ZfdG

Zeitschrift für digitale Geisteswissenschaften

Artikel aus:
Zeitschrift für digitale Geisteswissenschaften

Titel:
Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts

Autor/in:
Martin Fechner

Kontakt: fechner@bbaw.de
Institution: Berlin-Brandenburgische Akademie der Wissenschaften
GND: 1085650278 ORCID:

Autor/in:
Andreas Weiß

Kontakt: weiss@gei.de
Institution: Berlin-Brandenburgische Akademie der Wissenschaften
GND: 1103829955 ORCID:

DOI des Artikels:
10.17175/2017_005

Nachweis im OPAC der Herzog August Bibliothek:
89714368X

Erstveröffentlichung:
19.12.2017

Letzte Überprüfung aller Verweise:
18.12.2017

GND-Verschlagwortung:
Computerunterstütztes Verfahren | Geschichtswissenschaft | Methode |

Zitierweise:
Martin Fechner, Andreas Weiß: Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts. In: Zeitschrift für digitale Geisteswissenschaften. 2017. PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/2017_005.

Martin Fechner, Andreas Weiß

Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts

Abstracts

Das Digitale erlangt in den Geschichtswissenschaften zunehmend eine besondere Stellung. Es wird sowohl als Forschungsthema wahrgenommen, als auch als Möglichkeit, die Geschichtswissenschaften methodisch und technisch zu unterstützen. Trotz dieses breiten Interesses werden Projekte, die sich neuer, digitaler Forschungsmittel und -methoden wie Topic Modeling bedienen, weiterhin besonders kritisch hinterfragt. Der Beitrag will daher konkret an zwei Projektbeispielen zeigen, wie die Ansätze des Topic Modelings genutzt und weiterentwickelt werden können, um die historische Forschung zu unterstützen. Beide Projekte beschäftigen sich mit Wissensbeständen des 19. Jahrhunderts, zum einen im Kontext der wissenschaftlichen Durchsetzung der Spektralanalyse und zum anderen im Kontext des Geschichtsunterrichts im Wilhelminismus. Dabei wird deutlich, wo Topic Modeling noch verbessert werden kann und welche Rolle die klassische Hermeneutik als zentraler Bestandteil der Anwendung eines historischen Topic Modelings spielen muss. Der Beitrag zeigt so Grenzen und Reichweiten der Einsatzmöglichkeiten von Topic Modeling in den Geschichtswissenschaften auf.

Within the discipline of history, digital approaches, tools, and research methods are taking on increasing importance as both a research topic in its own right and as a way to support the field of history as a whole. In spite of this broad interest, projects that make use of new digital tools and methods such as topic modeling continue to be reviewed very critically. Using two projects as examples, this article therefore demonstrates how topic modeling can be used and further developed in order to support historical research. Both projects deal with bodies of knowledge from the 19th century, one in the context of the scholarly implementation of spectral analysis and the other within the context of historical education during the period of the German Emperor Wilhelm II. Through these examples, this article shows where topic modeling can still be improved and what role traditional hermeneutics must have as a central component in the application of topic modeling in the field of history, and thus the boundaries and capabilities of the use of topic modeling in historical research.

1. Einleitung

Medialer Wandel ist eines der Kennzeichen der europäischen Moderne. Mit seinem Entstehen und seinen Effekten beschäftigen sich sowohl die Kultur- wie die Geschichtswissenschaften. Doch neue Medien stoßen nicht immer sofort auf Begeisterung. So wird der digitale Wandel von Teilen der historisch forschenden Community immer noch als bedrohliche Entwicklung begriffen, auch wenn schon vor einiger Zeit sowohl auf die Vorteile wie auf die Notwendigkeit digitaler Geschichtswissenschaft hingewiesen wurde.[1] Denn digitale Geschichtswissenschaft muss sich nicht nur den Herausforderungen nachträglich digitalisierter Quellen stellen (wie es in den unten vorzustellenden Projekten der Fall ist), sondern zunehmend auch der Frage nach Digital-born-Data. Entwicklungen

[1] Schmale 2010.

zu rein elektronisch generierten Texten zeigen ganz eigene Konsequenzen, die sich nicht nur in der Debatte um scheinbar verloren gegangene Lesefähigkeiten bei Studierenden widerspiegeln sowie in der Kritik an einer vorgeblich stärker vorselektierten Quellenauswahl durch digital vorhandene Quellenkorpora, sondern auch darin, ob die neuen Methoden die Identität des Faches Geschichtswissenschaften in seinem Kern verändern oder insgesamt den Geisteswissenschaftlern verständlich, sprich interpretierbar, bleiben. Diese großen Fragen lassen sich wohl einfacher diskutieren, wenn sie in ihre einzelnen Bestandteile zerlegt werden. Dieser Artikel will daher anhand eines spezifischen Werkzeuges, des Topic Modelings, der Frage nachgehen, ob die Digital Humanities (im Weiteren DH) und ihre Untergruppe, die digitalen Geschichtswissenschaften, etwas revolutionär Neues, Umwerfendes für das Fach Geschichtswissenschaft oder nur ein weiteres Werkzeug sind. Zwar erfreut sich Topic Modeling vor allem in den Literaturwissenschaften und der Computerlinguistik zunehmender Beliebtheit, doch sind Anwendungsbeispiele aus den deutschsprachigen Geschichtswissenschaften noch selten, vor allem, wenn ein großes Korpus untersucht werden soll.[2] Beispielhaft soll hier ein Vergleich zwischen zwei Projekten vorgestellt werden: eine Analyse zur Diskussion über die 1859 erfundene Spektralanalyse in zeitgenössischen, naturwissenschaftlichen Publikationen und das Projekt »Welt der Kinder. Weltwissen und Weltdeutung in Schul- und Kinderbüchern zwischen 1850 und 1918«.[3] Zunächst wird jedoch in einem Abschnitt kurz auf Topic Modeling im historischen Kontext und die Wechselbedingungen von Fach und Technologie eingegangen. Danach werden in Abschnitt 1.2 die technischen Grundlagen von Topic Modeling und die verschiedenen Projekte vorgestellt sowie Spezifika und Unterschiede erwähnt. Im Hauptteil dieses Beitrags wird erst das Projekt »Spektralanalyse« erläutert, darauf folgt die Vorstellung des Projektes »Welt der Kinder«. Im dritten Abschnitt schließt sich der Vergleich und die Auswertung anhand der Arbeitsschritte an, die der hier vorgeschlagenen Vorgehensweise bei Topic Modeling entsprechen (Korpusbildung, Trainingsset, Topicerstellung, maschinelles Lernen, Überprüfung, Ergebnisse), bevor (ein vorläufiges) Urteil im Resümee den Artikel beschließt. Aus dieser vergleichenden Vorgehensweise anhand zweier Projekte ergibt sich eine gewisse Redundanz in der Darstellung und eine thematische Parallelisierung; es wurde aber versucht, dies so weit wie möglich zu vermeiden.

Grundvoraussetzungen für Topic Modeling ist eine gute OCR. Die Arbeit beginnt also schon vor der eigentlichen Berechnung; Grundlagen eines guten OCRs werden allerdings nicht Gegenstand dieses Artikels sein. Hier sollen vielmehr die Wechselwirkungen zwischen spezifischen Fragestellungen von Historikern und der Entwicklung und Verbesserung von Topic Modeling im Vordergrund stehen: welche Auswirkungen haben die unterschiedlichen Quellentypen auf die Erstellung von Topics (welche eignen sich hierfür mehr, und gibt es weniger geeignete)? Und wie lassen sich die Topics in eine geschichtswissenschaftliche Analyse einbinden?

[2] Für den amerikanischen Raum behauptete Ted Underwood das Gegenteil, allerdings beziehen sich alle amerikanischen Verweise auf dieselben Anwendungsbeispiele, v.a. »Mining the Dispatch« (http://dsl.richmond.edu/dispatch/); vgl. Underwood 2012. Zu Anwendungsbeispielen in den Literaturwissenschaften vgl. den Sammelband Erlin / Tatlock 2014.

[3] Zum Projekt »Welt der Kinder« vgl. http://welt-der-kinder.gei.de/; die Suchen können auf http://wdk.gei.de/ selbst wiederholt werden. Teilweise finden sich die zugrundeliegenden Daten unter https://github.com/UKPLab/corpus-explorer und https://github.com/ewomant/corpus-explorer. Die Abschnitte zur Spektralanalyse beruhen auf der Dissertation von Martin Fechner, vgl. Fechner 2016. Auf den Vergleich und die Unterschiede zwischen beiden Projekten wird in den einzelnen Abschnitten genauer eingegangen.

1.1 Historischer Teil

Klassischerweise entwickeln die Geschichtswissenschaften neue Fragestellungen im Wechselspiel mit neuen Methoden. Die Nutzung von Topic Modeling ist aber nur sinnvoll, wenn es eine konkrete Fragestellung gibt – von sich aus sagen die Daten nichts, sie sprechen nur begrenzt zum ›Nutzer‹. Die Annales-Schule, aber auch die historische Sozialforschung waren frühe Versuche, große Datenmengen und die Erschließung neuer Quellengattungen für historische Fragestellungen zu ermöglichen. Beide rekurrierten häufig auf alltagsgeschichtliche Themen. In die Computertechnik wurden dabei schon früh Hoffnungen gesetzt, die in den seltensten Fällen erfüllt wurden.[4] Die Wende zur Kulturgeschichte nach dem Ende der Sozialgeschichte brachte den Versuch auf den Weg, Hochkultur und Alltagsgeschichte miteinander zu vereinen. Da damit aber immer größere Quellenbestände erschlossen werden mussten, boten die DH einen Ausweg, eine Hoffnung, die besonders mit dem Konzept des »Distant Reading« verbunden wurde.[5] Die Erwartung war, man würde nun endlich auch die großen Masse an damals populärer Literatur schnell und einfach sichten können, die von der bisherigen, auf die Hochkultur konzentrierten Forschung ignoriert wurde, und so neue Erkenntnisse gewinnen. Gleichzeitig wurden Digitalisierungsbemühungen im Rahmen von Bestandssicherung, aber auch im politisch gewollten freien Zugang zu ›öffentlichen‹ Wissensbeständen gefördert. Doch nicht nur der scheinbar vergrößerte Datenbestand, aus dem Informationen gefiltert werden müssen, trieb die Digitalisierung voran. Durch die theoretischen Wenden seit den 1970er-Jahren wurden Diskurse Untersuchungsobjekte. Hier bietet Topic Modeling möglicherweise einen Ansatzpunkt, um Diskurse in den zu untersuchenden Texten auffinden zu können.[6] Topic Modeling wurde dabei entwickelt, um große Datenmengen nach zusammenhängenden Informationen zu durchsuchen; unabhängig davon, ob es sich um Text oder Desoxyribonukleinsäure (DNS) handelt – vielleicht einer der Gründe, warum viele Geisteswissenschaftler dem Verfahren mit einer gewissen Skepsis gegenüber stehen. Interessanterweise wird aber dieses Argument der vielseitigen Anwendbarkeit von Topic Modeling in den neueren geisteswissenschaftlichen Beiträgen hierzu nicht mehr betont; entweder ist dies ein Hinweis darauf, dass schon innerhalb kurzer Zeit die Ursprünge des Verfahrens verdrängt wurden oder darauf, dass man aus strategischen Gründen dieses Argument nicht mehr vorbringen wollte. Trotz dieser scheinbaren Erleichterung ist Topic Modeling in der derzeitigen historischen Forschung noch nicht ganz angekommen; dies zeigen aktuelle deutschsprachige Einführungen in die Thematik, obwohl auch diese von einem aktuellen Trend in der Forschung ausgehen.[7] Ein Großteil der Debatte über Sinn und Möglichkeiten des Einsatzes von Topic Modeling wird dabei in Blogs geführt.[8] Häufig genannte Gründe für den Widerstand von Historikern sind dabei

