 10 Prozent Zeichenfehlern im erkannten Text vor, was bereits recht gut lesbare Texte erzeugt, bei denen eine Korrektur deutlich weniger aufwändig ist als eine Transkription, die auf maschinelle Erkennung verzichtet. Bei der Zusammenstellung des Trainingsmaterials für die Entwicklung genereller Modelle darf die Menge an Schriftproben derselben Hand nicht zu groß sein. Ab mehreren Tausend Seiten einer Hand tendieren die Modelle dazu, sich darauf zu spezialisieren – eine Form des *overfitting* – und die Erkennung von unbekannten Handschriften verschlechtert sich. Aufgrund der zurzeit produzierten Menge an unterschiedlichen Trainingsmaterialien kann damit gerechnet werden, dass in den kommenden Jahren eine Vielzahl an generellen Modellen verfügbar wird.

Mit den Erörterungen über Trainingsmaterial, vortrainierte oder generische Handschriftenmodelle und Fehlerquoten ist die Leistungsfähigkeit der maschinellen Handschriftenerkennung nur scheinbar umrissen. Denn die möglichst genaue Wiedergabe eines handschriftlichen Textes als maschinell erstellte Transkription ist nur eine ‚Leistung' der Handschriftenerkennung, und zwar derjenigen, die sich am linearen Lesen eines Textes orientiert. Für eine Vielzahl von Anwendungen ist es aber ausreichend, wenn in einem Text möglichst alle Treffer für ein gesuchtes Wort gefunden werden. Auf den ersten Blick scheint das an die möglichst gute Erkennung des gesamten Textes gebunden zu sein. Es gibt aber Möglichkeiten, den Weg über den perfekten Text – der die Auffindbarkeit aller Instanzen einer Zeichenfolge garantieren würde – zu umgehen. Dabei spielen die oben angesprochenen *confidence matrices* eine Schlüsselrolle, die bei der Texterkennung mit neuronalen Net-

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u
0.00	0.00	0.97	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40
0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.81
0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.01	0.84	0.00	0.02
0.60	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00

Abbildung 1:
Eingefärbter Ausschnitt einer sogenannten *confidence matrix*, Ausgabe aus Transkribus, generiert aus der Erkennung des Wortes „causa", Schattierung aller Varianten mit einer Wahrscheinlichkeit ≥ 10 Prozent durch den Autor.
Jede Zeile gibt die Wahrscheinlichkeit eines Zeichens innerhalb eines Bildschnitts wieder. Links und rechts des hier präsentierten Tabellenausschnitts folgen weitere Zeichen wie Großbuchstaben, Satz- und Steuerzeichen. In der zweiten Zeile wird ein Vorschubzeichen ausgegeben, das bei der Textausgabe ersatzlos gestrichen wird.

zen angelegt werden. Denn auch bei Modellen mit schlechten Erkennraten (> 12 Prozent Zeichenfehlerquote) werden die korrekten Zeichen häufig zwar erkannt, jedoch nicht mit der höchsten Sicherheit (*confidence*) versehen. Eine confidence matrix lässt sich nur schwer lesen, da der Algorithmus an einer Vielzahl von Stellen ein sogenanntes „Vorschubzeichen" ausgibt. Der „Vorschub" bedeutet, dass der Algorithmus eine Erkennung bzw. eine Entscheidung zwischen verschiedenen möglichen Zeichen für noch nicht hinreichend gesichert hält und mit der Ausgabe abwartet.[18] Bei der Suche in der Variantentabelle (*confidence matrix*) wird die Zeichenkette jedoch aufgefunden und als mögliches Resultat ausgegeben, das überprüft werden kann.

Der Umweg über die Variantentabelle macht es möglich, ein sogenanntes „Keywordspotting" durchzuführen, ein Verfahren, das für das Auffinden von möglicherweise vorkommenden Zeichenketten höchst potent ist. Bereits bei Modellen, die für bestimmte Schriften nur mittelmäßige Resultate im Bereich von 15–18 Prozent Zeichenfehler ausgeben, werden mehr als 99 Prozent aller vorhandenen Instanzen einer Zeichenfolge aufgefunden (statistisch ausgedrückt ist der *recall* > 99 Prozent).

18 „Vorschub" sollte nicht verwechselt werden mit „Leerzeichen" oder ähnlichen visuell-sichtbaren Zeichen. Mit „Vorschub" wird einzig eine Unsicherheit ausgedrückt, die dazu führt, dass noch kein Zeichen als sehr wahrscheinlich vorkommend ausgegeben wird.

In diesem nahezu vollständigen Korpus an möglichen Treffern werden als Beifang auch nur vermeintliche Resultate (sogenannte *false-positives*) ausgegeben. Bei adäquater Instruktion der Suchenden ist dies jedoch kein Problem: Schließlich geht es bei einer Mehrzahl an Suchoperationen darum, alle oder zumindest möglichst viele Instanzen des Vorkommens aufzufinden. Der Ausdruck „Keywordspotting" gibt nicht wieder, dass diese Suchfunktion für beliebige Zeichenfolgen möglich ist. Ein bei der Entwicklung dieser Funktion unintendierter, aber nützlicher Nebeneffekt ist, dass auch unterschiedliche Schreibungen abgedeckt werden, da etwa Zeichenverdoppelungen (die in vormodernen Schreibungen häufig vorkommen) ebenso als möglich eingeschätzt werden wie Schreibungen ohne Doppelung.

