Nieländer, Maret

DiaCollo für GEI-Digital. Computerlinguistische Werkzeuge für die Analyse von mehr als 5.000 historischen deutschsprachigen Schulbüchern

Oberdorf, Andreas [Hrsg.]: Digital Turn und Historische Bildungsforschung. Bestandsaufnahme und Forschungsperspektiven. Bad Heilbrunn : Verlag Julius Klinkhardt 2022, S. 33-48

Kontakt / Contact:

pedocs
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Maret Nieländer

DiaCollo für GEI-Digital. Computerlinguistische Werkzeuge für die Analyse von mehr als 5.000 historischen deutschsprachigen Schulbüchern

Das Projekt „DiaCollo für GEI-Digital“ bereitete eine Sammlung von 5.036 retrodigitalisierten deutschsprachigen Schulbüchern des Publikationszeitraums 1648 bis 1921 für die Nutzung mit computerlinguistischen Werkzeugen für komplexe Suchanfragen, Frequenz- und diachrone Kollokationsanalysen auf und stellte dieses „GEI-Digital-2020“-Korpus in der Korpusmanagement-Umgebung D* als experimentelle Infrastruktur für die historische (Bildungs-)Forschung online zur Verfügung. Der vorliegende Beitrag stellt einige Spezifika und Hintergrundinformationen zu digitalen Datenbeständen und Werkzeugen vor und berichtet von den technischen Arbeiten und Initiativen zur Erhöhung der Nutzbarkeit durch Erstanwender:innen. Anhand von Beispielen wird gezeigt, wie die kombinierte Nutzung von Korpus und Werkzeugen Rückschlüsse auf deren Charakteristika, respektive Funktionsweisen sowie auf ihre gegenseitige Passfähigkeit ermöglicht.

1 Einleitung

Durch digitale Transformationsprozesse in den Geisteswissenschaften verlieren die dort ehemals klaren Abgrenzungen von Forschung und Forschungsinfrastruktur an Bedeutung und Kontur. Anbieter:innen digitaler Forschungsinfrastrukturen forschen zu Themen wie FAIR Data, zu semi- und vollautomatischer Aufbereitung, Anreicherung und Verknüpfung von Datenbeständen, zu digitalen Recherche- und Analysewerkzeugen und zu Fragen der Usability. Gleichzeitig werden kulturwissenschaftliche Forscher:innen vermehrt auch zu aktiven Kurator:innen und Gestalter:innen von Forschungsdaten, die digitale Werkzeuge für die Aufbereitung und Analyse gemäß ihrer individuellen Informationsbedürfnisse testen, auswählen, adaptieren und anwenden. Das Projekt „DiaCollo für GEI-Digital“ (https://diacollo.gei.de/) ist ein Beispiel für diese Entgrenzung und die konvergierenden Fragestellungen beider Sphären. Seine Datengrundlage bilden die historischen Quellen des langfristigen Forschungsinfrastrukturprojekts „GEI-Digital. Die digitale Schulbuch-Bibliothek“ (https://gei-digital.gei.de/) des Leibniz-Instituts für Bildungsmedien | Georg-Eckert-Institut (GEI). Ausgangspunkt des in diesem Beitrag vorgestellten

Projekts war der Wunsch, diese umfangreichen digitalen Bestände auch „im großen Stil" zum Beispiel auf inhaltliche und sprachliche Kontinuitäten und Brüche, Parallelen, Fort- und Umschreibungen untersuchen zu können. Besonders geeignet hierfür schienen die Werkzeuge für Vorverarbeitung und Analyse, die am Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) für die Arbeit mit zeitgenössischen und historischen Korpusdaten eingesetzt werden (https://www.bbaw.de/forschung/zentren/zentrum-sprache). Ob Daten und Werkzeuge zueinander passen würden und inwieweit ihre Kombination einen Mehrwert für Forschung und Forschungsinfrastrukturen ergäbe, war von allgemeinerem Interesse, so dass das experimentelle Projekt hausintern gefördert wurde.
Schulbücher sind eine attraktive Quelle für eine Vielzahl historischer und kulturwissenschaftlicher Fragestellungen. Im Folgenden werden zunächst die Quellengattung, die historische Sammlung des GEI und die digitale Infrastruktur „GEI-Digital" kurz vorgestellt. Das Potential einer computergestützten Analyse dieses Materials wird bei der anschließenden Vorstellung der wichtigsten Funktionen der Korpusmanagement-Umgebung D*, bzw. der darin verfügbaren Werkzeuge deutlich. Nach einigen Vorbemerkungen zum institutionellen Setting und den beteiligten Personen folgt dann die chronologische Beschreibung der wichtigsten technischen und konzeptionellen Arbeiten im Projekt. Die daraus entstandene Infrastruktur wird seitdem für die Exploration und für konkrete Forschungsfragen eingesetzt. Im letzten Teil dieses Beitrags wird anhand von Beispielen demonstriert, wie sich Nutzer:innen aus Feldern wie der historischen Bildungsforschung die beiden Black-Boxes *Korpusdaten* und *Tools* durch die reflektierte, aufeinander bezogene Nutzung gegenseitig erhellen können.

2 Die Datengrundlage: Die digitale Schulbuchbibliothek GEI-Digital

Die Forschungsbibliothek des GEI sammelt und erschließt schulische Bildungsmedien und Lehrpläne aus aller Welt. Schulbücher sind eine wertvolle Quelle für eine große Bandbreite kulturwissenschaftlicher und historischer Forschung, denn sie beinhalten Wissensbestände, Weltbilder und Wertvorstellungen, die verschiedene Gesellschaften ihren Kindern und Jugendlichen vermitteln woll(t)en. Als Massenmedien erreichen Schulbücher einen großen Anteil der Bevölkerung. Als Gattung erheben sie Anspruch auf Autorität und überzeitliche Gültigkeit ihrer Inhalte. Zugleich unterscheiden sich Inhalte und Präsentationen je nach Fächern, Ländern und Regionen und differenzieren zwischen Schulformen, Bildungsstufen und manchmal auch nach Konfession oder Geschlecht der Schüler:innenschaft. Hinzu kommen Veränderungen über die Zeit. Die Inhalte wandeln sich nach neuen wissenschaftlichen Erkenntnissen, Lehrmeinungen und Moden, und sie sind geprägt von den jeweiligen unterrichtspraktischen, verlegerischen und politischen Interessen ihrer Epoche.

doi.org/10.35468/5952-03

Schulbücher lassen sich als Gebrauchs-, aber auch als *Verbrauchs*literatur beschreiben, die selten gesammelt und erhalten wurde. Eine Gesamtbibliografie deutschsprachiger Schulbücher existiert nicht. Das GEI sammelt und erschließt vor allem Unterrichtswerke für die Fächer Geschichte, Geografie, Sozialkunde/Politik, Werteerziehung/Religion sowie deutschsprachige Lesebücher und internationale Fibeln. Die historische Sammlung der Schulbücher vor 1945 enthält vor allem deutschsprachige Werke: Knapp 3.100 Bücher bis ins Jahr 1870, knapp 13.000 aus der Zeit des Deutschen Kaiserreichs (1871–1918), etwa 5.880 aus der Weimarer Republik (1919–1932) und etwa 3.110 aus der Zeit des Nationalsozialismus (1933–1944).

