Travelogues: Fremdwahrnehmungen in Reiseberichten 1500–1876

EINE COMPUTERGESTÜTZTE ANALYSE¹

Doris Gruber [1], Martin Krickl [2], Lijun Lyu [3], Jan Rörden [4] & Arno Strohmeyer [5]

[1] Institut für Neuzeit- und Zeitgeschichtsforschung (ÖAW)
doris.gruber@oeaw.ac.at, <https://orcid.org/0000-0002-0512-100X>
<https://www.oeaw.ac.at/inz/personen/gruber-doris/>

[2] Österreichische Nationalbibliothek
martin.krickl@onb.ac.at
<https://www.onb.ac.at/>

[3] Institut L3S, Leibniz Universität Hannover
lyu@l3s.de
<https://www.l3s.de/~lyu/>

[4] AIT Austrian Institute of Technology GmbH
jan.roerden@ait.ac.at, <https://orcid.org/0000-0002-5824-8397>
<https://www.ait.ac.at>

[5] Institut für Neuzeit- und Zeitgeschichtsforschung (ÖAW), Universität Salzburg
arno.strohmeyer@oeaw.ac.at, <https://orcid.org/0000-0002-4077-8840>
<https://www.oeaw.ac.at/inz/personen/strohmeyer-arno/>

Keywords: Travelogues, OCR-Post-Correction, Genre Classification, Metadata Enrichment, Early Modern Prints

Abstract: This international and interdisciplinary project creates algorithms for the semi-automatic search for and evaluation of printed travelogues in German, published in the time 1500–1876 and preserved in the holdings of the Austrian National Library.

Einleitung

Der Beitrag liefert erste Einblicke in ein im April 2018 gestartetes Projekt, das sich mit gedruckten Reiseberichten der Zeit 1500–1876 beschäftigt. Der Quellentypus bietet sich aufgrund des Inhalts, der großen Zahl an überlieferten Werken und deren Materialität an, mit Hilfe neuer Methoden der Digital

¹ Das hier vorgestellte Projekt findet dank der Finanzierung durch den FWF (I 3795-G28) und die DFG (398697847) statt. Beteiligt sind: das Institut für Neuzeit- und Zeitgeschichtsforschung an der Österreichischen Akademie der Wissenschaften, das Austrian Institute of Technology, die Österreichische Nationalbibliothek und das Forschungszentrum L3S an der Universität Hannover. Weitere Details zum Projekt und künftige Publikationen werden über die Projekthomepage angekündigt: <http://www.travelogues-project.info/> (21.03.2019).

Humanities analysiert zu werden. Inhaltlich sind Reiseberichte besonders relevant, weil sie Wahrnehmungen von „Fremdheit“, wie geographische Regionen, Kulturen und Religionen, transportieren. Diese Fremdwahrnehmungen wurden stark von den an ihrer Produktion beteiligten Personen geprägt, was der Forschung zahlreiche Rückschlüsse erlaubt: Es kann etwa darauf geschlossen werden, wie diese Personen, respektive bestimmte Kulturkreise, mit „Fremdheit“ umgingen und welche Faktoren dafür verantwortlich waren, dass Stereotype und Vorurteile geprägt wurden und sich in diachroner Perspektive wandelten. Materiell gesehen stehen die gedruckten Texte der Reiseberichte im Fokus, deren inhaltliche Ausprägungen und materielle Formen (Typen- und Textsatz) zwar äußerst heterogen sind, aber deutlich besser und leichter in maschinenlesbare Form übersetzt werden können als andere Medienformen wie die ebenfalls in den gedruckten Reiseberichten erhaltenen Bilder oder auch Handschriften.