[4] Für ein frühes, hoffnungsfrohes Beispiel vgl. Tilly 1973. Der vor allem in der neueren Literatur erwähnte Jesuitenpater Roberto Busa spielt in dem hier vorgestellten Kontext keine Rolle, da seine Arbeiten vor allem wegweisend für digitale Editionsprojekte waren.
[5] Den Begriff und das Konzept »Distant Reading« prägte Franco Moretti; vgl. Moretti 2005.
[6] Allerdings müssen wir anmerken, dass es sich bei dem von Literaturwissenschaftlern verwendeten um einen sehr weiten Diskursbegriff handelt; wir würden im Deutschen den Begriff Themenkomplex bevorzugen; zum Diskurs vgl. Underwood 2012.
[7] Vgl. Koller 2016, der zwar zweimal konkret auf Topic Modeling eingeht, aber keine Anwendungsbeispiele deutschsprachiger Historiker nennt; ebenso Thomas Meyer, der auf aktuelle Anwendungen bei Twitter u.a. verweist, aber ebenso wenig auf deutsche Beispiele verweisen kann, Meyer 2016. Schon 2012 erschien ein Themenheft zu Topic Modeling im Journal of Digital Humanities; vgl. Journal of Digital Humanities 2 (2012), 1.
[8] Für den angelsächsischen Raum vgl. die genannten Beispiele; für den deutschsprachigen Raum ist sicher Hypotheses der wichtigste Metablog, http://hypotheses.org/, daneben der DHd-Blog.

eine »tiefsitzenden Tradition des Skeptizismus gegenüber quantitativen und empirischen Verfahren«, da jeder Geisteswissenschaftler die eigene Fragestellung am Quellenmaterial erproben will, dieses sich aber auch häufig gegen zu eindeutige Interpretationen sperrt.[9] Es geht also um die Frage, wie und welches Wissen generiert werden kann.

Die Frage von Wissensgewinnung über bekannte Texte hinaus steht im Mittelpunkt der beiden hier vorzustellenden Projekte. In beiden geht es um die Frage von Wissen: Welches Wissen wurde über die Spektralanalyse wann und in welcher Form verbreitet? Und was erfuhren Kinder des ausgehenden 19. Jahrhunderts aus ihren Schulbüchern über die Welt? Für beide Fragestellungen muss ein großes und diverses Quellenkonvolut erschlossen werden. Topic Modeling soll dabei helfen, schneller konkrete Dokumente aufzufinden und die Veränderung von Themen und Trends über den Untersuchungszeitraum hinweg herauszufiltern.

Insofern bieten die hier vorgestellten Ansätze eine differenzierte Vergleichsebene für die Darstellung der Vor- und Nachteile von Topic Modeling, da zwei unterschiedliche Herangehensweisen an zwei auch sprachlich differenten Korpora – und den daraus resultierenden unterschiedlichen Problemen – vorgestellt werden. Sie unterscheiden sich hierin auch von anderen aktuellen Publikationen zu diesem Themenbereich. So erschienen in den letzten zwei Jahren allein in dieser Zeitschrift zwei Artikel, die sich mit Topic Modeling auseinandersetzen. Peter Andorfer beschäftigt sich ebenfalls mit automatisierten Verfahren im Topic Modeling.[10] Er beschreibt wie eine manuelle Topic-Bildung und –zuordnung mit einer automatisierten verglichen werden kann und welche Vor- und Nachteile beide Methoden haben. In seinem Fazit präferiert er schließlich ein automatisiertes Verfahren. Der Artikel »Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)?« von Jörg Wettlaufer berührt zwar auch den Themenkreis des Topic Modeling, besitzt aber eine andere Ausrichtung. Dort wird diskutiert, wie sinnvoll verschiedene DH-Ansätze für die historische Forschung sind. Er untersucht drei Methodenbereiche – Handwritten Text Recognition, Visualisierungen und Netzwerkanalysen, Semantische Technologien – auf ihren Nutzen für die historische Forschung. Topic Modeling wird bei ihm kurz erwähnt, vor allem als Werkzeug in der Wissenschaftsgeschichte und für „Sprachkorpora in den alten Sprachen".[11]

Doch ein besonderes Problem soll überleiten zum technischen Teil dieser Einführung. Im 19. Jahrhundert waren weder die Unterrichts- noch die Universitätssprache normiert oder standardisiert, eine Herausforderung für die dem Topic Modeling zugrundeliegenden Modelle. Während beim Schulbuch aber eher das Problem ist, dass in sich wiederholenden sprachlichen Bildern auch die ›Ausreißer‹ gefunden werden sollen – und diese gab es –, schafft sich die wissenschaftliche Debatte um die Spektralanalyse ihren eigenen normierten Wortschatz. Dies betrifft einen Punkt, der besonders bei Verfahren wie dem Dynamic Topic

[9] So Kirschenbaum 2007.
[10] Andorfer 2017. Die Autoren des vorliegenden Artikels stimmen allerdings mit seinen weitreichenden Schlussfolgerungen nicht überein, die er aus einem relativ übersichtlichen, sehr gut aufbereiteten Corpusg ewonnen hat. Besonders wichtig scheint uns, dass man doch, auch wenn man Algorithmen und Werkzeuge nicht in aller Tiefe durchdringt, die »folgenden Konsequenzen« abschätzen sollte, ansonsten ist es riskant, eine Methode, die man nicht versteht, anzuwenden und die Ergebnisse als objektiv anzusehen.
[11] Wettlaufer 2016.

Modeling gern übersehen wird. Mit Hilfe von Dynamic Topic Modeling erkennt man an der Verschiebung der einzelnen Wörter im Ranking zum Beispiel wissenschaftliche Entwicklungen und Themensetzungen.[12] Das zugrundeliegende Verfahren ›verteilt‹ scheinbar die Topic-prägenden Begriffe über den untersuchten Zeitraum hinweg, da es diese ja im gesamten Zeitraum vermutet.[13]

1.2 Technischer Teil

Insgesamt gilt Topic Modeling als ein vergleichsweise technisch unaufwändiges, nachgeordnetes Verfahren. Der in der Forschung verbreitetste, weil einfachste Ansatz beruht auf Latent Dirichlet Allocation (LDA), das hauptsächlich von David M. Blei entwickelt wurde.[14] Es dient als Grundlage für den ›Werkzeugkasten‹ MALLET, der auch bei »Welt der Kinder« genutzt wurde.[15] Topic Modeling dient dazu, wie oben erwähnt, Themen oder Diskurse in Texten zu finden. Es bedient sich dabei statistischer Verfahren, die jedes Wort mit jedem weiteren Wort in Beziehung setzen, um so Häufigkeiten des Auftretens zu messen und daraus Beziehungen zwischen den Wörtern herzustellen. Die Wörter werden ohne Beachtung der Reihenfolge oder Auftreten gezählt.[16] Sowohl die einzelnen Topics wie jedes in ihnen enthaltene Wort können ›gerankt‹ werden. Je nach Textgattung und -art können sich dabei abstrakte Topics wie etwa »Liebe« ergeben, aber auch konkrete etwa zur »Französischen Revolution« als historischem Ereignis.

Allerdings hängt das Verfahren wesentlich von der Qualität der Rohdaten ab. OCR-bearbeitete Dokumente mit einer Korrektheit von mindestens 98 Prozent sind als Voraussetzung zu betrachten. Ein großes Hindernis sind bis heute die geringeren Erkennungsraten bei Fraktur-Texten.[17] Auch fehlen elektronische Wörterbücher und Online-

[12] Vgl. Blei 2012a.
[13] Schmidt macht dies in seiner lesenswerten kritischen Intervention an dem Beispiel »Kalter Krieg« fest, der ja nicht vor 1945 in den Quellen fassbar sein sollte; vgl. Schmidt 2012, S. 63. Allerdings sind manche Begriffe nicht so rezent, wie gerne angenommen wird; interessante Verwerfungen ergeben sich zum Beispiel beim Begriff Kolonialkrieg, der eine Bedeutung von einem europäischen Krieg um Kolonien (auf anderen Kontinenten) hin zur Eroberung von Kolonien erhält, um ein Beispiel aus unserer Forschung zu nennen.
[14] Blei et al. 2003. Es finden sich verschiedene Einführungen im Web, zwei mit zahlreichen Links sind http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/ sowie http://www.lisarhody.com/some-assembly-required/. Aus Letzterem stammt folgende Zusammenfassung: »In fact, text mining in general assumes that writers go to *great lengths* to make clear, unambiguous arguments. Computer scientists make that assumption because LDA was written to deal with large repositories and collections of non-fiction text. […] So […] the classic examples of topic models produce semantically and thematically coherent keyword distributions […] Returning once again to Blei's most accessible article for humanists, he writes: »The interpretable topic distributions arise by computing the hidden structure that likely generated the observed collection of documents.« Blei clarifies his statement in a footnote which reads: »Indeed calling these models »topic models« is retrospective – the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA« (Blei, »Introduction«, 79).« Rhody (und Blei) geht es hier darum darzulegen, dass Topics umso besser identifiziert werden können, je höher die Übereinstimmung zwischen sprachlichem Stil und Inhalt ist. Da der Algorithmus aber nicht zwischen wissenschaftlichen und literarischen Texten unterscheidet, kann dieses Verfahren, das von Blei nur an wissenschaftlichen Texten getestet wurde, auch auf poetische Texte übertragen werden. Wichtig ist immer nur Texte im Zusammenhang mit den anderen Texten des jeweiligen spezifischen Corpus zu lesen.
[15] Zu den technischen Grundlagen von »Welt der Kinder« vgl. Schnober / Gurevych 2015.
[16] Der sogenannte Bag-of-Words-Ansatz; die einzelnen technischen Voraussetzungen hierfür und die Aufbereitung des Korpus werden weiter unten dargestellt.
[17] Dem Problem der schlechten OCR-Daten bzw. dem Fehlen von OCR-Daten hat sich erfreulicherweise das DFG-Koordinierungsprojekt OCR-D unter Leitung von Thomas Stäcker, Alexander Geyken und Markus

Datenbanken für die Sprache des 19. Jahrhunderts, die bei der Bereinigung von OCR-Fehlern sowie bei Lemmatisierung, Stemming und Tokenisierung helfen würden.[18] Daher wurde im Projekt »Welt der Kinder« auf eine Schreibweisennormalisierung verzichtet.

