Extraction and Visualization of Toponyms in Diachronic Text Corpora

Adrien Barbaresi [12] & Hanno Biber [1]

[1] Austrian Academy of Sciences – [2] Berlin-Brandenburg Academy of Sciences

Context and text base

Work in progress between Austrian and Berlin-Brandenburg Academies of Sciences:

- Both centers provide digitalized historical corpora.
- *Academiae Corpora* (Austrian Academy of Sciences) is right-holder of the present study.
- The method is transferable to other diachronic corpora (Barbaresi, 2016).

- The AAC-FACKEL has been developed within the Austrian Academy Corpus project;
- Magazine originally published and almost entirely written by the satirist and language critic Karl Kraus in Vienna from 1899 to 1936;
- Free online access to 37 volumes, 415 issues, 922 numbers, comprising more than 22.500 pages and 6 million wordforms.
- Manually screened OCR and tokenization processes, and manual annotation of names of persons and institutions; most proper nouns which are not place names can be excluded.

Objectives

Current geographical databases, particularly gazetteers, lack coverage and detail due to extensive changes of names during the time of publication.

⇒ Objectives: Visualize the distribution of place names in *The Torch*, provide a general view of the magazine, and allow for conclusions on its content.

Method

Toponym resolution often relies on named-entity recognition and artificial intelligence (Leidner & Lieberman, 2011). However, knowledge-based methods using fine-grained data for example from Wikipedia have already been used with encouraging results (Hu, Janowicz, & Prasad, 2014).

→ Toponyms are extracted using a sliding window (for multi-word expressions up to three components).
The database we develop follows from a combination of approaches:

1. Gazetteers are curated in a supervised way to account for historical differences
2. Current geographical information is used as a fallback

- Place names with coordinates

  Geography (e.g. *mittelländisches Meer*)
  curated manually

  States (e.g. Austria-Hungary)
  curated manually

  Regions/Landscapes (e.g. Swabia)
  supervised curation (Wikipedia categories)

  Städte (e.g. Allenstein, Tschenstochau)
  supervised curation (gazetteers, Wikipedia, Wikidata)

  Geographical features (e.g. valleys or rivers)
  automatically gathered (categories on Wikipedia)

  Reference database Geonames (e.g. used by OpenStreetMap)
  comprehensive but also error-prone

- Since disambiguation is a major task (Leetaru, 2012), a combination of several indicators is used (Pouliquen et al., 2006):

  Distance to Vienna (Sinnott, 1984)

  Context last cities and countries seen

  Metadata place type, number of inhabitants

Visualization

TSV-data projected with TileMill and customized with CartoCSS:

Territories, regions, curated cities and data from Geonames, geographical features

