SFB 1199
Processes of Spatialization
under the Global Condition

Laura Rebecca Klettke

Lesen und lesen lassen – Ein interdisziplinäres Praktikum zur maschinellen Verarbeitung raumsemantischer Texte der Geographie.

Mit einer einleitenden Vorbemerkung von Ninja Steinbach-Hüther und Manuel Burghardt

Working paper series des SFB 1199 an der Universität Leipzig
Nr. 30

Collaborative Research Centre (SFB) 1199
„Processes of Spatialization under the Global Condition“
at Leipzig University

Funded by

DFG Deutsche Forschungsgemeinschaft

Laura Rebecca Klettke
Lesen und lesen lassen – Ein interdisziplinäres Praktikum zur maschinellen Verarbeitung raumsemantischer Texte der Geographie. Mit einer einleitenden Vorbemerkung von Ninja Steinbach-Hüther und Manuel Burghardt

This working paper is part of the Working Paper Series of the Collaborative Research Centre (SFB) 1199 “Processes of Spatialization under the Global Condition”. This working paper is also part of the Working Paper Series of ReCentGlobe, to which the SFB 1199 contributes since 2020.

Mein besonderer Dank gilt dem Teilprojekt C01 für die Möglichkeit, meinen Praktikumsbeitrag im Rahmen eines Working Papers zu veröffentlichen. Dirk Hänsgen, René Reuter und Ninja Steinbach-Hüther danke ich für ihre tatkräftige Unterstützung bei der redaktionellen Umsetzung dieses Working Papers. Laura Rebecca Klettke

12 / 2022

Vertrieb:
Leipziger Universitätsverlag GmbH, Oststraße 41, 04317 Leipzig
info@univerlag-leipzig.de

ISBN: 978-3-96023-455-5
ISSN: 2510-4845

Inhaltsverzeichnis

Vorbemerkung

Der vorliegende Bericht geht aus dem Praktikum hervor, welches Laura Rebecca Klettke als Studentin der Digital Humanities von Anfang Dezember 2020 bis Anfang Februar 2021 am Leibniz-Institut für Länderkunde (IfL) absolvierte. Es war das erste Praktikum, das im Rahmen des Forschungsprojektes „Raumsemantiken der Geographie im 19. und 20. Jahrhundert" in der Abteilung Theorie, Methodik und Geschichte der Geographie betreut wurde (Projektleitung: Ute Wardenga, weitere Wissenschaftler*innen im Projekt: Ninja Steinbach-Hüther, Dirk Hänsgen).[1] Darüber hinaus handelt es sich auch um das erste Praktikum einer Studentin aus dem Bereich der Digital Humanities überhaupt am IfL.

Das Praktikumsmodul (mit Manuel Burghardt als Studiengangs- und Modulverantwortlichem) ist Teil des im Wintersemester 2018/2019 erfolgreich gestarteten Masterprogramms im Bereich Digital Humanities,[2] welches konsekutiv zum bereits länger bestehenden Bachelor Digital Humanities konzipiert wurde. Während im Bachelor wichtige wissenschaftliche Grundlagen geschaffen werden, besteht im Master Digital Humanities ein stärkerer Praxisbezug, der sich u. a. in einem verpflichtenden Praktikumsmodul widerspiegelt. Ziel ist, im Rahmen dieses Moduls Studierende mit verschiedenen Institutionen, die sich der Gedächtnisarbeit und Pflege des Kulturerbes widmen (GLAM – galleries, libraries, archives and museums), sowie mit Forschungseinrichtungen zusammenzubringen, um ihnen so Einblicke in die praktische Projektarbeit und die Vernetzung mit einschlägigen Akteur*innen zu ermöglichen. Die Vermittlung erfolgt mithilfe einer Praktikumsbörse zum Semesterstart, bei der Projekte für das gemeinsame Arbeiten gewissermaßen einem Wettbewerbsverfahren im Sinne eines Eignungstests unterzogen werden.

Ninja Steinbach-Hüther und Dirk Hänsgen haben an der Praktikumsbörse im Oktober 2019 teilgenommen, um das Projekt vorzustellen und zunächst einmal die Interessenlage der Studierenden des Faches auszuloten. Im Projekt geht es auf Basis der Analyse internationaler geographischer Zeitschriften um die Frage nach „Raumsemantiken der Geographie im 19. und 20. Jahrhundert", ihre Etablierung, ihren Wandel im historischen Verlauf, um Resemantisierungen und sprachübergreifenden Transfer. Die methodische Vorgehensweise ist durch hermeneutische Verfahren in Kombination mit computergestützten Anwendungen gekennzeichnet. Den thematischen Überbau des Projektes bildet der SFB 1199 zu „Verräumlichungsprozesse[n] unter Globalisierungsbedingungen" (SFB 1199).[3] Der Sonderforschungsbereich startete im Jahr 2016 und befindet sich mittlerweile in seiner zweiten, bis Ende 2023 andauernden Förderphase. Dass es sich anbieten würde, Teile der Forschungsfragen computergestützt zu beantworten, war bereits klar und aufgrund von Vorerfahrungen aus anderen Forschungsprojekten[4] Teil der Projektplanung. Aber jedes Projekt unterliegt eigenen Herausforderungen, sodass wir speziell für die Analyse von Raumsemantiken der Geographie im Projekt C01 herausfinden wollten, ob es für Studierende der Digital Humanities überhaupt genügend praktische Bezüge

1 Es handelt sich um ein Teilprojekt des Sonderforschungsbereichs 1199 (SFB 1199) zu „Verräumlichungsprozesse[n] unter Globalisierungsbedingungen" und läuft unter dem Kürzel C01. Collaborative Research Center (SFB) 1199, „Section C01: Spatial Semantics of Geography in the 19th and 20th Centuries", https://research.uni-leipzig.de/~sfb1199/projects/project-c1/ (letzter Zugriff 17. Oktober 2022).

2 Universität Leipzig, „Digital Humanities (M. Sc.)", www.uni-leipzig.de/studium/vor-dem-studium/studienangebot/studiengang/course/show/digital-humanities-m-sc (letzter Zugriff 17. Oktober 2022).

3 Siehe Anmerkung 1; Collaborative Research Center (SFB) 1199, *Section C01*.

4 U. Wardenga u. a., „Von einer Geographie der Verräumlichung zu Geographien von Raumsemantiken: Digital Humanities als Schlüssel", in: M. Middell (Hrsg.), *Verräumlichungsprozesse unter Globalisierungsbedingungen*, Leipzig: Leipziger Universitätsverlag, 2021, S. 45–70; N. Steinbach-Hüther, *Bibliotheksdaten, Kulturtransfer und Digital Humanities: Zu einer Methodik bei der Untersuchung transregionaler Zirkulationen akademischer Literatur afrikanischer Autoren*, Leipzig: Leipziger Universitätsverlag, 2020; N. Steinbach-Hüther u. a., *Geographiegeschichtsschreibung und Digital Humanities: Neue Methoden für Zeitschriftenanalysen*, Working paper series des SFB 1199 an der Universität Leipzig 15, Leipzig: Universitätsverlag Leipzig, 2019; T. Efer und N. Steinbach-Hüther, „Quantitative Analyses in Global and Area Studies using Graph-based Filtering of Heterogeneous Catalogue Data", in: E. Plödereder u. a. (Hrsg.), *Informatik 2014: Big Data – Komplexität meistern; Tagung der Gesellschaft für Informatik, 22.–26. September 2014 in Stuttgart, Deutschland*, Bonn: Ges. für Informatik, 2014, S. 1027–1037.

bieten würde, die im Studienfach erworbenen Kenntnisse innerhalb der anberaumten Praktikumszeit exemplarisch auszutesten.

Dazu haben Ninja Steinbach-Hüther und Dirk Hänsgen anlässlich der Praktikumsbörse den forschungsperspektivischen Hintergrund des Projektes C01 kurz umrissen. Das bietet sich selbstverständlich auch an dieser Stelle an, um die im Praktikumsbericht von Frau Klettke dargelegten Tätigkeiten besser einordnen zu können: In der ersten Förderphase untersuchten wir in einem Kernteam von vier Personen (Ute Wardenga, Maximilian Georg, Ninja Steinbach-Hüther und Dirk Hänsgen), welches von zahlreichen studentischen Hilfskräften mit verschiedensten Sprachkenntnissen unterstützt wurde, zunächst wissenschaftliche Zeitschriften der internationalen Geographie seit 1821 (Gründungsjahr der ersten Geographischen Gesellschaft weltweit in Paris) bis 1914 (Beginn des Ersten Weltkrieges). Wir gingen orientiert an Ansätzen des Distant Reading[5] vor und entwickelten eine belastbare Methodik der Datenerfassung, -strukturierung und -generierung, die uns in späteren Arbeitsschritten weitere Analysemöglichkeiten ermöglichen sollte.[6]

Die betreffenden Zeitschriften wurden von Geographischen Gesellschaften weltweit herausgegeben, häufig im gegenseitigen Schriftentausch zugänglich gemacht und dienten der Information über geographische Fragen und Entwicklungen in verschiedensten Weltregionen. Die herausgebenden Geographischen Gesellschaften fungierten dabei nicht nur als mehr oder weniger aktive Akteure in Globalisierungsprozessen, sondern zeichneten sich durch ein hohes Maß an raumbezogenen Beobachtungen aus, welches in den Zeitschriften nicht nur demonstriert, sondern in den Texten auch konsolidiert und in die Öffentlichkeit getragen wurde. Dabei ist wichtig zu verstehen, dass Geographische Gesellschaften spätestens seit Mitte des 19. Jahrhunderts ein globales Phänomen waren und zahlreiche Expeditionen und explorative Forschungsreisen anregten. Diese führten insbesondere in die Polargebiete, nach Zentral-, Ost- und Südasien sowie nach Afrika. Geographische Gesellschaften sammelten häufig die Mittel zur Ausstattung der Reisen – besonders auch mit den neuesten verfügbaren Messinstrumenten, um qualitativ hochwertiges Datenmaterial aus den bis dahin für sie unbekannten Regionen zu erhalten – und übernahmen die Dokumentation, Verarbeitung und/oder kartographische Visualisierung der Ergebnisse dieser Reisen in den von ihnen herausgegebenen (wissenschaftlichen) Zeitschriften und Reihen.

Seit Anfang der 1870er Jahre waren sie es, die unter Nutzung ihrer vergleichsweise guten personellen und infrastrukturellen Ausstattung an wechselnden Orten internationale Geographenkongresse organisierten und so als Forum für die Herausbildung einer rasch wachsenden internationalen *Scientific Community* von Geographen fungierten. Zahlreiche europäische Gesellschaften setzten sich überdies für eine Hebung der Standards des Erdkundeunterrichts ein und nutzten ihre Verbindungen zu nationalen bzw. regionalen Eliten von Politik, Verwaltung und Wirtschaft, um die Einrichtung von geographischen Lehrstühlen an Universitäten voranzutreiben.

Im Ergebnis der Erhebungen der ersten Förderphase konnten strukturelle Spezifika der einzelnen Geographischen Gesellschaften, Verflechtungen zwischen den Standorten und die thematischen und regionalen Schwerpunktbildungen über das lange 19. Jahrhundert hinweg im internationalen Vergleich ermittelt werden.

Wir können anhand des existierenden Forschungsstandes, aber auch bezogen auf die Entwicklungen an Schulen und später an Universitäten, von einer „Geographisierung der Diskurse" ab dem 19. Jahrhundert ausgehen, was aber nicht dazu geführt hat, dass der Geographie allein, die sich als Universitätsdisziplin erst im letzten Drittel des 19. Jahrhunderts herausgebildet hat, Deutungskompetenz für weltweit verlaufene Prozesse der Neuverräumlichung zugeschrieben wird. Trotzdem haben diese Diskurse einen großen und wichtigen Anteil daran, wie Verräumlichungsprozesse

5 F. Moretti, *Distant Reading*, Konstanz: Konstanz University Press, 2016.

6 N. Steinbach-Hüther u. a., *Geographiegeschichtsschreibung und Digital Humanities*.

beschrieben wurden.[7] Es sind diese Diskurse, denen wir in der zweiten Förderphase des Projektes auf die Spur kommen wollen. So ist Gegenstand der Analysen in der zweiten Förderphase, die im Jahr 2020 begann und Ende 2023 abgeschlossen sein wird, das komplexe Geflecht von Verräumlichungslogiken, Formatierung von Räumen und der Herstellung von Raumordnungen. Hierzu haben wir den Untersuchungszeitraum ausgedehnt und den Beobachtungshorizont auf die Analyse des Wandels von Raumsemantiken und deren imaginationsgesteuerte Aufladungen gelegt.

Das Projekt ist in zwei Teilstudien eingeteilt. Teilstudie 1 arbeitet mit den zu den Geographischen Gesellschaften erhobenen Daten aus der ersten Förderphase; Teilstudie zwei nimmt zwei ergänzende Zeitschriftenkorpora für die Zwischenkriegs- und Kriegszeit (1924–1944) sowie für die (weitere) Nachkriegszeit (1949–2020) in den Blick. Den am Praktikum Interessierten verdeutlichten wir, dass sich die Arbeit im Praktikum nur auf die erste Teilstudie beziehen würde, die unter Nutzung von Methoden der Digital Humanities das in der ersten Förderphase erarbeitete Datenmaterial (1821–1914) analysiert, um dem synchronen und diachronen Wandel von Raumsemantiken im anglophonen, frankophonen und deutschsprachigen Bereich mit Blick auf Prozesse des Kulturtransfers systematisch auf die Spur zu kommen.