Die Erstellung der Topics wurde in den beiden Projekten unterschiedlich gehandhabt. Im Projekt »Welt der Kinder« mussten die Fachhistoriker auf externe Ressourcen zurückgreifen, um die Topics berechnen zu lassen. Somit war zwar eine höhere Verlässlichkeit auf der mathematisch-technischen Seite gewährleistet,[19] wiederholte Berechnungen mussten aber immer im Team abgesprochen und die einzelnen Arbeitsschritte vereinbart werden. Das Projekt »Spektralanalyse« hingegen konnte nur auf seine ›normalen‹ Ressourcen zurückgreifen – in diesem Fall ein eigener Rechner –, bewegt sich somit aber dichter an der Arbeitsrealität vieler Geisteswissenschaftler. Um dennoch eine hohe Qualität zu erreichen, wurde daher nicht auf eine elaborierte maschinelle Verarbeitung, sondern vor allem auf im Close Reading gewonnene Informationen für das Topic Modeling zurückgegriffen. Das heißt aber für beide Projekte, es konnten nicht beliebig viele Testläufe zu allen gewünschten Unterkorpora durchgeführt werden; auch bleibt das Verfahren bis zu einem gewissen Grade unüberwacht. Dieses Problem ergibt sich zum einen aus der technischen Ausstattung (Rechnerkapazität, Informatikkenntnisse), zum anderen aber auch aus den Projektkonstruktionen selbst (interdisziplinäres Arbeiten). Dies betrifft einen weiteren wichtigen Punkt disziplinären Unbehagens gegenüber Topic Modeling. Nicht nur überblicken viele Geisteswissenschaftler die mathematischen Modelle hinter Topic Modeling nicht, auch die Entwickler selbst tragen mit zweideutigen Aussagen nicht dazu bei, dieses Misstrauen zu verringern.[20]

Im Gegensatz zu vielen anderen Projekten und Analysen, die sich des Topic Modelings bedienten, verzichten die hier vorgestellten Anwendungen weitgehend auf die Nutzung von Visualisierung von Topic Modeling. Die hat vor allem zwei Gründe: Erstens verleitet Visualisierung zu vorschnellen Interpretationen von Nähe und Distanz der Topics, wenn schon die zugrundeliegenden Algorithmen nur begrenzt verstanden werden. Zweitens, und eng damit verbunden, ist Visualisierung nur ein weiteres simplifizierendes Hilfsinstrument, dank dem der Eindruck entstehen kann, man würde sein Material besser verstehen und interpretieren können als dies eigentlich der Fall ist.

Brantl angenommen, dass sich die Digitalisierung und Texterkennung aller deutschen Drucke bis ins 19. Jahrhundert vorgenommen hat.

[18] Vgl. z.B. Mimno 2012, der ebenfalls eine wissenschaftsgeschichtliche Untersuchung durchführt.

[19] Die Verknüpfung verschiedener Datensätze und die Implementierung in eine webbasierte Plattform, die Verknüpfungen zu verschiedenen anderen externen Datenbanken bietet, wäre ansonsten sicherlich nicht so einfach zu gewährleisten gewesen.

[20] Symptomatisch die Aussagen von David Blei: »Note that the statistical models are meant to help interpret and understand texts; it is still the scholar's job to do the actual interpreting and understanding. A model of texts, built with a particular theory in mind, cannot provide evidence for the theory. (After all, the theory is built into the assumptions of the model.) Rather, the hope is that the model helps point us to such evidence. Using humanist texts to do humanist scholarship is the job of a humanist. In summary, researchers in probabilistic modeling separate the essential activities of designing models and deriving their corresponding inference algorithms. The goal is for scholars and scientists to creatively design models with an intuitive language of components, and then for computer programs to derive and execute the corresponding inference algorithms with real data. The research process described above – where scholars interact with their archive through iterative statistical modeling – will be possible as this field matures. [...] Probabilistic models promise to give scholars a powerful language to articulate assumptions about their data and fast algorithms to compute with those assumptions on large archives. I hope for continued collaborations between humanists and computer scientists/statisticians. With such efforts, we can build the field of probabilistic modeling for the humanities, developing modeling components and algorithms that are tailored to humanistic questions about texts.«; Blei 2012b, S. 10–11.

Jedes andere Vorgehen führte zu einem je spezifisch anderem Evaluierungsprozedere. Für die Schulbücher wurden diese anhand von den Informationswissenschaftlern erstellten Excel-Tabellen beurteilt und diese Ergebnisse an die im Projekt beteiligten Computerlinguisten und Informatiker zur Weiterentwicklung des Topic Models zurückgespiegelt. Dieses gestufte Verfahren bedeutetet zwar für alle Seiten einen höheren Arbeitsaufwand, ermöglichte aber so, über die einzelnen Schritte immer wieder zu reflektieren und sie zu evaluieren.

2. Vorstellung der Einzelprojekte

2.1 Spektralanalyse

Ausgangspunkt des Projektes »Vom Labor in die Öffentlichkeit« (hier im Text »Spektralanalyse« genannt) war die Nachverfolgung wissenschaftlicher Themenkomplexe aus ihrem Entstehungskontext durch verschiedene Medien. Konkret wurde untersucht, in welcher Form wissenschaftliche Artikel und Buchpublikationen zur Spektralanalyse und zum Laser im Jahrzehnt nach ihrer Erfindung erschienen sind und welche thematischen Entwicklungen sich dort niederschlugen.[21]

Mitte des 19. Jahrhunderts wurde von dem Chemiker Robert Bunsen und dem Physiker Gustav Robert Kirchhoff das Problem der Identifizierung von chemischen Elementen untersucht. Resultat ihrer Forschung war eine Methode, die die optischen Spektren von chemischen Proben nutzte, um ihre Zusammensetzung zu bestimmen: die Spektralanalyse. Diese Methode stellte einen Meilenstein für die Chemie dar und gleichzeitig ermöglichte sie der Forschung, die Beschaffenheit von astronomischen Körpern aufgrund ihres Lichts zu untersuchen. So fand sich auch schnell eine Reihe von Anwendungen in der Industrie und in der Forschung, mit populärwissenschaftlichen Vorträgen und Publikationen wirkte die Erfindung in die Gesellschaft hinein. In einer Zeit, in der die Naturwissenschaften strukturell wuchsen und ihre Mathematisierung zunahm, unterstützten die Erfolge der Spektralanalyse die neuen methodischen Ansätze. Das hier vorgestellte Projekt untersuchte an den Publikationen zur Spektralanalyse unter anderem, wie sich mit zunehmender Forschung die thematischen Schwerpunkte verschoben, ob es wiederkehrende Muster gab und welchen Einfluss die wissenschaftlichen Publikationen auf die nachfolgende Forschung und populärwissenschaftliche Veröffentlichungen hatte.

Hierfür war es nicht nur notwendig die Quellentexte, zumeist wissenschaftliche Publikationen, zu den ausgewählten Themenkomplexen in den ausgewählten Medien zu ermitteln, sondern es war für die Analyse auch von Bedeutung, eine Binnendifferenzierung der Themenkomplexe vorzunehmen und ihr Vorkommen in den untersuchten Quellentexten zu dokumentieren.

[21] Vgl. dazu Fechner 2016, S. 157–280.

In Anlehnung an die Methode der qualitativen Inhaltsanalyse[22] aus der Kommunikationswissenschaft wurden daher für die Binnendifferenzierung der ausgewählten Themenkomplexe einzelne Topics definiert und deren explizites Vorhandensein in den untersuchten Quellentexten in entsprechenden Datensätzen festgehalten. Darüber hinaus war es sinnvoll, die Quellen jeweils als Ganzes einer Themenklasse zuzuordnen, um sie für die Analyse gruppieren zu können.

Im Projekt wurde zur Differenzierung und Gruppierung der Quellen ein Modell aufgestellt, welches gleichartige Kommunikationssituationen in Kommunikationsräumen zusammenfasst,[23] die durch konkrete Eigenschaften beschrieben werden können. Eine dieser Eigenschaften, die sich auf das Medium bezieht und an der Quelle selbst bestimmt werden kann, ist der in der Quelle behandelte Themenkomplex. Da die Quellen mit Blick auf einen konkreten Themenkomplex ausgewählt und untersucht wurden, war eine Binnendifferenzierung der Themen aus zweifacher Perspektive sinnvoll. Durch die Binnendifferenzierung wurden einzelne Topics definiert und in den Quellentexten identifiziert. Welche Anforderungen an dieses zweifache Topic Modeling gestellt werden mussten, wurde von der angestrebten Analyseposition bestimmt.

Den ersten Zugang zur Differenzierung stellte die Bildung von Themenklassen dar. Die Formulierung der Klassen basierte auf einem hierarchischen Zugang zum jeweiligen Themenkomplex. Die Quellentexte ließen sich entweder einem Teilbereich des untersuchten Themenkomplexes zuordnen oder sie thematisierten den gesamten Themenkomplex. Auch war es möglich, dass sie darüber hinaus gingen und auch noch andere Themen miteinbezogen. Die im Projekt verwendeten Themenklassen finden sich in der folgenden Liste:

- Allgemeine Themen: Diese Zuordnung umfasst Publikationen zur Wissenschaft, Physik, Chemie, Astronomie, biographische Schriften, u.a.
- Ausgewählter Themenkomplex: Hierzu zählen Publikationen, die sich in einführender, umfassender oder allgemeiner Weise mit der Spektralanalyse beschäftigen.
- Teilbereiche: Hierunter fallen Publikationen, die sich mit Teilbereichen der Spektralanalyse beschäftigen, vor allem mit Verbesserungen und neuen Spektralapparaten.
- Anwendungen: Bei einer großen Zahl von Publikationen steht nicht die Spektralanalyse selbst im Vordergrund, sondern ihre Anwendung. Dies betrifft vor allem Anwendungen in Astronomie und Forschung, etwa die Entdeckung und Identifizierung neuer Elemente.