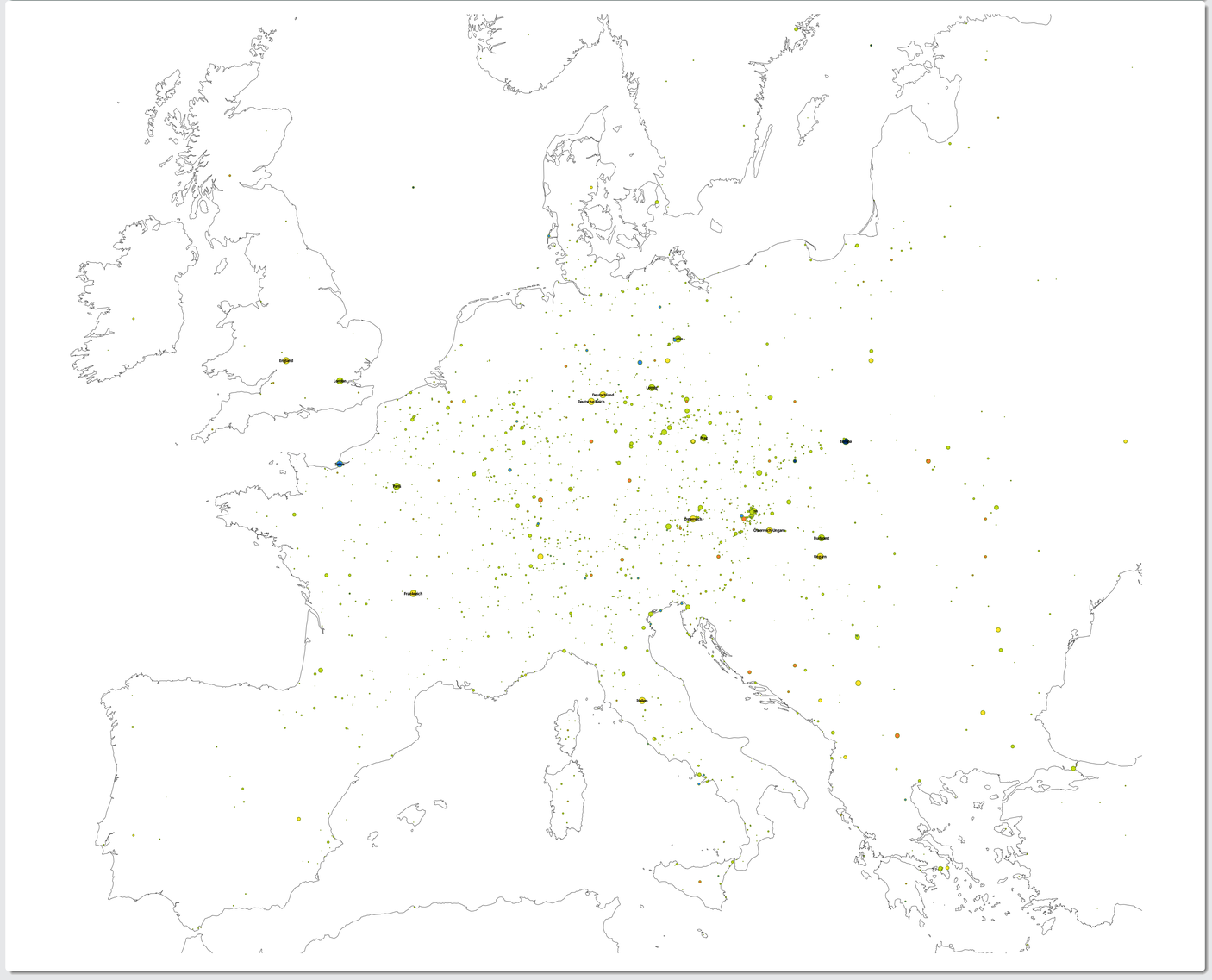
---

The human eye can make out differences and categories instinctively (Bertin, 1967), but a visualization stays a construct (Juvan, 2015) and a marking of difference (Wulfman, 2014).