Im Praktikum sollte es um die computergestützte Arbeit an ausgewählten Volltexten des Textkorpus von Zeitschriften Geographischer Gesellschaften gehen, um systematisch den Weg auf der Suche nach Raumsemantiken und ihrer Analyse zu verfolgen. Es handelte sich um ein Beispielsample aus Volltexten im PDF-Format in deutscher, englischer, französischer und spanischer Sprache, die in einem ersten Schritt maschinenlesbar gemacht werden sollten, um in einem zweiten Schritt erste analytische Versuche zu Raumbegrifflichkeiten vorzunehmen. Die Idee war, dass in einem Folgesemester an diesem Material tiefergehende Analysetechniken und -methoden in Bezug auf raumsemantische Fragen erprobt werden könnten. Kurzum, Ziel waren die Entwicklung eines prototypischen Workflows vom (Ursprungs-)Digitalisat (unterschiedlicher Qualität) zu einem korpus-/analysefähigen Produkt sowie erste raumbegriffliche Auswertungsansätze des Textmaterials (erster Ordnung = konkrete Raumbenennungen und zweiter Ordnung = raumbezogene Sprache[n]).

Laura Rebecca Klettke begeisterte sich bei der Praktikumsbörse für das Projekt und wir uns für die Arbeit der Studentin, sodass es letztlich zum Praktikum am IfL kam. Wir blicken auf rund zwei Monate Zusammenarbeit zurück, die von verschiedenen technischen und inhaltlichen Übersetzungsschritten geprägt waren. So mussten nicht nur die Texte in maschinenlesbares Material umgewandelt werden, sondern auch die Forschungsfrage und ihre Bearbeitung, die sich aus einer interdisziplinären Verknüpfung von Geographie, Global Studies und Kulturwissenschaften ergab, für das Feld der Digital Humanities „übersetzt" werden. Der Mehrwert, der für beide Seiten dadurch erfahrbar wurde, ist, dass wir gegenseitig erproben konnten, was funktioniert und was nicht, wo mehr oder weniger „Übersetzungsarbeit" geleistet werden sollte und welche Aussicht das alles für weiterführende Kooperationen im Rahmen von Praktika bzw. weiteren unterstützenden Tätigkeiten hat etc.

Zunächst einmal musste Frau Klettke verstehen, um was für Material es sich handelte, um was es in den Texten geht und warum computergestützte Arbeitsschritte von Vorteil waren. Wir wiederum mussten die Voraussetzungen, die unsere Praktikantin mitbrachte, besser einschätzen lernen. Gegenseitig war es notwendig, sich auf verschiedene Denkweisen einzulassen und letztlich eine gemeinsame Sprache zu entwickeln. Daher waren regelmäßige Absprachen und eine engmaschige Betreuung von Vorteil. Zur Dokumentation der Arbeitsschritte auch im Rahmen des im Projekt vollzogenen Datenmanagements legten wir Textdokumente und Tabellen an. Ziel war, die Tätigkeit für andere Kolleg*innen sowohl im Projekt selbst, aber auch darüber hinaus zu dokumentieren. So sollten Kenntnisse und das im Praktikum erworbene Wissen geteilt und auf Schwierigkeiten, Herausforderungen und Lösungswege beim Umgang mit Originaltexten auf dem Weg zu maschinenlesbaren Korpora hingewiesen werden.

7 U. Wardenga u. a., „Von einer Geographie der Verräumlichung zu Geographien von Raumsemantiken"; N. Steinbach-Hüther, *Bibliotheksdaten, Kulturtransfer und Digital Humanities*.

Wichtig zu erwähnen erscheint uns in diesem Zusammenhang auch, dass Laura Rebecca Klettkes Tätigkeit in Projekt C01 nicht mit dem Praktikum endete, sondern dass wir sie letztlich als studentische Hilfskraft im Projekt weiterbeschäftigen wollten und konnten. Auch Frau Klettkes Masterarbeit geht aus Vorarbeiten im Rahmen des Projektes hervor, die dann deutlich vertieft wurden. Sie wurde im Februar 2022 unter dem Titel „Anwendung von Optical Character Recognition auf die Zeitschriften Geographischer Gesellschaften sowie eine sprach- und zeitabhängige Analyse der semantischen Veränderung metageographischer Begriffe" erfolgreich eingereicht und von Manuel Burghardt als Erstbetreuer und Ninja Steinbach-Hüther als Zweitbetreuerin begleitet. So ergab sich aus dem für rund zwei Monate angesetzten Praktikum ein gemeinsamer Austausch, der auch nach dem Praktikumsabschluss im Rahmen der SHK-Beschäftigung und Vorbereitung auf die Masterarbeit fortwährte.

Ein Ergebnis dieses Austausches und dieser Arbeiten stellt Laura Rebecca Klettkes Praktikumsbericht dar, in dem sie auf die Anfänge dieses Arbeitsprozesses und seine Herausforderungen näher eingeht. Nicht erst nach einigen Rückfragen im Kreis der Kolleg*innen des SFB 1199, wie wir im Projekt auf das maschinenlesbare Material gekommen sind, welche Programme sich dafür besonders anbieten und vor welche Herausforderungen wir bei der Textbearbeitung gestellt waren, haben wir uns entschieden, den Praktikumsbericht im Rahmen der Working Paper Series des SFB 1199 zu veröffentlichen.

Wir erhoffen uns damit, zumindest ansatzweise in den Forschungsstand zum Thema einzuführen, Arbeitsschritte nachlesbar und reproduzierbar zu machen, dabei aber eine Reproduktion der Aufgaben in Praktika zu vermeiden und stattdessen für weitere Praktika in diesem interdisziplinären Umfeld aus traditionellen und digitalen Geisteswissenschaften bereits an anderer Stelle ansetzen zu können. Darüber hinaus ist unser Anliegen, den Text als Erfahrungsbericht mit Studierenden der Digital Humanities zu teilen und ihr Interesse für die praktische Umsetzung des theoretisch erworbenen Wissens noch zu wecken. Im Projekt C01 des SFB 1199 haben wir beispielsweise den Austausch zwischen Geographiegeschichtsschreibung und Global Studies mit den Digital Humanities nunmehr in einer neuen Konstellation aus Projektmitarbeitenden und studentischen Hilfskräften aus dem Bereich der Digital Humanities fortgeführt.

Die hier vorgestellte Praktikumsarbeit von Laura Rebecca Klettke verstehen wir als einen ersten, sehr erfolgreichen Testlauf in der Verbindung von Global Studies und Digital Humanities. Das Projekt profitierte dabei auch vom Beratungs- und Methodenangebot des seit 2021 im Aufbau befindlichen Digital Lab.[8] Das Digital Lab ist institutionell am ReCentGlobe angesiedelt, steht aber mit seinen vielfältigen Angeboten auch anderen DH-Interessierten aus dem Universitäts- und Akademiekontext zur Verfügung. Dies spiegelt sich auch in der jüngsten Verbindung des Labs mit dem schon länger bestehenden Forum für Digital Humanities Leipzig (FDHL) wider,[9] einem universitätsübergreifenden Forum für den Austausch und die Vernetzung von Digital Humanities-Akteur*innen in und um Leipzig. Gespannt sehen wir neuen Kooperationen zwischen Studierenden und Forschenden der Digital Humanities und Kolleg*innen anderer Disziplinen entgegen.

Ninja Steinbach-Hüther und Manuel Burghardt

8 Universität Leipzig, „Digital Lab", https://recentglobe.uni-leipzig.de/zentrum/infrastruktur/digital-sciences-lab (letzter Zugriff 15. November 2022)

9 Forum für Digital Humanities Leipzig, „Forum für Digital Humanities Leipzig", https://fdhl.info/ (letzter Zugriff 16. Oktober 2022).

1 Einleitung

Mein Praktikum absolvierte ich am Leibniz-Institut für Länderkunde (IfL). Das Institut – eine außeruniversitäre Forschungseinrichtung der Leibniz-Gemeinschaft – verfolgt das Ziel, geographische Sachverhalte zu untersuchen und zu analysieren, mit dem Zweck, den gesellschaftlichen Wandel über die Jahre hinweg darzustellen.[1] Verankert war das Praktikum im Projekt „Raumsemantiken der Geographie im 19. und 20. Jahrhundert" des Sonderforschungsbereichs „Verräumlichungsprozesse unter Globalisierungsbedingungen" (SFB 1199), angesiedelt an der Universität Leipzig, dem Leibniz-Institut für Geschichte und Kultur des östlichen Europa (GWZO) und dem IfL. Geleitet wird das Projekt von Ute Wardenga; Mitarbeitende sind Ninja Steinbach-Hüther und Dirk Hänsgen.[2]

Vom IfL lag eine Praktikumsbeschreibung und eine Einführung zum Projekt als Text vor. In dem Text wird auf die Zeitschriften der Geographischen Gesellschaften (GG) eingegangen, welche dem Projekt als Datengrundlage dienen. GG sind bereits seit 1821 in verschiedenen Ländern vertreten und publizierten regelmäßig Zeitschriften mit Reportagen, Dokumentation o. ä. zu fremden Ländern und Kulturen. Es liegen Zeitschriften in verschiedenen Sprachen vor. Darüber hinaus wurde im Praktikumsausschreiben die Tabelle mit Metadaten beschrieben, in der zu jedem Text relevante Informationen festgehalten werden. Diese Methodik geht aus der ersten Förderphase des Projekts hervor. Als Beschreibung der Anforderungen an das Praktikum wurden das Erstellen eines kleinen Korpus aus den Zeitschriften der GG und das Festhalten der Entwicklung vom vorliegenden Digitalisat zum erzeugten Korpus genannt. Mithilfe des Korpus sollten erste Analysen der Daten vorgenommen werden, um Raumbegrifflichkeiten zu finden und zu untersuchen.

Das Ziel, das die Ergebnisse des Praktikums unterstützen sollen, ist die Untersuchung, inwiefern sich der Einsatz von bestimmten geographischen Begriffen über die Zeit entwickelte. Zusätzlich soll die Nutzung von Sprache und Begrifflichkeiten in den einzelnen Ländern bzw. Sprachen hinsichtlich ihrer gegebenenfalls möglichen Unterschiede oder Gemeinsamkeiten analysiert werden.

Um die Aufgaben des Praktikums zu erfüllen, begann ich mit einer Literaturrecherche zum Thema *Optical Character Recognition* (OCR). Im Ergebnis entschied ich mich für das Tool *OCR4all*, um die Texte der Zeitschriften in eine maschinenlesbare Form zu bringen. Für die Texte der GG aus Berlin, London, New York, Paris und Madrid trainierte ich eigene Modelle, welche auf die jeweiligen Sprachen und Schriftarten angepasst waren.

Eine erste Analyse der raumbezogenen Sprache erfolgte anhand der Zeitschrift der GG Marseille. Zehn vollständige Jahrgänge wurden mit *OCR4all* in maschinenlesbare Form transferiert. Die Texte dienten als Grundlage, nach ausgewählten Begriffen zu suchen und alle Wörter, die häufig in Zusammenhang mit gewählten Begriffen auftreten, abzubilden. Die Ergebnisse wurden mittels Graphen visualisiert, die auch den Zeitverlauf, also die Entwicklung der genutzten Begriffe, darstellen.

Der Analyse liegt eine Literaturrecherche zum Thema *Natural Language Processing* (NLP) zugrunde.

1 Leibniz-Institut für Länderkunde (IfL), „Über das IfL", https://leibniz-ifl.de/institut/ueber-das-ifl (letzter Zugriff 23. März 2021).

2 IfL, „Sonderforschungsbereich ‚Verräumlichungsprozesse unter Globalisierungsbedingungen' (SFB 1199)", https://leibniz-ifl.de/forschung/forschungsthemen/verraeumlichungsprozesse-sfb-1199 (letzter Zugriff 16. Oktober 2022).

2 Organisatorisches

Das Praktikum lief vom 1. Dezember 2020 bis zum 7. Februar 2021 – mit zwei Wochen Weihnachtsferien. In diesem Zeitraum arbeitete ich 20 Stunden pro Woche.

In der Regel hielten Ninja Steinbach-Hüther und ich einmal die Woche ein Meeting ab und besprachen die letzten Arbeitsschritte sowie das weitere Vorgehen. Zusätzlich formulierte ich jeden Freitag eine E-Mail mit den Tätigkeiten der Woche und neuen Ergebnissen.

3 Projekt „Raumsemantiken der Geographie im 19. und 20. Jahrhundert“

Das Projekt „Raumsemantiken der Geographie im 19. und 20. Jahrhundert“ untersucht wissenschaftliche Zeitschriften, die sich geographischen Fragestellungen und Entwicklungen widmen. Es liegen Zeitschriften in gedruckter Form und als Digitalisate vor. Die Zeitschriften stammen von GG an verschiedenen Standorten weltweit und sind somit in verschiedenen Sprachen verfasst. Die Zeitschriften wurden in der Zeit von 1821 bis 1914 veröffentlicht. Die Abgrenzung des Untersuchungszeitraums von rund 100 Jahren ergibt sich aus dem Gründungsdatum der ersten GG in Paris und dem Beginn des ersten Weltkriegs. GG dokumentierten und visualisierten die Erforschung verschiedener Weltregionen, z. B. der Polargebiete, Ost-, Süd- und Zentralasien und Afrika. Sie finanzierten diverse Expeditionen, um Daten und Eindrücke zu sammeln und in Zeitschriften zu veröffentlichen.[3]

Innerhalb des Projekts bezeichnet der Begriff Raumsemantik „sprachlich und visuell vermittelte Bedeutungen und Sinnstrukturen von Raum und Räumlichkeit“.[4] Ziel des Projekts ist es, Veränderungen in der Nutzung und Entwicklung von Raumsemantiken aufzuzeigen. Dabei werden sowohl der Gebrauch von Raumsemantiken in den Veröffentlichungen der einzelnen GG betrachtet als auch die Unterschiede und Gemeinsamkeiten in der Nutzung von Raumsemantiken in verschiedenen Sprachen und durch individuelle GG.