Seit 2009 werden Teile dieses Bestands im Rahmen verschiedener DFG-geförderter Projekte digitalisiert und dabei teilweise mittels Leihgaben anderer Bibliotheken ergänzt. Um die Digitalisate weltweit allen Nutzer:innen mit Zugang zum Internet gemeinfrei (Lizenz CC0) zur Verfügung stellen zu können, werden bislang nur urheberrechtlich unbedenkliche Bücher bis etwa 1920 digitalisiert. Innerhalb dieser Vorgabe erfolgt die Auswahl in Abstimmung mit einem Digitalisierungsbeirat nach konservatorischen Kriterien und mit dem Ziel der Vollständigkeit oder Repräsentativität bestimmter Bestände. Angestrebt wird dabei die Digitalisierung der jeweils frühsten und spätesten nachgewiesenen Auflagen der Schulbücher sowie aller weiteren verfügbaren Auflagen, in denen signifikante Veränderungen festgestellt werden konnten (vgl. Hertling & Klaes 2018a, 23–25, 28f.).

Die Digitalisierung selbst erfolgt nach definierten Arbeitsschritten und Standards, die eine gleichbleibende und vergleichbare Datenqualität gewährleisten. Nach der Auswahl der Werke werden zunächst digitale Bilder erstellt, die eindeutige Kennungen (PPN und URN) erhalten und einer automatischen Volltexterkennung (OCR) unterzogen werden. Nur bei besonders alten Drucken sowie Atlanten und Fibeln wird darauf verzichtet, da die Fehlerraten hier bislang noch sehr hoch sind. Es folgt eine basale Annotation der Inhalte, bei der Strukturtypen wie Kapitelüberschriften, Abbildungen, Inhaltsverzeichnisse oder Werbung erfasst werden. Die bibliographischen Metadaten werden aus dem Bibliothekskatalog übernommen, der am GEI auch eine lokale Klassifikation mit Angaben zu Unterrichtsfach, Schulform und Bildungsstufe enthält.

Die Digitalisate, Volltexte und Metadaten werden Forschung und Öffentlichkeit über die Webseite „GEI-Digital – Die digitale Schulbuchbibliothek“ zugänglich gemacht. Dort finden sich zudem redaktionelle Inhalte zur Beschreibung des Projektes, verschiedene Präsentationsmodi, Suchfunktionen, Optionen für Downloads einzelner Werke oder Kapitel als PDF sowie Programmierschnittstellen (APIs), die auch genutzt werden, um die Werke im Katalog der Deutschen Digitalen Bibliothek nachzuweisen (vgl. Hertling & Klaes 2018 a, b).

doi.org/10.35468/5952-03

Für Forschende bedeutet die Forschungsinfrastruktur „GEI-Digital“ zunächst eine erhebliche Effizienzsteigerung gegenüber der analogen Recherche nach und in diesen Beständen. Solange für die weitere Nutzung nur die Lesbarkeit der Quellen durch den Menschen erforderlich ist, sind sie auch ‚kompatibel‘ mit analogen und digitalen Ressourcen anderer Provenienzen. Wenn jedoch die Daten mit digitalen Werkzeugen weiterverarbeitet und analysiert werden sollen, sind entsprechende informatische Fachkenntnisse, bzw. interdisziplinäre Zusammenarbeit erforderlich. Die Daten müssen – wie im Projekt „DiaCollo für GEI-Digital“ geschehen – mittels entsprechender Skripte heruntergeladen und den Anforderungen spezifischer Projekte entsprechend weiter kuratiert werden. Da diese Veränderungen nicht zwangsläufig bibliothekarischen Standards entsprechen, und zudem jeweils nur eine individuelle und statische Zusammenstellung von Daten betreffen, werden diese Daten üblicherweise nicht zurück in die Ursprungs-Infrastrukturumgebungen gespeist, sondern als Forschungsdaten ggf. separat verfügbar gemacht. Beispiele hierfür sind die nachnutzbaren Datenbestände und Benutzer:innenoberflächen von interdisziplinären Projekten des GEI wie „Welt der Kinder“ (http://wdk.gei.de/) oder „WorldViews“ (https://worldviews.gei.de/).

Das Projekt „DiaCollo für GEI-Digital“ verstand sich als Teil dieser Tradition und legte seinen Fokus auf die weitere Erschließung der digitalen Bestände für Nutzer:innen aus verschiedenen historischen Disziplinen sowie der Sprach- und Literaturwissenschaft. Die Motivation für die Einbindung und Nachnutzung von Infrastrukturen der Berlin-Brandenburgischen Akademie der Wissenschaften wird deutlich, wenn man den Funktionsumfang der dort eingesetzten Korpusmanagement-Umgebung D* und der damit verbundenen Werkzeuge betrachtet.

3 Die Werkzeuge: D* und DiaCollo

Die BBAW bildet das Dach für eine Vielzahl digitaler Editions- und Sammlungsprojekte (https://www.bbaw.de/bbaw-digital). Für korpusbasierte Projekte wird dabei die Korpusmanagement-Umgebung D* (dstar) eingesetzt. Das *Digitale Wörterbuch der deutschen Sprache (DWDS)* listet mehr als 30 verschiedene Korpora, die über D* oder die Nutzeroberflächen des DWDS teils ganz frei, teils nach Anmeldung online genutzt werden können (https://www.dwds.de/r). Für die historische Forschung besonders interessant ist das *Referenzkorpus des Deutschen Textarchivs* (DTA). Es bildet verschiedene Textgattungen vom frühen 16. bis zum frühen 20. Jahrhundert in möglichst repräsentativer Weise ab und umfasst neben einem Kern von ca. 1.500 Werken verschiedene Erweiterungen; derzeit sind 4.443 Werke recherchierbar. Andere Korpora, wie Parlamentsprotokolle, Zeitschriften- oder Zeitungskorpora zeugen von Sprachgebrauch und Debatten innerhalb ihrer Domänen und sind oftmals über viele Jahre, bzw. gesamte Publikationszeiträume

doi.org/10.35468/5952-03

vollständig erfasst. Alle diese Korpora basieren – wie „GEI-Digital“ – auf retrodigitalisierten Quellenbeständen und Sammlungen. Im Gegensatz zur digitalen Schulbuchsammlung wurden dabei jedoch zumeist die Volltexte händisch erfasst oder korrigiert und die Daten dann spezifischen Vorverarbeitungsschritten unterzogen, um sie mit maschinenlesbaren Informationen anzureichern. Der erste dieser Arbeitsschritte ist die Tokenisierung (also die Identifizierung verschiedener Zeichenketten als distinkte Wortformen, Nichtwörter, Satzzeichen), es folgen u. a. Normalisierung (Transliteration historischer Schreibweisen), die Erkennung von Wortarten sowie die Zuordnung zur Grundform des Wortes (Lemmatisierung). Als Korpusmanagement-Umgebung beinhaltet und verbindet D* verschiedene Zusammenstellungen (Indexierungen) dieser Daten sowie die grundsätzlich voneinander unabhängigen, jedoch interoperabel gestalteten Tools, die auf diese Indexe zugreifen.

