Die Geowissenschaftliche Analyse von großen Mengen historischer Texte: Die Visualisierung geographischer Verhältnisse in deutschen Familienzeitschriften

,
pmcisaac@umich.edu
Literature, Sciences and the Arts, Universitaet von Michigan, USA

,
sugih@umich.edu
Electrical Engineering and Computer Science, Universitaet von Michigan, USA

,
iibanez@umich.edu
School of Natural Resources, Universitaet von Michigan, USA

,
oskarsinger@gmail.com
Electrical Engineering and Computer Science, Universitaet von Michigan, USA

,
benrbray@umich.edu
Literature, Sciences and the Arts, Universitaet von Michigan, USA

In diesem Vortrag werden die Verarbeitung und Visualisierung von geowissenschaftlichen Daten in populären, in Deutschland zwischen 1853 und 1918 publizierten Familienzeitschriften wie *Die Gartenlaube*, *Deutsche Rundschau* und *Westermanns Illustrirte Deutsche Monatshefte* präsentiert. Nach einer einleitenden Diskussionen über die Fragestellungen, die eine geowissenschaftliche Analyse dieser Druckerzeugnisse aus älterer und „digitaler“ geisteswissenschaftlicher Sicht motivieren, werden unsere Herangehensweisen erläutert. Neben kuratorischen Aspekten behandelt die Präsentation die von uns entwickelten Techniken des maschinellen Lernens und einer historisierenden Visualisierung, die großen Mengen von historischen Texten gerecht werden. Darüber hinaus werden erste Ergebnisse der Visualisierung gezeigt, die neue Antworten auf noch ungelöste Fragen bieten.

Die Fragestellungen, die diesem DH-Projekt zugrunde liegen, entsprechen manchen zentralen Fragen der älteren geisteswissenschaftlichen Forschung. Diese interessierte sich für die Darstellung spezifischer geographischer Orte und Gebiete zunächst im Zusammenhang mit der Entwicklung einer modernen deutschen Nationalidentität, die als überregional und allen Deutschen gemeinsam verstanden wird (Belgum 1998: xi-xv). Familienzeitschriften befassten sich bekanntlich nicht nur programmatisch mit der Formulierung und Verbreitung der historischen, sprachlichen und geographischen Konturen einer solchen Nationalidentität (Barth 1975: 205-12), sie unternahmen dies als die ersten Druckerzeugnisse, deren Verbreitung ein annähernd nationales Ausmaß annahm (McIsaac 2014: 186-8; Belgum 1998: 1-27). Während ihre relativ erschwinglichen Preise und ihre breit angelegte inhaltliche Thematik ein unerhört zahlenreiches und breites Publikum ansprachen (McIsaac 2014: 186-8; Daum 2002), ermöglichten technische Entwicklungen die zeitgleiche wöchentliche bzw. monatliche Belieferung des gesamten geographischen Gebietes, das als territoriale Basis für Deutschland als politische Nation kritisch in Frage steht (Belgum 1998: 1-27). Innenpolitisch dürften diese Druckerzeugnisse also zum Nationalgefühl im Sinne von Benedict Andersons Begriff der Nation als „vorgestellte Gemeinschaft“ beigetragen haben (Anderson 2006). Zugleich war die Frage nach der geographischen Darstellung aber stets auch eine globale, indem die Familienzeitschriften Deutschlands Rollen als wichtiges Emigrationsland, später dann als aufstrebende Kolonial- und Weltmacht mit gezielten Beiträgen bewusst reflektierten (Belgum 1998: 142-82). Es geht bei diesen Fragen also um die lokalen und globalen territorialen Be-, Ein- und Entgrenzungen in ihrem Verhältnis zum deutschen Nationalgefühl.