Neben den Zeitschriften dienen Tabellen als Datengrundlage, die relevante Metadaten beinhalten (z. B. Autor, behandelte Region, Stichworte, Klassifizierung des Textes) und während der ersten Förderphase erstellt wurden.[5] Die Tabellen verwendete ich jedoch nicht für das Praktikum.

3 N. Steinbach-Hüther u. a., *Geographiegeschichtsschreibung und Digital Humanities*.

4 Leibniz-Institut für Länderkunde (IfL), „Projekt-Info: Raumsemantiken der Geographie im 19. und 20. Jahrhundert“, https://leibniz-ifl.de/forschung/forschungsthemen/historische-geographien/projekt/raumsemantiken-der-geographie-im-19-und-20-jahrhundert (letzter Zugriff 23. März 2021)

5 N. Steinbach-Hüther, *Beschreibung eines Praktikumsplatzes im Projekt C01 am Leibniz-Institut für Länderkunde (IfL)*, unveröffentlichtes Manuskript, 2020.

4 Optical Character Recognition

4.1 Literatur

Für das Praktikumsziel, ein Korpus zu erstellen, das mit computergestützten Methoden bearbeitbar ist, habe ich mich mit Literatur zu den Themen OCR und spezifischen OCR-*Tools* beschäftigt.

Holley und Tanner behandeln in ihren Texten die Historie von OCR und geben eine Einführung in das allgemeine Vorgehen eines OCR-*Workflows* – ein allgemeiner Ablauf findet sich auch bei Springmann.

Koistinen, Luscombe, Wehner und Lincke behandelten die Nutzung einer spezifischen OCR-Software, Koistinen und Luscombe widmen sich *Tesseract*, während Wehner und Lincke auf *OCR4all* eingehen.

Der Artikel „How Good Can It Get?" von Holley bietet einen Überblick über die gegenwärtigen Standards von OCR und befasst sich damit, welche Maßnahmen getroffen werden können, um die Texterkennung zu optimieren.[6] Zunächst gibt die Autorin die historische Entwicklung von OCR seit den 1990er Jahren wieder. Darüber hinaus wird der allgemeine Ablauf der Texterkennung behandelt, mit anschließendem Fokus auf mögliche Vorgehensweisen, um die Qualität der OCR zu verbessern. In Tabellenform werden einige Faktoren aufgelistet, die positiven Einfluss auf die Texterkennung haben, und auf ihre Anwendbarkeit überprüft. Die Auflistung der Faktoren ist für das Projekt von Relevanz, da sie eine gute, allgemein gültige Übersicht liefert, welche Merkmale eines Dokuments eine positive bzw. negative Auswirkung auf die Texterkennung haben. Ein Beispiel dafür ist, zu prüfen, welche OCR-Software und welches Bildbearbeitungsprogramm im entsprechenden Anwendungsfall die besten Ergebnisse liefern.

Tanner et al. beschäftigen sich in ihrem Text „Measuring Mass Text Digitization Quality and Usefulness" mit der Auswertung der Ergebnisse einer OCR.[7] Nach einer Betrachtung der Historie von OCR, angefangen im Jahre 1809, beschreiben die Autoren einen allgemeinen OCR-*Workflow*. Anschließend fokussieren sie sich auf das Ergebnis der Texterkennung, welches üblicherweise als Prozentsatz angegeben wird. Statt der Ermittlung eines einzelnen Ergebnisses stellen sie ihre Herangehensweise vor, neben der Genauigkeit von Wörtern und Buchstaben auch die Genauigkeit von signifikanten Wörtern und Wörtern, die mit einer Majuskel beginnen, sowie Zahlengruppen zu ermitteln. Diese Methode bietet eine bessere Übersicht als die Ermittlung eines einzelnen Wertes, der die Menge richtig erkannter Zeichen wiedergibt. In einem Beispiel werden die Genauigkeiten wie folgt aufgelistet: „Character accuracy = 83.6%, Word accuracy = 78%, Significant word accuracy = 68.4%, Words with capital letter start accuracy = 63.4%, Number group accuracy = 64.1%". Durch diese Aufstellung der Ergebnisse lässt sich zusätzlich ablesen, in welchem Bereich Fehler auftreten. Der Artikel wird von mir besonders als Quelle für die Historie von OCR sowie das allgemein übliche Vorgehen verwendet. Die Aufteilung der Ergebnisermittlung wird im Projekt nicht umgesetzt, stellt jedoch einen Ansatz dar, der einen besseren Aufschluss über den OCR-*Output* liefern kann.

Der Text „Ocrocis" von Springmann wird als Tutorial für einen OCR-Ablauf präsentiert.[8] Verwendet wurde hier das *Tool Ocropy* und mithilfe dieses *Tools* werden alle Schritte einer Texterkennung durchlaufen. Auch die Punkte „Umgang mit Annotationen" und „Training von OCR-Modellen" werden behandelt. Besonders für das Training ist dieser Text für das Praktikum hilfreich, da die Ausführungen allgemeine Standards benennen, die beim Training eingehalten werden sollten.

---

6 R. Holley, „How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs", *D-Lib Magazine* 15 (2009) 3/4.

7 S. Tanner, T. Muñoz und P. H. Ros, „Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive", *D-Lib Magazine* 15 (2009) 7/8.

8 U. Springmann, „Ocrocis: A High Accuracy OCR Method to Convert Early Printings Into Digital Text", A Tutorial (2015), http://cistern.cis.lmu.de/ocrocis/tutorial.pdf (letzter Zugriff 23. März 2021).

Gupta et al. liefern in ihrem Artikel „Automatic Assessment of OCR Quality in Historical Documents" einen eigenen Ansatz, um die Qualität der Bilddokumente, die als Grundlage für die Texterkennung dienen, zu verbessern.[9] Die Autoren listen mögliche Fehlerquellen in den Dokumenten auf und verfolgen das Ziel, Bereiche mit Unreinheiten aus den Texten zu entfernen. Dafür stellen sie das Konzept der *Bounding Boxes* vor. Die *Bounding Boxes* sind ein interessanter und vielversprechender Ansatz, um die Dokumente von Nicht-Text-Segmenten zu bereinigen. Sie werden in diesem Projekt aber nicht angewendet, weil die zur Verfügung stehenden Texte in der Regel in gutem Zustand sind und wenig bis keine Unreinheiten aufweisen, die von der OCR-Software fälschlicherweise als Text erkannt wurden. Daher würde die Verwendung der *Bounding Boxes* keine signifikante Verbesserung herbeiführen.

Der Text „How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source *Tesseract* OCR Engine" von Koistinen et al. liefert eine Einführung sowie Einsatzmöglichkeiten für OCR.[10] Die Autoren stellen des Weiteren Möglichkeiten vor, um die Texterkennung zu verbessern. Diese einzelnen Möglichkeiten sind in einem Prozess dargestellt, beginnend bei der Erstellung der Bilddatei, über die Verwendung der *Tesseract Engine*, die Entscheidung, welches Wort gewählt wird, bis zur Generierung des Outputdokuments. Zusätzlich gibt der Text einen kurzen Einblick in die Funktionsweisen der Software *Tesseract*.

Das Werk „The Text Recognition Algorithm Independent Evaluation (TRAIT)" von Godil et al. wird für diese Arbeit herangezogen, um Definitionen für *Precision* und *Recall* vorzustellen, die spezifisch die Resultate der Texterkennung auswerten.[11]

Ein weiterer Text, der sich bei meiner Recherche als nützlich erwies, ist der von Luscombe et al. veröffentlichte Aufsatz „Access to Information and Optical Character Recognition", wobei es sich um eine Anleitung zur Software *Tesseract* handelt.[12] Den Text verwendete ich, um alternative Texterkennungstools zu betrachten, die Entscheidung fiel jedoch auf die Anwendung *OCR4all*, da diese durch die grafische Benutzeroberfläche (*Graphical User Interface* [GUI]) und die Ausrichtung auf früh veröffentlichte Texte überzeugte.

Auch zu dem Tool *OCRopus* holte ich weitere Informationen ein. Nasarek gibt eine Zusammenfassung der Anwendungsgebiete und Benutzung der Software.[13]

Als Übersicht über *OCR4all* dient der Beitrag „*OCR4all* – Eine semiautomatische Open-Source-Software für die OCR historischer Drucke" im Tagungsband des Verbands Digital Humanities im deutschsprachigen Raum e.V. aus dem Jahr 2020, verfasst von den Entwicklern von *OCR4all*.[14] Neben den Einsatzmöglichkeiten von *OCR4all* gehen die Autoren auch auf einen grob umrissenen Ablauf der Texterkennung unter Verwendung von *OCR4all* ein.

In den Vortragsfolien von Lincke zum Thema „Coptic OCR: Even better models and improvements on user-friendliness" finden sich eine grafische Darstellung des OCR-*Workflows*, Übersichten zum

9 A. Gupta u. a., „Automatic Assessment of OCR Quality in Historical Documents", in: American Association for Artificial Intelligence (Hrsg.), *AAAI'15: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*: AAAI Press, 2015, S. 1735–1741.

10 M. Koistinen, K. Kettunen und J. Kervinen, „How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine", in: Z. Vetulani, P. Paroubek und M. Kubis (Hrsg.), *Human Language Technology: Challenges for Computer Science and Linguistics*, Language and Technology Conference 2017, Cham: Springer International Publishing, 2020, S. 17–30.

11 A. Godil, P. Grother und M. Ngan, „The Text Recognition Algorithm Independent Evaluation (TRAIT)" (2017), https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8199.pdf (letzter Zugriff 22. März 2021).

12 A. Luscombe u. a., „Access to Information and Optical Character Recognition (OCR): A Step-by-Step Guide to Tesseract: Part One of the CAIJ Computer Literacy Series", Winnipeg (2020), www.uwinnipeg.ca/caij/docs/caig-report-access-to-information-and-ocr.pdf (letzter Zugriff 31. März 2021).

13 R. Nasarek, „OCRopus – Hoffnungsträger der Frakturschrifterkennung – Digital Humanities selbst gestrickt", https://blogs.urz.uni-halle.de/strickdings/2017/05/ocropus-hoffnungstraeger-der-frakturschrifterkennung/ (letzter Zugriff 16. Oktober 2022).

14 M. Wehner u. a., „OCR4all – Eine semiautomatische Open-Source-Software für die OCR historischer Drucke", in: C. Schöch (Hrsg.), *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation: Konferenzabstracts*, 2020, S. 43–45.

Training von OCR-Modellen und eine Darstellung des *Workflows* von *OCR4all*.[15] Gerade diese Darstellung ist hilfreich für das Projekt, um die Funktionsweise und die Schritte von *OCR4all* klarer zu definieren.

4.2 Besonderheiten der Zeitschriftentexte

Als ersten Kontakt mit den Artikeln der Zeitschriften betrachtete ich jeweils einen Artikel der New Yorker, der Londoner, der Pariser und der Berliner GG. Die ursprünglich im Jahr 1859 veröffentlichten Artikel wurden zufällig ausgewählt. Später kamen Artikel aus Madrid und Marseille hinzu. Die Artikel sind bei JSTOR, DigiZeitschriften oder Gallica zu finden und frei zugänglich.

Häufig auftretende Besonderheiten der Texte sind Fußnoten, die mit unterschiedlichen Symbolen gekennzeichnet werden. Teilweise (z. B. in den Texten der Berliner GG) sind die Symbole nicht Teil der Sprache und müssen der OCR-Software „beigebracht" werden. Darüber hinaus befinden sich in der Regel Seitenzahl und Titel des Artikels oberhalb des Textes, während unten (z. B. bei den Texten der Londoner GG) ein *Copyright*-Vermerk von JSTOR eingefügt ist.

Da es sich um Texte mit geographischen Themen handelt, beinhalten sie viele allgemeine geographische bzw. topografische Bezeichnungen sowie konkrete Namen von Orten, Gebieten und Regionen (Topo- bzw. Choronyme), die teilweise mit Zeichen versehen sind, die nicht in der Sprache des Textes vorkommen.

Für diese und weitere Besonderheiten fertigte ich zwei Tabellen an. Die erste Tabelle umfasst alle Auffälligkeiten und beschreibt, mit welchem Verfahren und in welchem Bearbeitungsschritt das Problem gelöst werden kann. Einige der Textspezifika können mit der OCR-Software gehandhabt werden, andere in einer Bearbeitung der Bilddateien im Vorfeld oder einer Nachbearbeitung der Textdateien (siehe Anhang 1).

In der zweiten Tabelle sind die Spezifika der Texte nach Standorten unterteilt, damit Nutzer*innen, die die Texte einer bestimmten GG bearbeiten möchten, einsehen können, auf welche Besonderheiten bei dieser GG zu achten ist (siehe Anhang 2).