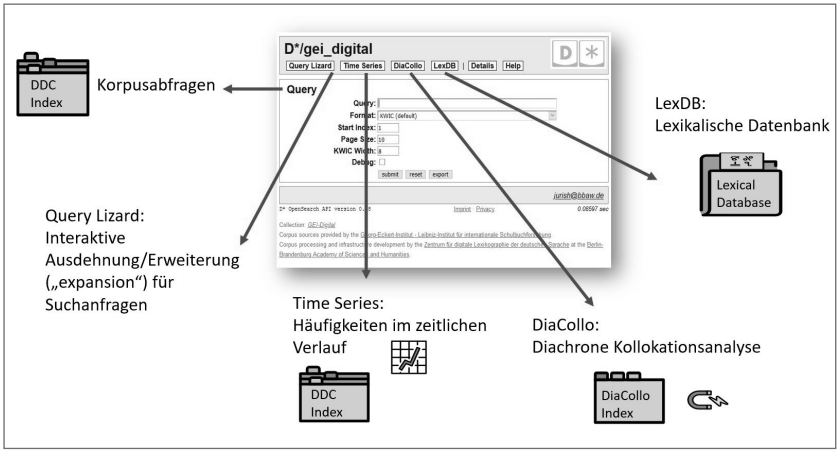


Abb. 1: Startseite der Benutzeroberfläche von D* (hier „D*/gei_digital“ für die Analyse des „GEI-Digital-2020“-Korpus) mit den dort verlinkten Werkzeugen. Quelle: Screenshot der D*-Instanz des GEI, Clipart von https://openmoji.org/, CC BY-SA 4.0

1. *Query/Suche:* Neben einfachen Stichwortsuchen sind auch komplexe Anfragen möglich, was allerdings die Kenntnis der eingesetzten Abfragesprache voraussetzt. Ermittelt werden können z. B. einzelne oder kombinierte Wortformen und Lemmata sowie unterschiedliche historische Schreibweisen. Die Ergebnisse können gezählt, nach bestimmten Kriterien sortiert und nach ihren Metadaten und deren Attributen gefiltert werden. Auch sogenannte ‚Reguläre Ausdrücke‘ können eingesetzt werden. Die Treffermengen werden standardmäßig im „Stichwort im Kontext“-Format (KWIC) angezeigt. Neben Anpassungen dieses

doi.org/10.35468/5952-03

Anzeigeformates und Export der Trefferlisten ist es durch die entsprechende Konfiguration von D* auch möglich, zu den zugrundeliegenden Digitalisaten auf externen Webseiten zu verlinken. So können Nutzer:innen den jeweiligen Fund im größeren Kontext der Quelle betrachten.

2. Mit dem Werkzeug *Query Lizard* können über interne oder eingebundene Tools wie „Germanet" zum Beispiel Synonyme, Ober- und Unterbegriffe zu Stichworten ermittelt, und diese dann für anschließende Anfragen mit dem *Query*-Werkzeug ausgewählt werden.
3. Das *Time Series*-Werkzeug nutzt eine Datenbank um Frequenzen und zeitliche Verteilung einzelner Stichwörter zu ermitteln und als Graph darzustellen.
4. *DiaCollo* ermöglicht die Analyse von Kollokationen, also typischen Wortverbindungen, über die Zeit. Das Open-Source-Werkzeug (https://metacpan.org/dist/DiaColloDB) ermittelt und visualisiert auffällig häufig auftretende Wortverbindungen zu frei wählbaren Stichworten in frei wählbaren Zeitabschnitten. Entwickelt wurde es ab 2015 an der BBAW im Rahmen des Verbundprojekts CLARIN-D von Bryan Jurish im Austausch mit Kolleg:innen sowie Thomas Werneke als Koordinator der CLARIN-D Facharbeitsgruppe „Zeitgeschichte".
5. Die Lexikalische Datenbank *LexDB* ist eine relationale SQLite-Datenbank. Sie enthält neben allen Rohdaten auch die während der Aufbereitung des Korpus erstellten Token-Attribute mit ihren jeweiligen Frequenzen, nicht jedoch die Metadaten des Korpus. Diese Datenbank kann mit SQL-Querys abgefragt werden, um das Vokabular eines Korpus zu untersuchen.

Die Algorithmen dieser Werkzeuge machen neben den digitalen Texten und ihren Metadaten auch die in der Vorverarbeitung generierten Zusatzinformationen nutzbar und ermöglichen so eine sehr viel ‚genauere' Suche innerhalb der Korpora. Dies ist nicht nur für Lexikograf:innen und Computerlinguist:innen interessant, sondern auch für historische Forschungsfragen, die sich auf die sprachliche Oberfläche von Texten beziehen oder durch deren Betrachtung unterstützt werden können: Welche Autor:innen drücken welche Sachverhalte, Ideen, Theorien zu welcher Zeit und in welchem Kontext auf eine bestimmte Weise aus? Wie unterscheidet sich dieser Sprachgebrauch von dem anderer Autor:innen, Diskursdomänen oder Zeiten?

Zusammen mit der Verlinkung zu den Digitalisaten der Quellen ermöglichen Werkzeuge wie Suche, Frequenzanalyse und DiaCollo eine Form des Arbeitens, die mitunter als ‚blended' oder ‚scalable reading' bezeichnet wird: Gemeint ist die neuartige Praxis, als Forscher:in mit digitalen Werkzeugen zu interagieren, und dabei sozusagen in die Datensammlungen hinein- und herauszuzoomen. So lassen sich computergestützte Analyse und Exploration gesamter Korpora im Hinblick auf spezifische Phänomene und übergreifende Muster auf fruchtbare Weise mit der individuellen Lektüre und Interpretation der Fundstellen im Kontext der jeweiligen Quellen verbinden (vgl. Burckhardt u. a. 2019; Jurish & Nieländer

doi.org/10.35468/5952-03

2020). Dieses Potential auch für die historischen Schulbuchquellen zu erschließen war die Motivation des im Folgenden näher beschriebenen Projektes.