In einigen wenigen Vorreiterprojekten wird *keywordspotting* bereits eingesetzt und kann mittels *post-processing* durch leistungsfähige Datenbanksysteme indexiert werden. Im Rahmen des Projekts READ (Recognition and Enrichment of Archival Documents) wurden mehrere Beispielsuchen über große Dokumentenmassen bereitgestellt, die für Websuchen aufbereitet wurden und innerhalb weniger Sekunden Resultate lieferten. Dazu gehören etwa die Schriften des britischen Staatsphilosophen Jeremy Bentham, die im Rahmen eines langjährigen Crowdsourcingprojektes in einer digitalen und gedruckten Edition präsentiert werden sollen. Über die bereitgestellte Websuche können bereits jetzt, mindestens zehn Jahre vor dem voraussichtlichen Abschluss der Editionsarbeit, eine Vielzahl der Schriften Benthams durchsucht werden.[19] Der Fall zeigt, wie interessant die beschriebene Suchfunktion für Erinnerungsinstitutionen sein kann, die gezielt ausgewählte Bestände preiswert aufbereiten und für alle Interessierten durchsuchbar machen können. Besonders serielle Quellen wie etwa Ratsprotokolle bieten sich für eine solche Aufbereitung an, wobei nachträglichen Korrekturen nichts im Wege steht, wenn sie von vornherein als Möglichkeit eingeplant werden.[20]

Bei der Erkennung handschriftlicher ebenso wie gedruckter Dokumente spielt das Layout oft eine wichtige Rolle, man denke nur an zweispaltige Texte, Seitenzahlen, Fußnoten oder Randnotizen. Sofern die Bestände ein einheitliches Layout haben, lässt sich die maschinelle Zeichenerkennung in einem vorbereitenden Schritt

19 *Louise Seaward*, Searching Jeremy Bentham's Manuscripts with Keyword Spotting, READ-Blog, 15.10.2018. Online: ‹https://read.transkribus.eu/2018/10/15/searching-benthams-manuscripts-kws/›.

20 Siehe dazu bspw. die Rezesse niederdeutscher Städte der Forschungsstelle für die Geschichte der Hanse und des Ostseeraums: ‹https://transkribus.eu/r/rezesse-niederdeutscher-staedtetage/#/›.

zusätzlich verfeinern. Auch dafür wird auf Technologien des maschinellen Lernens zurückgegriffen. Die Trainingsvorgänge beziehen sich dabei eben auf die Layouterkennung, also auf visuell abgesetzte Text- oder Bildsegmente, die identifiziert werden sollen. Regelmäßig auftretende Formen lassen sich in einem Modell zusammenfassen und werden in einem Schritt vor der Texterkennung identifiziert, was einer ersten semantischen Erkennung entspricht.[21] Die erkannten Dokumente lassen sich danach basierend auf den erkannten Schriftzeichen weiter anreichern. Auf unterschiedliche Bereiche können dabei unterschiedliche Texterkennungsmodelle angewandt oder nur bestimmte zentrale Teile korrigiert, durchsucht oder extrahiert werden.

Im Bereich der Layoutanalyse bestehen zurzeit allerdings auch noch Probleme. Um überhaupt den Schritt der Texterkennung zu erreichen, ist die einigermaßen korrekte Segmentierung der visuellen Seitenmerkmale nötig. Konkret müssen also die Zeilen bzw. die sogenannten *baselines*, die imaginären Linien, auf denen die Buchstaben liegen, richtig erkannt worden sein, damit überhaupt eine Erkennung von Zeichenketten erfolgen kann. Die Identifikation solcher texttragenden Teile eines Dokuments ist für Menschen, die in Kulturen mit hoher Schriftdurchdringung aufgewachsen sind, ausgesprochen simpel, selbst dann, wenn eine Zeile eher einem Bogen als einer Gerade folgt oder wenn die Seitenzahlen versehentlich von oben nach unten oder von der Mitte zum Rand springen. Für Computer ist dieser Schritt schwieriger. Einmal mehr erschwert wird das dadurch, dass die in diesem Arbeitsschritt eingesetzten neuronalen Netze kein Kontextwissen haben, sei es zur Farbe des Beschreibstoffs oder zum Aussehen von Materialschäden und Verunreinigungen, von Mäusefraß bis zu Kaffeeflecken. Wiederum treffen wir im Bereich der Lösungssuche auf trainierte neuronale Netze, die an der besseren Bewältigung der Aufgabe, nämlich der Identifikation von *baselines* und Textregionen tatkräftig mitwirken.[22] Seit der erfolgreichen Implementation von maschinellen Lernverfahren wird für eine Vielzahl von Dokumenten mit hoher Qualität das Layout, insbesondere die

---

21 Siehe dazu *Sofia Ares Oliveira/Benoît Seguin/Frédéric Kaplan*, dhSegment: A Generic Deep-Learning Approach for Document Segmentation, in: Umapada Pal/Lianwen Jin/Suresh Sundaram et al. (Eds.), Frontiers in Handwriting Recognition (ICFHR), 2018, 16th International Conference. Niagara Falls, NY 2018, 7–12, oder den Ansatz in *Lorenzo Quirós*, P2PaLA: Page to PAGE Layout Analysis toolkit, 2017. Online: ‹https://github.com/lquirosd/P2PaLA›.

22 *Tobias Grüning/Gundram Leifert/Tobias Strauss* et al., A Robust and Binarization-Free Approach for Text Line Detection in Historical Documents, in: Jean-Marc Ogier/Petra Gomez-Krämer/Iuliia Tkachenko et al.

baselines erkannt. Trotzdem kommt es in unregelmäßigen Abständen vor, dass ganze Zeilen übersprungen oder Randbemerkungen mit dem Haupttext zu einer Zeile zusammengefasst werden. Wiederum gilt, dass für eine Suche die Probleme zwar bekannt sein müssen, jedoch nicht zwangsläufig zu falschen Resultaten führen. Anders sieht es aus, wenn ein Text mit sehr hoher Textgüte extrahiert werden soll; in diesem Fall muss die maschinelle Zeilenerkennung überprüft werden.

Allen Fortschritten zum Trotz ist die Erkennung von Tabellen bislang eine ungelöste Herausforderung geblieben. Die an und für sich äußerst regelmäßigen Informationswiedergaben stellen sich für Maschinen, jedenfalls in Form handschriftlich ausgefüllter Tabellen, äußerst problematisch dar. Sogar wenn die Tabelle über vorgedruckte Rahmen oder Linien verfügt, ist die Unterscheidung von Druckvorlage und handschriftliche Eintragung eine noch nicht gelöste Aufgabe. In den Computerwissenschaften ist das Problem bekannt und wird zurzeit in Form von Wettbewerben erforscht.[23] Für Rechnungsbücher, die in großer Zahl in Archiven mit frühneuzeitlichen Beständen liegen und nicht nur über wirtschaftliche, sondern auch soziale- und kulturelle Verhältnisse Auskunft geben können, ist eine maschinell unterstützte Zeichenerkennung mit adäquater Wiedergabe von Tabellen, Spalten oder ähnlichen Layouts also noch nicht möglich. Die Lösung des Problems wird aller Voraussicht nach einmal mehr in der Anwendung von maschinellen Lernverfahren liegen.