4.3 Ablauf der *Optical Character Recognition*

Der Begriff OCR bezeichnet die Extraktion des Textes aus einem Bilddokument. Vor der Durchführung der Texterkennung liegt eine Bilddatei vor, die einen oder mehrere Textabschnitte enthält. Mittels einer OCR-Software ist es möglich, diesen Text in eine maschinenlesbare Form zu überführen. Der typische Ablauf einer Texterkennung beginnt mit der Vorbereitung, dem *Preprocessing*. Anschließend wird der Text segmentiert und in seine kleinsten Bestandteile zerlegt, es folgt eine Mustererkennung – häufig auf Buchstabenebene. Dafür sind der Software eine Vielzahl an Schriftarten bekannt, um trotz verschiedener Stile und Ausprägungen die einzelnen Zeichen richtig zu erkennen. Das Wort selbst kann schließlich geprüft werden, indem es mit einem Wörterbuch abgeglichen wird. Bei dem letzten Schritt handelt es sich um die Nachbereitung, das sogenannte *Postprocessing*.[16]

15 E.-S. Lincke, „Coptic OCR: Even Better Models and Improvements on User-Friendliness", http://kellia.uni-goettingen.de/digitalcoptic3/slides/CopticOCR_2020-12-07_Lincke.pdf (letzter Zugriff 23. März 2021).

16 Tanner, Muñoz and Ros, „Measuring Mass Text Digitization", S. 2–3.

4.3.1 Umgang mit Dokumenten vor der Texterkennung *(Preprocessing)*

Die Vorbereitung der Texterkennung beginnt mit dem Scannen der physischen Texte oder dem Bearbeiten der vorliegenden Bilddokumente. Es gibt einige Faktoren, die zu beachten sind, um bessere Ergebnisse bei der OCR zu erzielen. So sollte jede Seite möglichst vollständig und sauber sein, also keine Unreinheiten (sog. *Noise*), Notizen, usw. enthalten. Die Seiten sollten nicht in Farbe vorliegen, da bei bitonalen und grauen Dokumenten bessere Erkennbarkeit gegeben ist. Die Auflösung sollte so gut wie möglich sein, um viele Bildinformationen zu bewahren. Darüber hinaus sollten Bilder nicht schief gescannt werden oder – falls schief vorliegend – soweit gedreht, dass der Text horizontal dargestellt ist. Und schließlich sollte eine OCR-Software gewählt werden, die nicht nur allgemein gute Ergebnisse bringt, sondern auch auf das Anwendungsgebiet angepasst ist.[17]

Das *Preprocessing* besteht aus mehreren Schritten: Das Bild wird entzerrt und eine mögliche Neigung entfernt, sodass der Text horizontal steht. Der vorhandene *Noise* wird so gut es geht entfernt und die Seite segmentiert, sodass einzelne Textfelder erkannt und voneinander getrennt werden. Außerdem erfolgt eine Binarisierung: Bei der Texterkennung wird keine Farbe verwendet, sondern die Seite in schwarz-weiß gehalten.[18]

Tabelle 1[19]
Einflussfaktoren auf die Ergebnisse der OCR

Mögliche Fehlerquelle	Lösung (falls möglich)	Hilfsmittel
unvollständige Textsammlung	Originaldokument als Vorlage nutzen und fehlende Seiten scannen	
Flecken o. ä. auf dem physischen Dokument	Bereinigen des Dokuments	OCR-Software
farbige und kontrastschwache Dokumente	schwarz-weiße oder graue Version verwenden	OCR-Software
schlechte Auflösung	Originaldokument in besserer Qualität scannen	
geneigte Schrift	Text drehen	Bildbearbeitungsprogramm (z. B. GIMP)
Wahl der OCR-Software	auf Anwendungsgebiet angepasste Software wählen	

Bei der Segmentierung einer Seite werden möglichst alle Textbausteine einzeln herausgearbeitet. Dieser Schritt wird auch *Zoning* genannt, da die Seite in einzelne Zonen geteilt wird. Diese Zonen können ganze Abschnitte sein sowie Überschriften, Zeilen usw.[20]

Die Segmentierung wird von der OCR-Software ausgeführt.

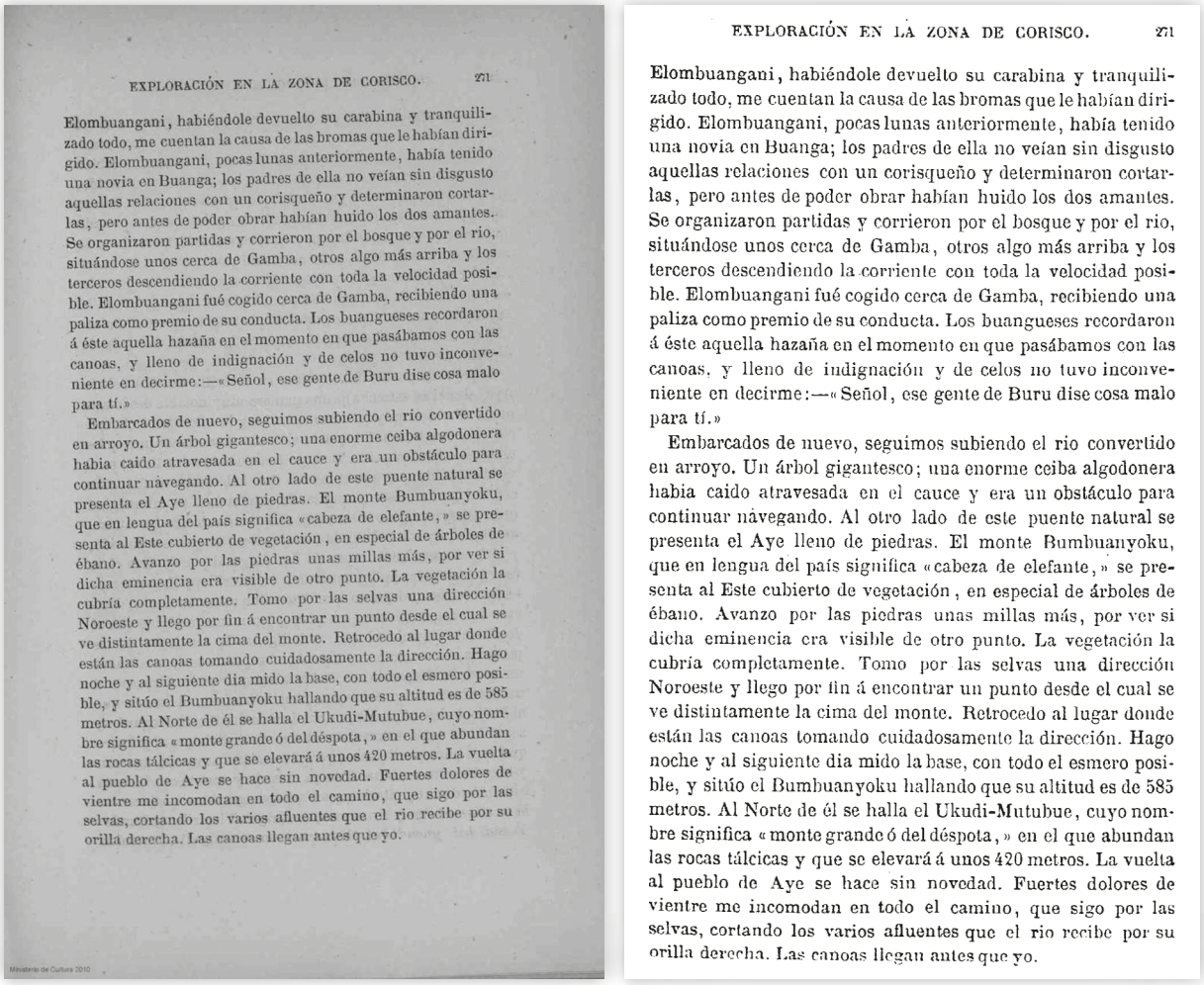
17 Holley, „How Good Can It Get?", S. 3–5.
18 U. Springmann, *Ocrocis*, S. 8.
19 In der Tabelle habe ich alle Punkte zusammengefasst, die die Ergebnisse der OCR beeinflussen können und wie und mit welchen Hilfsmitteln die einzelnen Punkte bearbeitet werden.
20 U. Springmann, *Ocrocis*, S. 9.

4.3.1.1 ScanTailor

ScanTailor ist eine freie Software zur Nachbearbeitung von Bildern und zur Vorbereitung der Bilder für die Texterkennung. Nacheinander können folgende Schritte mit *ScanTailor* ausgeführt werden: Seite drehen (90° oder 180°), Seiten aufteilen (ein oder zwei Textabschnitte einer Seite markieren), Ausrichten (Seite drehen, sodass der Text möglichst horizontal dargestellt ist), Inhalt auswählen (nur Textbereich der Seite markieren), Stege/Ränder (Größe der Ränder der Outputdatei festlegen), Ausgabe (Auflösung, Farbe, Dicke der Schrift, Seitenkrümmung und Grad der Bereinigung festlegen).[21]

Die Beispielseite stammt aus einem Zeitschriftenartikel aus dem Jahr 1878 der Sociedad Geográfica de Madrid: links das gescannte Bild, rechts der *Output*, den *ScanTailor* liefert.

Abbildung 1–1, 1–2
links: gescannte Seite, rechts: Outputdokument von ScanTailor



EXPLORACIÓN EN LA ZONA DE CORISCO. 271

Elombuangani, habiéndole devuelto su carabina y tranquilizado todo, me cuentan la causa de las bromas que le habían dirigido. Elombuangani, pocas lunas anteriormente, había tenido una novia en Buanga; los padres de ella no veían sin disgusto aquellas relaciones con un corisqueño y determinaron cortarlas, pero antes de poder obrar habían huido los dos amantes. Se organizaron partidas y corrieron por el bosque y por el rio, situándose unos cerca de Gamba, otros algo más arriba y los terceros descendiendo la corriente con toda la velocidad posible. Elombuangani fué cogido cerca de Gamba, recibiendo una paliza como premio de su conducta. Los buangueses recordaron á éste aquella hazaña en el momento en que pasábamos con las canoas, y lleno de indignación y de celos no tuvo inconveniente en decirme:—«Señol, ese gente de Buru dise cosa malo para tí.»

Embarcados de nuevo, seguimos subiendo el rio convertido en arroyo. Un árbol gigantesco; una enorme ceiba algodonera habia caido atravesada en el cauce y era un obstáculo para continuar navegando. Al otro lado de este puente natural se presenta el Aye lleno de piedras. El monte Bumbuanyoku, que en lengua del país significa «cabeza de elefante,» se presenta al Este cubierto de vegetación, en especial de árboles de ébano. Avanzo por las piedras unas millas más, por ver si dicha eminencia era visible de otro punto. La vegetación la cubría completamente. Tomo por las selvas una dirección Noroeste y llego por fin á encontrar un punto desde el cual se ve distintamente la cima del monte. Retrocedo al lugar donde están las canoas tomando cuidadosamente la dirección. Hago noche y al siguiente dia mido la base, con todo el esmero posible, y sitúo el Bumbuanyoku hallando que su altitud es de 585 metros. Al Norte de él se halla el Ukudi-Mutubue, cuyo nombre significa «monte grande ó del déspota,» en el que abundan las rocas tálcicas y que se elevará á unos 420 metros. La vuelta al pueblo de Aye se hace sin novedad. Fuertes dolores de vientre me incomodan en todo el camino, que sigo por las selvas, cortando los varios afluentes que el rio recibe por su orilla derecha. Las canoas llegan antes que yo.

Ministerio de Cultura 2010

EXPLORACIÓN EN LA ZONA DE CORISCO. 271

Elombuangani, habiéndole devuelto su carabina y tranquilizado todo, me cuentan la causa de las bromas que le habían dirigido. Elombuangani, pocas lunas anteriormente, había tenido una novia en Buanga; los padres de ella no veían sin disgusto aquellas relaciones con un corisqueño y determinaron cortarlas, pero antes de poder obrar habían huido los dos amantes. Se organizaron partidas y corrieron por el bosque y por el rio, situándose unos cerca de Gamba, otros algo más arriba y los terceros descendiendo la corriente con toda la velocidad posible. Elombuangani fué cogido cerca de Gamba, recibiendo una paliza como premio de su conducta. Los buangueses recordaron á éste aquella hazaña en el momento en que pasábamos con las canoas, y lleno de indignación y de celos no tuvo inconveniente en decirme:—«Señol, ese gente de Buru dise cosa malo para tí.»

Embarcados de nuevo, seguimos subiendo el rio convertido en arroyo. Un árbol gigantesco; una enorme ceiba algodonera habia caido atravesada en el cauce y era un obstáculo para continuar navegando. Al otro lado de este puente natural se presenta el Aye lleno de piedras. El monte Bumbuanyoku, que en lengua del país significa «cabeza de elefante,» se presenta al Este cubierto de vegetación, en especial de árboles de ébano. Avanzo por las piedras unas millas más, por ver si dicha eminencia era visible de otro punto. La vegetación la cubría completamente. Tomo por las selvas una dirección Noroeste y llego por fin á encontrar un punto desde el cual se ve distintamente la cima del monte. Retrocedo al lugar donde están las canoas tomando cuidadosamente la dirección. Hago noche y al siguiente dia mido la base, con todo el esmero posible, y sitúo el Bumbuanyoku hallando que su altitud es de 585 metros. Al Norte de él se halla el Ukudi-Mutubue, cuyo nombre significa «monte grande ó del déspota,» en el que abundan las rocas tálcicas y que se elevará á unos 420 metros. La vuelta al pueblo de Aye se hace sin novedad. Fuertes dolores de vientre me incomodan en todo el camino, que sigo por las selvas, cortando los varios afluentes que el rio recibe por su orilla derecha. Las canoas llegan antes que yo.

