Extraction and Visualization of Toponyms in Diachronic Text Corpora

Adrien Barbaresi, Hanno Biber

domestic wares. It produced crocks, jugs, bowls, jars, lids, flower pots, planters, architectural terra cotta and other types of ceramics. Ceramic technology developed rapidly during this period to include semi-mechanized means of forming pots, known as "jigging" or "jollying," which allowed for mass production of vessels. LPW also exhibits the latest innovations in nineteenth-century kiln design in the form of downdraft kilns. When the LPW was founded, Lincoln was a prosperous and rapidly growing city with a population of c. 13,000 in 1880 which had increased to 55,000 in 1890 (Bleed and Schoen, 1990: 34). After the factory closed, most of the property became part of a housing development. The site was excavated by Peter Bleed in 1986-1987, and the results were published in 1993. The LPW collections are currently in the Nebraska State Historical Society.

This is a pilot project and a work in progress. The digital recording of representative types of LPW ceramics through 3D interactive models is the first step in the creation of an online exhibit and digital resource. Currently, we are documenting the most common types of ceramics produced by LPW. The LPW collection is extremely large. For example, the estimated number of bowls is 5,633 vessels and represent 38.45% of the assemblage. Thus, our goal is to document a representative sample of the collection, we do not plan to document the collection fully. Once the pilot phase is completed, we will make the 3D models, along with other content, available on a website (in collaboration with the Nebraska State Historical Society).

The digitization project will facilitate different kinds of analysis that build on the original publication of the results; for example, the distribution of LPW products in other Midwestern states (the records of the business have not been preserved), socioeconomic aspects, food ways, a gendered perspective, etc. The first step, however, is to create a digital resource that will draw attention to this collection and engage other researchers and the public. By utilizing the latest technology to create and present accurate models of LPW representative products, we hope to bring wider attention to this important assemblage, which is an integral part of the history of nineteenth century Nebraska and the industrial archaeology of the Great Plains region.

Bibliography

Bleed, P. and Schoen, Ch. (1990). The Lincoln Pottery Works: A Historical Perspective. *Nebraska History*, **71**: 34-44.

Schoen, Ch. and Bleed, P. (1993). The Archaeology of the Lincoln Pottery Works. *Central Plains Archaeology*, **3**(1): 1-240.

Extraction and Visualization of Toponyms in Diachronic Text Corpora

Adrien Barbaresi
adrien.barbaresi@oeaw.ac.at
Austrian Academy of Sciences, Austria; Berlin-Brandenburg Academy of Sciences, Germany

Hanno Biber
hanno.biber@oeaw.ac.at
Austrian Academy of Sciences, Austria

Introduction

This paper focuses on the extraction of German and Austrian place names in historical texts. It is part of a cooperation between the Berlin-Brandenburg and the Austrian Academies of Sciences. The latter is the holder of the text basis for this investigation, the digitized version of the satirical literary magazine "Die Fackel" ("The Torch"). It has been originally published and almost entirely written by the satirist and language critic Karl Kraus in Vienna from 1899 until 1936, and contains a considerable variety of toponyms (Biber, 2001).

Gazetteers and lists

Toponym resolution often relies on named-entity recognition and artificial intelligence (Leidner and Lieberman, 2011). However, knowledge-based methods using fine-grained data – for example from Wikipedia – have already been used with encouraging results (Hu et al., 2014).During the 20th century there have been significant political changes in Central Europe that have severely affected toponyms, so that geographical databases lack coverage and detail. Consequently, the database we develop follows from a combination of approaches: gazetteers are curated in a supervised way to account for historical differences,and current geographical information is used as a fallback.

First, a cascade of filtersis used: (1) current and historical states (e.g. Austria-Hungary); (2) regions, important subparts of states, and regional landscapes (e.g. Swabia); (3) populated places; (4) geographical features (e.g. valleys). Wikipedia's API is used to navigate in categories and to retrieve coordinates, which are completed by hand for states and regions.Second, current information is also compiled from the Geonames database[1]: data for European countries are retrieved and preprocessed(variants and place types).

In order to exclude common and proper nouns, the German version of the Wiktionary serves as a reference[2], and registers of frequent surnamesand family names, as well as well-known persons (especially writers) are built using Wikipedia and Wikidata.

