Netzwerkmodellierung mit NER und NEL. Schulbuchforschung an der Schnittstelle zur Infrastruktur

Dombrowski, Fabian
fabian.dombrowski@gei.de
Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut, Deutschland

Klaes, Sebastian
klaes@gei.de
Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut, Deutschland

Leitgeb, Johannes
johannes.leitgeb@stud-mail.uni-wuerzburg.de
Zentrum für Philologie und Digitalität (ZPD), Universität Würzburg, Deutschland

Reul, Christian
christian.reul@uni-wuerzburg.de
Zentrum für Philologie und Digitalität (ZPD), Universität Würzburg, Deutschland
ORCID: 0000-0002-1776-1469

Forschungsgegenstand

Schulbücher kommunizieren einen Wissenskanon, den eine Gesellschaft als grundlegend erachtet. Der historischen Forschung geben sie jedoch nicht nur Einblicke in die vermittelten Inhalte, sondern auch in die Sinnwelten und Werte einer Zeit. Mit der Einführung der Schulpflicht im langen 19. Jahrhundert erlangte das Schulbuch – bereits vorher ein zentrales Massenmedium europäischer Gesellschaften – eine noch prominentere Rolle. Es verbreitete die Geschichtsbilder der entstehenden Nationalismen, stiftete Identität und stabilisierte die sozialen Zusammenhänge im Staat. Dieses Ziel erreicht es aber nicht allein durch seine Inhalte, sondern darüber hinaus durch deren zeitgenössische pädagogische Aufbereitung und flächendeckende Verbreitung. Zu diesem Zweck wurden in Deutschland zwischen 1800 und 1945 nach aktuellem Stand etwa 2.200 Geschichtsschulbücher veröffentlicht (Fuchs et al 2014; Jacobmeyer 2011).

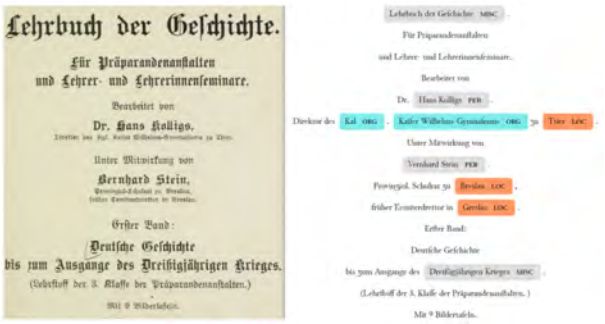
Trotz dieser bedeutenden Rolle bleibt vieles über die Entstehung und Produktion von Schulbüchern unklar. Die Forschung neigt dazu, sich hauptsächlich auf die Analyse der Inhalte zu konzentrieren und dabei die Netzwerke der beteiligten Personen, Verlage und staatlichen Institutionen zu vernachlässigen (Otto 2018). Fragen wie „Wer schrieb die Schulbücher?“, „Wer stellte sie her und verbreitete sie?“ und „Welche Netzwerke waren daran beteiligt?“ wurden bisher nur für Einzelbeispiele beantwortet (z.B. Keiderling 2002; Kreusch 2008), eine umfassende und strukturierte Annäherung blieb aus.

Methoden und Daten

Das vorzustellende Poster zeigt einen Weg auf, diese Lücken zu füllen: Die Modellierung eines Netzwerkes der Schulbuchproduktion im deutschen Kaiserreich mit den Methoden der soziohistorischen Netzwerkanalyse (Ahnert 2021). Grundlage hierfür sind bibliographische Daten aus dem Bibliothekskatalog des Leibniz-Instituts für Bildungsmedien | Georg-Eckert-Institut (GEI) sowie digitalisierte Schulbuchsammlungen aus GEI-Digital. [1] Zwei zentrale Technologien, die dabei zum Einsatz kommen, sind die Named Entity Recognition (NER) und das Named Entity Linking (NEL). Die primären Datenquellen für diese Untersuchung umfassen bibliographische Datensätze des GEI und digitalisierte Geschichtsschulbücher aus der Zeit von 1871 bis 1918. Dieses Korpus umfasst 1.751 Schulbücher mit über 300.000 Seiten. Aktuell werden die Schulbuch-Digitalisate und Volltexte aus dem Kaiserreich durch das Zentrum für Philologie und Digitalität (ZPD) [2] der Universität Würzburg aufbereitet. Im Vordergrund steht dabei die Extraktion von maschinenverarbeitbaren Volltexten aus den Digitalisaten mittels Verfahren der automatischen Texterkennung, wobei ausschließlich Open Source Werkzeuge, hauptsächlich OCR4all [3] (Reul et al. 2019; Nöth et al. 2024) in Kombination mit OCR-D [4] (Neudecker et al. 2019), zum Einsatz kommen. Die so vorverarbeiteten Texte werden im Schritt der Named Entity Recognition (NER) automatisch annotiert. Von Interesse sind dabei insbesondere die auf dem Titel sowie im Vorwort explizit genannten Personennamen sowie die davon abhängigen Koreferenzen (Orte, organisatorische Entitäten). Um die Nachnutzbarkeit der auf diese Weise extrahierten Entitäten zu gewährleisten, werden im anschließenden Named Entity Linking (NEL) zugehörige Normdaten identifiziert, die die Basis für eine ontologische Netzwerkdarstellung bilden (Menzel et al. 2021; Wintergrün 2019).