4 Das Projekt: Organisation, technische Arbeiten und Bemühungen um Usability

Das Projekt „DiaCollo für GEI-Digital" arbeitete mit frei zugänglichen Datenbeständen und fast ausschließlich mit Open-Source-Software und wäre somit grundsätzlich auch von einzelnen und/oder externen Forscher:innen durchführbar gewesen. Durch die institutionellen Anbindungen an GEI und BBAW und die fachlichen Voraussetzungen der Beteiligten konnte das Projekt jedoch mit einem erheblich reduzierten Aufwand an Zeit und Kosten realisiert werden. Die *Finanzierung* erfolgte über den hausinternen Seed Fonds des GEI, der einen Werkvertrag für den externen Partner und den Einsatz eines gewissen Stundenkontingentes der beteiligten GEI-Mitarbeiter:innen ermöglichte. Das *Projektteam* bestand neben der Autorin (Konzeption, Koordination, Testen) aus Bryan Jurish vom Zentrum Sprache an der BBAW (Computerlinguistik, Infrastruktur) und Christian Scheel aus der Abteilung für Digitale Forschungs- und Informationsinfrastrukturen des GEI (Data Science). Beteiligt an einzelnen Arbeitsschritten waren zudem Kolleg:innen der Forschungsbibliothek und der IT des GEI. Da sich das Projektteam bereits aus dem Kontext des Infrastruktur-Verbundprojektes CLARIN-D kannte, die Mitglieder DH-Projekterfahrung besaßen und mit den Schulbuchdaten, respektive mit D* und den damit verbundenen Werkzeugen, vertraut war, verkürzte sich die sonst bei DH-Projekten übliche *Onboarding- und Orientierungsphase* auf ein Minimum. Zwei vorgesehene Präsenztreffen wurden aufgrund des Pandemiegeschehens durch Videomeetings ersetzt. Die *Kommunikationswege und Dokumentation* konnten durch die geringe Projektgröße flexibel und nach den Vorlieben der Beteiligten gestaltet werden. Für Dokumentation und Handling von Daten und Tools nutzten die Entwickler eine git-Instanz der BBAW. Die Autorin dokumentierte zusätzlich den Projektverlauf, sowie ihre Fragen und Beobachtungen zur Nutzung der Werkzeuge in Textform. Anfragen und Aufgabenverteilungen wurden per E-Mail geteilt und entweder direkt beantwortet, oder als Tickets und Kommentare im git hinterlegt. *Zeitlich* lief das Projekt ab Mai 2020 parallel zu anderen Daueraufgaben und Projekten der Teammitglieder. Externe Deadlines waren nicht vorgegeben. Dass das Projekt trotzdem vergleichsweise zügig realisiert werden konnte, erklärt sich gleichermaßen aus der hohen Motivation der Beteiligten und aus der Unterstützung ihrer Heimatinstitutionen. Der Ablauf der technischen Arbeiten war folgender:

- Zusammenstellung und Formatanpassung: Die zu Projektbeginn im Mai 2020 in „GEI-Digital" verfügbaren Volltexte und Metadaten (mit Ausnahme der Strukturannotationen) wurden erfasst und nach TEI konvertiert; Datentransfer zur BBAW.

doi.org/10.35468/5952-03

- Vorverarbeitung: Die Volltexte wurden mit den an der BBAW für historische deutschsprachige Texte eingesetzten und optimierten Natural Language Processing-Werkzeugen bearbeitet.
- Indexierung: Die aufbereiteten Daten wurden auf unterschiedliche, auf die Funktionsweise der später darauf zugreifenden Werkzeuge angepasste Weise indexiert (DDC-Index, DiaCollo-Index, LexDB-Einträge).
- Test-Instanz: Eine Instanz der D* Korpusmanagement-Umgebung mit den obengenannten Tools und den bislang generierten Daten wurde aufgesetzt und webbasiert zur Verfügung gestellt.
- Testphase: Ende Mai bis Dezember 2020 wurden Bugs eliminiert sowie weitere Desiderata etwa zur Usability identifiziert. Ein Beispiel: In „GEI-Digital" werden Schulbuchreihen als eigene, den Einzelwerken übergeordnete Datensätze angelegt. Sie enthalten u. a. den Reihentitel, so dass die Titel-Information der einzelnen dazugehörigen Bände dann z. B. nur „Band 2" oder „Für das 8. bis 10. Schuljahr" lauten. Deshalb mussten für die Metadaten-Anzeige in D* die entsprechenden Angaben der Datensätze miteinander kombiniert werden.
- Erweiterung der Datengrundlage: In „GEI-Digital" werden laufend weitere Werke erschlossen und bereitgestellt. Zusätzlich wurden für dieses Projekt von Mai bis Dezember 2020 eine Reihe bislang nur im Bild-Digitalisat bereitgestellter Texte aus dem 17. und 18. Jahrhundert einer automatischen Volltexterkennung unterzogen.
- Hardware und digitale Zugänge: Am GEI wurde eine virtuelle Maschine aufgesetzt und die Zugriffsrechte konfiguriert. Eine Projektwebseite als virtuelle Startseite wurde konzipiert und zunächst als HTML-Eigenbau, dann als Wordpress-Instanz aufgesetzt. Eine basale Auswertung des Traffics wird unter Einhaltung der Datenschutzrichtlinien durch die Anbindung an die Analytiksoftware Matomo erreicht, zusätzlich zählt der Server die Anzahl der Anfragen.
- Erstellung und Bereitstellung des „GEI-Digital-2020"-Korpus: Mitte Dezember 2020 wurden die Daten der 5.036 zu diesem Zeitpunkt in „GEI-Digital" mit Volltext verfügbaren Werke erfasst, konvertiert, einem NLP (Natural Language Processing) unterzogen und indexiert und in eine D*-Instanz integriert. Diese wurde in einem Docker-Container auf der virtuellen Maschine des GEI installiert und ist seitdem über die Projektwebseite https://diacollo.gei.de/zu erreichen. Gleichzeitig wurde das Korpus auch über das Zentrum Sprache der BBAW als Bestandteil der dortigen historischen Korpora verfügbar gemacht (https://www.dwds.de/d/gei_digital).