V. Anwendungschancen in einer digitalisierten Geschichtswissenschaft

Die maschinell gestützte Texterkennung und insbesondere die Handschriftenerkennung hält für die Geschichtswissenschaft eine ganze Reihe von neuen Möglichkeiten bereit, die von der Suche über die Erschließung bis hin zur Auswertung von historischen Dokumenten reicht.

(Eds.), 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017. Kyoto 2017, 236–241. Online: ‹https://doi.org/10.1109/ICDAR.2017.47›.

23 Siehe bspw. den Wettbewerb an der ICDAR 2019 zu Tabellen: *Hervé Déjean/Jean-Luc Meunier/Liangcai Gao* et al., ICDAR 2019 Competition on Table Detection and Recognition (cTDaR), 23.04.2019. Online: ‹https://doi.org/10.5281/zenodo.3239032›.

Bei der raschen Durchsicht, dem *skimming* von Texten, das gerade bei engen Fragestellungen, die auf ein weites Korpus angewandt werden, üblich ist, lässt sich nicht nur Zeit sparen, sondern mithilfe angepasster Suchfunktionen auch der Ertrag an potentiell einschlägigen Treffern erhöhen. Das gilt insbesondere, wenn der Ansatz der Texterkennung auf umfangreiche Bestände angewandt wird, die mittels Volltextsuche in korrekt erkanntem Text oder dem technisch ausgefeilten *keyword-spotting* erfasst werden. Bei Dokumenten, die mit einer Fehlerquote von weniger als 5 Prozent erkannt werden, kann eine (herkömmliche) Volltextsuche eingerichtet werden, die mit kleinen „Levensthein-Distanzen" operiert.[24] Diese auch als Editierdistanz bezeichnete Größe beschreibt die Anzahl an Schritten, die vorgenommen werden, um von einer zur nächsten Zeichenkette zu kommen. Kleine Distanzen, also die Ersetzung, Einfügung oder Löschung von ein bis zwei Zeichen, werden mittlerweile von diversen Datenbanksystemen unterstützt und sind ausgesprochen hilfreich, wenn Schreibungen variabel sind oder die Erkennung nicht perfekt ist. Wer nach „Not" sucht, findet dann auch „Noth", und die Suche nach „Geld" bringt auch „Gelt" ans Licht. Somit lässt sich über große Textkonvolute hinweg mit einer hinreichend kleinen Bandbreite nach Zeichenfolgen suchen, auch wenn Begriffe oder Namen variabel geschrieben sind. Mit dem Einbau solcher Variabilitäten werden in gewissem Umfang, wie schon erwähnt, auch *false-positives*, also nur vermeintliche Treffer gemeldet, was in den meisten Fällen unproblematisch sein dürfte. Allerdings kann die Suchfunktion bei gleichermaßen ähnlichen wie häufigen Zeichenfolgen – man denke an „Hans", „Haus" und „Hand" – zu sehr umfangreichen Trefferlisten mit so vielen Fehltreffern führen, dass die Bereinigung der Resultate einigen Aufwand bedeutet.

Die Variantenerkennung und der Einbezug unsicherer Lesungen im *keyword-spotting* über die *confidence matrices* ermöglichen es, für die häufigsten europäischen Schrifttypen (lateinisches Alphabet sowie Kurrentschrift und kyrillische Schriften) innerhalb von kurzer Zeit ein System aufzusetzen, das mit sehr hoher Wahrscheinlichkeit alle Treffer eines einzelnen Begriffs bzw. einer Zeichenkette auffindbar macht. Bei unregelmäßigen Händen oder Wörtern mit sehr ähnlichem Buchstabenbestand wird die höhere Zahl von *false-positives* zu einem gewissen Nachteil. Entspre-

24 *Vladimir Iosifovich Levenshtein*, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, in: Soviet Physics Doklady 10/8, Feb. 1966, 707–710.

chende Schwierigkeiten sind bei der Konzeption von Forschungsvorhaben frühzeitig abzuklären und bei der Planung zu bedenken.

Von den kurz skizzierten Suchfunktionen in digitalisierten handschriftlichen Beständen ausgehend eröffnet sich eine Reihe von Anwendungsszenarien.

Keywordspotting bzw. Volltextsuche dürfte bald auch in (virtuellen) Archiven und Bibliotheken mit digitalisierten Handschriftenbeständen möglich sein und wird eine zusätzliche Strategie der Heuristik werden, der Recherche nach möglichen Quellen für eine historische Untersuchung.

Ganz neue Chancen bieten die neuen digitalen Techniken für eine Erneuerung der Begriffsgeschichte. Durch die Erweiterung der leicht zugänglichen Quellenbasis lässt sich das Vorkommen von Begriffen sehr viel genauer prüfen als bisher. Außerdem können ausgewählte Korpora nun auch wesentlich besser quantifizierend und vergleichend untersucht werden. Wenn gefragt wird, welche Begriffe in welchen Dokumententypen überhaupt gebraucht wurden, lässt sich die ‚alte' Begriffsgeschichte mit neueren Ausrichtungen der historischen Semantik in Beziehung bringen.[25] Es wird davon ausgehend die Arbeit der zukünftigen Historiker*innen in Zusammenarbeit mit Linguist*innen sein, Mittel und Wege zu finden, kreative Suchanfragen zu formulieren, die auch unterschiedliche Schreibungen berücksichtigen und somit zu Treffern führt, die die Varietäten berücksichtigen. Dadurch werden Überlegungen notwendig, die die Verknüpfung von Konzepten und Wörtern offenlegen.

