Bearbeitung großer digitaler Korpora mit Topic Modelling

Bei „Welt der Kinder“ handelt es sich um den geschichtswissenschaftlichen Versuch, große Bestände digitaler Korpora für die historische Arbeit nutzbar und zugänglich zu machen. Durch die Stärkung einer engen Zusammenarbeit zwischen Historikern, Informationswissenschaftlern und Informatikern zielt es darauf, neue Erkenntnisse über die Periode zwischen 1850 bis 1918 zu gewinnen: Eine Zeit beschleunigter Wissensproduktion, die geprägt war von gleichzeitigen Prozessen der Globalisierung und der Nationalisierung.

Die Forschung will Zugang zu deutschsprachigen Massenquellen aus der Zeit zwischen 1850 und 1918 ermöglichen. Dieses Material spiegelt auf der einen Seite zeitgenössische Interpretationsmuster der Welt sowie Elemente eines kulturellen Gedächtnisses wieder, formte sie aber gleichzeitig auf der anderen Seite. Aber schon aufgrund ihrer reinen Menge können diese Quellen nicht mit klassisch heuristischen Methoden bearbeitet werden. Daher werden in interdisziplinärer und explorativer Arbeit digitale Werkzeuge entworfen, welche eine Analyse großer (digitaler) Korpora ermöglichen. Dieser Entwicklungsprozess implementierte User-zentrierte Methoden, um die Forschungsfragen der Historiker zu unterstützen. Die so bereitgestellten Werkzeuge helfen, semantische Strukturen und Muster in einer Vielzahl von Bildungsmedien des 19. Jahrhunderts zu erkennen. Dies ermöglicht den Historikern einen innovativen Ansatz zur Analyse digitalen Quellenmaterials umzusetzen; vorher musste sich gewissermaßen auf Volltextsuche beschränkt werden. Als Grundstock dazu dienen annähernd 3500 Schulbücher aus dem Bestand des Georg-Eckert-Instituts für Internationale Schulbuchforschung in Braunschweig mit einem Erscheinungsdatum vor 1919. Schulbücher wurden gewählt, da diese zum einen den quasi-offiziellen Diskurs des Kaiserreiches widerspiegeln und es sich zum anderen um eine weitverbreitete und viel rezipierte Quellengattung handelt. Allerdings geraten beim Umfang des Quellenmaterials (momentan über 600.000 Seiten) die klassischen historischen Herangehensweisen schnell an ihre Grenzen. Für die Zukunft des Projektes ist diese Methodenfrage umso bedeutsamer, da der digitale Quellenbestand noch ausgebaut werden wird. Hier wendet nun das Projekt für die Geschichtswissenschaft neue digitale Methoden an, um die Menge an Quellenmaterial bewältigen zu können.

Drei Projektziele spielen für die hiesige Präsentation eine wesentliche Rolle:

1.) Historische Forschung über Repräsentationen und Interpretationen der Welt in der oben stehenden Periode, in der Wissen über die Welt normalerweise nicht durch eigene Anschauung wie Reisen oder durch audiovisuelle Medien gesammelt werden konnte. Daher sind Schulbücher und andere gedruckte Medien die Hauptinformationsquelle für junge Erwachsene.
2.) Die Erforschung eines spezifischen Quellentypus (Schulbücher), die Millionen von späteren Bürgern prägten, die aber bisher noch nicht mit einem Ansatz, der Medientyp, Zirkulation und Wissenstransformation zusammenbringt, untersucht wurden.
3.) Grundlagenforschung in Computerlinguistik; die Entwicklung und Adaptierung verschiedener Methoden der semantischen Analyse und Opinion Mining, die an die Sprache des 19. Jahrhunderts und diesen spezifischen Quellentyp angepasst werden. Wir werden die Daten und Methoden, wie sie bisher im Projekt genutzt wurden, präsentieren sowie Herausforderungen, Erfahrungen und Probleme, die sich im Vorlauf der bisherigen Arbeit ergaben, vorstellen.