So bestätigt sich bei ersten Auswertungen, dass Personen und Orte den größten Teil der Entitäten ausmachen. Nicht alle identifizierten Personen sind dabei jedoch für die Netzwerkanalyse relevant ("Friedrich der Große", "Caesar"). Auch gilt es zu beachten, dass im Zuge der NER ein zweistufiger Entscheidungsprozess stattfindet: Zunächst geschieht die grundsätzliche Erkennung der Named Entity. In diesem kann es wie auch im anschließenden zweiten Schritt der Klassifizierung von Entitäten in die Kategorien PER, LOC oder ORG zu Falschannahmen kommen. Insbesondere die Genauigkeit der Ergebnisse im zweiten

Schritt ist für das anschließende NEL von Bedeutung. Dennoch zeigen erste Tests eine erstaunlich hohe Präzision für die Identifikation der Entitäten. Das FlairNLP Framework (Akbik et. al., 2019) liefert beispielsweise für eine Goldstandard-Menge von 79 Entitäten eine Precision von 0,94. Diese ist von leicht abgeschwächter Aussagekraft, da auch nicht vollständige Entitäten als Treffer angesehen wurden, solange ein Teil richtig identifiziert wurde.



Links eine Abbildung einer Schulbuch-Titelei, rechts die passend extrahierten Entitäten.

Die beiden eingesetzten Technologien NER und NEL helfen, die bibliographischen Personendaten aus dem Bibliothekskatalog im Sinne der Netzwerkmodellierung zu erweitern und anzureichern ; dies wiederum erlaubt die Modellierung der Netzwerke zwischen u.a. Autor:innen, Herausgeber:innen, Verlagen und staatlichen Institutionen. Darüber hinaus können diese Daten aus anderen Projekten am GEI erweitert werden. So z. B. durch die Ergebnisse des Projektes SchulbuchEvolution . [5]

Ergebnisse

Das Plakat soll zum einen diesen Workflow verdeutlichen, zum anderen aber auch sowohl anhand einzelner Beispiele wie auch systematisch aufzeigen, dass die Probleme bei der Ermittlung des Netzwerkes nicht nur die Frage nach einem historischen Phänomen betreffen, sondern zugleich auch Herausforderungen für Infrastrukturoperationen herausstellen (Hertling u. Klaes 2022). Es muss z.B. bedacht werden, dass Verfahren der Sozialen Netzwerkanalyse (SNA) oft nicht auf die Inkonsistenz und Unvollständigkeit historischer Daten eingerichtet sind, welche aus der grundsätzlichen Überlieferungssituation erwächst. Normalerweise streben Netzwerkanalysen eine Vollständigkeit an, wie sie aber für Historiker:innen im Grunde nie verfügbar ist. Sie können maximal einen „Sättigungsgrad“ (Ferihumer, nach Dombrowski u. Haslam 2024) der Daten realisieren, um repräsentative Aussagen treffen zu können. Ein solches Projekt nutzt die Daten von Infrastrukturen nach, während es gleichzeitig auf diese zurückwirkt. Und zwar nicht nur weil Verknüpfungen zwischen Bibliothekskatalog und Normdatei geschaffen werden, die die Qualität einer oder mehrere Datenbanken verbessert, sondern auch weil das Bild der Sammlungen selbst klarer wird durch ein Verständnis der bestehenden Lücken und Inkonsistenzen aus Perspektive eines Netzwerkes. Letztlich entscheidet sich an diesen Leerstellen, wie weit Netzwerkanalysen im Speziellen und digitalen Methoden im Allgemeinen DH-Projekte in der Schulbuchforschung tragen.

Fußnoten

1. https://gei-digital.gei.de/ (zuletzt besucht 4.12.2024).
2. https://www.uni-wuerzburg.de/zpd (zuletzt besucht 4.12.2024).
3. https://www.ocr4all.org (zuletzt besucht 4.12.2024).
4. https://ocr-d.de (zuletzt besucht 4.12.2024).
5. https://www.gei.de/forschung/projekte/die-evolution-des-schulbuchs (zuletzt besucht 4.12.2024).