Anhand einiger Beispiele lassen sich die neuen Untersuchungsmöglichkeiten gut aufzeigen. Der feste Ausdruck „Schutz und Schirm", der in den 1990er Jahren in der deutschsprachigen Mediävistik aufbauend auf Otto Brunners „Land und Herrschaft" äußerst kontrovers diskutiert wurde[26], könnte auf weit breiterer Quellenbasis und vor allem mit Rückgriff auf Kontexte im Sinne von umgebenden Wörtern

25 *Kathrin Kollmeier*, Begriffsgeschichte und Historische Semantik, Version 2.0, in: Docupedia-Zeitgeschichte, 29.10.2012. Online: ‹http://docupedia.de/zg/kollmeier_begriffsgeschichte_v2_de_2012›. Als Beispiel einer Arbeit, die das analog in Bezug auf die Begriffe „brief" und „schrift" für einen großen Urkundenbestand umgesetzt hat, siehe *Joseph Morsel*, Brief und Schrift: Überlegungen zu den sozialen Grundlagen schriftlichen Austauschs im Spätmittelalter am Beispiel Frankens, in: Ludolf Kuchenbuch/Uta Kleine (Hrsg.), Textus im Mittelalter. Göttingen 2006, 285–323.

26 Siehe insbesondere *Gadi Algazi*, Herrengewalt und Gewalt der Herren im späten Mittelalter. Herrschaft, Gegenseitigkeit und Sprachgebrauch. Frankfurt am Main 1996; *Christine Reinle*, Bauernfehden. Studien zur Fehdeführung Nichtadliger im spätmittelalterlichen römisch-deutschen Reich, besonders in den bayerischen Herzogtümern. Stuttgart 2003.

neu interpretiert werden. Mit einfachen Auswertungsmethoden lassen sich nämlich häufige Wörter im Umfeld eines Begriffs identifizieren, die dessen Bedeutung oder semantische Färbung genauer zu erkennen geben. Auch der formelhafte Sprachgebrauch in spezifischen Satzkonfigurationen, wie er etwa in Abgaben- oder Steuerverzeichnissen auftritt, lässt sich mit Unterstützung digitaler Techniken breiter analysieren.[27] Ein weiteres Feld, das die maschinelle Handschriftenerkennung fruchtbar bereichern kann, betrifft das Spektrum von Gebrauchssituationen, in denen einzelne Wörter auftreten können. Ansatzweise geben das manche Wörterbücher heute in einem Set exemplarischer Zitate an – als Vorzeigebeispiel genannt sei nur das Deutsche Rechtswörterbuch. Künftig werden sich solche Vorhaben viel umfassender realisieren lassen. Für das historische Verständnis und die Kontextualisierung von ‚Bedeutung' stehen also neue Analysemethoden bereit, die sich auf wesentlich umfassendere Überlieferungsbestände abstützen können.[28]

Etwas anders gelagert sind Forschungen, die zwar einen Begriff ins Visier nehmen, sich aber nicht nur für dessen Bedeutung, sondern ebenso für dessen spezifische Gebrauchsumstände interessieren. Diese Vorgehensweise der Umfeldsuche ist insbesondere in der Computerlinguistik beliebt, kann aber auch für historische Fragestellungen genutzt werden, insbesondere wenn es um die Konnotationen und den Einsatz einzelner Wörter geht. Aufbauend auf einer digital vorliegenden, äußerst umfangreichen Quellensammlung[29] haben das Tim Geelhaar und Silke Schwandt für die zwei lateinischen Begriffe „Christianitas" und „Virtus" demonstriert.[30] Beide haben mit dem in Frankfurt entwickelten Tool HSCM (Historical Semantics Corpus Management) gearbeitet, das auf Dokumente zugreift, die als manuelle Transkription erschlossen oder mittels OCR erkannt und anschließend nachbearbeitet wur-

27 Siehe dazu beispielhaft *Ludolf Kuchenbuch*, Die dreidimensionale Werk-Sprache von Theophilus presbyter: „Arbeits"-semantische Untersuchungen am Traktat *De diversis artibus*, in: Ders., Reflexive Mediävistik. Textus, Opus, Feudalismus. Frankfurt am Main 2012, 341–401.

28 Programmatisch aufgezeigt in *Alain Guerreau*, L'avenir d'un passé incertain: Quelle histoire du Moyen Âge au XXIe siècle? Paris 2001.

29 Neben ausgewählten Beständen der MGH ist die Basis des Korpus die „Patrologia": *Jacques-Paul Migne*, Patrologiæ cursus completus: sive bibliotheca universalis, integra, uniformis, commoda, oeconomica, omnium SS. patrum, doctorum, scriptorumque ecclesiasticorum. Series latina. Paris 1844.

30 *Tim Geelhaar*, Christianitas. Eine Wortgeschichte von der Spätantike bis zum Mittelalter. (Historische Semantik, Bd. 24.) Göttingen 2015; *Silke Schwandt*, Virtus. Zur Semantik eines politischen Konzepts im Mittelalter. (Historische Politikforschung, Bd. 22.) Frankfurt am Main 2014.

den, so dass das Quellenkorpus in zwar nicht fehlerfreier, aber sehr hoher Textgüte zur Verfügung steht.[31]

Analog zu diesem einen Korpus sind weitere Sammlungen und Korpora am Entstehen, die ein *text mining* erlauben, das zwar eine gute Texterkennung, aber nicht den perfekt erkannten Texten voraussetzt. Solche Korpora können auf intertextuelle Bezüge hin untersucht werden, auf Abschreibe- und Textübernahmepraktiken (analog zu stemmatischen Auswertungen in der Editorik).[32] Solche Ansätze sind von zentraler Bedeutung für rezeptionsgeschichtliche Forschungen von der Geschichtsschreibung bis hin zur Rechtsgeschichte und Diplomatik – sei es in Bezug auf Vorlagennutzung, Privilegienbestätigungen oder Urkundenfomular. Maschinell erkannte Volltexte können auch stilistisch analysiert werden, so dass sich Verfasser*innen identifizieren und vorliegende Zuschreibungen besser prüfen oder Verfasserkreise umreißen lassen. Praktiken der Textgenese, aber auch Transfers von Denkfiguren von einem Handlungs- oder Wissensfeld hin in ein anderes oder auch andere Formen kultureller Approbation lassen sich auf dieser Basis erweitert oder methodisch neu untersuchen.

