Visualisierung von Ortsnamen im Deutschen Textarchiv

,
adrien.barbaresi@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Österreich; Berlin-Branbenburgische Akademie der Wissenschaften, Deutschland

Textarchiv und Umfang der Studie

Das DFG-geförderte Projekt „DeutschesTextarchiv“ (DTA) der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) stellt deutschsprachige Drucke als Bilddigitalisate und TEI-XML-annotierte Volltexte aus mehr als 300 Jahren, vom Beginn des 17. bis zum frühen 20. Jahrhundert, über das Internet zur freien Nutzung bereit. Das DTA mit seinen Erweiterungskorpora umfasst derzeit knapp 2800 Dokumente mit mehr als 630 000 digitalisierten Seiten und ca. 1,1 Mrd. Zeichen (Stand: 7.10.2015). Neben dem Anspruch, vielseitig nutzbare und qualitativ hochwertige Primärquellen frei verfügbar zumachen, liegt der Fokus des DTA-Projekts auf der korpus- bzw. computerlinguistischen Analyse der elektronischen Volltexte. Alle Quellen stehen in verschiedenen Formaten zum Herunterladen und auch zum „Harvesten“ über eine API bereit (BBAW 2007-2015).

Im Rahmen des DTA wurden in den letzten beiden Jahren verschiedene automatische und semiautomatische Ansätze zur Erkennung von Personen- und Ortsnamen evaluiert. Immer wieder bedeuteten dabei die Heterogenität des Korpus und die große sprachliche Varianz innerhalb des Textkorpus eine Herausforderung für die Tools: Je früher die Entstehungszeit eines Textes liegt, desto größer werden sprachliche und sachliche Differenz bei der Benennung von Eigennamen. Der Fokus wird im Folgenden auf Ortsnamen in historischen Texten des DTA liegen.

Das Ziel der Studie besteht darin, die Verteilung der im DTA erwähnten Ortsnamen darzustellen, um ein synthetisches Bild der Sammlung zusammenzustellen und gleichzeitig Rückschlüsse auf den Inhalt zu ermöglichen. Sie erfolgt im Rahmen einer Kooperation zwischen der Berlin-Brandenburgischen Akademie der Wissenschaften (Zentrum Sprache) und der Österreichischen Akademie der Wissenschaften (ICLTT, Institut für Corpuslinguistik und Texttechnologie), beide Zentren verfügen über digitalisierte historische Textkorpora.

Erkennung von Ortsnamen

Spezialisierte Werkzeuge aus dem Gebiet der Computerlinguistik werden im Rahmen dieser Studie eingesetzt. Erstens wird für die Tokenisierung (Segmentierung in Wortformen) die Software WASTE (Jurish / Würzner 2013) benutzt, die speziell für Texte verschiedener Epochen im Rahmen des DTA entwickelt worden ist. So lassen sich Sprachqualitäten besser annähern, die von den heutigen Standards abweichen.

Die deutsche Version des „Wikiwörterbuchs“ Wiktionary der Wikimedia Stiftung, das von Internetnutzern gepflegt wird, wird verwendet, um lexikalische Informationen über Wörtern zu sammeln. Ziel dieses Vorgehens ist es unter anderem, solche Token zu erkennen, die mit Sicherheit keine Eigennamen sind. Ein weiteres mögliches Problem besteht bei Eigennamen, die keine Ortsnamen sind, jedoch aus verschiedenen Gründen als solche ausgezeichnet worden sind, u. a. Namen von bekannten Autoren so wie fiktive Namen und Vornamen. Listen werden also benutzt, um bereits bekannte Eigennamen auszugrenzen.

Die Erkennung von Ortsnamen beruht oft auf Verfahren aus der künstlichen Intelligenz sowie named-entity recognition (Leidner / Lieberman 2011). Wissensbasierte Methoden zeigen jedoch auch vielversprechend Ergebnisse, so wie zum Beispiel anhand von Datenbanken aus Wikipedia (Hu et al. 2014).

Unsere Erkennung der Ortsnamen erfolgt über Datenbanken, die die Vorteile von geisteswissenschaftliche Sorgfalt und Opportunismus aus Big Data Herangehensweisen kombiniert. Über ein gleitendes Fenster wird nach Treffern (einschließlich Mehrwortausdrücken) gesucht. Aus der passenden Datenbank werden Koordinaten und gegebenenfalls weitere geographisch relevante Informationen extrahiert, diese Daten werden wiederum in einer weiteren für das gesamte Verfahren angelegten Datenbank zusammengefasst. Falls mehrere Möglichkeiten bestehen, ist ein Disambiguierungsverfahren nötig, das Informationen wie Distanz, Kontext und aktuelle Bevölkerungszahlen benutzt.