Trotz des grundsätzlich experimentellen Charakters gehörte es zu den Zielen des Projektes, die Nachnutzung von D*-Instanz und „GEI-Digital-2020"-Korpus für die historische (Bildungs-(medien-))Forschung zu ermöglichen. Hierzu wurden im Projektverlauf konkrete Ideen entwickelt, die teils verworfen, teils realisiert wurden:

doi.org/10.35468/5952-03

- Informationen zu Funktionen und Parametern der Werkzeuge: Um Anfänger:innen den Einstieg in die Korpusanalyse mit computerlinguistischen Tools in D* zu erleichtern, war ursprünglich geplant, die bereits existierenden Dokumentationen auf Deutsch und Englisch um kurze Videotutorials zu ergänzen. Aus den Fragen der Autorin und den Antworten von Bryan Jurish erwuchs jedoch eine umfangreiche Beispielsammlung, die zunächst als PDF-Tutorial veröffentlicht wurde (vgl. Nieländer & Jurish 2021).
- Informationen zur Zusammensetzung des Korpus: Um potentiellen Nutzer:innen eine schnelle Übersicht zu ermöglichen, passte Christian Scheel die Visualisierungen und Filterfunktionen von „GEI-Digital Visualized" an, die erstmals 2017 in einer Kooperation mit der FH Potsdam und dem dortigen Urban Complexibility Lab erstellt worden waren (https://diacollo.gei.de/gei-digital-2020/visualized/). Auf der Projektwebseite wurde zudem eine Tabelle mit den bibliographischen Angaben aller Werke im Korpus zum Download bereitgestellt.
- Zugangspunkt und Nutzer:innenführung: Die D*-Korpusmanagement-Umgebung bildet ein in sich geschlossenes System, in dem die Werkzeuge untereinander verlinkt sind. Ein Klick auf „Home" führt dort somit jeweils zurück zur Abfrage-Startseite in D*, nicht etwa zu einer Projektbeschreibung auf den Webseiten der Anbieter:innen dieser Infrastruktur. Zunächst war deshalb angedacht, Linkstruktur und auch visuelle Gestaltungselemente zu verändern. Da dies einerseits zu aufwändig, und andererseits für bereits erfahrene Nutzer:innen eher verwirrend erschien, wurde stattdessen die bereits erwähnte Webseite zum Projekt eingerichtet. Als vorgeschaltete Startseite verlinkt sie auf die D*-Instanz und stellt Informationen zu Korpus, Werkzeugen und Projekt bereit. Innerhalb von D* befindet sich der Link zu dieser Projektstartseite im Footer der Benutzeroberfläche. Nutzer:innen erkennen innerhalb von D* letztlich nur an der URL, ob sie in der Instanz des GEI oder der BBAW mit dem „GEI-Digital-2020"-Korpus arbeiten.
- Exportmöglichkeiten der Rechercheergebnisse: Um auch technisch wenig versierte Nutzer:innen in die Lage zu versetzen, die Treffermengen ihrer Korpusabfragen z. B. in ein Tabellenkalkulationsprogramm zu exportieren, um sie dort weiter zu bearbeiten oder archivieren zu können, wurde zusätzlich zu den bereits existierenden, u. a. auch maschinenlesbaren Formaten eine zusätzliche Exportmöglichkeit im KWIC/CSV-Format implementiert.
- Öffentlichkeitsarbeit: Durch Beschreibung und Verlinkung auf den Webseiten von GEI und DWDS, durch Schulungsangebote, Vorträge, Poster-Präsentationen, Social Media, Blog- und Newsletterbeiträge werden potentielle Nutzer:innen auf die Existenz und Besonderheiten des „GEI-Digital-2020"-Korpus aufmerksam gemacht.

doi.org/10.35468/5952-03

Im Verlauf der Testphase und auch bei der späteren Nutzung der finalisierten Infrastruktur wurde immer wieder deutlich, dass die Operationalisierung geisteswissenschaftlicher Forschungsfragen und die Formulierung zielführender Korpusabfragen für Nutzer:innen ohne computerlinguistischen Fachhintergrund nicht trivial ist. Die Ergebnisse der Abfragen lassen jedoch Rückschlüsse sowohl auf die Korpusdaten als auch die Funktionsweisen der Werkzeuge zu, also auf beide sich potentiellen Nutzer:innen zunächst als Black Boxes darstellenden Unbekannten. Im letzten Teil dieses Beitrages wird dies anhand einiger Beispiele demonstriert, die gleichzeitig die Frage beantworten, wie gut Tools und Korpus zueinander „passen".

5 DiaCollo für GEI-Digital in der Anwendung: Lessons learned

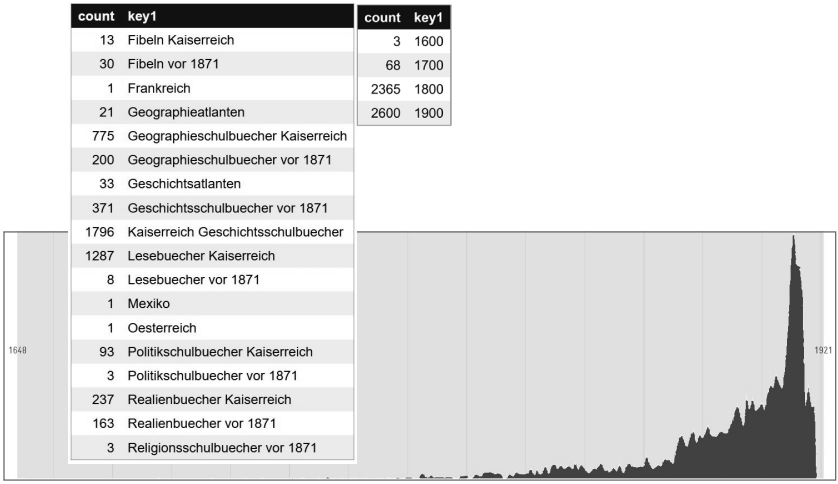
Bei der obigen Beschreibung von D* und den an der BBAW damit verfügbar gemachten Korpora wurden bereits einige Aspekte genannt, die für eine erkenntnisgewinnbringende Weiterverarbeitung und Analyse der Daten wichtig sind. Dies ist zunächst eine rechtlich und ethisch nutzbare, langfristig und eindeutig referenzierbare, ausreichend große Datengrundlage. Für statistische Analysen ist dabei eine möglichst gleichmäßige Verteilung über beabsichtigte Analysekriterien wie z. B. Publikationszeiträume wichtig. Das Korpus sollte möglichst aussagekräftige, vollständige und korrekte Metadaten (also bibliographische Informationen und weitere Annotationen bzw. generierte Token-Attribute) und möglichst originalgetreue Volltexte besitzen.
Im Falle der „GEI-Digital"-Sammlung ist dies in unterschiedlichem Maße gegeben. Sie ist vergleichsweise groß, jedoch ist die Verteilung der Daten über die Zeit durchaus heterogen und die Volltexte sind durch die rein automatische Erkennung mehr oder weniger fehlerhaft. Mit der lokalen Klassifikation, die schulbuchspezifische Eigenschaften beschreibt, und mit der Annotation bestimmter Strukturmerkmale besitzt die Sammlung wertvolle gattungs- und werkspezifische Metadaten. Die Auswirkungen dieser Voraussetzungen sollen im Folgenden genauer erläutert werden.