Reiseberichte wurden von der Forschung zwar schon vielfach untersucht, bisherige Studien beschränkten sich aber meist auf ein relativ kleines Korpus von selten mehr als einem Dutzend Reiseberichten.² Obwohl sich heuristische Ansätze als fruchtbar erwiesen,³ fehlt es an quantitativen Analysen. Diese können einerseits bisherige Forschungsthesen überprüfen und andererseits dominante Aspekte (wie führende Verlage, Produktionsspitzen, regionale und thematische Schwerpunkte) aufzeigen, die durch rein qualitative Zugänge nicht ersichtlich sind.⁴

Deshalb wird im Zuge dieses Forschungsprojekts der Frage nachgegangen, wie neue digitale Analyseverfahren genutzt werden können, um gedruckte Reiseberichte und insbesondere darin erhaltene Fremdwahrnehmungen quantitativ und qualitativ zu untersuchen. Das Projekt setzt hierfür auf mehreren Ebenen an, die mitunter parallel bearbeitet werden. Bislang beschäftigten wir uns vorwiegend mit der Korpusbildung, der Aufbereitung der Metadaten zu den Drucken und der Aufbesserung der mit Optical Character Recognition (OCR) ausgelesenen Texte.

Korpusbildung und Metadatenaufwertung

Die Korpusbildung ist bereits weit vorangeschritten und erfolgt durch etablierte Methoden der Geisteswissenschaften, die durch neue digitale Erhebungsverfahren ergänzt werden. Das Korpus beschränkt sich grundsätzlich auf deutschsprachige Reiseberichte im Bestand der Österreichischen Nationalbibliothek (ÖNB), die zwischen 1500 und 1876 gedruckt wurden. Obwohl durch die Begrenzung auf die Bestände einer Bibliothek nicht die Gesamtheit der Reiseliteratur untersucht werden kann, eignen sich jene der ÖNB ausgesprochen gut für exemplarische Analysen mit Methoden der Digital Humanities, da die Bestände im Rahmen des Projektes Austrian Books Online (ABO) zwischen 2011 und 2018 mehrheitlich digitalisiert (ca. 600.000 Objekte) und die Texte mittels OCR-Software ausgelesen wurden (verfügbar via SACHA⁵).

2 Beispielhaft für eine weitreichende, wenn auch unvollständige Bibliographie: Carlo SALZANI / Steven TÖTÖSY DE ZEPETNEK, Bibliography for Work in Travel Studies, in: Library Series, CLCWeb: Comparative Literature and Culture (Version 14.09.2016), online unter: <https://docs.lib.purdue.edu/clcweblibrary/travelstudiesbibliography> (21.03.2019).

3 Beispielhaft: Bekim AGAI / Stephan CONERMANN, Hg., „Wenn einer eine Reise tut, hat er was zu erzählen“: Präfiguration – Konfiguration – Refiguration in muslimischen Reiseberichten, Berlin 2013; Almut HÖFERT, Den Feind beschreiben. „Türkengefahr“ und europäisches Wissen über das Osmanische Reich 1450–1600, Frankfurt am Main / New York 2003; Michiel VAN GROESEN, The Representations of the Overseas World in the De Bry Collection of Voyages (1590–1634), Boston / Leiden 2008.

4 Franco MORETTI, Distant reading, London / New York 2013.

5 ÖNB, Hg., Simple Access to Cultural Heritage Assets (SACHA), Online-Datenbank, online unter: <https://iiif.onb.ac.at/> (21.03.2019).

Die Korpusbildung im klassischen Sinne fand im Katalogsystem über kombinierte Suchanfragen zu Titelinformationen und Schlagwörtern⁶ statt, deren Ergebnisse fachlich kuratiert und durch Erhebungen in Bibliographien zum Thema laufend ergänzt werden.⁷ Das Korpus selbst wird direkt im Bibliothekssystem ALMA erstellt, wodurch die Daten zum einen systemintern effizient weiter bearbeitet werden können und zum anderen – im Gegensatz zu Stand-Alone-Datenbanken – nachhaltig gesichert werden. Außerdem stehen dadurch sowohl im User-Interface (Primo) als auch über die LOD (*Linked Open Data*)-Schnittstellen stets aktuelle Datensätze zur Verfügung.