Abbildungsnachweis: Boletin de la Sociedad Geográfica de Madrid, Tomo IV Año III Número 4 – 1878 Abril, https://realsociedadgeografica.com/publicaciones/boletin

21 ScanTailor, „ScanTailor", https://scantailor.org/ (letzter Zugriff 23. März 2021).

4.3.2 Umgang mit Dokumenten nach der Texterkennung *(Postprocessing)*

Beim *Postprocessing* werden die Ergebnisse ausgewertet. Dabei können verschiedene Gruppen untersucht werden, um genauer Aufschluss über die einzelnen Fehlerraten zu erhalten. Geprüft werden z. B. falsch dargestellte Buchstaben, Wörter, Wörter, die mit einem Großbuchstaben beginnen oder – für das Projekt besonders relevant – falsche Topo- und Choronyme oder andere geographische Begriffe bzw. Bezeichnungen. Üblicherweise sind gerade diese speziellen Wörter anfälliger für Fehler, da sie seltener im Text auftreten.[22]

Als Indikatoren für die Richtigkeit des Outputs dienen *Precision* und *Recall*. Der Wert der *Precision* berechnet sich aus den richtig erkannten Zeichen geteilt durch alle richtig oder falsch erkannten Zeichen. Der Wert des *Recalls* ist die Anzahl der richtig erkannten Zeichen geteilt durch alle Zeichen, die im Eingangstext enthalten sind. Hier zählen auch die Zeichen mit hinein, die von der OCR-Software gar nicht erkannt wurden.[23] Häufig wird die Evaluation des Ergebnisses auch von der Software selbst angeboten.

$$\text{Precision} = \frac{\text{richtig erkannte Zeichen}}{\text{richtig erkannte Zeichen} + \text{falsch erkannte Zeichen}}$$

$$\text{Recall} = \frac{\text{richtig erkannte Zeichen}}{\text{alle Zeichen des Ursprungstextes}}$$

4.4 Training eines Modells

Um ein möglichst erfolgreiches Ergebnis zu erzielen, muss ein Modell trainiert werden. Dafür werden Daten als Trainingsgrundlage benötigt, die einerseits optisch und qualitativ den Texten entsprechen, die in eine maschinenlesbare Form gebracht werden sollen, und andererseits bereits transkribiert vorliegen oder transkribiert werden. Die für dieses Praktikum händisch transkribierten Texte werden als *Ground Truth* bezeichnet. Der Begriff stammt ursprünglich aus dem Bereich der Kartographie, bei dem es darum ging, durch einen sogenannten „Feldvergleich" die Boden- bzw. Geländewirklichkeit und deren Kartendarstellung auf ihre Übereinstimmung zu prüfen. Für die wissenschaftliche Verlässlichkeit der durch OCR-Technologien gewonnenen Datenbestände ist die Schaffung einer solchen Trainingsgrundlage ein überaus wichtiger Arbeitsschritt.

Die Trainingsdaten werden verwendet, um mit der OCR-Software eine Texterkennung auszuführen. Anschließend hilft die bereits transkribierte Version, den gelieferten *Output* abzugleichen und Fehler zu erkennen. Durch diesen Abgleich kann die Software lernen, Buchstaben und Wörter richtig zu erkennen und zu übertragen. Aus diesem Grund müssen die Trainingsdaten sehr umfangreich sein und möglichst alle Stile, Schreibweisen und Zeichen enthalten, welche auch in den späteren Textdokumenten vorkommen. Zeichen, die üblicherweise in den Trainingsdaten unterrepräsentiert sind, sind Großbuchstaben, Fragezeichen/Ausrufezeichen und Zahlen.[24]

22 Tanner, Muñoz and Ros, „Measuring Mass Text Digitization", S. 9–10.

23 A. Godil, P. Grother und M. Ngan, *The Text Recognition Algorithm Independent Evaluation*, S. 14.

24 U. Springmann, *Ocrocis*, S. 10–12.

4.5 OCR mit *OCR4all*

Für die Anwendung von OCR informierte ich mich zu den Tools *Tesseract*, *OCRopus* und *OCR4all*.[25] Alle drei Anwendungen sind *Open Source*. Für *OCR4all* entschied ich mich, da weder *Tesseract* noch *OCRopus* eine GUI besitzen. Die GUI von *OCR4all* ermöglicht einen leichteren Einstieg und kann auch mit nur kurzer Einarbeitung von anderen Personen bedient werden. Darüber hinaus wurde *OCR4all* speziell für historische Dokumente entwickelt. Es erfolgt kein Wortabgleich mit einem Lexikon, sondern die Texterkennung funktioniert über *Machine Learning*. Da die Texte der GG viele Fremdwörter (vorrangig Ortsbezeichnungen, s. o.) enthalten, ist dieser Ansatz erfolgsversprechender bei der OCR.

OCR4all wurde im Jahre 2019 publiziert. Die Entwickler verfolgten das Ziel, ein OCR-Programm für ein breites Anwenderspektrum zu entwickeln. Der Ablauf der OCR umfasst insgesamt zehn Schritte.

Er beginnt mit Schritt 1 – der Konvertierung der Seite in die gewünschte Form (PNG-Dokument, einheitliche Benennung, bitonal oder grau). Sollte ein PDF-Dokument in den Input-Ordner gelegt werden, werden die einzelnen Seiten in PNG-Dateien umgewandelt, was – je nach Größe des Dokuments – einige Zeit in Anspruch nehmen kann. Anschließend erfolgt Schritt 2 – das *Preprocessing* – gefolgt von Schritt 3, einem *Noise Removal*. Die Seiten können schließlich einzeln eingesehen werden, wobei entfernte Regionen oder Punkte rot gekennzeichnet sind.

Die Segmentierung als Schritt 4 kann entweder mithilfe des Tools *Dummy* durchgeführt werden, wobei Nutzer*innen keinen direkten Einblick in die Ergebnisse haben, oder mithilfe von *LAREX*, welches den Nutzer*innen die Möglichkeit bietet, eigene Änderungen bei der Segmentierung, der Klassifizierung der Segmente (z. B. als Fließtext, Überschrift oder Grafik) und der *Reading Order* (Reihenfolge der einzeln erkannten Textbausteine) vorzunehmen.

In Kapitel 4.2 habe ich die Seitenzahl samt Titel am oberen Rand und den *Copyright*-Vermerk unterhalb des Textes als Besonderheiten der Texte aufgeführt. Um diese Bereiche von der Texterkennung auszuschließen, kann in *LAREX* festgelegt werden, welche Teile der Seiten ignoriert werden. Somit werden weder Seitenzahl noch Titel oder *Copyright* im *Output* enthalten sein.

Anschließend erfolgen der 5. Schritt, die *Line Segmentation*, und der 6. Schritt, die *Recognition*. Bei der *Recognition* werden händisch beliebig viele Modelle ausgewählt.

Im Schritt 7 – *Ground Truth Production* – kann in einem Editor der erkannte Text eingesehen und korrigiert werden. Anhand dieser Korrektur ist es in den nächsten Schritten möglich, das Ergebnis zu evaluieren. Auch neue Modelle können auf Grundlage der Daten trainiert werden.

Die Ausgabe der Ergebnisse erfolgt entweder als XML-Dokument oder als Textdatei.[26]

In Abbildung 2 ist der modulare Aufbau von *OCR4all* grafisch dargestellt.[27]

4.5.1 Trainieren eigener Modelle

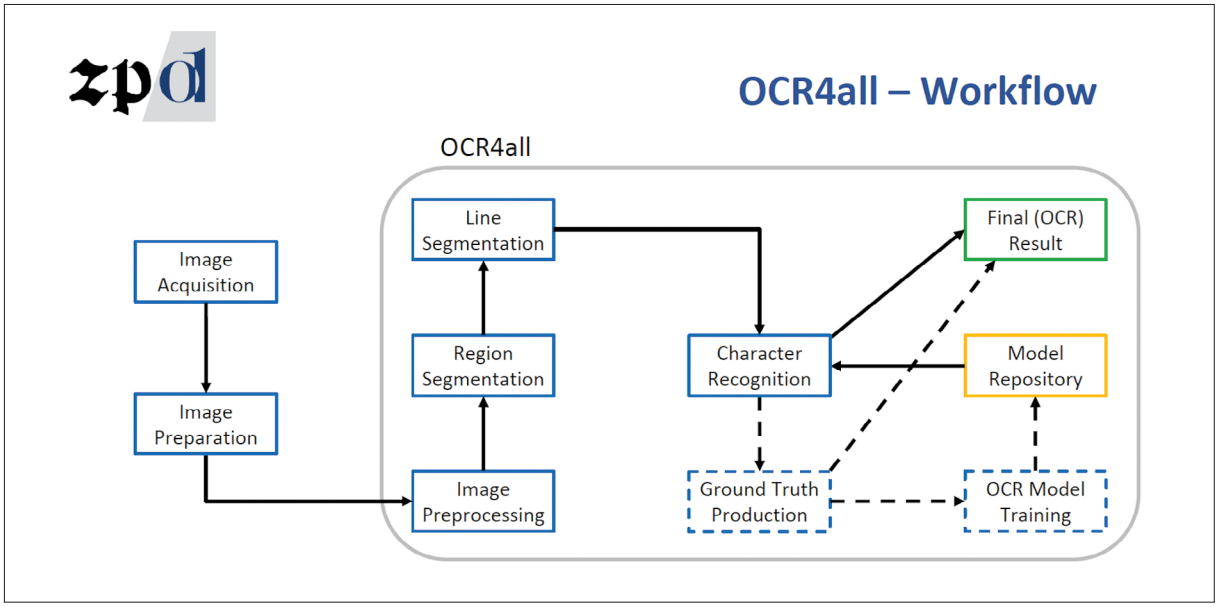
Für das Training eigener Modelle wendete ich zunächst die von *OCR4all* angebotenen Modelle auf die ersten drei Seiten eines Textes an. Die Ergebnisse waren noch nicht zufriedenstellend und die Korrektur der *Ground Truth Data* dementsprechend zeitaufwendig. Trainiert wurden Modelle für die Zeitschriften der Berliner, der Londoner, der New Yorker, der Madrider und der Pariser GG. Für die Umwandlung der Texte der GG Marseille konnten die Modelle, welche anhand von Seiten aus den Veröffentlichungen der GG Paris trainiert wurden, genutzt werden, da sowohl *Layout* als auch Sprache übereinstimmten. Die händisch kontrollierten und korrigierten Texte dienten nun als Grundlage für den ersten Trainingsschritt. Bei einem Durchlauf erstellte ich jeweils fünf neue Modelle für die entsprechende Sprache. Diese neuen Modelle wurden anschließend für das Trainieren weiterer Modelle

25 Nasarek, *OCRopus – Hoffnungsträger der Frakturschrifterkennung – Digital Humanities selbst gestrickt*.

26 M. Wehner u. a., „OCR4all – Eine semiautomatische Open-Source-Software", S. 43–45.

27 OCR4all, „User Guide – Introduction", www.ocr4all.org/guide/user-guide/introduction (letzter Zugriff 17. Oktober 2022).

Abbildung 2
Workflow OCR4all



Abbildungsnachweis: OCR4all: User Guide – Introduction, https://www.ocr4all.org/guide/user-guide/introduction

verwendet. Mit der Anwendung der eigenen Modelle waren die Ergebnisse der OCR deutlich verbessert und die Korrektur der *Ground Truth Data* benötigte weniger Zeit. Die Gesamtheit aller trainierten Modelle kann zukünftig für die Texte der GG verwendet werden. Die Genauigkeit der Ergebnisse bei einer Anwendung auf den Londoner Text „The Lake Regions of Central Equatorial Africa" zeigt die Abbildung 3.[28] Das Ergebnis wurde im Evaluationsschritt von *OCR4all* ausgegeben.

Abbildung 3
Evaluation der trainierten Modelle für das Anwendungsbeispiel London

```
Got mean normalized label error rate of 0.22% (27 errs, 12313 total chars, 27 sync errs)
GT       PRED     COUNT    PERCENT
{a}      {}           3     11.11%
{*}      {'}          2      7.41%
{B }     {}           2     14.81%
{"}      {'}          1      3.70%
{e}      {}           1      3.70%
{W}      {}           1      3.70%
{e}      {c}          1      3.70%
{h}      {l }         1      7.41%
{I}      {1}          1      3.70%
{T}      {}           1      3.70%
The remaining but hidden errors make up 37.04%
```

Eigene Berechnungen

28 R. F. Burton, „The Lake Regions of Central Equatorial Africa with Notices of the Lunar Mountains and the Source of the White Nile, Being the Results of an Expedition Undertaken under the Patronage of Her Majesty's Government … in the Years 1857–1859", *Journal of the Royal Geographical Society of London* 29 (1859).

Eine Möglichkeit für bessere Ergebnisse: Sollte beispielsweise ein deutscher Text viele Eigennamen spanischer Herkunft enthalten, können bei der Texterkennung neben den deutschen Modellen auch spanische Modelle verwendet werden, um zu garantieren, dass der Software alle verwendeten Zeichen bekannt sind.

4.5.2 Ergebnisse der Anwendung

Als Eigenheiten der Texte der Marseiller GG identifizierte ich die Seitenzahlen und Titel der Artikel, die sich über dem Textabschnitt befinden (vgl. Anhang 2). Außerdem sind in den Zeitschriften Grafiken enthalten, die ich vor der Nutzung von *OCR4all* händisch entfernte.