Am anderen Ende der Bearbeitungsskala steht die Auswertung des Textes beziehungsweise der Zugriff auf große Textmengen, aber der Methodenapparat des *distant reading* kann helfen, um mehr oder weniger große Dokumentenkorpora zu sondieren, zu untersuchen und miteinander in Beziehung zu setzen.

Dazu bietet sich *topic modeling* an, das spezifische Lösungen für bestehende Bedürfnisse, etwa alternative Zugänge zu großen Dokumentenserien, bereitstellt. Algorithmen eruieren dabei Themenfelder aus Dokumenten beziehungsweise den darin enthaltenen Zeichenketten. Es handelt sich dabei um einen Vorgang, der versucht, häufig vorkommende Zeichenketten zusammenzuführen und die Wahrscheinlichkeit errechnet, dass zwei Zeichenketten aufeinanderfolgen bzw. inner-

31 *Silke Schwandt*, Inducing Linguistic Networks from Historical Corpora. Towards a New Method in Historical Semantics, in: Paul Bennett / Martin Durrell / Silke Scheible / Richard J. Whitt (Hrsg.), Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache. (New Methods in Historical Corpora, Bd. 3.) Tübingen 2013, 257–274.

32 Siehe als Ansatz: *Marco Büchler / Philip R. Burns / Martin Müller* et al., Towards a Historical Text Re-use Detection, in: Chris Biemann / Alexander Mehler (Eds.), Text Mining. From Ontology Learning to Automated Text Processing Applications. (Theory and Applications of Natural Language Processing.) Cham 2014, 221–238. Online: ‹https://doi.org/10.1007/978–3-319–12655–5_11›.

halb einer vordefinierten Kette an Wörtern vorkommen.[33] In ähnlichen Dokumenten häufig vorkommende Zeichenketten werden am Ende des Prozesses als *keywords* (nicht zu verwechseln mit *keywordspotting*) einem „Themenfeld" zugeschlagen. Parallel dazu wird errechnet, welche Themenfelder in welchen Dokumenten mit welcher Häufigkeit auftauchen. In aller Kürze aufzeigen lässt sich dies anhand einer Auswertung der Zürcher Regierungsratsprotokolle, die zwischen 1803 und 1883 verschriftlicht wurden.[34] Ein generiertes Themenfeld fokussiert auf Kriegsaktivitäten, die aus *keywords* wie *Kriegsrathe* oder *Infanterie* und *Truppen* besteht. Das Themenfeld zeigt sich erwartungsgemäß im Korpus vorwiegend zu Zeiten der Mobilmachung oder bei tatsächlichen kriegerischen Handlungen (etwa vor und während des Sonderbundskriegs 1847).[35]

Die Interpretation, welche Themen tatsächlich ein „Themenfeld" ausmachen, bleibt den Forschenden überlassen. Die erfolgversprechenden Topic Modeling-Anwendungen basieren wie die zuvor genannten Technologien auf maschinellem Lernen, wobei es sich nicht um eine *supervised* – also auf Trainingsmaterial basierende – Anwendung handelt, sondern um eine *unsupervised* Clusteringform, bei der aus den vorgegebenen Korpora eine Struktur identifiziert werden soll. Da die letztendliche Auswertung oder besser gesagt die Zuschreibung von häufigen Wörtern zu Themen ein hermeneutischer Prozess ist, bleibt die Deutungshoheit bis zu einem hohen Grad beim Menschen – wobei natürlich auch die Mischung der automatisierten Texterkennung mit *topic modeling* Ansätzen denkbar ist, um große unkorrigierte Datenmengen auf unterschiedliche Arten zugänglich zu machen.[36]

---

33 Zu *topic modeling* siehe einführend *David M. Blei*, Introduction to Probabilistic Topic Models, in: Communications of the ACM 55/4, 2012, 77–84. Online: ‹https://doi.org/10.1145/2133806.2133826›; *Christof Schöch*, Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, in: Digital Humanities Quarterly 11/2, 2017. Online: ‹http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html›. Kritisch eingeordnet in *Ben Schmidt*, Words Alone: Dismantling Topic Models in the Humanities, in: Journal of Digital Humanities 2/1, 2013. Online: ‹http://journalofdigitalhumanities.org/2–1/words-alone-by-benjamin-m-schmidt/›.

34 Zum Projekt siehe online: ‹https://archives-quickaccess.ch/search/stazh/rrb›.

35 Siehe dazu *Hodel*, Supervised and Unsupervised (wie Anm. 14), 165–167.

36 Erste Erkundungen in diesem Bereich führten noch nicht zu Erfolgen, siehe dazu *Stephen Mutuvi/Antoine Doucet/Moses Odeo* et al., Evaluating the Impact of OCR Errors on Topic Modeling, in: Milena Dobreva/Annika Hinze/Maja Žumer (Eds.), Maturity and Innovation in Digital Libraries. 20th International Conference on Asia-Pacific Digital Libraries, ICADL 2018, Hamilton, New Zealand, November 19–22, 2018, Proceedings. (Lecture Notes in Computer Science, Vol. 11279.) 2018, 3–14. Online: ‹https://doi.org/10.1007/978–3-030–04257–8_1›.

Die Ansätze können sequentiell angewandt werden, beginnend bei der Identifikation von Textstruktur über Texterkennung zu *keywordspotting* und *topic modeling*, lassen sich Dokumente aufbereiten und auf unterschiedliche Arten durchsuchbar machen. Allen Ansätzen ist die Nutzung von *machine learning* gemein. Ohne Einsatz der selbstlernenden Technologie wäre keiner der Ansätze denkbar. Alle zuvor entwickelten Algorithmen waren in der Leistungsfähigkeit nicht annähernd so erfolgreich. Gleichzeitig bedeutet dies jedoch, dass die Probleme des maschinellen Lernens mannigfaltig im gesamten Aufbereitungsprozess gespiegelt werden.