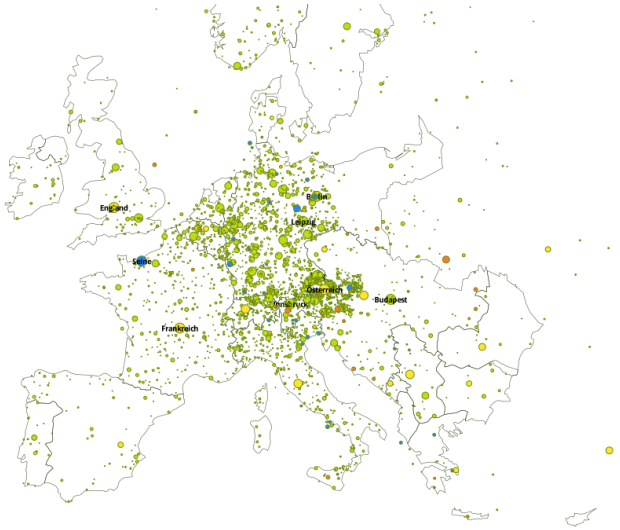
Extraction

The texts were digitized, manually corrected as well as manually annotated with respect to the names of persons and institutions, so that most proper nouns which are not place names can be excluded from the study.

The tokenized files of works to be analyzed (Jurish and Würzner, 2013) are filtered and matched with the databaseby finite-state automatons: toponyms are extracted using a sliding window (for multi-word names up to three components). Disambiguation being a critical component (Leetaru, 2012), an algorithm similar to Pouliquen et al. (2006), who demonstrated that an acceptable precision can be reached that way,guesses the most probable entry based on distanceto Vienna (Sinnott, 1984), contextual information(closest-country, last names resolved), and importance (place type, population count).

Visualization

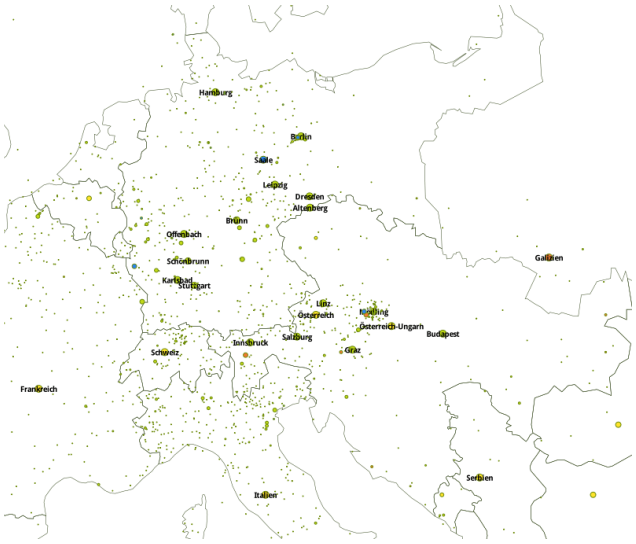
The results are projected on a map of Europe with boundaries of 1914[3] using TileMill[4]. They are customized with CartoCSS: multiple trial-and-error iterations are performed concerning both data quality and graphical output. The two experimental maps belowground on the same data, they result from different settings.



Map 1– Experiment on European scale with boundaries of 1914 (yellow: sovereign territories; orange: regions; green: populated places; blue: geographical features)

Discussion

Potential conceptual caveats include previous times as well as fictitious places, especially names which can refer to mythological and actual places of Ancient Greece or Rome. Practical caveats are for instance false localizations due to disambiguation errors (e.g. Brünn/Brno on map 2). We plan to bypass the disambiguation for a hand-picked list of places. As big datais an entanglement of implicit theoretical assumptions (Crawford et al., 2014), the difference between a mere data collection project and a digital humanities study resides precisely in the number and diversity of filters used. The code and listings produced for this study are available online.[5] We plan to integrate corpora of greater variety and scope and to include more specific metadata in order to design versatile visualizations.



Map 2– Central Europe, experiment with a restrictive filtering

Conclusion

A map is a discrimination, a marking of difference (Wulfman, 2014): our maps highlight the linguistic and cultural ties of Kraus and his contemporaries with Bohemia and Northern Italy, where there are more numerous toponyms to be found than in Hungary.Beyond that, "Die Fackel" is (at least) a European phenomenon; Kraus' vision of Europe is more inclined towards cultural centers (Prague, Munich, Paris, Berlin).

It is our hope that visualization studies based upon mixed methods contribute to a greater awareness of the potential of digital heritage as well as literary studies in the digital age. Although the maps seem immediately interpretable, they are not an objective result but a construct (Juvan, 2015), the result of a filtering. The "human" interventions on the map as well as the technical competence to do so replace this study in the hermeneutic circle of the philological tradition.

Bibliography

AAC-FACKEL. Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936. In Biber, H., Breiteneder, E., Kabas, H., Mörth, K.; Graphic Design: Burdick, A.(eds), *AAC Digital Edition 1*, http://www.aac.ac.at/fackel.

Biber, H. (2001). In Wien, in Prag und infolgedessen in Berlin - Ortskonstellationen in der "Fackel". In Marten-Finnis, S., Uecker, M. (ed.) *Berlin-Wien-Prag. Moderne, Minderheiten und Migration in der Zwischenkriegszeit*, Peter Lang, pp. 15-26.