5.1 Korpuszusammensetzung und (statistische) Repräsentativität der Daten

Inwieweit ein Korpus für die darin enthaltene(n) Quellenart(en) oder in Bezug auf konkrete Forschungsfragen repräsentativ oder gar vollständig ist, kann nicht durch die Analyse des Korpus selbst, sondern nur mit Blick auf den Entstehungskontext der Quellen entschieden werden. Im Falle der Schulbücher wären dies z. B. Informationen zu Bevölkerungsentwicklung und Alphabetisierungsraten, zu Akteur:innen und Institutionen des Unterrichtswesens, Genese schulischer Fächer und wissenschaftlicher Fachdisziplinen, zu Buchmarkt und -handel usw. *Innerhalb* eines Korpus können durch entsprechende Korpusabfragen zumindest die Verteilungen der Datenmengen in bestimmten Kategorien sichtbar gemacht werden. So ergeben sich Hinweise auf ihre jeweilige (statistische) Aussagekraft.

doi.org/10.35468/5952-03

Abbildung 2 listet links die Ergebnisse der Suchabfrage COUNT(* #in file) #by[geiclass], also die Anzahl der im „GEI-Digital-2020"-Korpus enthaltenen Werke pro Schulbuchgruppe (Metadatum „geiclass"). Rechts ist eine Gliederung der vorhandenen Mengen über die Zeit, grob nach Jahrhunderten als Ergebnis der Abfrage COUNT(* #in file) #by[date/100] zu sehen. Darunter ist dies noch mittels einer Zeitleiste verdeutlicht. Diese Abfragen belegen die sehr heterogene Verteilung der Korpusdaten. Nutzer:innen müssen sich vergegenwärtigen, dass es zwar möglich ist, im „GEI-Digital-2020"-Korpus beispielsweise einige französischsprachige oder vor 1700 publizierte Schulbücher zu finden und auszuwerten. Aufgrund der für diese Klassen extrem geringen Datengrundlage können hier aber kaum verallgemeinernde Aussagen getroffen werden. Insbesondere kann bei Frequenzanalysen kaum entschieden werden, ob es sich bei den Daten vor 1800 um Ausreißer oder Muster im Sprachgebrauch handelt; ein direkter Vergleich von Worthäufigkeiten und -verteilungen früher Publikationsjahre mit solchen aus der datenreichen Zeit des Deutschen Kaiserreichs kann deshalb zu falschen Schlüssen führen.



count	key1
13	Fibeln Kaiserreich
30	Fibeln vor 1871
1	Frankreich
21	Geographieatlanten
775	Geographieschulbuecher Kaiserreich
200	Geographieschulbuecher vor 1871
33	Geschichtsatlanten
371	Geschichtsschulbuecher vor 1871
1796	Kaiserreich Geschichtsschulbuecher
1287	Lesebuecher Kaiserreich
8	Lesebuecher vor 1871
1	Mexiko
1	Oesterreich
93	Politikschulbuecher Kaiserreich
3	Politikschulbuecher vor 1871
237	Realienbuecher Kaiserreich
163	Realienbuecher vor 1871
3	Religionsschulbuecher vor 1871

count	key1
3	1600
68	1700
2365	1800
2600	1900

Abb. 2.: Zusammensetzung des „GEI-Digital-2020"-Korpus: Links: Anzahl Werke pro Untersammlung; rechts oben: Anzahl Werke pro Jahrhundert, rechts unten: Visualisierung der Anzahl der Werke/Jahr.

5.2 Metadaten

Die digitalen Werkzeuge in D* ermöglichen es den Nutzer:innen, sich alle mit dem Korpus verfügbaren, d. h. alle indexierten Metadaten (wie *bildungslevel, dokumenttyp, editor, geiclass, land, place, ppn, publisher, schulform, unterrichtsfach*) und deren jeweilige Attribute anzeigen und auflisten zu lassen. Diese können

doi.org/10.35468/5952-03

dann wiederum für Suchen und für das Filtern und Sortieren konkreter Suchanfragen genutzt werden, wie z. B. *Suchbegriff* #HAS[geiclass,'Realienbuecher Kaiserreich']. Da die Metadaten nicht denselben Vorverarbeitungsschritten unterzogen werden wie die Volltexte, können ihre Elemente nicht wie diese einzeln analysiert werden. Mittels ‚Regular Expressions' können jedoch bestimmte Zeichenfolgen darin gesucht werden (vgl. Nieländer & Jurish 2021, 33, 40–42). Im Rahmen der maximalen Länge von Suchanfragen ist es auch möglich, frei gewählte Werke und somit ein selbst zusammengestelltes Subkorpus zu untersuchen, wenn man die Identifier (hier: PPNs) zur Auflistung nutzt: *Suchbegriff* #HAS[ppn,{*Ziffernfolge*, *Ziffernfolge*, *Ziffernfolge*}].

Die Metadaten in „GEI-Digital" wurden für alle Werke nach bibliothekarischen Kriterien und Standards erfasst und haben ein Qualitätssicherungsverfahren durchlaufen. Sie sollen die Quellengattung für möglichst viele Forscher:innen möglichst gut erschließen. Die folgenden drei Szenarien zeigen Beispiele dafür, dass diese Daten möglicherweise trotzdem – oder gerade deshalb – keine perfekte Passung für einzelne Forschungsfragen haben können:

- ‚Fehlende' Metadaten(-Kategorien): Für bestimmte Forschungsinteressen wären zusätzliche Metadaten, bzw. darauf basierende Such- und Filterfunktionen nützlich. Sprachwandel oder Fragen zum Verlagswesen ließen sich möglicherweise besser untersuchen, wenn nach Erstausgaben oder den frühesten verfügbaren Ausgaben gefiltert werden könnte. Für die Untersuchung kultur- und sozialgeschichtlicher Fragen wäre es hingegen nützlich, Werke zu unterscheiden, die explizit für den Unterricht von Jungen, von Mädchen oder von gemischten Klassen, für bestimmte Konfessionen oder aber für die Ausbildung angehender Lehrkräfte vorgesehen waren.
- Kategoriale Zuordnung: In „GEI-Digital" werden die Werke durch die lokale Klassifikation derjenigen Untersammlung zugeordnet, der ihr Inhalt zum *überwiegenden Teil* entspricht. So wird das Genre der „Kinderfreunde" unter „Lesebücher" gruppiert, auch wenn ihre Lesestücke inhaltlich oft auch als Realien- oder Landeskunde beschreibbar wären.
- Nutzung der vorhandenen Metadaten: Nicht alle Metadaten können von jeder Art Werkzeug gewinnbringend weiterverarbeitet werden. Die LexDB enthält beispielsweise nur Tokens und deren Attribute, aber keine Metadaten. So lässt sie sich zwar nutzen, um z. B. die häufigsten Wörter im Korpus zu bestimmen – nicht aber, um die häufigsten Wörter in einzelnen Werken oder Gruppen zu bestimmen. Manche Metadaten müssen neu kombiniert oder arrangiert werden, damit man sie nutzen kann. Ein Beispiel ist der bereits erwähnte Fall der auf verschiedene Datensätze verteilten Reihen- und Werktitel, ein weiteres Beispiel ist die in „GEI-Digital" derzeit getroffene zeitliche Unterteilung mancher Fächergruppen, wie etwa „Fibeln vor 1871" vs. „Fibeln Kaiserreich". Diese Unterteilung ist für Abfragen im Gesamtkorpus irrelevant, da hier die Publi-

doi.org/10.35468/5952-03

kationsjahre ausschlaggebend sind. Für Abfragen in der betreffenden Fächergruppe müssen sie hingegen kombiniert werden, um durchgängige Zeitreihen untersuchen zu können.

