DiaCollo für GEI-Digital

Ein experimentelles Projekt zur weiteren Erschließung digitalisierter historischer Schulbuchbestände

Nieländer, Maret
nielaender@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Germany

Jurish, Bryan
jurish@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Scheel, Christian
scheel@leibniz-gei.de
Georg-Eckert-Institut - Leibniz-Institut für internationale Schulbuchforschung, Germany

Seit 2009 digitalisiert die Forschungsbibliothek des Leibniz-Institut für Bildungsmedien | Georg-Eckert-Institut (GEI) mit Förderung der DFG seine historischen Bestände. Die „Digitale Schulbuchbibliothek GEI-Digital“[1] umfasst mittlerweile etwa 6000 vor 1920 erschienene Werke, v.a. deutschsprachige Realienkundebücher, Fibeln und Lesebücher sowie Bücher für die Fächer Geographie, Geschichte, Politik und Religion. Die Digitalisate, Metadaten und OCR-generierten Volltexte können online genutzt werden und stehen in verschiedenen Formaten unter der Lizenz CC0 zum Download zur Verfügung (Hertling / Klaes 2018a, 2018b).

Das Projekt

Das experimentelle, GEI-intern geförderte Projekt „DiaCollo für GEI-Digital“[2] zielte darauf ab, den aktuellen digitalen Bestand mit etablierten computerlinguistischen Werkzeugen zu verbinden und dabei die Passfähigkeit von Daten, Werkzeugen und Bedarfen der Nutzer:innen zu testen und ggf. zu erhöhen. Es schließt damit an frühere Projekte an, bei denen mit verschiedenen Partnern unterschiedliche Ansätze für die weitere digitale Erschließung der Bestände erprobt und entwickelt wurden: so etwa zur Visualisierung der Metadaten im Projekt „GEI-Digital Visualized“[3] und zur Nutzung von Volltexten und Metadaten für Filterung, Gruppenvergleiche, Suchen in Verbindung mit Topic Modells u. ä. im Projekt „Welt der Kinder“.[4] Das Projekt ist somit Bestandteil von Bedarfserhebung und Benchmarking für Tool-Entwicklungen am GEI (De Luca et.al. 2019).

Vorgehen

Ende 2020 wurden die Daten aller Werke, die bis zu diesem Zeitpunkt mit automatisch generierten Volltexten zur Verfügung standen über die bestehenden APIs gesammelt, nach TEI konvertiert und zum *GEI-Digital-2020* Korpus zusammengefasst.

Für die maschinelle Vorverarbeitung und Indexierung wurden Werkzeuge und Workflows genutzt, die am Zentrum Sprache an der Berlin-Brandenburgischen Akademie der Wissenschaften speziell für historische deutschsprachige Texte entwickelt und genutzt werden. Diese Werkzeuge sind dafür optimiert, möglichst vorlagengetreue digitale Volltexte um Zusatzinformationen anzureichern, um so z. B. Frequenz- und diachrone Kollokationsanalysen und komplexe Suchen unter Einbeziehung von Wortarten zu ermöglichen.

Hierfür wurde eine Instanz der ebenfalls am Zentrum Sprache genutzten und entwickelten D*- und DiaCollo-Software für das GEI aufgesetzt. Das open-source Werkzeug DiaCollo wurde von Bryan Jurish in Zusammenarbeit mit Historiker:innen entwickelt, um den Wortgebrauch über die Zeit sowohl im Distant Reading zu untersuchen und zu visualisieren, als auch die Ergebnisse jederzeit am konkreten Beleg in der Quelle überprüfen zu können (Jurish 2018; Jurish / Nieländer 2020). Üblicherweise wird DiaCollo mit historischen Referenzkorpora oder mehrere Jahrgänge umfassenden Zeitungs- und Zeitschriften-Korpora eingesetzt.

Abb. 1: Startseite von „DiaCollo für GEI-Digital“

Usabilty und Nachnutzbarkeit

- Korpus und Werkzeuge sind über die Webseite des Projektes nutzbar.
- Die Benutzeroberflächen von D* und DiaCollo wurden für Benutzer:innen mit einem gewissen Maß an Vorerfahrung mit korpuslinguistischen Methoden und Terminologie entwickelt. Um die Usability für Nutzer:innen aus anderen Disziplinen und für Laien zu erhöhen wurde ein umfangreiches Tutorial erstellt, das die Benutzeroberflächen, einige der anpassbaren Parameter sowie Beispielabfragen präsentiert. Auch Vorverarbeitung, Indexierung und einige Besonderheiten des Korpus werden im Tutorial vorgestellt (Nieländer / Jurish 2021).
- Um den Nutzer:innen intuitivere Einblicke in die Zusammensetzung des Korpus‘ zu ermöglichen, wurden die Visualisierungen und Filterfunktionen des o.g. Projektes „GEI-Digital Visualized“ nachgenutzt, die 2017 in einer Kooperation mit der Fachhochschule Potsdam entwickelt worden waren. [5] Zudem stehen die bibliographischen Metadaten aller Werke des GEI-Digital-2020 Korpus zum Download als Excel-Liste bereit. [6]
- Die verfügbaren Exportmöglichkeiten für Treffermengen wurden um ein KWIC/CSV Format ergänzt, um auch technisch wenig versierte Nutzer:innen in die Lage zu versetzen, diese z. B. in ein Tabellenkalkulationsprogramm zu exportie-