VI. Maschinelles Lernen als Herausforderung des 21. Jahrhunderts

Der Umgang mit neuronalen Netzen und dem Bereich des *deep learning* ist in den Geschichts- wie auch den Geisteswissenschaften allgemein noch nicht wirklich Teil des Methodenapparats. Eine Beschäftigung mit der imitierenden Selbstlernstrategie von Computern ist jedoch nicht nur aufgrund des erfolgreichen Einsatzes in den Geisteswissenschaften angebracht, sondern auch dadurch, dass die Technologie vermehrt in alltägliche Entscheidungsprozesse eingebunden wird. Von Suchmaschinen über Onlineshopping bis zu Bewerbungsprozessen, Algorithmen des maschinellen Lernens und vor allem vermehrt aus dem Bereich des *deep learning* werden mit Trainingsdaten (etwa vergangenen Entscheidungen) gefüttert, damit Modelle entwickelt werden können, die Nutzerverhalten voraussagen und scheinbar effiziente Entscheidungen treffen. Da die Systeme daran gemessen werden, ob sie vergangene Muster bestätigen und an exakt solchen Mustern trainiert werden, entsteht eine Verstärkung der menschlichen Voreingenommenheit (*human bias*).

Daraus generierte ‚Fehler', etwa die Identifikation von Menschen dunkler Hautfarbe als Gorillas[37], wird mitunter auf problematische Daten geschoben, also die Begründung vorgeschoben, dass offenbar zu wenig Trainingsmaterial mit Bildern von Schwarzen vorlagen (analog von zu wenig transkribierten Vorlagen zum Training eines Handschriftenmodells), die korrekt zugeordnet waren. Die Begründung selbst zeugt jedoch vom Missverständnis, dass die Technologie an sich als ‚neutral' verstan-

37 *Noble*, Algorithms of Oppression (wie Anm. 2), 6–11.

den wird und erst durch menschlichen Input die Vorurteile beziehungsweise vorherrschende Machtverhältnisse replizieren. Die Sammlung und speziell auch die Strukturierung von Daten erfolgt zwar ohne bösartige Absicht, widerspiegelt aber im Kern nicht nur die Vorurteile, sondern bewegt sich in kulturell geprägten Kategorienkatalogen, die reproduziert und durch die Algorithmen verstärkt werden. Das Beispiel der Identifikation als Menschenaffen lässt sich dadurch nachvollziehen, dass neben Bildmaterial von Menschen mit vorwiegend weißer Hautfarbe und der Einspeisung von Zootieren in dasselbe Netz, vereinfach gesagt, zur maschineninternen Logik „Gesicht + hell = Mensch" und „Gesicht + dunkel = Menschenaffe" führte.

Wie die „All Models"-Initiative treffend ausführt, sind eben alle Modelle politisch und benötigen menschlichen Input.[38] Da die Modelle sich an Quantitäten orientieren bzw. die Ausrichtung der Neuronen nur durch Eingaben mit ähnlichen, aber nicht identischen Formen erreicht wird, ist die Zusammenstellung von Trainings- und Validierungskorpora zu einer heiklen Aufgabe geworden mit teilweise ungeahnten Konsequenzen, werden doch die Trainingskorpora häufig wiederverwendet, weil die Erstellung der Datensätze aufwändig – da kontrolliert – geschehen muss.

Für das Feld der Texterkennung ist das größtenteils unproblematisch. Einzig überbrachte Vorstellungen von „Text" als Folge von Zeichen wird übernommen und Sets von Zeichen als gegeben vorausgesetzt. Aus Sicht der Geschichtswissenschaft, die sich mit Texten aus der Vormoderne beschäftigt, beinhaltet dies ein kleines Gefahrenpotential, da teilweise unklar ist, was für Zeichen als bekannt bzw. genutzt angenommen werden können. Die paläographische Diskussion, was ein Zeichen ist und was eine Abwandlung eines Zeichens, wird etwa implizit über die Publikation von Editionen und darin genutzten Zeichensets geführt. Je nach Editionsform wird somit zwischen Stab- und Rund-s-Formen unterschieden, wobei sich die literaturwissenschaftlich und linguistisch orientierte Forschung häufiger auf genauere Wiedergaben konzentriert, während historische Editionen Verallgemeinerungen vornehmen.[39] Aus technischer Perspektive spielt es keine Rolle, ob Stab- und Rund-s zu einer Form zusammengefügt oder aber als individuelle Zeichen definiert werden.

38 Siehe ‹https://allmodels.ai/index.html›. Beim menschlichen Input bezieht sich die Initiative vorwiegend auf versteckte Arbeit (etwa im Rahmen des Amazon Mechanical Turk Programms) zur Aufbereitung von riesigen Datensets, die zum Training von neuronalen Netzen benötigt werden.

39 Als Beispiel, das großen Wert auf Zeichengenauigkeit legt, siehe das Projekt Ambraser Heldenbuch: ‹https://www.uibk.ac.at/projects/ahb/›.

Neuronale Netze sind darauf ausgelegt, dass sie imitierend Ausgaben machen, was gleichzeitig auch bedeutet, dass ein identifiziertes Zeichen in keiner Weise mit einem graphischen Zeichen in Verbindung steht. Erst wenn bei der Dekodierung mehr als ca. 200 unterschiedliche Zeichen erlernt werden sollen, führt das bei den aktuell eingesetzten Systemen zu massiv schlechteren Resultaten. In Konsequenz bedeutet dies, dass etwa chinesische Schriftzeichen mit aktuellen Systemen für die Handschriftenerkennung nicht erkannt werden können.

Eine deutlich wahrnehmbare Stufe problematischer und weitaus folgenreicher wird die Übernahme der Technologie bei der Identifikation von Entitäten, also Personen und Orten, die in Texten als Zeichenfolgen identifiziert werden können. Die Frage, was eine Person ist, etwa was Name und was Zuschreibung einer Person und somit welche Merkmale einer Person zugeschrieben werden in und durch einen Text, wird verstärkt zurückgespiegelt. Mit dem Einsatz des *deep learning* für solche Vorgehensweisen expliziert und verstärkt sich das Problem zunehmend.