Das Ergebnis dieser Erhebungen ist ein Basiskorpus von circa 3.200 Bänden an Reiseberichten (16. Jh.: 77, 17. Jh.: 180, 18. Jh.: 780, 19. Jh.: 2.200). Allerdings sind die Titel- und Schlagwortsuchen nur begrenzt aussagekräftig, da sich die Qualität der historisch gewachsenen Metadaten (wie Angaben zu Titeln, AutorInnen oder Druckorten) im Bibliothekssystem als äußerst heterogen erwies und Fehler ans Tageslicht traten. Bei der Migration vom handschriftlichen Zettelkatalog zum digitalen Bibliothekssystem durchliefen diese diverse Bearbeitungsprozesse, die zu unzähligen Dubletten, verkürzten Titeln, mangelnden oder oberflächlichen Beschlagwortungen und fehlenden Genre-Klassifizierungen führten. Die in ihrem Aufwand nicht zu unterschätzende Datenbereinigung (Metadata-Cleaning) und die darauf aufbauende systematische Datenaufwertung (Metadata-Enrichment) kann im Zuge des Travelogues-Projekts zwar nicht für den Gesamtbestand der ÖNB erfolgen, wohl aber für das erhobene Korpus an Reiseberichten. Bis Ende 2019 werden qualitativ hochwertige, kritisch geprüfte Metadaten im User-Interface des Bibliothekskataloges⁸ und über LOD-Schnittstellen in diversen Formaten (RDF, Bibframe, RDA) zu unserem Korpus abrufbar sein. Für weitere Analysen, Visualisierungen und Annotationen der Metadaten stehen die ÖNB-LibraryLabs zur Verfügung.⁹

Classifier

Aufgrund der generellen Limitierungen durch Titel- und Schlagwortsuchen, die durch die heterogene Qualität der Metadaten im Bibliothekssystem der ÖNB weiter verstärkt werden, und weil bisher keine Gesamtbibliographie zu deutschsprachigen Reiseberichten der Zeit 1500–1876 vorliegt, können durch die oben beschriebenen Vorgehensweisen unmöglich alle relevanten Werke im Bestand der ÖNB identifiziert werden. Deswegen wird im Rahmen dieses Projektes ein weitgehend automatisches Verfahren zur Auffindung bislang unbekannter Reiseberichte entwickelt, das sich an inhaltlichen Ähnlichkeiten der OCR-ausgelesenen Texte orientiert und von uns als „Classifier“ bezeichnet wird.

Zur Implementierung von Methoden des Machine Learning wurde zunächst eine Groundtruth erstellt, die sich aus dem wie oben beschrieben intellektuell identifizierten Korpus von ca. 3.200 Reiseberichten sowie derselben Zahl an Negativbeispielen (d.h. keine Reiseberichte) zusammensetzt. Letztere wurden durch ein randomisiertes Verfahren erhoben und ebenfalls intellektuell überprüft. Anhand der Unterschiede zwischen diesen beiden Klassen („Reisebericht“ und „Andere“) wurde eine SVM (Support Vector Machine) an einem Trainingsset trainiert. Um sprachlichen und inhaltlichen Wandlungsprozessen im Laufe des Untersuchungszeitraums zumindest teilweise gerecht zu werden, wurde das Korpus in vier Untersets zu je einem Jahrhundert gegliedert, zu denen jeweils eigene Modelle trainiert wurden. Mit

6 Gesucht wurde nach unterschiedlichen, trunkierten Schreibweisen der Begriffe „Reise“ und „Fahrt“, sowie nach den Schlagworten „Reisebericht“ und „Forschungsreise“.

7 Beispielhaft: Eutiner Landesbibliothek, Hg., Bibliografie und Katalog Reiseliteratur, online unter: <https://lb-eutin.kreis-oh.de/index.php?id=275> (21.03.2019).