Für die Anwendung von *OCR4all* auf die Texte der Marseiller GG wurden die Modelle, die für die Pariser GG trainiert wurden, genutzt. Neben der Sprache ist auch die Schriftart ähnlich, sodass – obwohl als Datengrundlage eine andere Zeitschrift diente – die Ergebnisse gut sind. Probleme, die häufiger bei der Nutzung von *OCR4all* auftraten, waren die Erkennung der Überschriften, wenn diese ausschließlich aus Großbuchstaben bestanden. Außerdem war die *Reading Order* oft fehleranfällig. Die Textseiten, die *OCR4all* als Ergebnis ausgab, waren besonders dann durcheinander, wenn viele verschiedene Segmente erkannt wurden: die Überschrift konnte unter dem Absatz stehen oder die Fußnoten nicht in der richtigen Reihenfolge untereinander. Der Großteil der Seiten besteht allerdings aus Fließtext, der als ein Textsegment erkannt wurde. In dem Fall können keine Probleme bei der *Reading Order* auftreten.

Für *Distant Reading*-Anwendungen sind die Resultate ausreichend, da im Gesamtüberblick die Reihenfolge der Seiten gewahrt bleibt und die oben genannten Fehler in der Reihenfolge der Anordnung einzelner Textbausteine nur einen kleinen Teil ausmachen, sodass das finale Ergebnis nicht merklich beeinträchtigt wird.

Der *Workflow* vom Digitalisat zum maschinenlesbaren Text beinhaltet folgende Schritte:

- Prüfen der digital vorliegenden Zeitschriften: Teilweise sind Zeitschriftentexte bereits in maschinenlesbarer Form. Dies traf auf die untersuchten Texte der Londoner und New Yorker GG zu. Die Ergebnisse sind aber von schlechter Qualität und können – sofern die Beispieltexte die durchschnittliche Qualität wiederspiegeln – nicht für Analysezwecke verwendet werden. Bei der Untersuchung der Texte reicht ein „Durchklicken" durch die vorliegenden Seiten. Geprüft werden sollte:
  - Befinden sich Notizen, Flecken, o. ä. auf den Seiten?
  - Ist der Text stark geneigt und nicht horizontal?
  - Ist der Kontrast zwischen Text und Hintergrund schwach?
  - Befinden sich zwei Buch- oder Zeitschriftenseiten auf einer digitalen Seite?
  - Befinden sich Überreste des Textes der anderen Buch- oder Zeitschriftenseite auf der digitalen Seite?

  Können alle Fragen mit Nein beantwortet werden, ist eine Bearbeitung mit ScanTailor nicht notwendig. Sollte eine Frage mit Ja beantwortet werden, kann ScanTailor eingesetzt werden, um das Problem zu beheben und die Texte für OCR vorzubereiten.

- Anwendung von ScanTailor

- Anwendung von *OCR4all:* Um *OCR4all* zu installieren und zu benutzen, existieren Handbücher, die detailliert alle Schritte umfassen. Darüber hinaus ist die Software benutzerfreundlich und intuitiv aufgebaut, sodass kein umfangreiches informatisches Vorwissen benötigt wird. Die Texte müssen in einen vorgegebenen *Input*-Ordner gelegt werden. PDFs können ebenfalls

verwendet werden; sie werden im ersten Schritt in PNG-Dateien umgewandelt, was einige Zeit in Anspruch nehmen kann. Die von mir trainierten Modelle müssen im *Models*-Ordner gespeichert sein, um für die OCR genutzt werden zu können. Es sollten alle Modelle ausgewählt werden, die zu der Sprache bzw. zum Standort gehören.
Anschließend werden die Ergebnisse (als Text- oder XML-Datei) im *Result*-Ordner hinterlegt. Es werden sowohl eine Gesamtdatei als auch jede Seite einzeln gespeichert.

5 Natural Language Processing

5.1 Literatur

Für das Projektziel, das erstellte Korpus mit computergestützten Methoden zu bearbeiten und raumbezogene Sprache zu erkennen und zu analysieren, habe ich mich mit Literatur zum Thema *Natural Language Processing* (NLP) beschäftigt. NLP wird in den Texten von Indurkhya, Nadkarni, Schmidt und vom Unternehmen IBM im Gesamtüberblick behandelt, während sich Wevers auf *Word Embeddings* und Schumacher auf *Named Entity Recognition* (NER) fokussieren.

Indurkhya et al. beschreiben das Vorgehen von NLP, beginnend mit dem *Text-Preprocessing*. Anschließend folgt eine *Lexical Analysis* – hier kann entweder *Stemming* angewendet (nur der Wortstamm eines Wortes wird verwendet) oder das Lemma des Wortes gebildet werden. Danach wird die Syntax der Sätze untersucht, die Semantik und abschließend die Pragmatik.[29]

Nadkarni et al. geben ebenfalls eine Einführung zu NLP, angefangen mit der Historie. Sie gehen auf die Herausforderungen bei der Nutzung von NLP und die *Low-Level* und *High-Level-Tasks* ein.[30]

Anwendungsbeispiele werden ebenfalls genannt, unter anderem *Statistical NLP* und *Supervised* und *Unsupervised Learning*. Während beim *Supervised Learning* jede Eingabe mit einer korrekten Antwort versehen ist, wodurch die Maschine den „Lösungsweg" zu der Antwort erlernen kann, wird beim *Unsupervised Learning* keine korrekte Antwort vorgegeben und die Maschine soll selbstständig Muster erkennen. Diese Anwendungsbeispiele wurden im Praktikum nur theoretisch betrachtet und für das Projekt nicht angewendet.

Schmidt et al. beschreiben in ihrem Text eine *Hatespeech Detection* mit NLP.[31] Sie geben eine kurze Übersicht über verschiedene *Features*, z. B. *Sentiment Analysis*, womit die Stimmung einzelner Wörter, Sätze oder eines vollständigen Textes bestimmt wird, *Lexical Resources* – die Nutzung eines Wörterbuchs, um Wörter, die in diesem Fall typisch für *Hatespeech* sind, zu identifizieren – und *Dependency Relationships*, wo Zusammenhänge zwischen Wörtern (auch über kurze Entfernungen innerhalb eines Textes hinweg) erkannt werden.

Schumacher behandelt NER am Beispiel norddeutscher Kriminalromane des 20. Jahrhunderts, welche als Grundlage zur Untersuchung weiblicher Romanfiguren herangezogen werden.[32] *Named Entities* können Personen, Orte oder Organisationen sein, die automatisch in einem Text erkannt und

29 N. Indurkhya und F. J. Damerau, *Handbook of Natural Language Processing*, Chapman & Hall/CRC Machine Learning & Pattern Recognition Series, Boca Raton (FL): Taylor & Francis, 2010.

30 P. M. Nadkarni, L. Ohno-Machado und W. W. Chapman, „Natural Language Processing: An Introduction", *Journal of the American Medical Informatics Association: JAMIA* 18 (2011) 5.

31 A. Schmidt und M. Wiegand, „A Survey on Hate Speech Detection using Natural Language Processing", in: L.-W. Ku und C.-T. Li (Hrsg.), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Stroudsburg (PA): Association for Computational Linguistics, 2017, S. 1–10.

32 M. Schumacher, „Named Entity Recognition (NER)", *forTEXT. Literatur digital erforschen* (2018).

klassifiziert werden sollen. Der Text beschreibt die Erfolgsaussichten bei NER, die technischen Grundlagen und wie die Technik für *Distant Reading* eingesetzt werden kann.

Wevers et al. widmen sich den *Word Embedding* Modellen.[33] Sie stellen die Historie, Entwicklung und Einsatzmöglichkeiten vor. *Word Embeddings* zeigen die Zusammenhänge der Semantik zweier Wörter mittels geographischer Darstellung der Wörter. Sie werden von den Autoren zur Untersuchung historischer Texte verwendet.

Der Text „Natural Language Processing" beinhaltet eine Definition und stellt die Bereiche vor, aus denen sich NLP zusammensetzt.[34] Die Einsatzmöglichkeiten und Probleme bei der Nutzung von NLP werden erläutert und mögliche Tools genannt, unter anderem *NLTK* in *Python* und *Statistical NLP*.

5.2 Vorgehen

Da zum Ende des Praktikums nicht genügend Zeit für eine umfangreiche Anwendung von NLP blieb, bestand das alternative Vorgehehen darin, ein Programm zu erstellen, das eigenständig eingegebene Begriffe in den Texten der Marseiller GG sucht, die Häufigkeit des Begriffs pro Text ausgibt und alle Wörter findet, welche gemeinsam mit dem eingegebenen Begriff auftreten. Als Grundlage dienten die Ausgaben der Zeitschrift der Marseiller GG von zehn Jahren, jeweils jahresweise zusammengefasst. Der erste Schritt bestand darin, alle Wörter in den zehn Ausgaben zu zählen, die Texte von *Stopwords* zu bereinigen und anschließend ein Dokument auszugeben, das die Wörter nach Häufigkeit sortiert auflistet. Das Dokument schickte ich anschließend an Ninja Steinbach-Hüther, die die Begriffe sichtete und jene auswählte und in eine kürzere Liste übertrug, die für das Projekt besonders von Interesse waren.

Für jedes der Wörter wurden die sogenannten begleitenden Begriffe ermittelt: Bei jedem Auftreten des Wortes wurden die 15 Wörter vor und nach dem gefundenen Wort gespeichert und anschließend ebenfalls nach Häufigkeit sortiert ausgegeben. Dabei entstand eine Übersicht von Begriffen, die häufig gemeinsam mit dem Ausgangswort genannt werden und somit thematische Ähnlichkeit aufweisen. Vor der Ausgabe wurden *Stopwords* entfernt.

Der Begriff von Interesse kann bei der Ausführung des Programms eingetragen werden, um zu zeigen, in welchem Zusammenhang verschiedene Wörter mit geographischem Bezug genutzt werden. Es werden alle gewählten Jahre einzeln betrachtet, damit erkennbar ist, inwiefern sich die Verwendung der begleitenden Begriffe in den unterschiedlichen Publikationsjahren unterscheidet.

5.2.1 Visualisierung

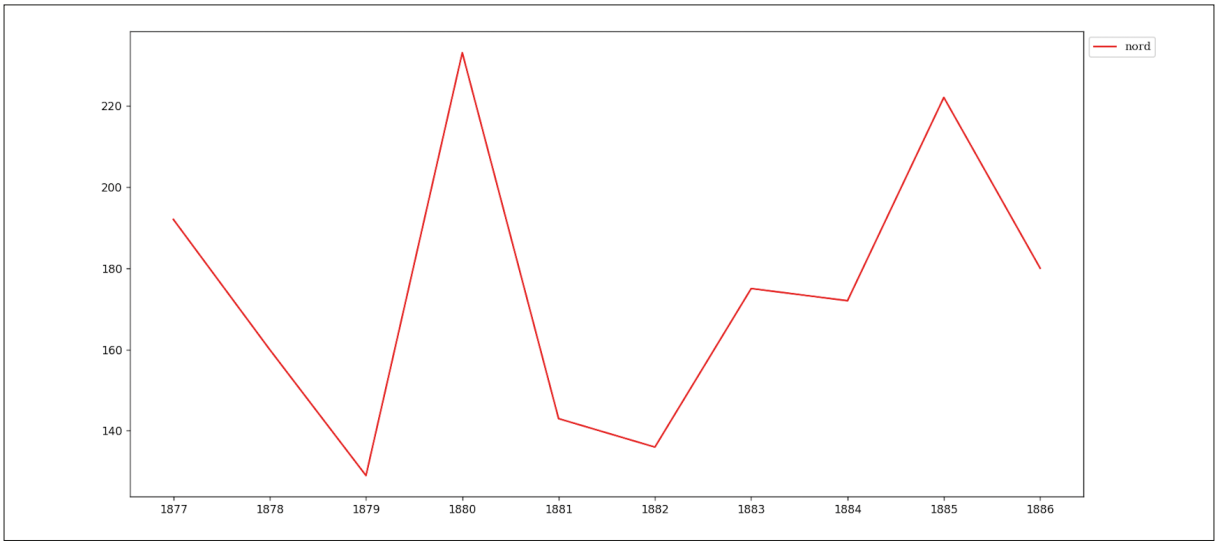
Der letzte Schritt des Praktikums bestand darin, die Ergebnisse der Wörtersuche zu visualisieren. Auch hier muss bei Ausführung des Programms ein Wort von Interesse eingegeben werden. Voraussetzung ist, dass zu dem Wort bereits die Textdateien vorliegen, die die begleitenden Begriffe auflisten.

Abbildung 4 zeigt am Beispiel des Begriffs *nord* nur das Wort selbst und die Anzahl der Nennungen von *nord* in den Zeitschriften der Marseiller GG für jedes Jahr. Dabei stellt die x-Achse das jeweilige Publikationsjahr in der gewählten Zeitspanne von zehn Jahren dar, während auf der y-Achse die absolute Häufigkeit aufgeführt ist, mit welcher das Wort in den Zeitschriftenausgaben auftritt.

33 M. Wevers und M. Koolen, „Digital begriffsgeschichte: Tracing Semantic Change Using Word Embeddings", *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53 (2020) 4.

34 IBM Cloud Education, „Natural Language Processing", www.ibm.com/cloud/learn/natural-language-processing (letzter Zugriff 23. März 2021).