Bibliographie

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019, „FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP.“ Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) , 54–59. https://doi.org/10.18653/v1/n19-4010 .

Ahnert, Ruth, Sebastian E. Ahnert, Catherine Nicole Coleman, und Scott B. Weingart. 2021. The Network Turn. Changing Perspectives in the Humanities . Elements in Publishing in the Humanities. Cambridge: Cambridge University Press.

Dombrowski, Fabian, und Rachael Haslam. 2024. „Conference Report: Our Interlocked Universe. Sociohistorical Network Analysis: Methods, Applications, and New Directions“. H-Soz-Kult , Juli. https://www.hsozkult.de/conferencereport/id/fdkn-145347 .

Fuchs, Eckhardt, Inga Niehaus, und Almut Stoletzki, Hrsg. 2014. Das Schulbuch in der Forschung: Analysen und Empfehlungen für die Bildungspraxis . 1. Aufl. Eckert. Expertise 4. Göttingen: V&R unipress.

Hertling, Anke, und Sebastian Klaes. 2022. „Volltexte für die Forschung. OCR partizipativ, iterativ und on Demand“. o-bib. Das offene Bibliotheksjournal 9 (3): 1–11.

Jacobmeyer, Wolfgang. 2011. Das deutsche Schulgeschichtsbuch 1700–1945. Die erste Epoche seiner Gattungsgeschichte im Spiegel der Vorworte . 3 Bde. Geschichtskultur und historisches Lernen. Berlin: Lit Verlag.

Jacobmeyer, Wolfgang. 2023. Hungerleider werden Bildungsbürger. Preußische Gymnasiallehrer 1820 – 1914. Profile einer Profession . Geschichtsdidaktik diskursiv. Public History und Historisches Denken. Berlin: Peter Lang GmbH, Internationaler Verlag der Wissenschaften.

Keiderling, Thomas. 2002. „Der Schulbuchverleger und sein Autor. Zu Spezialisierungs- und Professionalisierungstendenzen im 19. und frühen 20.

Jahrhundert“. In Die Rolle von Schulbüchern für Identifikationsprozesse in historischer Perspektive , herausgegeben von Heinz-Werner Wollersheim, Hans-Martin Moderow, und Cathrin Friedrich. Leipziger Studien zur Erforschung von regionenbezogenen Identifikationsprozessen 5.

Kreuch, Julia. 2008. Der Verlag der Buchhandlung des Waisenhauses als Schulbuchverlag zwischen 1830 und 1918 die erfolgreichen Geografie- und Geschichtslehrbücher und ihre Autoren . 2008. Hallesche Forschungen 25. Tübingen: Niemeyer Harrassowitz.

Menzel, Sina, Hannes Schnaitter, Josefine Zinck, Vivien Petras, Clemens Neudecker, Kai Labusch, Elena Leitner, und Georg Rehm. 2021. „Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten“. In Qualität in der Inhaltserschließung , herausgegeben von Michael Franke-Maier, Anna Kasprzik, Andreas Ledl, und Hans Schürmann, 229–58. Bibliotheks- und Informationspraxis 70. Berlin: De Gruyter.

Neudecker, Clemens, Konstantin Baierer, Maria Federbusch, Matthias Boenig, Kay-Michael Würzner, Volker Hartmann and Elisa Herrmann. 2019. OCR-D: An End-to-End Open Source OCR Framework for Historical Printed Documents. In Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage , 53–58. DATeCH2019. ACM, 8–10 May 2019.

Nieländer, Maret, und Ernesto William De Luca, Hrsg. 2018. Digital Humanities in der Internationalen Schulbuchforschung. Forschungsinfrastrukturen und Projekte . Vandenhoeck & Ruprecht.

Nöth, Maximilian, Herbert Baier and Christian Reul. 2024. OCR4all 1.0 – Flexible open-source OCR/HTR based on various single-step Solutions. 16th IAPR International Workshop On Document Analysis Systems (DAS) (September).

Otto, Marcus. 2018. „Textbook Authors, Authorship, and Author Function“. In The Palgrave Handbook of Textbook Studies , herausgegeben von Eckhardt Fuchs und Annekatrin Bock, 95–102. New York: Palgrave Macmillan.

Reul, Christian, Dennis Christ, Alexander Hartelt, Nico Balbach, Maximilian Wehner, Uwe Springmann, Christoph Wick, Christine Grundig, Andreas Büttner und Frank Puppe. 2019. OCR4all – An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings. Applied Sciences 9, Nr. 22.

Wintergrün, Dirk. 2019. „Netzwerkanaylse und semantische Datenmodellierung als heuristische Instrumente für die historische Forschung“. Erlangen / Nürnberg: Friedrich-Alexander-Universität.