Die Quellentexte, die sich auf einen Teilbereich des Themenkomplexes konzentrierten, benötigten jeweils eigene passend deklarierte Klassen. Die Einteilung der Klassen geschah passend zum Quellenmaterial. Bei der Beschreibung der Klassen wurde darauf geachtet, dass die Klasse

[22] Die qualitative Inhaltsanalyse wird besonders im Bereich der Kommunikationswissenschaft eingesetzt, etwa um die argumentativen Verschiebungen in der publizistischen Berichterstattung nachzuweisen, vgl. Früh 2006 und Mayring 2010.
[23] Die untersuchten Medien mit gleichen Eigenschaften spannen einen entsprechenden Kommunikationsraum auf. Die bestimmbaren Parameter sind etwa dort behandelte Themenkomplexe, benutzte Sprache, mediales Objekt. Eine genaue Behandlung dazu findet sich in Fechner 2016, S. 85–96.

- die Quelle passend beschrieb,
- eine hohe Reliabilität aufwies, die Zuordnung also eindeutig war, und
- von anderen Klassen abgegrenzt werden konnte.

Ausschlaggebend für die Zuteilung der Quellen zu den Klassen war die wissenschaftliche Expertise über den gesamten Themenkomplex. Die Klassen selbst sowie die Zuordnung und Schwierigkeiten der Zuordnung wurden während der Forschung mitdokumentiert.[24] Ein maschinengestütztes Topic Modeling und eine automatische Zuordnung der Quellentexte zu den Klassen waren im Projekt aufgrund der geringen technischen Ressourcen und der ungenügenden OCR-Qualität der Texte nicht umsetzbar.

Der zweite Ansatz zur Binnendifferenzierung des Themenkomplexes in einzelne Topics setzte an anderer Stelle bei Betrachtung der Quellentexte an. Um genauer festzustellen, wie die Quellen den Themenkomplex behandeln, in welcher Länge und Reihenfolge sie auf einzelne Topics eingehen, wurden kleinere Abschnitte eines Quellentextes separat behandelt. In den Abschnitten konnten die Topics dann jeweils einzeln nachgewiesen werden. Damit ergab sich aus der Binnenstruktur der Quellentexte, aus den Kapiteln und Absätzen, die Formulierung konkreter Topics. Da die Länge der betrachteten Abschnitte deutlich kürzer war als die Quellentexte, unterschieden sich die so gebildeten Topics auch deutlich von den oben betrachteten Themenklassen. Es handelte sich bei ihnen um einzelne konkrete Themen im Gegensatz zu den allgemeineren Klassen. Diese konkreten Topics, die in den Quellen zum Themenkomplex der Spektralanalyse vorkamen, wurden im Vorfeld der Analyse der Zeitschriftenartikel anhand von zeitgenössischer Hand- und Lehrbuchliteratur identifiziert und dokumentiert. Dann wurde für alle Abschnitte in den Zeitschriftenartikeln erarbeitet, welche der dokumentierten Topics dort jeweils behandelt wurden. Einige Topics ließen sich anhand des Vorhandenseins von einzelnen Schlüsselworten oder von bestimmten Wortkombinationen direkt nachweisen. Sie konnten den Quellen daher leicht zugeordnet werden. Es gab auch Abschnitte in den Quellen, die ein hohes Textverständnis erforderten, also Fachkenntnisse zur Spektralanalyse, um zu identifizieren, welches Topic an diesen Stellen behandelt wurde. Die Zuordnung war in diesen Fällen schwieriger, jedoch nicht weniger eindeutig. Der Übergang von einfachen Topics, die durch bestimmte Schlüsselworte identifizierbar waren, hin zu abstrakteren Topics, für deren Identifizierung ein tieferes Textverständnis benötigt wurde, war dabei fließend. Wenn sich ein Abschnitt keinem der bekannten Topics zuordnen ließ, behandelte er unter Umständen ein noch nicht dokumentiertes Thema. Dieses wurde dann der Dokumentation hinzugefügt und für die weitere Zuordnung verwendet. Im Abschnitt 3 werden das technische Vorgehen und die Analyse aus dem Projekt »Spektralanalyse« unter Einbezug der Erfahrungen aus dem Projekt „Welt der Kinder“ weiterentwickelt und in Einzelschritten erläutert.

2.1.2 Ergebnisse für die historische Forschung

[24] Fechner 2016, S. 135f.

Anhand der Quellen konnte belegt werden, dass die Forschung zur Spektralanalyse einem auch in anderen Kontexten vorhandenen Muster folgte. Nach der Entwicklung der Spektralanalyse als neuem Instrument für die Forschung erschienen zunächst einige Schriften, welche die Spektralanalyse umfassend behandelten. Weiterhin stieg die Zahl der Artikel zu dem Themenkomplex an, wobei sich der Fokus zunehmend auf die Anwendung verschob. Dabei reduzierte sich die Zahl der angesprochenen Themen und viele Texte fokussierten sich lediglich auf einen oder zwei Anwendungsbereiche. Auffällig ist die große Bandbreite angesprochener Themen in den frühen Texten der Forscher, die auch in der Folgezeit viel veröffentlichten. Auch diese konzentrierten sich mit der Zeit auf immer weniger Themen. Auch die im Projekt aufgestellte These von unterschiedlichen Textformen im wissenschaftlichen Forschungsdiskurs, die sich auf die Ergebnisse einer Zitationsanalyse stützte, konnte mit der Analyse nach Themen bekräftigt werden. So gab es einerseits Forschungsberichte, die neue Forschungen schildern, und andererseits zusammenfassende Artikel, was sich in Unterschieden in der behandelten Themenauswahl, der Zitationen und der Rezeption niederschlug. Die Entwicklung der Spektralanalyse schlug sich in den Artikeln oft in gleicher Weise nieder. Einige Themen traten sehr häufig zusammen auf und formten den Hintergrund, vor dem die Entdeckung der Spektralanalyse wahrgenommen wurde. So wurde die Spektralanalyse besonders in jenen Texten als beeindruckende Erfindung bezeichnet, die gleichzeitig auch auf die Möglichkeit Elemente in der Sonne nachzuweisen eingingen. Diese Argumentationslinien finden sich entsprechend auch in den späteren populären Texten und der Lehrbuchliteratur wieder.

2.2 Schulbücher – das Projekt Welt der Kinder

Das 19. Jahrhundert war eine Zeit der Bildungsexpansion. Am Ende des Jahrhunderts besuchten über 90 Prozent der Kinder zumindest eine Grundschule. Für viele Kinder waren Schulbücher der erste (und oft einzige) reguläre Lesestoff, den sie nutzten, um sich Wissen über die Welt anzueignen. Um dieses Wissensbedürfnis zu befriedigen, gab es eine Vielzahl von Büchern und Reihen, die sich, wie schon Titelhinweise wie etwa »für Schule und Haus« anzeigen, auf die Wissensvermittlung im Grenzbereich von Schule und Zuhause konzentrierten. Schulbücher eignen sich daher in besonderer Weise, ›anerkannte‹, sprich verbreitete und populäre Wissensbestände, die dazu weitgehend als so von der Obrigkeit gewünschte Interpretationen politischer Ereignisse anzusehen sind, nachzuvollziehen. Dem Projekt »Welt der Kinder« stand dank der Vorarbeit des Georg-Eckert-Instituts mit GEI-Digital die umfassendste Quellensammlung deutschsprachiger Schulbücher ausgewählter Fächer zur Verfügung.[25] Für die Testläufe mit Topic Modeling wurden etwa 3.340 Werke des Erscheinungszeitraumes 1850 bis 1920 ausgewählt, um sie über eine Solr-Plattform webbasiert zugänglich zu machen. Allerdings hatte dieses Vorgehen den Nachteil, dass eine Nachbearbeitung des Korpus nur begrenzt möglich war.

Die sehr umfangreiche Fragestellung nach für Heranwachsende über Schulbücher zugänglichem populären Wissen des 19. Jahrhunderts sollte mit Topic Modeling handhabbar gemacht werden. Zunächst wurden unkontrolliert und ohne gewichtete Worte Topics

[25] http://gei-digital.gei.de/viewer/.

generiert. Darauf wurden verschiedene Versuchsreihen gewählt, um Plausibilität, Kohärenz und Aussagekraft der Topics zu überprüfen. Damit sollten auch Forschungsmeinungen zur inhaltlichen Gewichtung von Schulbüchern des 19. Jahrhunderts überprüft werden können. Dazu wurde unter anderem für die Topic-Überprüfung der Themenkomplex »Krieg« zur Evaluation genutzt, denn Kriege und Kriegsdarstellungen waren ein wichtiges Thema in den ereignisgeschichtlich orientierten Darstellungen der Schulbücher. Hier soll ein konkretes Beispiel aus der Forschung zur vorgeblichen Militarisierung des Deutschen Kaiserreiches herangezogen werden, um die Möglichkeiten, aber auch die Grenzen anzudeuten, die Topic Modeling in einem interdisziplinären Projekt zur historischen Schulbuchforschung an einem großen Korpus bietet.