ren um sie dort weiter zu bearbeiten oder archivieren zu können.
- Das *GEI-Digital-2020* Korpus wurde auch über das Zentrum Sprache zugänglich gemacht, wo es z.B. von der Community der Sprachwissenschaft und Germanistik nachgenutzt wird. Die historischen Schulbücher können dort vergleichend oder gemeinsam mit weiteren historischen Quellensammlungen der Jahre 1465–1969 untersucht werden. [7]

Befunde und Ausblick

Das Projekt verdeutlicht einmal mehr die Vorteile der Nachnutzung und der offenen, interoperablen Gestaltung von Datenbeständen und digitalen Werkzeugen. Als Anwendungsfall bereits erprobter Abläufe war es verhältnismäßig ressourcenschonend realisierbar und half gleichzeitig, diese Abläufe weiter zu testen und optimieren. Die Möglichkeiten für digital gestützte Analysen historischer Schulbücher wurden erheblich erweitert auch wenn die Aussagekraft computerlinguistischer Analysen durch die Fehlerquote der automatischen Texterkennung einschränkt bleibt. Einige Charakteristika von Schulbüchern und der Korpuszusammensetzung haben sich als nicht optimal kompatibel mit den Logiken der Analysewerkzeuge erwiesen. Dies ist zum Teil durch die Nutzung von Filterfunktionen mit der DDC-Abfragesprache kompensierbar. Grundsätzlich zeigten sich im Projekt nutzer:innenseitige Bedarfe für die ausführlich dokumentierte und auf unterschiedliche Zielgruppen abgestimmte Gestaltung von Forschungsinfrastrukturen, die ggf. auch projektspezifische Korpuszusammenstellungen und modulare, individuell durchführbare Datenkuration erlauben.

Fußnoten

1. http://gei-digital.gei.de/
2. https://diacollo.gei.de/
3. http://gei-digital.gei.de/visualized
4. http://wdk.gei.de/
5. https://diacollo.gei.de/gei-digital-2020/visualized/
6. https://diacollo.gei.de/wp-content/uploads/2021/04/gei-digital-2020.xlsx
7. https://www.dwds.de/d/korpora/dtaxl

Bibliographie

Hertling, Anke / Klaes, Sebastian (2018): „Historische Schulbücher als digitales Korpus für die Forschung: Auswahl und Aufbau einer digitalen Schulbuchbibliothek“ in: Nieländer, Maret / De Luca, Ernesto William (eds.): *Digital Humanities in der internationalen Schulbuchforschung.* Göttingen: V&R unipress 21–44. DOI: 10.14220/9783737009539.21

Hertling, Anke / Klaes, Sebastian (2018): „»GEI-Digital« als Grundlage für Digital-Humanities-Projekte: Erschließung und Datenaufbereitung“ in: Nieländer, Maret / De Luca, Ernesto William (eds.): *Digital Humanities in der internationalen Schulbuchforschung.* Göttingen: V&R unipress 45–68. DOI: 10.14220/9783737009539.45

Jurish, Bryan (2018): „Diachronic Collocations, Genre, and DiaCollo“ in: Whitt, R. J. (ed.): *Diachronic Corpora, Genre, and Language Change*. Amsterdam: John Benjamins 42–64.

Jurish, Bryan / Nieländer, Maret (2020): “Using DiaCollo for historical research” in: Simov, Kiril / Eskevich, Maria (eds.), *Selected Papers from the CLARIN Annual Conference 2019, Linköping Electronic Conference Proceedings* 172:5: 33–40. DOI: 10.3384/ecp2020172005

Nieländer, Maret / Jurish, Bryan (2021): *D* für Anfänger:innen: Ein Tutorial. Einfache und komplexe Suchanfragen, Frequenzanalysen und diachrone Kollokationsanalysen in der D*-Korpusmanagement-Umgebung.* urn:nbn:de:0220-2021-0088.

De Luca, Ernesto William / Fallucchi, Francesca / Ligi, Alessandro / Tarquini, Massimiliano (2019): “A Research Toolbox: A Complete Suite for Analysis in Digital Humanities” in: Garoufallou E. / Fallucchi F. / William De Luca E. (eds): *Metadata and Semantic Research. MTSR 2019. Communications in Computer and Information Science*, vol 1057. Springer, Cham: 385–397. DOI: 10.1007/978-3-030-36599-8_35