Im Rahmen korpusbasierter Forschungs- und Editionsprojekte ist immer zu klären, welche Metadaten zusätzlich erhoben bzw. annotiert werden müssen, und welche der eventuell bereits durch Forschungsinfrastrukturen bereitgestellten Metadaten in welcher Form und mit welchem Aufwand für das konkrete Vorhaben nutzbar gemacht werden können. Im vorliegenden Fall hätten die in „GEI-Digital" vorhandenen Metadaten und Strukturannotationen genutzt werden können, um das Korpus nach Maßgabe spezifischer Forschungsfragen zu bereinigen, also je nach Interessenlage z. B. fremdsprachliche Werke, oder Elemente wie Werbung, Register oder Kapitelüberschriften herauszufiltern. Nicht optimal genutzt wurden die vorhandenen Informationen zu Seitenzahlen: Für die Trefferanzeigen im „GEI-Digital-2020"-Korpus werden nur die fortlaufende Nummerierung der Digitalisate, nicht aber die gedruckten Seitenzahlen der physischen Quelle angezeigt. Für etwaige Neuindexierungen wäre auch zu erwägen, Strukturannotationen mitzukonvertieren, in das Korpus zu integrieren und so für Korpusabfragen mit XPath für entsprechende Filterungen nutzbar zu machen.[1]

5.3 Texttreue der automatisch erstellten Volltexte (OCR)

Bei allen Unterschieden ihrer Forschungsgebiete sehen sich Historiker:innen bis dato fast immer mit einer gewissen Kargheit der Quellenlage konfrontiert und entsprechend dazu gezwungen, sowohl möglichst viele der erhaltenen Quellen zu ermitteln als diese dann auch sehr sorgfältig zu lesen und zu kontextualisieren, um aus ihnen Rückschlüsse auf Lücken in der Überlieferung ziehen zu können. Insofern dürfte vielen Forschenden die Aussicht, 5.036 Schulbücher auf einmal durchsuchen zu können, attraktiv erscheinen. Zudem verspricht man sich Genauigkeit: Denn ein Computer kann nichts ‚übersehen' und wird *sämtliche* erfragten Zeichenfolgen auflisten und vergleichen können, die im Index enthalten sind. An dieser Stelle wird allerdings die Problematik der oft mangelnden Originaltreue digitalisierter Volltexte relevant. Anbieter:innen digitaler Infrastrukturen müssen entscheiden, ob sie ihre vorhandenen Ressourcen einsetzen, um große Datenmengen in begrenzter Qualität, oder erheblich geringere Mengen in hoher Qualität anbieten zu können.

Wenn Fehler bei der Texterfassung nicht vor der Datenaufnahme in die Werkzeug-Kaskaden der automatischen Sprachverarbeitung behoben wurden, werden sie von Schritt zu Schritt ‚weitergereicht'. Sie werden dabei potenziert, da auch die zu ihnen generierten Token-Attribute fehlerhaft sind und die grammatische Struktur größerer Einheiten für die Werkzeuge unkenntlich oder uneindeutig

1 Ich danke Frank Wiegand für diesen Hinweis.

doi.org/10.35468/5952-03

wird. Wenn Analysewerkzeuge wie Suchmaschinen oder DiaCollo mit dieser Art ‚schmutziger' Daten arbeiten, sind die Ergebnisse nur in Bezug auf die indexierten Daten, nicht aber in Bezug auf die tatsächlichen Originalquellen vollständig und korrekt (vgl. Nieländer & Jurish 2021, 10). Die OCR-Fehler der „GEI-Digital"-Texte verringern also die Aussagekraft der Funde und statistischen Vergleiche im „GEI-Digital-2020"-Korpus. Ein wenig aufgefangen werden kann das Problem durch die Verlinkung zum Digitalisat der Quelle, die eine manuelle Prüfung und Kontextualisierung einzelner Ergebnisse durch Close Reading ermöglicht. Eine andere Möglichkeit besteht darin, in der LexDB nach ähnlichen Oberflächenformen zu suchen. Um auch falsch erkannte Formen z. B. von „Asien" zu finden, können dort Vorkommen von Zeichenfolgen wie „Asia" gesucht und die Treffer ggf. händisch ausgewertet werden. Dass sich „Asien" jedoch auch hinter falsch erkannten Zeichenfolgen wie „&AG'a" verbergen kann, dürfte allerdings kein:e Nutzer:in antizipieren können.

5.4 Funktionsweisen digitaler Analysewerkzeuge

Anbieter:innen (digitaler) Forschungsinfrastrukturen und kulturwissenschaftliche Forscher:innen sehen sich bei der Auswahl geeigneter Werkzeuge für die digitale Aufbereitung und Analyse ihrer Quellenbestände mit ähnlichen Problemlagen konfrontiert. Entweder müssen Werkzeuge bedarfsgerecht programmiert oder angepasst werden, was viel Zeit und Aufwand erfordert, aber neben der Passgenauigkeit auf die jeweiligen Anforderungen den Vorteil hat, dass die unmittelbar Beteiligten deren Funktionsweisen genau verstehen und erklären können. Oder es werden existierende Werkzeuge nachgenutzt, was zunächst weniger Aufwand bedeutet, aber eben nicht in jedem Fall ideal auf die Bedürfnisse spezifischer Quellengattungen bzw. individueller Forschungsdesigns passt und eine intensive Einarbeitung in die Funktionsweisen erfordert. Die folgenden Beispiele zu *Frequenzanalysen* und *diachronen Kollokationsanalysen* mögen illustrieren, dass eine solche Lernkurve im Zuge des ‚Digital Turn' in keinem Fall umgangen werden kann.