Crawford, K., Gray, M. and Miltner, K. (2014). Big Data | Critiquing Big Data: Politics, Ethics, Epistemology | Special Section Introduction. *International Journal of Communication*, **8**: 1663–72.

Hu, Y., Janowicz, K. and Prasad, S. (2014). Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia. *Proceedings of the 8th Workshop on Geographic Information Retrieval*, ACM, pp. 8-16.

Juvan, M. (2015). From Spatial Turn to GIS-Mapping of Literary Cultures. *European Review* **23**(1): 81-96.

Jurish, B. and Würzner, K.-M. (2013). Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, **28**(2): 61–83.

Leidner, J. L. and Lieberman, M. D. (2011). Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language. *SIGSPATIAL Special*, **3**(2): 5-11.

Pouliquen, B., Kimler, M., Steinberger, R., Ignat, C., Oellinger, T., Blackler, K., Fluart, F., Zaghouani, W., Widiger, A., Forslund, A.-C. and Best, C. (2006). Geocoding multilingual texts: Recognition, disambiguation and visualisation. *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 53-58.

Sinnott, R. W. (1984). Virtues of the Haversine. *Sky and Telescope* **68**(2): 159.

Wulfman, C. E. (2014). The Plot of the Plot: Graphs and Visualizations. *The Journal of Modern Periodical Studies*, **5**(1): 94-109.

Notes

[1] http://www.geonames.org/

[2] Thanks to Kay-Michael Würzner (BBAW) for his extraction script.

[3] http://dev2.dariah.eu/geoserver/web/

[4] https://www.mapbox.com/tilemill/

[5] https://github.com/adbar/toponyms

Named Entity Extraction from digitized texts of Mongolian Historical Documents in Traditional Mongolian Script

Biligsaikhan Batjargal
biligsaikhan@gmail.com
Research Organization of Science and Engineering,
Ritsumeikan University, Japan

Garmaabazar Khaltarkhuu
garmaabazar@gmail.com
Mongolia-Japan Center for Human Resources Development,
Mongolia

Akira Maeda
amaeda@media.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

In this paper, we demonstrate a named entity extraction method for digitized ancient Mongolian documents by using features of traditional Mongolian script. In the field of humanities, getting knowledge by analyzing various historical documents is an important task. There are increasing demands from Mongolian humanities researchers to perform text analysis at massive scale with prompt and accurate results. A few ancient Mongolian historical manuscripts including 1) the "Qad-un ündüsün-ü quriyangγui altan tobči neretü sudur (The Altan Tobchi or the Golden Summary: Short history of the Origins of the Khans)" a.k.a "Little" Altan Tobchi, and 2) the "Asaraġci neretü-yin teüke or Asragch nėrtiin tüükh (The Story of Asragch)", which were written in traditional Mongolian script have been converted to digital texts and made publicly available through the traditional Mongolian script digital library (TMSDL) (Batjargal et al., 2013). Figure 1 shows a page of the "Little" Altan Tobchi in the TMSDL. The demands from Mongolian humanities researchers, as well as the lessons learned from the TMSDL have encouraged us to conduct further research in developing a new method for extracting named entities from ancient Mongolian historical documents. However, there has been little research on text mining or named entity extraction for Mongolian language and none of the research has considered text mining on ancient Mongolian historical documents due to the lack of research in those areas. Thus, we want to propose a named entity extraction method for ancient historical documents in traditional Mongolian script by employing machine learning techniques for aiming to reduce the labor-intensive analysis on historical text.

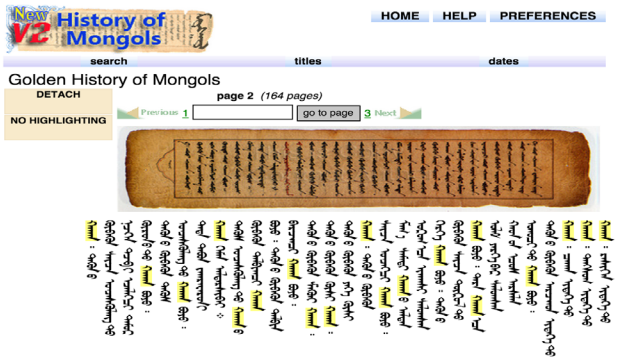


Figure 1. Screenshot of the TMSDL

In the proposed approach, an ancient Mongolian corpus gets tokenized, each token gets annotated and gold standard annotations are prepared for inputting into computer system for learning. The proposed method learns the extraction rules of personal names and place names from annotated training corpora, and then extracts named entities from ancient Mongolian texts by employing machine learning techniques (Batjargal et al., 2015).

We use the IOB2 (Ramshaw and Marcus, 1995) format for tagging tokens. Because of some unique features of