Kaiser Wilhelm II. forderte ab 1889 verschiedentlich, der neueren Geschichte seit den Befreiungskriegen im Schulunterricht Vorrang vor der antiken Geschichte zu geben. Diese Schwerpunktverschiebung im Unterricht sollte der Herausbildung treuer, auf das Haus Hohenzollern fokussierender Untertanen dienen. Insofern befand er sich in gewissem Gegensatz zu den Gymnasiallehrern, die sich zwar durchaus als Vertreter eines auf das Kaiserhaus bezogenen Nationalismus sahen, aber weiterhin der humanistischen Bildung Vorrang beim Erwerb des Abiturs einräumen wollten. Um diese Frage zu klären, wurden 1890 und 1900 zwei Schulkonferenzen einberufen, die diese Themen unter verschiedenen Gesichtspunkten diskutierten. Am Ende einigte man sich darauf, den kaiserlichen Wünschen entgegenzukommen. Zwar bedeutete dies nicht das Ende humanistischer Themen, doch sollten nun verstärkt die Kriege der Neuzeit thematisiert werden; zumindest nahm dies die bisherige bildungsgeschichtliche Forschung an und folgerte daraus weitgehend eine zunehmende Militarisierung des Schulunterrichts in den Dekaden vor dem Ersten Weltkrieg.[26]

Die Annahme zur Überprüfung, sprich Validierung oder Falsifizierung einer bekannten historischen Forschung bestand nun darin, zum einen die generelle Verbreitung klar dem Komplex »moderne Kriege« und »Antike« zuordenbare Topics (falls vorhanden) zu finden und dann deren Entwicklung über den Zeitverlauf 1890–1900 nachzuvollziehen.[27] Dies wurde für diesen Artikel anhand der Sammlung Geschichtsschulbücher getestet, da auf dem Geschichtsunterricht das besondere Augenmerk des Kaisers lag. Für die Überprüfung, ob der Anteil antiker Geschichte in den Schulbüchern nach 1890 anteilsmäßig gegenüber der sogenannten ›Vaterländischen Geschichte‹ zurückging, wurden jeweils mehrere Topics pro Themenkomplex ausgewählt und ihre Entwicklung über den Zeitverlauf des Untersuchungszeitraums hinweg nachgezeichnet. Um diese Entwicklung nachvollziehen zu können, mussten zuerst die Topics identifiziert werden, die sich dem Themenbereich »Krieg« zuordnen ließen. Neben eindeutigen Topics, in denen Wörter wie »Krieg« oder »Schlacht« explizit auftauchten, fanden sich Topics, die für konkrete kriegerische Konflikten stehen, zum Beispiel dem Peloponnesischen oder dem Dreißigjährigen Krieg. Bei einer Grobsortierung nach den Sammlungen von GEI-Digital (vor 1871 vs. Kaiserreich) zeigte sich folgendes

[26] Vgl. Arbeitsgruppe Lehrer und Krieg 1987; Schubert-Weller 1991, v.a. S. 503–505; Berg 1991.
[27] Die Überprüfung wurde in gemeinsamer Teamarbeit bei verschiedenen Projektsitzungen des Projektes »Welt der Kinder« erarbeitet; daher sei an dieser Stelle explizit Lisa Brunkhorst, Maik Fiedler, Ben Heuwing, Carsten Schnober und Mark Frederik Winter gedankt. Die Topics zu »antiken« und »modernen« Kriegen lassen sich am Einfachsten identifizieren, da hier Schlüsselbegriffe wie Peloponnes, Sparta oder Napoleon auftauchen; sie wurden daher für die kontrastierende Vergleichsstudie herangezogen.

Ergebnis: Bei einer Topic-Liste von 50 Topics ließen sich für die Geschichtsschulbücher vor 1871 drei Topics Napoleonischen Kriegen und vier antiken Kriegen (von insgesamt 11 eindeutig kriegsbezogenen Topics) zuordnen, für das Kaiserreich drei Topics ohne Blick in die Dokumente den Kriegen nach 1789 und zwei antiken Kriegen (ebenfalls 11). Für das Grobraster scheint also auf den ersten Blick die oben angeführte These auch durch das Topic Modeling belegt. Auch bei Probeläufen über verschiedene Subkorpora (katholische und protestantische Schulbücher; Schulbücher für verschiedene Schulformen, zum Beispiel für Mädchenschulen) überwiegen Topics zu den Napoleonischen Kriegen. Erhöht man die Zahl der Topics (Listenlänge), finden sich mehr und spezifischere Topics zu »Krieg« (zu den Ergebnissen vgl. den folgenden Abschnitt). Allerdings scheint es so, dass bei kürzeren Topic-Listen diese zwar ungenauer sind, die zugeordneten Dokumente aber mehr den Themen entsprechen, während bei längeren Topic-Listen die Topics feiner sind, sie aber gelegentlich verschiedene Themen aufgrund Wörter, die in diesen auftauchen, zusammenfassen, obwohl die Themen nicht zusammengehören. Auch gilt es zu beachten, dass sich bei längeren Topic-Listen thematische Doppelungen ergeben können. Somit zeigte sich, dass die Vaterländische Geschichte schon vor 1890 einen größeren Anteil hatte, zusätzlich aber noch an Bedeutung zunahm, allerdings nicht in exorbitantem Ausmaß.

2.2.2 Ergebnisse für die Forschung

Die Ergebnisse zeigen, dass politische Interventionen sich nur zeitverzögert und abgeschwächt in den Schulbüchern wiederfinden. Doch gab es Unterschiede in der Themensetzung in den verschiedenen Regionen und Schulformen des Kaiserreiches. Einzelnen Bundesstaaten des Reiches tradierten durchaus ihre eigenen Narrative der gewaltsamen Reichseinigung »von oben«, die nur langsam und teilweise vom kleindeutsch-preußischen Narrativ einer von allen Deutschen gewollten Vereinigung 1871 überlagert wurden. Konfessionelle Orientierung behielt seine Bedeutung für die Interpretation der Kriege, ebenso gab es durchaus unterschiedlich politisch gefärbte Darstellungen, die linksliberale oder sozialdemokratische Vorstellungen nahestanden.

Wie schon oben angedeutet, sieht man einen Unterschied in der Verteilung kriegsbezogener Topics vor und nach 1871. Berechnet man die Topics für Geschichtswerke nach Kategorien getrennt (Geschlecht, Religion, Schultyp), so findet man im Durchschnitt nach 1871 13,5 kriegsbezogene Topics (bei einer Liste von 50 Topics), und diese sind relativ gleichmäßig verteilt (von 10 bis 16 Topics), während sich vor 1871 ein Durchschnitt von 10,8 Topics ergibt, hier aber die Spannbreite zwischen 6 und 18 Topics je Sammlung liegt. Für die Gymnasialbücher im Spezifischen, und auf sie zielte ja die oben entwickelte Fragestellung, sieht man zwar keine Abnahme antikenspezifischer Topics, eher eine leichte Zunahme (von 6 auf 7), während die Anzahl der Topics mit Bezügen zu Kriegen seit 1789 gleichbleibt (2). Topics, die sammlungsübergreifend gebildet wurden, und Topics, die auf den Einzelsammlungen basieren, zeichnen so ähnliche Entwicklungen ab – mit Ausnahme der Gymnasialschulbücher.

3. Vorgehen

Im Folgenden wird ein neues kontrolliertes Verfahren eines historisch-methodischen Topic Modeling entwickelt, welches die bisher beschriebenen Ansätze kombiniert. Ziele sind, Quellen aufzufinden, die sich nicht in bisher bekannte Muster einordnen lassen, sowie mithilfe von Topics diskursive Argumentationsverläufe in den Quellen nachzuweisen. Das Vorgehen unterscheidet sich in großen Teilen von der klassischen historischen Forschung, insbesondere besitzt ein im Vorfeld formuliertes Studiendesign, welches explizit festlegt, wie die Thesen und Forschungsfragen mit den Mitteln des Topic Modeling beantwortet werden können, eine hohe Bedeutung dafür, welche Ergebnisse und Analysen im Anschluss an dieses Vorgehen möglich sind. Das Vorgehen besteht darin, nicht allein ein automatisiertes Verfahren einzusetzen, welches Topics generiert und Textteilen zuordnet, sondern es ist mit Blick auf die Aussagekraft der Analysen sinnvoller einen iterativen Ansatz zu nutzen, wie er bei Verfahren des Machine Learnings verwendet wird. Es gibt hier bereits Verfahren, die die Texte mit semantischen Analysen für ein Topic Modeling vorbereiten.[28] Dabei kann der Computer auf das Erkennen von Topics trainiert werden. Bevor ein Algorithmus in der Lage ist, selbst Topics vorzuschlagen, sollte er jedoch auch in die Lage versetzt werden, vorgegebene Topics in Texten zu erkennen.

Bisherige Topic Modeling-Verfahren zeigten ihre Stärken darin, auf einem inhaltlich sehr heterogenen Korpus die einzelnen Texte voneinander zu differenzieren. So ließen sich im Projekt »Welt der Kinder« Topics aus Geschichtsschulbüchern relativ schnell identifizieren, obwohl der Corpus aus Geschichts- und Geographiebüchern zusammengestellt ist. Diese Form der Korpusstrukturierung ist für die historischen Wissenschaften dann sinnvoll, wenn aus einem großen Korpus einzelne Teile, die für die Forschung von Interesse sein können, herausgefiltert werden sollen. Wenn sich die historische Forschung wie in den hier vorgestellten Projekten für die Entwicklung und die Bezüge einzelner Topics selbst interessiert, kommt dem Topic Modeling eine ungleich größere Bedeutung zu und ein solches Verfahren zur Filterung des Korpus ist nicht mehr ausreichend. Denn bei einer solchen historischen Textanalyse weisen die untersuchten Themenkomplexe selbst eine größere Vielschichtigkeit auf, als dass sie etwa mit einer Wortwolke allein erfasst werden könnten. Computergestützte Verfahren in Kombination mit einem geeigneten Modell können aber die Komplexität erstmals vollständig sichtbar machen. Mit einer Erhöhung der Granularität der untersuchten Themenkomplexe in ihrer Binnenstruktur, durch eine tiefere Differenzierung in einzelne Topics kann die Komplexität abgebildet und erkannt werden.

Daher erscheint vor dem Einsatz von Algorithmen eine Analyse der Quellentexte hinsichtlich ihrer Topics anhand eines Beispielsets notwendig. Aufbauend auf den Verfahren des Machine Learnings, des Topic Modelings und der qualitativen Inhaltsanalyse wird Folgendes vorgeschlagen:

- Bildung eines Quellenkorpus
- Bildung von Trainings- und Testset
- Erarbeitung von Topics
- Machine Learning
- Überprüfung der Resultate am Testset

[28] Corcoglioniti et al. 2016.

- Historische Analyse der Resultate

Wenn die Überprüfung der Resultate des Machine Learning-Prozesses eine niedrige Aussagekraft ergibt, können die definierten Topics nochmals in Hinblick auf die Resultate verändert werden. Die Schritte zwei bis fünf können iterativ wiederholt werden, bis eine hohe Aussagekraft erreicht ist. Erst auf dieser Grundlage ist es dann sinnvoll zur Auseinandersetzung mit den Resultaten überzugehen.

3.1 Korpus

Zuerst muss für das konkrete Forschungsvorhaben ein Quellenkorpus zusammengestellt bzw. ausgewählt werden; die Kriterien hierfür müssen sich aus der geschichtswissenschaftlichen Fragestellung und dem Kontext der Quellen begründen lassen. Je größer der zu analysierende Quellenkorpus ist, desto eher kommt das hier vorgeschlagene Vorgehen in Frage, da es eine Zeitersparnis in der Analyse bedeuten kann. Das hier vorgeschlagene Verfahren bietet sich auch an, wenn für die Forschung eine Sichtbarmachung der Topics und von Argumentationslinien gewünscht wird. Weiterhin wird mit dieser Form der Analyse zum Korpus ein neuer Forschungsdatensatz generiert, der die Forschung selbst transparenter werden lässt und für weitere Forschungen nachgenutzt werden kann.