Die Erkennung erfolgt über die Durchsuchung von Listen unterschiedlichen Ranges: als Erstes wird nach aktuellen sowie ehemaligen Ländern und vergleichbaren Hoheitsgebieten gesucht (z. B. Österreich-Ungarn), dann wird die Suchanfrage um Regionen oder regionale Landschaften erweitert (z. B. Schwaben), bei einem negativen Ergebnis wird anschließend nach Städten und schließlich nach geographischen Merkmalen wie Flüssen oder Bergen gesucht. Die dafür nötigen Informationen wurden zum Teil händisch (Staaten und Regionen) und zum Teil automatisch gesammelt und händisch zusammengefasst oder überprüft (Städte und Geographie). Da mancherorts die Staatsgrenzen bis ins 20. Jahrhundert instabil geblieben sind und da gewisse Staaten sich durchaus als multinational verstehen lassen, wurden insbesondere für Mitteleuropa Listen einschließlich der

aktuellen oder ehemaligen deutschen Namen erstellt, u. a. anhand von bereits im Web auffindbaren Listen wie zum Beispiel Kategorien oder Listen von Wikipedia.

Jedem eindeutigen Ortsnamen wurden dann Koordinaten hinzugefügt, entweder durch automatische Abfrage von Wikipedia und Wikidata oder händisch unter Heranziehungen historischer Beschreibungen oder Atlanten. Bei politischen Entitäten wurde bisher Europa im 19. und 20. Jahrhundert in Betracht gezogen. Die Listen werden regelmäßig erweitert, sie umfassen derzeit 78 Hoheitsgebiete, 858 Regionen, 9.846 Städte und 13.962 geographische Merkmale.

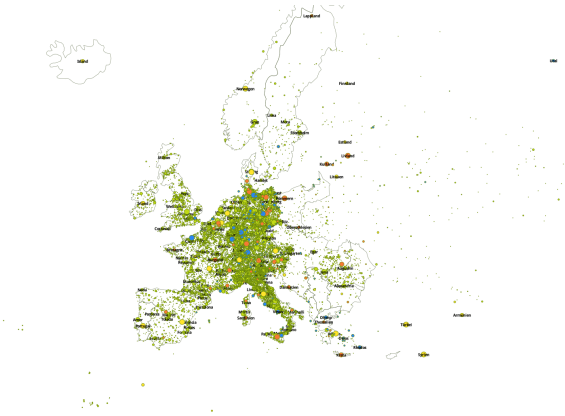
Wenn kein Treffer in den Listen gefunden wird, werden größere, automatisch erstellte Ortsregister in Betracht gezogen. Geographische Informationen über Orte stammen dann aus den Geonames-Datenbanken, die zum Beispiel von dem Openstreetmap Projekt benutzt werden, und dessen Creative Commons Attribution Lizenz eine Wiederverwendung der Daten ermöglicht. Alle Datenbanken für aktuelle europäische Länder sind gesammelt und verarbeitet worden: gewisse Ortstypen (nämlich Region und bewohnter Ort) sind ausgewählt worden, und existierende Varianten in diversen europäischen Alphabeten sind extrahiert worden, um mögliche Änderungen im Laufe der Geschichte zu reflektieren.

Projektion auf einer Karte

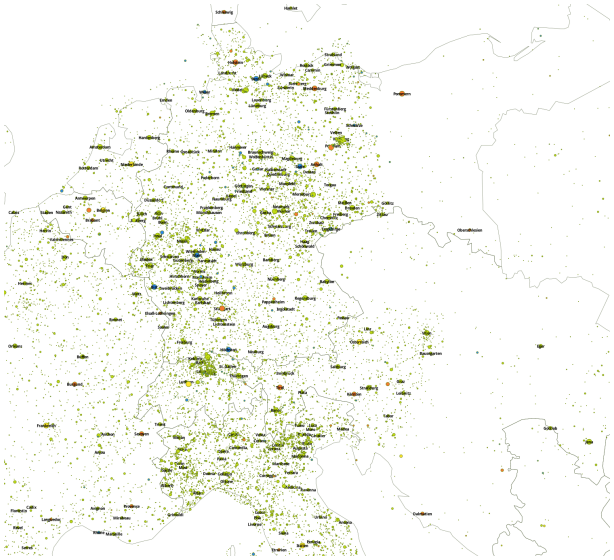
Schließlich werden die Ergebnisse auf eine Karte Europas projiziert, die die tatsächliche politische Lage dieser Zeit spiegelt. Dafür werden die Grenzen von 1914 gezeigt. Der quantitative Schwerpunkt des Korpus liegt nämlich auf dem 19. Jahrhundert liegt und der Stand vor dem ersten Weltkrieg gibt ein vernünftiges Bild von Europa während des 'langen 19. Jahrhunderts'. Die Qualität der Daten so wie des graphischen Resultats wurde in mehreren Durchläufen geprüft, dabei wurden jeweils verbliebene Fehler eliminiert: die Karte bzw. die Projektion der Daten wird so sukzessive verbessert und feiner justiert.