Abbildung 4
Auftreten des Begriffs *nord*

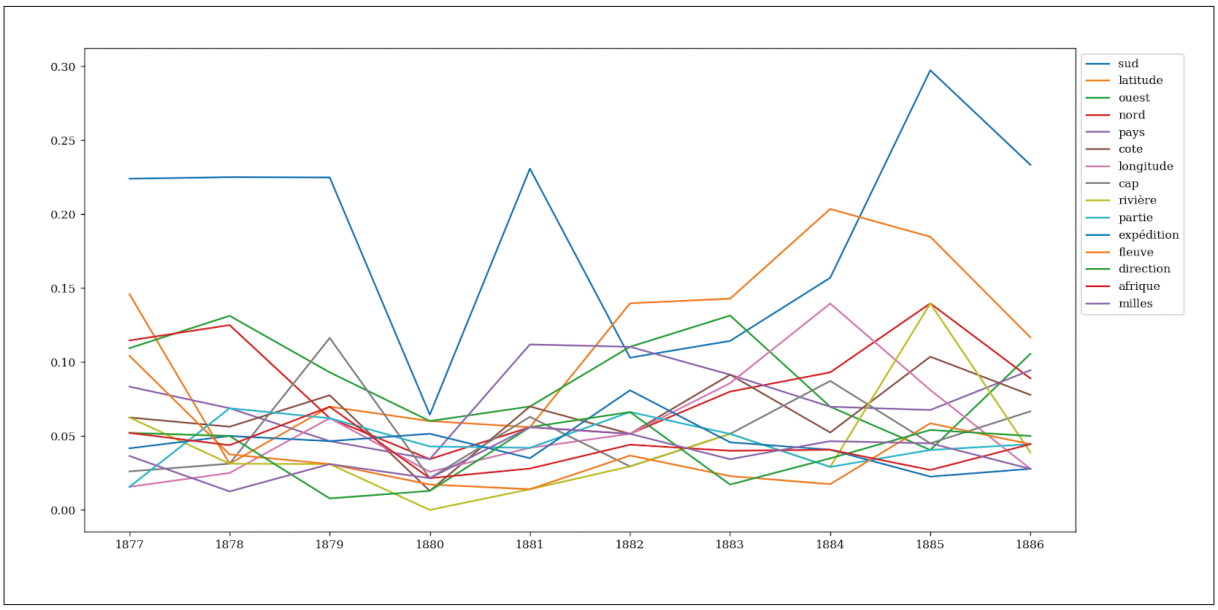


Eigene Darstellung

Abbildung 5 zeigt die 15 Wörter, die – in allen Ausgaben der Jahre 1877 bis einschließlich 1886 zusammengenommen – am häufigsten gemeinsam mit dem gesuchten Wort (in diesem Beispiel *nord*) auftreten. Der Begriff wird in der Grafik nicht selbst genannt, jedoch bei der Speicherung der Visualisierung automatisch in den Dateiennamen eingefügt.

Die x-Achse stellt erneut die Publikationsjahre dar, die y-Achse zeigt die Häufigkeit, mit der die verschiedenen Begriffe in Zusammenhang mit dem Wort *nord* genannt werden.

Abbildung 5
Auftreten der 15 häufigsten Begriffe, die gemeinsam mit *nord* genannt werden



Eigene Darstellung

Alle Ergebnisse sind im Rahmen des Praktikums am IfL entstanden. Mögliche Ansätze für eine Fortführung des Projektes unter Nutzung der Ergebnisse werden im folgenden Abschnitt aufgeführt.

5.3 Denkbare Fortführungen oder alternative Ansätze

Gegen Ende des Praktikums diskutierten wir alternative Ideen zur Fortführung bzw. Ausführung bisheriger Ergebnisse. So wäre eine ausführlichere Möglichkeit der Visualisierung die Erstellung eines *Stream-Graphen* mit Interaktionsoption gewesen, um durch *Mouseover* die häufigsten drei begleitenden Begriffe in dem entsprechenden Jahr anzuzeigen. Des Weiteren könnten mittels NER Ortsbezeichnungen oder raumbezogene Begriffe gefunden werden. NER wäre ein denkbarer Ansatz, diese Begriffe für weitere Untersuchungen zu identifizieren.

Eine weitere Möglichkeit war, anstelle der Auslistung der Begriffe, die gemeinsam mit einem gesuchten Wort genannt werden, mit *Word Embeddings* zu arbeiten. Die *Word Embeddings* würden das Verhalten verschiedener Begriffe zueinander visualisieren und könnten dazu beitragen, die Nutzung der raumbezogenen Sprache zu verdeutlichen und den Kontext, in welchem ein Wort genutzt wurde, aufzuzeigen. Letztlich bin ich diesem Ansatz in meiner Masterarbeit nachgekommen und habe auf Basis der Ausgangsfragen, die das Forschungsprojekt C01 begleiten, *Word Embeddings* genutzt, um aufzulisten, welche Begriffe mit diversen gewählten Ausgangswörtern in einem gemeinsamen Kontext auftreten. Im Rahmen einer Hiwi-Stelle überführte ich weitere Zeitungsartikel verschiedener GG in maschinenlesbare Form und nutzte die Daten für die Anwendung von *Word Embeddings*.

6 Zusammenfassung

Das Praktikum eignete sich für mich sowohl, um meine Kenntnisse in Bezug auf wissenschaftliches Arbeiten, Recherchieren und Schreiben zu vertiefen, aber auch – wie vorgesehen – praktische Erfahrungen für das Studium der Digital Humanities zu sammeln. Gerade da der Studiengang erst seit einigen Jahren an der Universität Leipzig angeboten wird, war es für mich interessant zu erfahren, welche Einrichtungen welche Projekte vorstellten und anschließend am Leibniz-Institut für Länderkunde einen tieferen Einblick zu erlangen, inwiefern die dortige Forschung von Methoden der Digital Humanities profitieren kann.

In der Zeit meines Praktikums widmete ich mich vorrangig der Recherche und Anwendung von OCR. Ich stieg mit Literatur zum generellen Ablauf vom gescannten Dokument zum maschinenlesbaren Text ein und informierte mich über die einzelnen Schritte auf diesem Weg, angefangen beim *Preprocessing* samt *Noise Removal*, gefolgt von der *Segmentation*, der *Character Recognition* und schließlich dem *Postprocessing* samt *Evaluation*. Anschließend informierte ich mich zu verschiedenen OCR-*Tools* und entschied mich für *OCR4all* als Anwendungssoftware, da *OCR4all* in meinen Augen gut für das Projekt geeignet ist sowie leicht in der Handhabung und das Projekt auch in Zukunft von anderen Mitarbeitenden fortgeführt werden könnte.

Mittels *OCR4all* erstellte ich Modelle für die Texte der Pariser, Londoner, New Yorker, Berliner und Madrider GG. Diese Modelle verbessern die Ergebnisse der Texterkennung deutlich im Vergleich zu den Standardmodellen, die *OCR4all* mitbringt. Anschließend folgte die Anwendung für zehn Jahrgänge der Zeitschrift der GG aus Marseille. Die Ergebnisse der OCR sind für verschiedene Aufgaben des *Distant Reading* geeignet. Zur Vorbereitung auf die computergestützte Auswertung der Texte führte ich erneut eine Literaturrecherche zum Thema NLP durch. Es sind verschiedene Methoden denkbar, um die Texte zu analysieren und die Ergebnisse zu präsentieren.

Um Ergebnisse auf die Frage, inwiefern sich die Nutzung raumbezogener Sprache bzw. geographischer Begrifflichkeiten über die einzelnen Jahre entwickelte, zu liefern, entwickelte ich ein Programm, das die sogenannten begleitenden Begriffe zu einem selbstgewählten Wort findet. Es wird eine Liste

der nach Häufigkeit sortierten Begriffe ausgegeben, die gemeinsam mit dem Ausgangswort (es werden jeweils 15 Wörter davor und danach betrachtet) auftreten. Die Ergebnisse des Programms wurden anschließend mithilfe einer Anwendung, erstellt in *Python*, visualisiert. Diese Anwendungen machten jedoch nur die letzten 2,5 Wochen des Praktikums aus. Den Großteil der Zeit widmete ich mich dem Thema OCR und der Überführung der vorliegenden Bilddokumente in maschinenlesbaren Text.

Wie bereits beschrieben, war die Untersuchung der Zeitschriften Geographischer Gesellschaften, die Arbeit mit OCR sowie NLP auch Bestandteil meiner Masterarbeit. Nach Ende des Praktikums arbeitete ich als studentische Hilfskraft am Leibniz-Institut für Länderkunde. Bei dieser Beschäftigung stand die Beschaffung, Vorbereitung und Umwandlung der Bild- zu Textdokumenten im Vordergrund. Die Fragestellung der Masterarbeit entwickelte sich aus meinen bisherigen Untersuchungen und den Tätigkeiten, die ich als studentische Hilfskraft ausübte. Schlussendlich ergaben sich aus dem Praktikum nicht nur die erforderlichen Leistungspunkte, sondern auch eine anschließende Beschäftigung am Institut sowie der Untersuchungsgegenstand und die Betreuung meiner Masterarbeit.

7 Anhang

Anhang 1

Stichwort	Spezifika im Text	Verfahren	Zeitpunkt der Anpassung
Seitenzahl, Verweis etc.	Seitenzahl und Titel am oberen Rand	Darf nicht in Inhalt einfließen	Segmentierung
Fußnoten	Fußnoten in kleiner Schrift unter dem Text	Schriftart und -größe richtig erkennen	Texterkennung
Namen und Eigenbegriffe	Eigennamen in fremder Sprache beinhalten Zeichen, die nicht Teil der Textsprache sind	Zeichen dennoch übernehmen (müssen vorher bekannt sein)	Training
Leerzeichen	Wichtige Ortsnamen oder Teilüberschriften sind im Original durch Sperrsatz bzw. Sperrschrift hervorgehoben	Wörter richtig erkennen trotz gesperrter Schrift	Training
Schreibweise	An manchen Stellen modernes s, an anderen Stellen altdeutsches f (wird zusammen als ss oder ß verwendet)	Beide Schreibweisen erkennen	Training und Texterkennung
Schreibweise	Alte Schreibweise mancher Wörter	In heutige Schreibweise überführen oder beibehalten	Spätere Nachbearbeitung
Zahlen, Sonderzeichen etc.	Zahlenbrüche	Richtig erkennen und darstellen	Texterkennung
Zahlen, Sonderzeichen etc.	Römische Zahlen	Richtig erkennen und darstellen	Texterkennung
Abkürzungen	Verwendung von Abkürzungen (z. B. Nachm.)	Beibehalten oder ausschreiben	Spätere Nachbearbeitung
Abkürzungen	Teilweise sind Ortsnamen nach erster Nennung nur mit Buchstaben abgekürzt	Ortsnamen immer ausschreiben	Spätere Nachbearbeitung
Schriftgröße	Variierende Schriftgröße	Gesamten Text erkennen	Texterkennung
Seitenzahl, Verweis etc.	Dokument enthält Verweis zur Download-Seite	Darf nicht in Inhalt einfließen	Segmentierung
Fußnoten	Fußnoten nicht nummeriert, sondern mit Symbolen als Verweise	Symbole kennen und richtig darstellen	Training und Texterkennung
Fußnoten	Kurze Fußnoten nebeneinander statt untereinander	Getrennt voneinander erkennen	Spätere Nachbearbeitung

Stichwort	Spezifika im Text	Verfahren	Zeitpunkt der Anpassung
Layout	Änderung des Layouts einiger Zeilen	Richtig erkennen	Segmentierung
Abkürzungen	Häufige Verwendung von Himmelsrichtungen, werden unterschiedlich abgekürzt (z. B. SW und S. W.)	Beibehalten oder einheitliche Abkürzungen	Spätere Nachbearbeitung
Grafiken etc.	Zeichnungen	Als Bilddokumente erkennen oder entfernen	Texterkennung oder vorherige Bearbeitung
Grafiken etc.	Darstellung von Karten	Trotz Beschriftungen der Karte als Bilddokumente erkennen	Texterkennung oder vorherige Bearbeitung
Grafiken etc.	Tabelle mit Übersetzungen	Textrichtung richtig erkennen	Segmentierung
Grafiken etc.	Tabelle mit Zahlen	Richtig erkennen und wiedergeben	Segmentierung und Texterkennung
Zahlen, Sonderzeichen etc.	Zitierte Geschichte mit Anführungszeichen vor jeder Zeile dargestellt	Zeichen richtig erkennen	Texterkennung
Layout	Zwei Spalten	Textrichtung richtig erkennen	Zeilensegmentierung
Layout	Grau gescannt	In bitonales Dokument umwandelt	Preprocessing
Layout	Schief gescannte Seiten	Möglichst horizontal darstellen	Preprocessing

Anhang 2

Name der GG	Jahr	Name des Artikels	Text- oder Bildformat
Berlin (Zeitschrift für allgemeine Erdkunde)	1859	Itinerar der kleinasiatischen Reise P. v. Tschichatschef's im Jahre 1858	Bild
London (The Journal of the Royal Geographical Society of London)	1859	The Lake Regions of Central Equatorial Africa, with Notices of the Lunar Mountains and the Sources of the White Nile; Being the Results of an Expedition Undertaken under the Patronage of Her Majesty's Government and the Royal Geographical Society of London, in the Years 1857–1859	Text
	1836	Observations on the Coast of Arabia between Rás Mohammed and Jiddah	Text
	1846	A Description of the Province of Khúzistán	Text
	1866	An Overland Expedition from Port Denison to Cape York; Under the Command of F. andA. Jardine, Esqrs	S. 1–7 Bild, ab 7 Text
	1876	Description of the Country and Natives of Port Moresby and Neighbourhood, New Guinea	Text