Das Besondere des hier vorgestellten Vergleichs ist die unterschiedliche Zusammenstellung der beiden Test-Korpora. »Welt der Kinder« arbeitete mit einem Korpus, an dessen Zusammenstellung die Projektmitarbeiter nicht beteiligt waren, der aber durchaus den Anspruch auf ›Vollständigkeit‹ erhebt. Grundlage ist der inzwischen für die Zeit vor 1918 in seiner Korpusbildung abgeschlossene, schon eingangs erwähnte Bestand von GEI-Digital. Dieses Projekt versammelt OCR-bearbeitet Digitalisate aller deutschsprachigen Geographie- und Geschichtsschulbücher, die in deutschen Bibliotheken gesammelt wurden (plus im Aufbau befindlicher Sammlungsbestände, die aber nicht für das hier besprochene Projekt herangezogen wurden). Im Projekt »Spektralanalyse« wurde hingegen das Korpus kontrolliert zusammengestellt. Für die hier vorgestellten Teile des Projekts »Spektralanalyse« wurde eine komplette Sichtung der Zeitschrift *Annalen der Physik und Chemie* durchgeführt. Diese im 19. Jahrhundert bedeutende Zeitschrift für Physik, nach ihrem Herausgeber auch *Poggendorff's Annalen* genannt, enthielt eine ganze Reihe von wissenschaftlichen Artikeln verschiedener Autoren, die sich mit der Entwicklung der neuen Methode der Spektralanalyse und ihren Folgen befassten. Diese Artikel wurden identifiziert und bildeten zusammen das zu analysierende Korpus.

3.2 Bildung von Trainingsset und Testset

Die Verfahren des Machine Learnings basieren auf dem Trainieren des Algorithmus, was durch wiederholte Durchläufe und iteratives Anpassen von Parametern geschieht. Um festzustellen, wie gut die Ergebnisse sind, die ein Algorithmus mit bestimmten Parametern

erreicht, ist es notwendig zu wissen, wie optimale Ergebnisse auszusehen haben.[29] Aus dem Gesamtkorpus muss also eine Teilmenge – das Trainingsset – zunächst ohne Machine Learning analysiert werden. Dann kann der Algorithmus mit dem Verfahren des Machine Learnings auf die Erkennung der vorgegebenen Muster trainiert werden. Es gibt die Gefahr des Overfittings, die beim Machine Learning-Verfahren vermieden werden sollte. So kann es sein, dass der Algorithmus nach dem Training zwar die Texte im Trainingsset wie gewünscht analysiert, aber für den Rest des Korpus keine guten Ergebnisse liefert, dass also die Auswahl der Trainingssets nicht repräsentativ für den Gesamtkorpus war. Um dieses Verhalten zu vermeiden, können zwei Maßnahmen ergriffen werden. Zunächst kann das Verhalten des trainierten Algorithmus an einem zweiten, im Vorfeld analysierten Korpus – dem Testset, welches kleiner als das Trainingsset sein kann – überprüft werden. Die Aufteilung in ein Trainingsset und ein Testset hilft Verzerrungen, die durch ein einseitiges Trainingsset entstehen können, zu erkennen. Eine zweite Maßnahme, die der einseitigen Auswahl der Texte für das Trainingsset und das Testset vorbeugen soll, ist die zufällige Auswahl der Texte aus dem vorher gebildeten Gesamtkorpus, um eine repräsentative Teilmenge zu erhalten.

3.3 Erarbeitung von Topics: manuell versus maschinell

Für das Trainingsset und das Testset ist eine manuelle Analyse der Texte notwendig. Hierfür bietet sich das Vorgehen der qualitativen Inhaltsanalyse an.[30]

Da es bei der Bearbeitung der Texte oft subjektive Unterschiede gibt, ist es sinnvoll, die Analyse auf verschiedene Personen zu verteilen, wobei ein Text von mehreren Personen bearbeitet werden sollte, damit die Reliabilität überprüft werden kann. Ist die Reliabilität, also die Zuverlässigkeit der Analyse, zu schlecht, ist der eindeutige Nachweis des formulierten Topics schon für die Forscher schwierig, wird sich das Topic daher auch für Machine Learning-Verfahren nur wenig eignen. In einem solchen Falle muss die Formulierung des Topics im Dialog zwischen den Forschern neu gestaltet und dokumentiert werden. Mit dem Erreichen einer hohen Reliabilität ist ein guter Ausgangspunkt für das Machine Learning-Verfahren geschaffen. Vorteil einer computergestützten Textanalyse gegenüber einer manuellen Analyse ist in der Folge zusätzlich die Reproduzierbarkeit der Ergebnisse.

Im Projekt »Spektralanalyse« wurden die Topics anhand von zeitgenössischen Hand- und Lehrbüchern erarbeitet. Dadurch stand zu Beginn der Analyse der Texte bereits eine umfassende Liste mit möglichen Topics zur Verfügung. Dennoch tauchten auch in den analysierten Texten weitere, bisher nicht dokumentierte Topics auf, die dann in die Dokumentation übernommen wurden. Die Erkennung der Topics in den Texten erforderte meist ein großes Textverständnis und setzte auch Hintergrundwissen zum Themenkomplex der Spektralanalyse voraus. Wichtig ist, dass es immer die geisteswissenschaftliche Expertise braucht, um den ›Sinn‹ in Topics zu erkennen, der natürlich je nach Fragestellung

[29] Auch Blei et al. 2003 verwenden einen iterativen Prozess, damit das von ihnen vorgeschlagene Verfahren LDA auf dem Beispielkorpus gute Ergebnisse liefert. Sie versuchen dabei eine möglichst eindeutige Zuordnung herzustellen. Da sie unbekannte Topics in den Texten identifizieren möchten, geben Sie aber keine zu erkennenden Topics vor, was wiederum die Überprüfung der Resultate erschwert.
[30] Mayring 2010.

unterschiedlich ausfallen kann. Ein maschinelles Topic Modeling wird eine so tief gehende Strukturierung vermutlich auch in näherer Zukunft nicht leisten können. Daher wird hier vorgeschlagen, darauf zu verzichten, komplexe Topics zu formulieren. Erhöht man stattdessen die Granularität, können große Textmengen leichter erfasst werden. Die komplexe Analyse verschiebt sich auf die Interpretation der Ergebnisse. Dort können aus den einzelnen einfachen Topics, die im Folgenden »Small Topics« genannt werden, mit dem nötigen Hintergrundwissen und dem Textverständnis komplexere Argumentationen sichtbar gemacht werden.

Die hier und weiter unten vorgeschlagenen Arbeitsschritte führen den manuellen Ansatz und den maschinellen aus den beiden vorgestellten Projekten zusammen. Im Projekt »Welt der Kinder« wurden die Topics, wie eingangs erwähnt und anders als im Projekt »Spektralanalyse«, mit LDA erstellt und im Nachhinein auf Sinnhaftigkeit geprüft. Nach mehreren Durchläufen, die allein von den ›Technikern‹ des Projektes betreut wurden und die vor allem der Optimierung von LDA dienten, wurden den Fachwissenschaftlern mehrere Topic-Listen mit verschiedenen Längen vorgelegt (50 Topics, 100 Topics, 200 Topics) und diese von Letzteren nach ihrem offensichtlichen Aussagewert (Kohärenz) und möglichen Fehlern beurteilt. Diese Evaluationen wurden nun in das Modell zurückgespiegelt und die hier angewandte Variante von LDA schrittweise verbessert.[31] Trotz des großen Corpus (Gesamtkorpus etwa 800.000 Seiten, Untersuchungszeitraum allein 646.171 Seiten) erwies sich die Berechnung auf Satzeinheit als sinnvoll.

Schulbücher des 19. Jahrhunderts stellen spezifische Herausforderungen an Topic Modeling, abgesehen von unterschiedlichen Schreibweisen von Namen (Orts- und Personennamen), einer nicht normierten Orthographie und dem generellen Problem mit Fraktur-Schriften in der OCR-Bearbeitung. So dominieren ohne Filter schnell die zahlreichen Abkürzungen die Topics. Ebenso stellen die kleinteilige Gliederung und Textelemente, also Inhaltsverzeichnisse, (Zwischen-)Überschriften, Tabellen und Bildunterschriften, ein Spezifikum dar. Um dieses Problem handhabbar zu machen, können zwei Kategorien für die Einstellung des Junkfilter gewählt werden: »mindestens 50 Wörter in mindestens 3 Sätzen pro Seite bei möglichst geringem Anteil von Sonderzeichen pro Wort« und »mindestens fünf Worteinheiten pro Satz«; beide Einstellungen verhindern, dass die Topics aus zu kleinen Satzfragmenten oder -einheiten berechnet werden, die die Ergebnisse verfälschen können. Werden diese eingeschaltet, reduziert sich der Korpus auf 645.141 Seiten. Der Filter »Medientyp« hat sich nur begrenzt als geeignet erwiesen, um Wörter in Tabellen, Bildunterschriften, etc. aus der Topic-Berechnung zu filtern, da man sich hier auf die Qualität der von externen Dienstleistern ausgezeichneten Metadaten verlassen muss, diese aber nicht einheitlich ist.

Ein weiteres Problem, die Verbindung oder Vermischung von zwei Themenkomplexen in einem Topic, lässt sich allerdings durch das Trainieren des Topic Models einigermaßen kontrollieren. Allerdings stellen hier der Umgang mit allgemeinen Begriffen wie »Schlacht« und »Krieg« oder Personennamen wie »Friedrich« und »Ludwig« weiterhin eine bisher im Projekt ungelöste Aufgabe dar, da diese in den Schulbüchern epochenübergreifend (sprich für Mittelalter wie Neuzeit) gebraucht werden. Hier wären weitere Versuche mit gewichteten

[31] Schnober / Gurevych 2015; Heuwing / Womser-Hacker 2015; Heuwing et al. 2015.

Topics, Named Entity Recognition oder Ontologien hilfreich, die bisher nur ansatzweise im Projekt »Welt der Kinder« getestet wurden.

Die Kriegstopics sind erstaunlich eindeutig, wenn auch gelegentlich nicht immer auf ersten Blick erkennbar ist, welcher preußische Krieg z.B. um Schlesien oder gegen Frankreich gemeint ist. Hier kann nur anhand einer Überprüfung der damit verbundenen Dokumente eine Klarstellung erreicht werden. Die Solr-Webplattform erlaubt es im Projekt »Welt der Kinder« sich die Topic-Entwicklung über den Zeitverlauf hinweg in Kombination mit Metadaten (unter anderem Sammlungen, Schultyp/Schulform, Bildungsstufe, Region, Konfession, Geschlecht) anzeigen zu lassen. Mit der Funktion »Gruppenvergleich« kann man sich die Zahlen in einer Excel-Tabelle ausgeben lassen. Die Funktion ermöglicht den Vergleich zwischen zwei Kategorien; die Veränderungen werden sowohl in absoluten wie Prozentzahlen angezeigt und die größten negativen und positiven Abweichungen farblich markiert.