Deutlich wurde das Problem im Umgang mit Suchmaschinen benannt. In ihrem aufschlussreichen und luziden Werk, „Algorithms of Oppression" macht Safiya Umoya Noble auf die Deutungsmacht von Algorithmen bei der Gewichtung von Suchresultaten aufmerksam.[40] Obwohl die Algorithmen nicht zur Aufbereitung von Text verwendet werden, ist die dahinterliegende Logik vergleichbar und die genutzte Technologie identisch. Ziel kommerzieller Suchmaschinen ist es nicht nur, die für einen individuellen Nutzer als zentral erachteten Informationen darzustellen, sondern gleichzeitig auch auf Produkte aufmerksam zu machen, die damit im Zusammenhang stehen oder auf Webseiten zu verweisen, die für eine Platzierung im oberen Bereich der Resultate bezahlen (und aus Gründen der Transparenz – zumindest leicht – markiert werden).

Im Gegensatz zu Suchmaschinen sind geisteswissenschaftliche Anwendungen aufgrund des verhältnismäßig beschränkten Nutzerkreises wenig interessant für kommerzielle Anbieter, digitale Infrastrukturen wie Bibliothekskataloge oder Archivfindmittel mal ausgenommen. Dadurch ist die Gefahr einer Konfrontation mit problematisch geprägten Algorithmen kleiner. Wegen der guten Resultate (wie oben in III gezeigt) wird die Technologie zukünftig vermehrt nicht nur in den hochdigitalisierten Teilen der Geschichtswissenschaft eingesetzt, sondern auch zumin-

40 *Noble*, Algorithms of Oppression (wie Anm. 2).

dest in mediatisierter Form über entsprechende Recherchetools und Plattformen für alle Historikerinnen und Historiker relevant.[41] Da die Zunft sich bereits seit längerer Zeit kritisch mit Entstehungsumständen und Überlieferungswegen von Dokumenten beschäftigt[42], wird die kritische Reflexion digitaler Korpora analog dazu Einzug halten in den Methodenapparat der Geschichtswissenschaft.

Aus Sicht der Informatik und der Mathematik ist die Anwendung neuronaler Netze auf geisteswissenschaftliche Probleme ebenfalls von großem Interesse, da in den Geisteswissenschaften hochinteressante Probleme formuliert werden, die sich nicht über statistische Anwendungen und streng formulierte Lösungsprozesse bearbeiten lassen.

Das beschriebene Potential der Anwendung von Algorithmen und neuronalen Netzen auch und insbesondere in den Geisteswissenschaften konnte eindrücklich demonstriert werden, solange mit datenschutzrechtlich unbedenklichem Material gearbeitet wird. Die Technologie eröffnet einerseits Zugang und Auswertungsansätze unter Einbezug großer Quellenmengen und andererseits handelt es sich bei historischen Quellen vielfach um verhältnismäßig unproblematische Dokumente, die als Spielwiese dienen, um mit einer an sich äußerst problematischen Technologie umgehen zu lernen. Erfahrungen mit dieser noch jungen Technologie verhelfen nicht nur im Hinblick auf zukünftiges Forschungsdesign zu Aufschlüssen, sondern auch zu Einsichten, wie Software strukturiert und gebaut ist, die das Leben auf dem ganzen Planeten aktuell und zukünftig stark beeinflusst.

VII. Schluss

Maschinelles Lernen basierend auf neuronalen Netzen ist dabei, die Welt umzukrempeln, größere und kleinere Firmen optimieren ihre Entscheidungsvorgänge mit Hilfe von Algorithmen, die auf maschinell erlernten Modellen basieren. Auch in den Geschichtswissenschaften lässt sich dies beobachten und anhand der Handschriftenerkennung direkt erleben. Bei der Handschriftenerkennung wird es jedoch

41 Siehe dazu etwa das Projekt Impresso, mit Zugriff auf Themenfelder: ‹https://impresso-project.ch/app/topics/tm-fr-all-v2.0_tp97_fr›.

42 Siehe als Prototyp *Arnold Esch*, Überlieferungs-Chance und Überlieferungs-Zufall als methodisches Problem des Historikers, in: HZ 240, 1985, 529–570.

auch in der Disziplin nicht bleiben, sondern es werden vielmehr Aufbereitungs- und Entscheidungsprozesse zur Aufbereitung großer Dokumentenmengen damit zukünftig unterstützt. Als nächstes etwa bei der Identifikation von Eigennamen oder der Vernetzung von Schlagwörtern mit Konzepten oder ganzen Ontologien.

Dabei bleibt durchweg problematisch, dass *deep learning* in der Tendenz und basierend auf Trainingsmaterial menschliche Vorurteile verstärkt und dadurch Einschätzungen als vermeintlich objektivierte, algorithmisierte Ausgabe taxiert. Der Umgang mit Algorithmen des maschinellen Lernens erlaubt daher, praktische Probleme, die die Nutzung von *deep learning* mit sich bringen, anhand von bekanntem und kontrollierbarem Material auszutesten und profund zu kritisieren. Die Kenntnisse der Methode verhelfen entsprechend zu einer Auseinandersetzung mit virulenten Problemen einer Gesellschaft in der Digitalisierung.