Visualisierungen stellen per Definition eine Reduzierung von Komplexität dar; ihr Zweck ist es, bestimmte Phänomene sichtbar(er) zu machen, indem andere ausgeblendet werden. Bei der Visualisierung von *Frequenzanalysen* werden Häufigkeit und zeitliche Verteilung von Stichworten auf x-y-Koordinatensystemen abgetragen und somit vermeintlich objektiv vergleichbar gemacht. Dass man hierbei zwischen absoluten Häufigkeiten und Häufigkeiten relativ zur Größe der jeweiligen Grundgesamtheit eines Zeitabschnitts zu unterscheiden hat, darf sicher als Allgemeinwissen gelten. Seit der Berichterstattung zur Corona-Pandemie sind auch Sinn und Funktionsprinzip von Glättungen mittlerweile geläufig: Statt absoluter Zahlen wird ein gleitender Mittelwert betrachtet, für den z. B. Tagesdaten zu Wochen zusammengefasst werden, um reguläre Schwankungen, aber auch Ausreißer auszugleichen und den Vergleich größerer Zeiträume zu erleichtern.

doi.org/10.35468/5952-03

Innerhalb von D* können Frequenzanalysen mit unterschiedlichen Werkzeugen durchgeführt und dabei eine Reihe von Parametern angepasst werden. Dabei empfiehlt es sich, unterschiedliche Arten der Glättung explorativ zu testen und ggf. die Frequenzen eines Begriffs in verschiedenen Korpora und/oder anderen Begriffen im selben Korpus zu vergleichen, um sich ein fundierteres Urteil zu Häufigkeit und Verteilung (und dann erst im nächsten Schritt über deren Bedeutung für Interpretationen von ‚Gewöhnlichkeit' oder ‚Bedeutsamkeit' bestimmter Suchbegriffe) bilden zu können. Dabei ist zu beachten, dass die Werkzeuge, wie eingangs erwähnt, auf unterschiedliche Indexe und damit unterschiedlich große Grundgesamtheiten zugreifen, und dass speziell beim „GEI-Digital-2020"-Korpus die Datengrundlage heterogen ist und OCR-Fehler beinhaltet.

Auch das *DiaCollo*-Werkzeug arbeitet mit der Analyse von Worthäufigkeiten. Es ermittelt, vereinfacht dargestellt, die Häufigkeit der im Umfeld eines Suchbegriffs vorkommenden Worte und vergleicht sie mit ihrer durchschnittlichen Häufigkeit im Korpus, bzw. den ausgewählten Zeiträumen. Dabei sind die Entscheidungen darüber, was als ‚starkes Kollokat' gilt, in die Algorithmen eingeschrieben und können über wählbare Parameter der Abfragen justiert werden. Allerdings ist auch dieses Werkzeug für die Nutzung mit originalgetreuen Volltexten konzipiert und natürlich nicht dazu in der Lage, zwischen OCR-Fehlern und richtig erkannten Wörtern zu unterscheiden. Das führt dazu, dass DiaCollo OCR-Fehler als ‚starke Kollokate' für bestimmte Suchbegriffe interpretiert, weil sie insgesamt sehr selten, aber eben durch Zufall ein- oder zweimal mit diesen Suchbegriffen vorkommen. Zu statistischen Auffälligkeiten bei Kollokationsanalysen im „GEI-Digital-2020"-Korpus führt außerdem die Tatsache, dass viele Texte mehrfach im Korpus enthalten sind – oft im Wortlaut, aber teils auch gekürzt oder leicht adaptiert. Dabei handelt es sich um Lesestücke wie Gedichte, Prosa, Briefe oder politische Reden. Es sind explizit zitierte Quellen, aber auch ohne Autor:innenangaben abgedruckte ‚nachgenutzte' Texte. Solche Dopplungen sind typisch für das Schulbuchkorpus, das ja wie eingangs dargelegt unterschiedliche Auflagen, Regionalausgaben, Bücher für verschiedene Schulformen usw. enthält. DiaCollo, das als Werkzeug für die Beobachtung von Sprachwandel konzipiert wurde, wird hier zum Werkzeug zum Auffinden von Wiederverwendungen von Texten; die ermittelten Wortverbindungen sind damit zwar nicht unbedingt typisch für den Sprachgebrauch der Zeit, wohl aber typisch für bestimmte, häufig abgedruckte Texte. Unabdingbar für eine ‚Plausibilitätsprüfung' ermittelter Kollokationen ist auch hier die Stichwort-im-Kontext-Ansicht, durch die identische Texte schnell zu identifizieren sind. Bei diesen Funden ist dann weniger das einzelne Wort bedeutsam durch hohe Frequenz, als vielmehr der gesamte Text(abschnitt), der offenbar Gegenstand einer Art Kanonisierung ist. Auch kann der Vergleich mit den glücklicherweise vorhandenen Referenz- oder Spezialkorpora sehr erhellend sein.

doi.org/10.35468/5952-03

6 Resümee

Die Nutzung und Anpassung digitaler Quellensammlungen und Analysewerkzeuge findet zunehmend Eingang in universitäre Curricula geisteswissenschaftlicher Fächer. Bis die souveräne Nutzung und entsprechende digitale Quellen- und Methodenkritik jedoch ‚Main Stream' geworden sind, kann der ‚Digital Turn' in diese Richtung am besten von interdisziplinären Teams genommen werden. Die Beispiele in diesem Beitrag zeigen, dass es für Nutzer:innen notwendig ist, sich Kenntnisse über die digitale Kuration einer Quellensammlung sowie die Funktionsmechanismen der für Analysen eingesetzten digitalen Werkzeuge anzueignen, um digitale Quellen- und Methodenkritik betreiben, und diese auch in ihre Interpretation digital gewonnener Ergebnismengen einbeziehen zu können. Dies geschieht am besten in der Praxis, durch die Nutzung entsprechender Infrastrukturen. Deren Anbieter:innen können diese Prozesse durch die Bereitstellung entsprechender Dokumentation erleichtern und profitieren ihrerseits von einem aktiven Austausch über die verschiedenen Nutzer:innenbedürfnisse.

Literatur

Burckhardt, D.; Geyken, A.; Saupe, A.; Werneke, Th. (2019): Distant Reading in der Zeitgeschichte. Möglichkeiten und Grenzen einer computergestützten Historischen Semantik am Beispiel der DDR-Presse. In: Zeithistorische Forschungen/Studies in Contemporary History. DOI: https://doi.org/10.14765/zzf.dok-1345

Jurish, B. (2018): Diachronic Collocations, Genre, and DiaCollo. In: R. J. Whitt (ed.): Diachronic Corpora, Genre, and Language Change. Amsterdam: John Benjamins, 42–64. DOI: https://doi.org/10.1075/scl.85.03jur

Jurish, B. & Nieländer, M. (2020): Using DiaCollo for historical research. In: K. Simov & M. Eskevich (Hrsg): Selected Papers from the CLARIN Annual Conference 2019, Linköping Electronic Conference Proceedings 172, 33–40. DOI: https://doi.org/10.3384/ecp2020172005

Hertling, A. & Klaes, S. (2018a): Historische Schulbücher als digitales Korpus für die Forschung: Auswahl und Aufbau einer digitalen Schulbuchbibliothek. In: M. Nieländer & E. W. De Luca (Hrsg.): Digital Humanities in der internationalen Schulbuchforschung. Göttingen: V&R, 21–44. DOI: https://doi.org/10.14220/9783737009539.21

Hertling, A. & Klaes, S. (2018b): »GEI-Digital« als Grundlage für Digital-Humanities-Projekte: Erschließung und Datenaufbereitung. In: M. Nieländer & E. W. De Luca (Hrsg.): Digital Humanities in der internationalen Schulbuchforschung. Göttingen: V&R, 45–68. DOI: https://doi.org/10.14220/9783737009539.45

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

doi.org/10.35468/5952-03