8 ÖNB, Hg., Bibliothekskatalog (Primo), online unter: <https://search.onb.ac.at/> (21.03.2019).

9 ÖNB, Hg., Labs, online unter: <https://labs.onb.ac.at/de/> (21.03.2019).

dieser Methode konnten die Werke aus unserem Testset mit einer Genauigkeit zwischen 88 und 94 Prozent korrekt zugeordnet werden.

Um weitere Werke zu identifizieren, die unserer Definition von Reisebericht entsprechen, klassifizierten wir mit Hilfe dieser Modelle zudem alle bisher von uns nicht evaluierten, OCR-ausgelesenen Texte im Bestand der ÖNB. Für jedes Werk erhielten wir neben der binär definierten Klasse auch die Konfidenzwerte des Modells, sodass ein Ranking entstand. Um die Effektivität unserer Methode zu evaluieren, wurden für jedes Jahrhundert die jeweils ersten 100 als vermeintliche Reiseberichte klassifizierten Werke intellektuell überprüft. Zwischen 10 und 21 entsprachen jeweils unserer Definition („true positives“) und insgesamt konnten 56¹⁰ uns zuvor unbekannte Reiseberichte aufgefunden werden. Die vergleichsweise geringe Zahl an „true positives“ spricht einerseits für die Effektivität der klassischen geisteswissenschaftlichen Erhebungsmethoden, andererseits gründet sie in der großen Zahl an Texten, die als Reiseberichte klassifiziert wurden, ihnen inhaltlich auch stark ähneln, aber eben keine Reiseberichte sind (wie historiographische und geographische Werke oder Reiseführer). Deshalb arbeiten wir bereits an einer feineren Klassifikation, um beispielsweise mittels Named Entity Recognition und Topic Modeling unsere Modelle zu verbessern.

OCR-Nachbearbeitung

Neben der Korpusbildung und Metadatenbereinigung werden im Projekt derzeit die maschinenlesbaren Texte aufgebessert. Wenngleich sich die OCR-Software in den letzten Jahren zunehmend verbesserte, sind die damit erzeugten Transkriptionen weiterhin stark fehlerhaft, wenn auch in unterschiedlicher Ausprägung. Die Fehlerquote ist etwa abhängig von der Textstruktur im Digitalisat (z.B. Ein- und Mehrspaltigkeit) oder dem Vorhandensein grafischer Elemente. Weitere Herausforderungen sind der Einfluss des diachronen Sprachwandels und der Mangel an stilistischer und orthographischer Standardisierung während der Untersuchungsperiode (z.B. die Varianzen u/v, u/j und s/l).

Die Nachbearbeitung der OCR-ausgelesenen Texte in der ÖNB erfolgt im Travelogues-Projekt mit Hilfe Neuronaler Netzwerke (NN). Die Groundtruth hierfür stammt aus einem Testset von 13 Büchern, deren Volltexte sowohl in der OCR-ausgelesenen Form in der ÖNB vorliegen als auch im Zuge des Projekts „Deutsches Textarchiv“ (DTA)¹¹ manuell transkribiert wurden. Aus diesen Texten wurden N-Gram-Wortsequenzen extrahiert, mit Hilfe von Locality Sensitive Hashing (LSH) fehlerhafte Stellen in den OCR-Texten identifiziert und mit den sauberen Transkriptionen gepaart (insgesamt ca. 800.000 Paare). An dieser Groundtruth wurden verschiedene Encoder-Decoder-Modelle gemeinsam trainiert (basierend auf einem gewöhnlichen Encoder-Decoder-Modell wurde vor allem mit Sequence Representation und Weighted Attention experimentiert) und auf Character-, N-Gram-Character und BPE-Sequence-Level getestet. Die besten Ergebnisse wurden auf dem Character-Level erzielt. Weitere Experimente mit Global Attention und Hierarchical Attention führten dazu, dass unser Character Global Attention Modell die Word Error Rate (WER) der OCR-gelesenen Texte im Schnitt um 30 Prozent reduzieren kann und ein Multiple Attention Modell die Ergebnisse sogar weiter verbessert. Dennoch kamen wir zum Schluss, dass kein einzelnes Modell alle Probleme lösen kann, weswegen weitere Experimente stattfinden müssen, bevor die Neuronalen Netzwerke das gesamte Textkorpus des