Zur Projektion wird die Kartographieumgebung TileMill benutzt, die eine Anpassung anhand der Stylesheet-Sprache CartoCSS ermöglicht. So können wichtige Punkte im Graphen hervorgehoben werden. Die Wahl verschiedener Farben erleichtert den Überblick über das visualisierte Ergebnis und dessen Interpretation, da im Feld der Visualisierungsstudien bekannt ist, dass das menschliche Auge instinktiv unterscheiden und klassifizieren kann (Bertin 1967).

Karte



Karte von Europa in den Grenzen von 1914 (Hoheitsgebiete in Gelb, Regionen in Orange, Städte sowie aus Geonames extrahierte Ortsnamen in Grün, geographische Merkmale in Blau).



Zoom: Karte von Mitteleuropa in den Grenzen von 1914 (Hoheitsgebiete in Gelb, Regionen in Orange, Städte sowie aus Geonames extrahierte Ortsnamen in Grün, geographische Merkmale in Blau).

Diskussion

Wir hegen die Hoffnung, dass solche Visualisierungsstudien den Weg nach einer größeren Sichtbarkeit von digitalem Kulturerbe und von literarischer Forschung im digitalen Zeitalter ebnen. Genauer betrachtet glauben wir, dass detailreiche Annäherungsweisen gefragt werden, die sowohl auf technischer Kompetenz als auch auf

historisches und literarisches Wissen aufbauen. In diesem Sinne planen wir, mehr Metadaten einzubeziehen sowie vielseitige Visualisierungen zu erzeugen.

Es sollte immer berücksichtigt werden, dass die linguistischen Korpora, die als Basis für die Karte benutzt werden, immer schon ein Konstrukt sind, woraus folgt, dass die auf diesen Daten basierenden Projektionen ebenso Konstrukte sind: Auch wenn sie unmittelbar interpretierbar scheinen, spielen Qualität der Daten, Spezialisierungsgrad der Verarbeitungskette und Qualitätsprüfung eine maßgebende Rolle. Deswegen sind wir der Meinung, dass eine gewisse Dekonstruktion des Prozesses nötig ist, im Sinne einer Öffnung der black box, die dem Betrachter das originelle Moment der Entzückung vielleicht wegnimmt, aus wissenschaftlicher Sicht jedoch wünschenswert ist. So möchten wir keine Fehler kaschieren, eventuelle Verzerrungen nicht verschweigen, und für die Reproduzierbarkeit des gesamten Prozesses sorgen, einerseits durch detailreiche Dokumentierung des Prozesses, andererseits durch die Herausgabe möglichst aller dabei verwendeten Tools und Komponenten als Open Source Software.

So können unmögliche, in diesem Kontext falsche Verbindungen vermieden werden: es gibt zum Beispiel einen Ort namens "Hermann" in Norwegen, was Fakt und Datum zugleich ist. Es ist jedoch nötig, sich nicht allein auf diese Datengrundlage zu verlassen, wenn man nach Orten sucht, sonst wird das Endprodukt – d. h. die Karte – verfälscht.

Interessanterweise bietet die Karte für einen möglichen Hermeneuten genau diese Fälle in Perspektive, die Existenz eines möglicherweise falschen Knotens bleibt nicht unbemerkt und wirft Fragen auf. Auf dieser Weise ist uns zum Beispiel ein systematischer Fehler mit den Vornamen aufgefallen. Durch diese hermeneutische Schleife wird die Analyse nach und nach verschärft.

Die Säuberung der Daten ist in dieser Hinsicht entscheidend, die Anzahl und Vielfalt von Filtern, die eingesetzt werden, erheben unsere Arbeit von einer massiven Datensammlung und -Analyse auf das Niveau einer Studie in Digital Humanities, die Rücksicht auf Besonderheiten einer Sprache und einer Epoche nimmt.

Bibliographie

BBAW (2007-2015): *Deutsches Textarchiv*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften http://www.deutschestextarchiv.de/ [letzter Zugriff 30. Januar 2016].

Bertin, Jacques(1967): *Sémiologie graphique*. Paris: Mouton / Gauthier-Villars.

Hu, Yingjie / Janowicz, Krzysztof / Prasad, Sathya (2014): „Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia“, in: *Proceedings of the 8th ACM Workshop on Geographic Information Retrieval, Dallas, Texas* 8–16.

Jurish, Bryan / Würzner, Kay-Michael (2013): „Word and Sentence Tokenization with Hidden Markov Models“, in: *Journal for Language Technology and Computational Linguistics* 28, 2: 61–83.

Leidner, Jochen L. / Lieberman, Michael D. (2011): „Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language“, in: *SIGSPATIAL Special* 3, 2: 5–11.