3.4 Machine Learning

Der Machine Learning-Prozess wird auf dem gebildeten Trainingsset gestartet. Ziel ist es, die veränderlichen Parameter des Algorithmus solange sukzessive zu verändern, bis die Resultate des Prozesses den vorgegebenen Kriterien möglichst nahekommen.

Die Erfahrungen aus den geschilderten Projekten zeigen, dass sich zahlreiche Topics anhand von bestimmten Schlüsselworten nachweisen lassen. Damit können die üblichen Topic Model-Verfahren, die auf Wortebene ansetzen, an dieser Stelle eingesetzt werden. Wie die Erfahrungen aus der Entwicklung von Werkzeugen für die Korpuslinguistik nahelegen haben, gibt es eine Reihe weiterer Schritte, die sinnvoll sind, um dem Algorithmus das Topic Modeling zu erleichtern. Die bekannten Verfahren belegen, dass etwa die Nutzung von Lemmatisierung die Erkennungsquote von Schlüsselwörtern erheblich verbessert. Erfreulicherweise wurde in der Korpuslinguistik schon eine ganze Reihe von Werkzeugen entwickelt, die hier als Technik eingesetzt werden können.[32]

Um den Argumentationsverlauf auch innerhalb der Quellentexte sichtbar machen zu können, ist eine Aufteilung der Quellentexte notwendig. Hierfür bietet sich eine Aufteilung nach Absätzen an, die oft eine argumentative Einheit eines Textes darstellen. Werden die bisherigen Topic Model-Verfahren eingesetzt, so besteht für den Algorithmus ein Topic aus einer Sammlung von Wörtern, die dort jeweils mit einer spezifischen Wahrscheinlichkeit vorkommen. Die Übereinstimmung, die ein so verstandenes Topic mit den vorhanden Wörtern einer Analyseeinheit, also eines Absatzes, erzielt, kann dann als Quote (zwischen 0 und 1) ausgedrückt werden. Zu beachten ist, dass in einem Absatz mit unterschiedlichen Übereinstimmungsquoten auch mehrere Topics vorhanden sein können. Dieser Umstand deutet auf die inhaltliche Nähe solcher Topics hin und kann später in der Analyse genutzt werden.

[32] An dieser Stelle soll nur auf bestehende Services wie etwa das bei Clarin-D angesiedelte WebLicht oder das Textanalysewerkzeug GATE (general architecture for text engineering) sowie die zahlreichen R-Module zum Text-Mining verwiesen werden.

3.5 Überprüfung der Resultate am Testset

Nach dem Vorbereiten des Algorithmus am Trainingsset kann dieser mit den bestimmten Parametern wiederum überprüft werden. Sind sowohl Trainingsset als auch Testset repräsentativ aus dem Korpus ausgewählt worden, sollte auf dem Testset eine ähnlich hohe Übereinstimmung von erkannten mit manuell zugeordneten Topics erreicht werden wie auf dem ursprünglichen Set. Weicht die Übereinstimmung mit beiden Sets deutlich voneinander ab, wurde der Algorithmus zu einseitig trainiert. Entweder wurden in diesem Fall also die Sets nicht repräsentativ ausgewählt oder sind zu klein, die Topics wurden zu unscharf formuliert oder der Algorithmus müsste nochmals überarbeitet werden. Wenn die Resultate zu dem gewünschten Ergebnis führen, kann zur weiteren Analyse übergegangen werden. Die Übereinstimmungsquoten, man spricht hierbei von Recall und Precision, also das Verhältnis von korrekt zugeordneten Topics zu falschen Zuordnungen, geben dabei an, wie aussagekräftig die Resultate des Topic Model-Verfahrens sind.

Wenn keine Überprüfung der Resultate mit den von der Forschung im Vorfeld gemachten Vorgaben stattfindet, besteht generell das Risiko, dass die Fragestellung im Nachhinein zu sehr an die Resultate und damit auch an die mathematischen Modelle des benutzten Verfahrens angepasst wird. Eine Unkenntnis gegenüber den technischen Bedingungen ohne Möglichkeit der Überprüfung kann nur schwer unter Fachhistorikern Akzeptanz erfahren. Damit blieben aber bestimmte Fragestellungen zu Unrecht aus der Forschung ausgeschlossen. Nur mit der Reflexion und Prüfung der Ergebnisse und einem erhöhten methodischen Verständnis lässt sich dies vermeiden.

3.6 Analyse/Resultate

Das Folgende ist ein Vorschlag zur Analyse der aus dem oben beschriebenen Verfahren gewonnenen Resultate. Die erkannten Topics in den Quellentexten geben Aufschluss über den Aufbau der einzelnen Texte. So wird von den Topics eines Textes ein Netzwerk aufgespannt, in welchem die Topics die Knoten bilden und die Nähe der Topics zueinander innerhalb der Quelle anzeigen.[33] Für verschiedene Texte ergeben sich unterschiedliche Netzwerke. Wenn die untersuchten Themenkomplexe auf ihre historische Entwicklung etwa in zeitlich aufeinander folgenden Texten untersucht werden, kann sich diese entweder in einer Veränderlichkeit der entsprechenden Netzwerke von festen, unveränderlichen »Small Topics«, die sich in allen Texten wiederfinden, oder in einer Veränderlichkeit der Topic-Observablen bei einem gleichbleibendem Netzwerk äußern. Eine Änderung von Topic-Observablen bedeutet dabei, dass zwar das gleiche Topic in verschiedenen Texten erkannt wird, es aber anders oder neu formuliert wird, sich also etwa die Zusammensetzung der damit einhergehende Wortgruppe für die verschiedenen Quellentexte unterscheidet. Dabei muss beachtet werden, dass bei der Bildung von Topics über Wortgruppen die Reihenfolge der Wörter im Satz unberücksichtigt

[33] Es lassen sich verschiedene Bestimmungsverfahren für die Nähe von Topics denken. Der Abstand zweier Topics könnte etwa die Zahl der Abschnitte oder Absätze sein, die sich zwischen den Absätzen der Topics befinden. Man könnte aber auch die Zahl der Topics, welche sich zwischen den beiden zu betrachtenden liegen, als Kennzahl nehmen.

bleibt, also nur der Satzinhalt erkannt wird, aber nicht das semantische Gebilde. Gibt es eine Veränderlichkeit sowohl vom Topic-Netzwerk als auch der Topic-Observablen, ist die Aussagekraft einer solchen Beobachtung nur gering. In einem solchen Fall verändern sich sowohl die Begriffe als auch das Wissensnetzwerk. Damit ist der Korpus zu heterogen und die Granularität der Topics ist zu groß oder klein gewählt.

In den Texten aus dem Projekt »Spektralanalyse« finden sich wiederkehrende Argumentationen.[34] So sind es besonders die Texte, die auf die astronomische Bedeutung der Spektralanalyse eingehen und speziell auf die Möglichkeit mit der neuen Methode die atomaren Bestandteile der Sonne zu identifizieren, die auch betonen, dass die Spektralanalyse eine beeindruckende Erfindung sei. Diese Form der Argumentation, die wissenschaftliche Fortschritte mit einer Wertung versieht, findet sich auch später in anderen Publikationen wieder. Von den vielen neuen Möglichkeiten, die die Spektralanalyse der Wissenschaft bot, kann die hier vollzogene Analyse also aufzeigen, dass die Identifizierung der solaren Elemente entscheidend für die breite Wahrnehmung in der Öffentlichkeit und Wissenschaft war.

Bei der Auswertung des Kriegs-Themas im Projekt »Welt der Kinder« erweist sich, dass Kriegs-Topics zwar relevant, aber nicht dominant sind. Zu den Wichtigsten zählen, wie in diesem Beitrag skizziert, Topics zu den Napoleonischen Kriegen, dem Siebenjährigen Krieg, den Perserkriegen und dem Peloponnesischen Krieg. Eine Schlussfolgerung hieraus könnte lauten, dass Kinder im Untersuchungszeitraum Europa (und die Welt) in den Geschichtsschulbüchern eher über Schlachten und Kriege als über Handelsbeziehungen und Erfindungen kennenlernten, da sich für diese auf den ersten Blick weniger eindeutige Topics finden.

4. Resümee und Vorschläge für die Weiterentwicklung

Das vorgeschlagene Verfahren unterstützt die Forschung, indem es zur Schärfung und zum besseren Verständnis der selbst formulierten Topics beiträgt. Hierdurch können auch bisher nicht berücksichtigte Topics erkannt werden. Die Struktur der Texte kann durch die absatzgenaue Identifikation von Topics offenbar gemacht werden. Dadurch könnte zum Beispiel in nachfolgenden Untersuchungen der Textaufbau verglichen werden, aber auch nach Plagiaten, Zitaten und Abschriften gesucht werden. Dieses Verfahren versteht sich so auch als Weiterentwicklung bestehender Ansätze. Peter Andorfer, zum Beispiel, geht unserer Meinung nach nicht weit genug und bietet noch keine neuen Ansätze für die zukünftige Forschung. Die von ihm präsentierten Verfahren wurden im Rahmen der hier vorgestellten Forschungen ebenfalls getestet und führen, wie beschrieben, zu folgenden Herausforderungen: eine rein manuelle Topicvergabe besitzt nicht die Reproduzierbarket, Transparenz und Effizienz wie eine maschinelle Verarbeitung, umgekehrt kann eine rein maschinelle Verarbeitung keine komplexen Themen erfassen wie die manuelle, was die Bewertung der Ergebnisse

[34] Mithilfe von Matrix-Clustering-Verfahren lässt sich die Nähe von Topics bestimmen und anschließend können die Argumentationsmuster von Texten, hier Topic Graphen genannt, miteinander verglichen werden. Historisch interessant ist dabei, wie sich Topic Graphen über die Zeit ändern oder auch die Texte zu identifizieren, wo Narrationen erstmals auftauchen.

erschwert, und stellt erhöhte Anforderungen an die Qualitätssicherung. Um diesen Problemen zu begegnen wurde in dem hier gemachten Vorschlag darauf gesetzt, beide Verfahren miteinander zu kombinieren, indem iterativ mit Trainings- und Testsets gearbeitet wird. Hiermit und mit der Konzentration auf »Small Topics« lassen sich transparent und reproduzierbar sinnvolle Ergebnisse erreichen. Folgt man den Wettlauferschen Zuordnungen lassen sich die hier vorgestellten Verfahren dem Bereich Semantischer Technologien zuordnen. Die Verfahren sind, wie beschrieben, auf verschiedene Textsorten anwendbar und können, entsprechend aufbereitet, für Semantic Web-Anwendungen nachgenutzt werden.