¹⁰ Diese Zahl betrifft die bisherigen Evaluierungen vom 16. bis zum 18. Jahrhundert und umfasst auch Reiseberichte, die bei der Evaluierung der Negativbeispiele zufällig gefunden wurden; Ohne diese sind es 47.

¹¹ Berlin-Brandenburgischen Akademie der Wissenschaften, Hg., Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache, Berlin 2019, online unter: <http://www.deustextarchiv.de/> (21.03.2019).

Travelogues-Projekts aufbessern werden. Wir erwarten, dass durch diese Vereinheitlichung des Textmaterials die oben beschriebene Klassifikation, wie auch alle künftigen Analysen, bessere Ergebnisse erzielen.

Resümee und Ausblick

Durch den Einbezug von Machine Learning Verfahren wurden bereits vielversprechende Zwischenergebnisse bei der Korpusbildung und der OCR-Nachbearbeitung erzielt: Wir konnten Reiseberichte identifizieren, die durch klassische Erhebungsmethoden der Geisteswissenschaften nicht auffindbar waren und die Qualität der OCR-gelesenen Texte wurde an einem Testset deutlich verbessert. Die Arbeiten sind jedoch in beiden Bereichen noch nicht abgeschlossen und bedürfen weiterer Analysen.

Darüber hinaus werden wir uns künftig der semi-automatisierten Analyse der textuellen Inhalte der Reiseberichte zuwenden. Im Zentrum werden hierbei intertextuelle Relationen zwischen den einzelnen Reiseberichten (v.a. in-/direkte Zitate, Übernahmen von Textstellen und Allusionen) sowie die in den Reiseberichten tradierten Fremdwahrnehmungen stehen.

Literaturverzeichnis

- Bekim AGAI / Stephan CONERMANN, Hg., „Wenn einer eine Reise tut, hat er was zu erzählen“: Präfiguration – Konfiguration – Refiguration in muslimischen Reiseberichten, Berlin 2013.
- Berlin-Brandenburgischen Akademie der Wissenschaften, Hg., Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache, Berlin 2019, online unter: <http://www.deutschestextarchiv.de/> (21.03.2019).
- Eutiner Landesbibliothek, Hg., Bibliografie und Katalog Reiseliteratur, online unter: <https://lb-eutin.kreis-oh.de/index.php?id=275> (21.03.2019).
- Almut HÖFERT, Den Feind beschreiben. „Türkengefahr“ und europäisches Wissen über das Osmanische Reich 1450–1600, Frankfurt am Main / New York 2003.
- Franco MORETTI, Distant reading, London / New York 2013.
- ÖNB, Hg., Bibliothekskatalog (Primo), online unter: <https://search.onb.ac.at/> (21.03.2019).
- ÖNB, Hg., Labs, online unter: <https://labs.onb.ac.at/de/> (21.03.2019).
- ÖNB, Hg., Simple Access to Cultural Heritage Assets (SACHA), Online-Datenbank, online unter: <https://iiif.onb.ac.at/> (21.03.2019).
- Carlo SALZANI / Steven TÖTÖSY DE ZEPETNEK, Bibliography for Work in Travel Studies, in: Library Series, CLCWeb: Comparative Literature and Culture (Version 14.09.2016), online unter: <https://docs.lib.purdue.edu/clcweblibrary/travelstudiesbibliography> (21.03.2019).
- Michiel VAN GROESEN, The Representations of the Overseas World in the De Bry Collection of Voyages (1590–1634), Boston / Leiden 2008.