Im Zeitalter der Globalisierung und Massenmigration haben diese Fragen nach der nationalen Identität in lokalen und internationalen Kontexten nichts an Brisanz eingebüßt, auch wenn (oder gerade weil) ihre Beantwortung mittels traditioneller Methoden nur in Ansätzen gelungen ist. Dass dies mit herkömmlicher Analyse nicht mehr zu erreichen ist, hängt im großen Maße mit der Fülle an Lesematerial zusammen, der mit normaler Lektüre nicht beizukommen ist (McIsaac 2014: 185). Erst mit der Digitalisierung ganzer Zeitschriftenauflagen, wie dies Google in Zusammenarbeit mit dem US-amerikanischen HathiTrust-Consortium unternommen hat, ist es möglich geworden, mit computerbasierten Techniken an diese Fragen heranzugehen. Diese Techniken bergen insbesondere die Möglichkeit einer kartographischen Visualisierung der geowissenschaftlichen Daten in den Familienblättern in sich, und zwar eine, die das langjährige Erscheinen der Blätter in regelmäßigen Zeitabständen historisch zu verwerten trachtet. In Bezug auf die angestrebte historisierende Visualisierung

geowissenschaftlicher Daten gibt es allerdings technische, finanzielle und methodische Probleme, deren Lösung für große Mengen von historischen Texten weder trivial noch vollkommen ist.

Auf welche Weise diese Probleme sich bewältigen lassen, wird Gegenstand des Vortrags anhand von einem Korpus (*Deutsche Rundschau* 1873-1918) sein. Geschildert werden zunächst Techniken, die nicht nur zur Behebung von Problemen historischer und kuratorischer Natur (z. B. Verbesserung der optischen Zeichenerkennung bei Fraktur; Algorithmen zur passenden Gliederung der Zeitschriften) dienten, sondern auch zur Entwicklung einer skalierbaren Datenbank beitrugen. Diese ist so konzipiert worden, dass Metadaten und Annotationen verschiedener Art mit den jeweiligen Korpora assoziiert werden können und als die Basis für Anwendungen des maschinellen Lernens verwendet werden. Bei diesen Anwendungen geht es uns besonders um eine automatisierte Auflösung von Ortsnamen (eine Form von automated toponym resolution) im Zusammenhang von Named-Entity-Recognition (NER), die die vorkommenden Ortsnamen mit hoher Präzision in großen Mengen von Texten identifizieren. Um unseren Beitrag klarer darzustellen, werden unsere Anwendungen von Methoden und Programmbibliotheken anderer Forschunggruppen (allen voran Statros et al/ Kim; Wing & Baldrige; Speriosu & Baldrige; DeLosier) erläutert.

Um der historischen Spezifizität unserer Texte gerecht zu werden, werden die Ortsnamen aus einer speziell von uns zusammengestellten Datenbank von geokodierten historischen Ortsnamen gespeist (Datenquelle: Mini-Gov Datenbank 2015). Zum Schluss wird mittels eines Open-Source-Plug-Ins der Omeka-Plattform (neatline) eine Visualisierung der geowissenschaftlichen Daten ermöglicht, die nicht nur synchronische geographische Verhältnisse zwischen Zeitschriftentext, Thema und Ort bzw. Region darstellen, sondern auch diachronische. Die Grenzen dieser Methode im Vergleich zu jenen eines GIS-Systems werden kurz besprochen. Somit wird eine solide Basis für die Möglichkeit neuen geisteswissenschaftlichen Wissens gestellt, die dann zuschließend mit ersten Ergebnissen gezeigt wird.

Bibliographie

Anderson, Benedict (2006): *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. New York: Verso.

Barth, Dieter (1975): "Das Familienblatt — ein Phänomen der Unterhaltungspresse des 19. Jahrhunderts. Beispiele zur Gründungs- und Verlagsgeschichte", in: *Archiv für Geschichte des Buchwesens* 15: cols. 205-12.

Belgum, Kirsten (1998): *Popularizing the Nation: Audience, Representation, and the Production of Identity in Die Gartenlaube, 1853-1900*. Omaha: U Nebraska P.

Daum, Andreas (2002): *Wissenschaftspopularisierung im 19. Jahrhundert: bürgerliche Kultur, naturwissenschaftliche Bildung und die deutsche Öffentlichkeit, 1848-1914*. Munich: Oldenbourg.

McIsaac, Peter (2014): "Rethinking Non-Fiction: Distant Reading the Nineteenth-Century Science-Literature Divide," in: Tatlock, Lynne / Erlin, Matt (eds.): *Distant Readings: Topologies of German Literature in the Long Nineteenth Century*. Rochester: Camden House 185-208.

Mini-Gov Datenbank (2015): *Mini-GOV (Genealogisches Ortsverzeichnis)*. Daten des Genealogischen Ortsverzeichnisses GOV , Verein für Computergenealogie e. V. http://wiki-de.genealogy.net/GOV/Mini-GOV [letzter Zugriff 28. Dezember 2015].