	Spezifika des Textes	To Do für Texterkennung	Notizen
	Seitenzahl und Titel am oberen Rand	Darf nicht in Inhalt einfließen	Im Preprocessing entfernen
	Fußnoten in kleiner Schrift	Schriftart und -größe richtig erkennen	
	Eigennamen in fremder Sprache beinhalten Zeichen, die nicht Teil der Textsprache sind	Zeichen dennoch übernehmen	Müssen im Vorfeld bekannt sein
	Wichtige Ortsnamen oder Teilüberschriften sind im Original durch Sperrsatz bzw. Sperrschrift hervorgehoben	Wörter richtig erkennen trotz gesperrter Schrift	
	An manchen Stellen modernes s, an anderen Stellen altdeutsches ſ (wird zusammen als ss oder ß verwendet)	Beide Schreibweisen erkennen	
	Alte Schreibweise mancher Wörter		In heutige Schreibweise überführen oder beibehalten?
	Zahlenbrüche	Richtig erkennen und darstellen	
	Römische Zahlen	Richtig erkennen und darstellen	
	Verwendung von Abkürzungen (z. B. Nachm.)		Entscheiden, wie damit verfahren
	Teilweise sind Ortsname nach erster Nennung nur mit Buchstaben abgekürzt	Ortsnamen immer ausschreiben	
	Variierende Schriftgröße	Gesamten Text erkennen	
	Dokument enthält Verweis zur Download-Seite	Darf nicht in Inhalt einfließen	Im Preprocessing entfernen
	Seitenzahl und Titel am oberen Rand	Darf nicht in Inhalt einfließen	Im Preprocessing entfernen
	Fußnoten in kleiner Schrift	Schriftart und -größe richtig erkennen	
	Fußnoten nicht nummeriert, sondern mit Symbolen als Verweise	Symbole kennen und richtig darstellen	
	Änderung des Layouts einiger Zeilen	Richtig erkennen	
	Eigennamen in fremder Sprache beinhalten Zeichen, die nicht Teil der Textsprache sind	Zeichen dennoch übernehmen	Müssen im Vorfeld bekannt sein

Name der GG	Jahr	Name des Artikels	Text- oder Bildformat
Paris (Bulletin e la Société de géographie)	1859	Les lacs de Tanganyika et Nyanza d'Ukerewe	Bild
New York (Journal of the American Geographical and Statistical Society)	1859	Micronesia. The Ruins on Ponape, or Ascension Island	Text
		On the Manner of Taking a Census. Part 2	Text
Madrid (Boletin de la Sociedad Geográfica de Madrid)	1878	Un Diaro de Viajes de Exploración en la Zona de Corisco	Bild
Marseille (Bulletin de la Société de géographie de Marseille)	1877	Séance d'Inauguration	Bild

	Spezifika des Textes	To Do für Texterkennung	Notizen
	Kurze Fußnoten nebeneinander statt untereinander	Getrennt voneinander erkennen	
	Häufige Verwendung von Himmelsrichtungen, werden unterschiedlich abgekürzt (z. B. SW und S. W.)		Wie damit im Postprocessing verfahren?
	Zeichnungen	Als Bilddokumente erkennen und behandeln oder vorher entfernen	
	Darstellung von Karten	Trotz Beschriftungen der Karte als Bilddokumente erkennen oder vorher entfernen	
	Tabelle mit Übersetzungen	Textrichtung richtig erkennen	
	Zitierte Geschichte mit Anführungszeichen vor jeder Zeile dargestellt	Zeichen erkennen	Wenn überflüssig im maschinenlesbaren Text entfernen
	Fußnoten in kleiner Schrift	Schriftart und -größe richtig erkennen	
	Layout: zwei Spalten	Textrichtung richtig erkennen	
	Sehr kleine Schrift	Gesamten Text erkennen	
	Dokument enthält Verweis zur Download-Seite	Darf nicht in Inhalt einfließen	Im Preprocessing entfernen
	Seitenzahl und Titel am oberen Rand	Darf nicht in Inhalt einfließen	Im Preprocessing entfernen
	Fußnoten in kleiner Schrift	Schriftart und -größe richtig erkennen	
	Tabellen mit Zahlen im Text	Richtig erkennen und wiedergeben	
	Grau gescannt	In bitonales Dokument umwandeln	Kann im Vorfeld mit Bildbearbeitungsprogramm oder von OCR-Tool selbst umgesetzt werden
	Seitenzahl und Titel am oberen Rand	Darf nicht in Inhalt einfließen	Im Preprocessing entfernen
	Einige Seiten schief eingescannt	Text möglichst horizontal darstellen	
	Seitenzahl und Titel am oberen Rand (innerhalb Umrandung)	Darf nicht in Inhalt einfließen	Im Preprocessing entfernen

8 Literaturverzeichnis

American Association for Artificial Intelligence (Hrsg.), *AAAI'15: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*: AAAI Press, 2015.

A. Gupta u. a., „Automatic Assessment of OCR Quality in Historical Documents", in: American Association for Artificial Intelligence (Hrsg.), *AAAI'15: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*: AAAI Press, 2015, S. 1735–1741.

R. F. Burton, „The Lake Regions of Central Equatorial Africa with Notices of the Lunar Mountains and the Source of the White Nile, Being the Results of an Expedition Undertaken under the Patronage of Her Majesty's Government … in the Years 1857–1859", *Journal of the Royal Geographical Society of London* 29 (1859), S. 1–454.

Collaborative Research Center (SFB) 1199, „Section C01: Spatial Semantics of Geography in the 19th and 20th Centuries", https://research.uni-leipzig.de/~sfb1199/projects/project-c1/ (letzter Zugriff 17. Oktober 2022).

T. Efer und N. Steinbach-Hüther, „Quantitative Analyses in Global and Area Studies using Graph-based Filtering of Heterogeneous Catalogue Data", in: E. Plödereder u. a. (Hrsg.), *Informatik 2014: Big Data – Komplexität meistern; Tagung der Gesellschaft für Informatik, 22.–26. September 2014 in Stuttgart, Deutschland*, Bonn: Ges. für Informatik, 2014, S. 1027–1037.

Forum für Digital Humanities Leipzig, „Forum für Digital Humanities Leipzig", https://fdhl.info/ (letzter Zugriff 16. Oktober 2022).

A. Godil, P. Grother und M. Ngan, „The Text Recognition Algorithm Independent Evaluation (TRAIT)" (2017), https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8199.pdf (letzter Zugriff 22. März 2021).

R. Holley, „How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs", *D-Lib Magazine* 15 (2009) ¾.

IBM Cloud Education, „Natural Language Processing", www.ibm.com/cloud/learn/natural-language-processing (letzter Zugriff 23. März 2021).

IfL, „Sonderforschungsbereich ‚Verräumlichungsprozesse unter Globalisierungsbedingungen' (SFB 1199)", https://leibniz-ifl.de/forschung/forschungsthemen/verraeumlichungsprozesse-sfb-1199 (letzter Zugriff 16. Oktober 2022).

N. Indurkhya und F. J. Damerau, *Handbook of Natural Language Processing*, Chapman & Hall/CRC Machine Learning & Pattern Recognition Series, Boca Raton (FL): Taylor & Francis, 2010.

M. Koistinen, K. Kettunen und J. Kervinen, „How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine", in: Z. Vetulani, P. Paroubek und M. Kubis (Hrsg.), *Human Language Technology: Challenges for Computer Science and Linguistics*, Language and Technology Conference 2017, Cham: Springer International Publishing, 2020, S. 17–30.

L.-W. Ku and C.-T. Li (Hrsg.), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Stroudsburg (PA): Association for Computational Linguistics, 2017.

Leibniz-Institut für Länderkunde (IfL), „Projekt-Info: Raumsemantiken der Geographie im 19. und 20. Jahrhundert", https://leibniz-ifl.de/forschung/forschungsthemen/historische-geographien/projekt/raumsemantiken-der-geographie-im-19-und-20-jahrhundert (letzter Zugriff 23. März 2021).

Leibniz-Institut für Länderkunde (IfL), „Über das IfL", https://leibniz-ifl.de/institut/ueber-das-ifl (letzter Zugriff 23. März 2021).

E.-S. Lincke, „Coptic OCR: Even Better Models and Improvements on User-Friendliness", http://kellia.uni-goettingen.de/digitalcoptic3/slides/CopticOCR_2020-12-07_Lincke.pdf (letzter Zugriff 23. März 2021).

A. Luscombe u. a., „Access to Information and Optical Character Recognition (OCR): A Step-by-Step Guide to Tesseract: Part One of the CAIJ Computer Literacy Series", Winnipeg (2020), www.uwinnipeg.ca/caij/docs/caig-report-access-to-information-and-ocr.pdf (letzter Zugriff 31. März 2021).

M. Middell (Hrsg.), *Verräumlichungsprozesse unter Globalisierungsbedingungen*, Leipzig: Leipziger Universitätsverlag, 2021.

F. Moretti, *Distant Reading*, Konstanz: Konstanz University Press, 2016.

P. M. Nadkarni, L. Ohno-Machado und W. W. Chapman, „Natural Language Processing: An Introduction", *Journal of the American Medical Informatics Association: JAMIA* 18 (2011) 5, S. 544–551.

R. Nasarek, „OCRopus – Hoffnungsträger der Frakturschrifterkennung – Digital Humanities selbst gestrickt", https://blogs.urz.uni-halle.de/strickdings/2017/05/ocropus-hoffnungstraeger-der-frakturschrifterkennung/ (letzter Zugriff 16. Oktober 2022).

OCR4all, „User Guide – Introduction", www.ocr4all.org/guide/user-guide/introduction (letzter Zugriff 17. Oktober 2022).

E. Plödereder u. a. (Hrsg.), *Informatik 2014: Big Data – Komplexität meistern; Tagung der Gesellschaft für Informatik, 22.–26. September 2014 in Stuttgart, Deutschland*, Bonn: Ges. für Informatik, 2014.

ScanTailor, „ScanTailor", https://scantailor.org/ (letzter Zugriff 23. März 2021).

A. Schmidt und M. Wiegand, „A Survey on Hate Speech Detection using Natural Language Processing", in: L.-W. Ku und C.-T. Li (Hrsg.), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Stroudsburg (PA): Association for Computational Linguistics, 2017, S. 1–10.

C. Schöch (Hrsg.), *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation: Konferenzabstracts*, 2020.
M. Schumacher, „Named Entity Recognition (NER)", *forTEXT. Literatur digital erforschen* (2018).
U. Springmann, „Ocrocis: A High Accuracy OCR Method to Convert Early Printings Into Digital Text", A Tutorial (2015), http://cistern.cis.lmu.de/ocrocis/tutorial.pdf (letzter Zugriff 23. März 2021).
N. Steinbach-Hüther u. a., *Geographiegeschichtsschreibung und Digital Humanities: Neue Methoden für Zeitschriftenanalysen*, Working paper series des SFB 1199 an der Universität Leipzig 15, Leipzig: Universitätsverlag Leipzig, 2019.
N. Steinbach-Hüther, 2020, *Beschreibung eines Praktikumsplatzes im Projekt C01 am Leibniz-Institut für Länderkunde (IfL)*.
N. Steinbach-Hüther, *Bibliotheksdaten, Kulturtransfer und Digital Humanities: Zu einer Methodik bei der Untersuchung transregionaler Zirkulationen akademischer Literatur afrikanischer Autoren*, Leipzig: Leipziger Universitätsverlag, 2020.
S. Tanner, T. Muñoz und P. H. Ros, „Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive", *D-Lib Magazine* 15 (2009) 7/8.
Universität Leipzig, „Digital Humanities (M. Sc.)", www.uni-leipzig.de/studium/vor-dem-studium/studienangebot/studiengang/course/show/digital-humanities-m-sc (letzter Zugriff 17. Oktober 2022).
Universität Leipzig, „Digital Lab", https://recentglobe.uni-leipzig.de/zentrum/infrastruktur/digital-sciences-lab (letzter Zugriff 15. November 2022).
Z. Vetulani, P. Paroubek and M. Kubis (Hrsg.), *Human Language Technology: Challenges for Computer Science and Linguistics*, Language and Technology Conference 2017, Cham: Springer International Publishing, 2020.
U. Wardenga u. a., „Von einer Geographie der Verräumlichung zu Geographien von Raumsemantiken: Digital Humanities als Schlüssel", in: M. Middell (Hrsg.), *Verräumlichungsprozesse unter Globalisierungsbedingungen*, Leipzig: Leipziger Universitätsverlag, 2021, S. 45–70.
M. Wehner u. a., „OCR4all – Eine semiautomatische Open-Source-Software für die OCR historischer Drucke", in: C. Schöch (Hrsg.), *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation: Konferenzabstracts*, 2020, S. 43–45.
M. Wevers und M. Koolen, „Digital begriffsgeschichte: Tracing Semantic Change Using Word Embeddings", *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53 (2020) 4, S. 226–243.

Leipzig Research Centre Global Dynamics

Working paper series des SFB 1199 an der Universität Leipzig No. 30

ISBN: 978-3-96023-455-5
ISSN: 2510-4845

Universität Leipzig
SFB 1199

E-Mail: sfb1199@uni-leipzig.de

http://research.uni-leipzig.de/~sfb1199

UNIVERSITÄT LEIPZIG
Leipzig Research Centre Global Dynamics

Leibniz-Institut für Länderkunde

Leibniz-Institut für Geschichte und Kultur des östlichen Europa

TECHNISCHE UNIVERSITÄT DRESDEN

Funded by
DFG Deutsche Forschungsgemeinschaft