Die Algorithmen selber werden dabei als rhetorische Maschinen verstanden[43], deren Aussagen zwar ernst genommen, jedoch gleichzeitig kritisch analysiert werden müssen. Nur so lassen sich die durchaus erfolgversprechenden Auswertungsformen des *distant reading* von *topic modeling* über Autorattribution bis hin zu semantischen Auswertungsformen verantwortungsvoll nutzen und das Potential einer digitalen Geschichtswissenschaft ansatzweise ausschöpfen. Ein Blick auf das Feld der *software and technology studies* (STS), die die wissenschaftliche Erforschung der Konzeption und des Verhaltens von Software sowie dem Umgang damit zum Ziel hat, lohnt sich. Die STS gehen von der Erkenntnis aus, dass jede Art von (Computer-)Programm auf epistemologischen und methodologischen Vorannahmen beruht und mit Blick darauf untersucht werden muss, ähnlich wie jede andere Form der Verarbeitung oder Interpretation. Das Feld ist seit wenigen Jahren stark am Wachsen und bedient sich hermeneutischer Methoden ebenso wie Analysen aus Programmiersicht und Projektmanagementperspektiven.[44]

Im Gegensatz zu effekthascherischen Voraussagen, die künstliche Intelligenz vorwiegend mit dem Verlust von Arbeitsplätzen gleichsetzen und gar mit der Übernahme der Entscheidungsgewalt durch Computersysteme, erlaubt die Kooperation

43 Nach *Estee Beck*, A Theory of Persuasive Computer Algorithms for Rhetorical Code Studies, in: enculturation. journal of rhetoric, writing, and culture, 02.11.2016. Online: ‹http://enculturation.net/a-theory-of-persuasive-computer-algorithms›.

44 Einführend siehe *Tara Andrews*, 'Software and Scholarship' – Editorial, in: Interdisciplinary Science Reviews 40/4, 02.10.2015, 342–348. Online: ‹https://doi.org/10.1080/03080188.2016.1165456›.

von Mensch und Maschine eine sinnvolle Arbeitsteilung, die die Stärken symbiotisch zusammenführt.

Die Orientierung hin zu Auswertungsmechanismen und den dahinter liegenden Algorithmen befreit den Wissenschaftsbetrieb gleichzeitig aus der Unmündigkeit, die durch oberflächlich simple, häufig kommerzielle Systeme induziert werden. Eine Blackbox wie etwa Google Books können somit nicht nur umgangen, sondern auch Aussagen zur Qualität und damit verbunden zu Auswertungsmöglichkeiten und -grenzen gemacht werden, wobei diesbezüglich noch weitere Erfahrungen gesammelt werden müssen. Analog zur gut eingebundenen Methodenkritik, ist eine Reflexion der genutzten digitalen Helfer in den Geisteswissenschaften aktuell und zukünftig mehr denn je notwendig, wie prominent von Andreas Fickers gefordert.[45] Die Anwendung von Algorithmen wird somit zum Teil der angewandten Methode, die im Rahmen jeder geschichtswissenschaftlichen (und geisteswissenschaftlichen) Auseinandersetzung kritisch eingeordnet werden muss.

Der großflächige Einsatz digitaler Technologien führte insbesondere in den Informationswissenschaften, aber auch anderen geisteswissenschaftlichen Feldern in den vergangenen Jahren zu einer Sensibilisierung im Umgang mit Algorithmen und damit verbundenen Infrastrukturen. Untersucht man die Rhetorik von Algorithmen, spielt der Akt der Überzeugung eine zentrale Rolle, der parallel zu den Argumentationslinien von Fachwissenschaftler*innen nachvollzogen, akzeptiert und unter Umständen auch verworfen werden kann. Passenderweise wird so die Handlungsmacht, die *agency*, von Dingen auch im Sinne Latours mitgedacht[46], wie die Notwendigkeit zur Analyse der intern ablaufenden Vorgänge. Maschine und Mensch stehen somit in einer kooperativen und vor allem mehr oder minder gleichberechtigten Stellung. Aus dieser Perspektive wird es interessant, die Ausgabe von Algorithmen nicht nur auf ihr (bisweilen subjektiv interpretierte) Richtigkeit hin zu untersuchen, sondern mit den Ausgaben zu arbeiten und die Algorithmen als Agierende ernst zu nehmen.

Wenn wir aktuell mit quantifizierenden Aussagen wie Zeichen- oder Wortfehler-

---

45 *Andreas Fickers*, Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?, in: Zeithistorische Forschungen 1, 2020, 157–168. Online: ‹https://doi.org/10.14765/ZZF.DOK-1765›.

46 *Bruno Latour*, Eine neue Soziologie für eine neue Gesellschaft. Einführung in die Akteur-Netzwerk-Theorie. Frankfurt am Main 2010.

quoten operieren, ist nur beschränkt fassbar, was dies bedeutet. Vielmehr sind griffige Angaben erstrebenswert, die nach Verlusten bei der Suche nach Zeichenketten (Wörtern) und darauf aufbauend der Belastbarkeit von Aussagen zu Worthäufigkeiten machen. Die digitale Geschichtswissenschaft bewegt sich so innerhalb eines kontrollierbaren und methodisch abgesicherten Raums. Damit kontern wir auch die Ansicht, dass *big data* zu einer theoriefreien Wissenschaft geführt hätte, die aus Mustern Schlüsse ziehen könne ohne eingehende Diskussion eingebrachter Perspektiven. Eine Geschichtswissenschaft wird daraus entstehen, die gar noch stärker auf theoretische Fundierung und methodische Diskussion abstellt.

Zusammenfassung

Das Lesen historischer Handschriften gehört seit Jahrhunderten zum hilfswissenschaftlichen Methodenapparat, die automatisierte Erkennung entspricht somit einer potentiellen Arbeitserleichterung, die den Geschichtswissenschaften neue Perspektiven eröffnet. Um die Technologie und ihr Potential einzuschätzen, wird erläutert, welche Faktoren für die erfolgreiche Erkennung entscheidend sind und wie die Resultate weiterverarbeitet werden können. Mittels *machine* oder *deep learning* wird der Methodenapparat aber auch um ein Lernverfahren erweitert, das auf große Datenmengen angewiesen ist und aus den zugrundeliegenden Daten (Be-)Wertungen übernimmt, die häufig zu unerwünschten Nebeneffekten führt. Anhand der Handschriftenerkennung kann so eine Technologie erfahren und kritisch eingeordnet werden, die aktuell in diversen Wissenschaftsbereichen und unserem täglichen Leben Einzug hält.

Vielen Dank an Claudia Moddelmog für die kritische Durchsicht und die Diskussionen. Alle Links wurden letztmals am 17.01.2022 abgerufen.

Prof. Dr. *Tobias Hodel*, Universität Bern, Digital Humanities, Walter Benjamin Kolleg, Mittelstrasse 43, 3012 Bern