Beim momentanen Stand der Ausbildung von Historikern und Informatikern ist es unerlässlich, sich vor dem Beginn eines Projektes klar auf Arbeitsweise, Vorgehensweise und Ziele zu verständigen. Wir unterscheiden Arbeits- und Vorgehensweise, da erstere mehr die Absprachen über Werkzeuge und Methoden meint, letztere die einzelnen Arbeitsschritte betrifft. Topics allein liefern noch keine Antwort, sie bedürfen immer der Interpretation durch die jeweilige Fachdisziplin. Wie die oben angeführten Beispiele belegen, ist für die wissenschaftliche Interpretation von Topics fachliche Expertise unersetzlich. Zwar verändern sich Fragestellung und Erkenntnisinteresse je nach disziplinärer Zuordnung, doch zeigen die Beispiele der Vorgehensweise Topic Modeling zu verbessern klar, dass mindestens Kontextwissen erforderlich ist, um Topics aussagekräftiger zu machen. Für die Erkennung von Topics braucht man interpretationsfähige Schlüsselwörter. Diese können Fachbegriffe sein, aber auch mehr oder weniger bekannte Personen, Dynastien oder Territorien. Die aufgelistete Sekundärliteratur hat sich breit mit diesem Thema beschäftigt und vielfältige Ansätze getestet, so gibt es schon Versuche mit automatisiertem Labeling von Topics.[35]

Die Unterschiede in den verschiedenen Quellentypen bedingen auch Unterschiede in der Analyse. Während es zum Beispiel für die Schulbuchanalyse auf einer ersten Ebene durchaus aussagekräftig ist, allein die Verteilung von Themen zu extrapolieren, braucht ein wissenschaftshistorischer Zugang Modelle, die stärker Veränderungen und Trends widerspiegeln. Allerdings zeigen diese unterschiedlichen Quellengattungen sowie die zahlreichen, unterschiedlichen Beispiele in der Sekundär- und ›grauen‹ Literatur, dass prinzipiell jede Textsorte geeignet für Topic Modeling ist, solange man die jeweiligen Spezifika in der Modellbildung berücksichtigt. Es handelt sich um eine neue Technik, eine neue Methode, doch der grundlegende Erkenntnisweg, eine Themenzuordnung und ein Topic Modeling über die Inhalte der Quellen zu erreichen, bleibt.

Abschließend bleibt festzuhalten, dass Topic Modeling durchaus geeignet ist für eine breitere Nutzung in den Geistes- und Geschichtswissenschaften, doch sollte man die hierfür nötigen Mühen, den Aufwand und die Kosten nicht unterschätzen. Das gilt auch für den hier gemachten Vorschlag, der die Vorteile der Ansätze beider Projekte zusammenführt. Wie die oben angeführten Beispiele zeigen, bedürfen beide Ansätze, kontrollierte Topicbildung anhand ausgewählter Dokumentausschnitte sowie unkontrollierte Verfahren bei großen Korpora, aufeinander aufbauender Arbeitsschritte, wobei die Qualität der Rohdaten in den seltensten Fällen von den Forschern beeinflusst noch nachträglich grundlegend verbessert werden kann.

[35] Vgl. zum Beispiel die Arbeiten von Jey Han Lau: Lau et al. 2011; Aletras et al. 2014.

Beide Ansätze unterscheiden sich auch in ihrer Zielstellung. Während es im Projekt »Welt der Kinder« vornehmlich darum ging, große Textmengen zu filtern und nach Überraschungen zu suchen, war es dem Projekt »Spektralanalyse« eher daran gelegen, semantische Entwicklungen nachzuvollziehen. Beide Zielstellungen können als erfüllt betrachtet werden. Ohne eine historische Fragestellung und eine Interpretation durch den Fachwissenschaftler allerdings wären die Befunde an sich wenig aussagekräftig.

Bibliographische Angaben

Nikolaos Aletras / Timothy Baldwin / Jey Han Lau / Mark Stevenson: Representing Topics Labels for Exploring Digital Libraries. PDF. [online] In: 2014 IEEE/ACM Joint Conference on Digital Libraries. Proceedings. (JCDL 2014, London, 8.-12.9.2014). Piscataway, NJ. 2014, S. 239-248. [Nachweis im GBV]

Peter Andorfer: Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich. In: Zeitschrift für digitale Geisteswissenschaften. 2017. DOI: 10.17175/2017_002

Lehrer helfen siegen. Kriegspädagogik im Kaiserreich mit Beiträgen zur NS-Kriegspädagogik. Hg. von Arbeitsgruppe »Lehrer und Krieg«. Berlin 1987. [Nachweis im GBV]

Christa Berg: Einleitung zu Militär und Militarisierung. In: Handbuch der deutschen Bildungsgeschichte. Hg. von ders. Band IV: 1870–1918. München 1991, S. 501–503. [Nachweis im GBV]

David M. Blei / Andrew Y. Ng / Michael I. Jordan: Latent Dirichlet allocation. PDF. [online] In: Journal of Machine Learning Research 3 (2003), S. 993–1022. [online]

David M. Blei (2012a): Probabilistic Topic Models. DOI: 10.1145/2133806.2133826 In: Communications of the ACM 55 (2012), H. 4, S. 77-84. [online]

David M. Blei (2012b): Topic Modeling and Digital Humanities. [online] In: Journal of Digital Humanities 2 (2012), H. 1, S. 8–11. [online]

Francesco Corcoglioniti / Marco Rospocher / Alessio Palmero Aprosio: Frame-based Ontology Population with Pikes. PDF. [online] In: IEEE Transactions on Knowledge and Data Engineering 28 (2016), H. 12, S. 3261-3275. [Nachweis im GBV]

Distant Readings. Topologies of German Culture in the Long Nineteenth Century. Hg. von Matt Erlin / Lynne Tatlock. Rochester, NY 2014. [Nachweis im GBV]

Martin Fechner: Kommunikation von Wissenschaft in der Neuzeit: Vom Labor in die Öffentlichkeit. Eine Untersuchung zum Wandel des Publikationsverhaltens erfolgreicher Wissenschaft am Beispiel der Spektralanalyse und des Lasers. Berlin 2016. (= MPIWG Preprint Series, 477) [Nachweis im GBV]

Werner Früh: Inhaltsanalyse. Theorie und Praxis. 6., überarbeitete Auflage. Konstanz 2006. (= UTB, 2501) [Nachweis im GBV]

Ben Heuwing / Christa Womser-Hacker: Zwischen Beobachtung und Partizipation – nutzerzentrierte Methoden für eine Bedarfsanalyse in der digitalen Geschichtswissenschaft. In: Information – Wissenschaft & Praxis 66 (2015), H. 5–6, S. 335–344. DOI: 10.1515/iwp-2015-005810.1515/iwp-2015-0058

Ben Heuwing / Thomas Mandl / Christa Womser-Hacker: Projekt Welt der Kinder. Überblick über die informationswissenschaftliche Bedarfsanalyse. In: HiER 2015 – Proceedings des 9. Hildesheimer Evaluierungs- und Retrievalworkshop. Hrsg. von Stefanie Elbeshausen / Gertrud Faaß / Joachim Griesbaum / Ben Heuwing / Julia Jürgens. (HIER 2015, Hildesheim, 9.-10.7.2015). Hildesheim 2015, S. 11–18. DOI: 10.18442/337

Journal of Digital Humanities: JDH 2 (2012), H. 1. Hg. von Roy Rosenzweig Center for History and New Media. Fairfax, VA. 2012. ISSN 2165-6673. [online]

Matthew Kirschenbaum: The Remaking of Reading: Data Mining and the Digital Humanities. (NGDM: 2007, Baltimore, MD, 10.-12.10.2007). Baltimore 2007. PDF. [online]

Guido Koller: Geschichte digital. Historische Welten vermessen. Stuttgart 2016. [Nachweis im GBV]

Jey Han Lau / Karl Grieser / David Newman / Timothy Baldwin: Automatic Labelling of Topics. PDF. [online] In: 9th Annual Meeting of the Association for Computational Linguistics 2011. ((ACL: 49, Portland, OR, 19.-24.06.2011). Vol. 1: Human Language Technologies. Red Hook, NY. 2011, S. 1536–1545. PDF. [online] [Nachweis im GBV]

Philipp Mayring: Qualitative Inhaltsanalyse. 11., aktualisierte und überarbeitete Auflage. Weinheim 2010. [Nachweis im GBV]

Thomas Meyer: Digitale Werkzeuge. In: Clio Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften. Hg. von Laura Busse / Wilfried Enderle / Rüdiger Hohls / Gregor Horstkemper / Thomas Meyer / Jens Prellwitz / Annette Schuhmann. Berlin 2016, S. 1–42. PDF. [online]

David Mimno: Computational Historiography: Data Mining in a Century of Classics Journals. DOI: 10.1145/2160165.2160168 In: ACM Journal of Computing in Cultural Heritage 5 (2012), H. 1, S. 1–19. PDF. [online]

Franco Moretti: Graphs, Maps, Trees: Abstract Models for Literary Theory. London, New York 2005. [Nachweis im GBV]

Charles Tilly: Computers in Historical Analysis. In: Computers and the Humanities 7 (1973), 6, S. 323–335. [Nachweis im GBV]

Wolfgang Schmale: Digitale Geschichtswissenschaft. Wien u.a. 2010. [Nachweis im GBV]

Benjamin M. Schmidt: Words Alone: Dismantling Topic Models in the Humanities. [online] In: Journal of Digital Humanities 2 (2012), H. 1, S. 49–65. [online]

Carsten Schnober / Iryna Gurevych: Combining Topic Models for Corpus Exploration: Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project. DOI: 10.1145/2809936.2809939 In: Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. (TM 2015, Melbourne, VIC, 19-23.10.2015). New York 2015, S. 11–20. [online]

Christoph Schubert-Weller: Vormilitärische Jugenderziehung. In: Handbuch der deutschen Bildungsgeschichte. Hg. von Christa Berg. Band IV: 1870–1918. München 1991, S. 503–515. [Nachweis im GBV]

Ted Underwood: Topic Modeling made just simple enough. In: The Stone and the Shell. Using large digital libraries to advance literary history. Blogeintrag vom 07. April 2012. [online]

Jörg Wettlaufer: Neue Erkenntnisse durch digitalisierte Geschichtswissenschaft(en)? Zur hermeneutischen Reichweite aktueller digitaler Methoden in informationszentrierten Fächern. In: Zeitschrift für digitale Geisteswissenschaften. 2016. DOI: 10.17175/2016_011