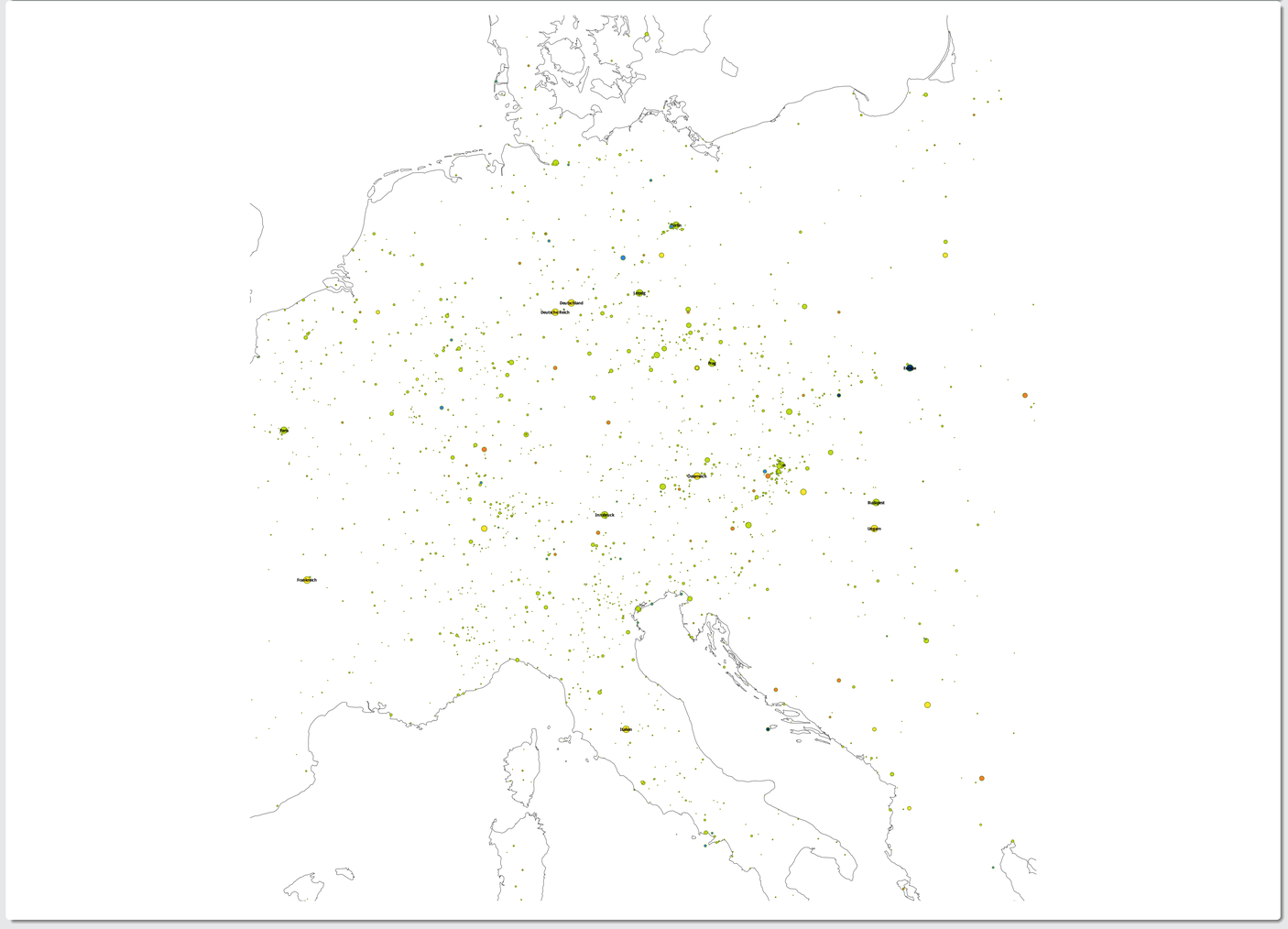
The statements behind the maps:

- hide neither errors nor potential distortions
- ensure the reproducibility of processes

Europe – restrictive filtering



Central Europe – opportunistic setting



Future steps

Several iterations needed in order to improve and fine-tune extraction and projection:

- include more metadata
- try different visualization techniques
- problem solving
  - person names
  - mythical and mythological places
  - fictional toponyms
  - representation from *Topoi*
  - evaluation of disambiguation processes

⇒ Technological expertise as well as critical knowledge in history and literature is called for!

Sources and additional information

Barbaresi, A. (2016). Visualisierung von Ortsnamen im Deutschen Textarchiv. In *Proceedings of DHd 2016* (pp. 264–267).
Bertin, J. (1967). *Sémiologie graphique : les diagrammes, les réseaux, les cartes*. Paris: Bordas.
Hu, Y., Janowicz, K., & Prasad, S. (2014). Improving Wikipedia-Based Place Name Disambiguation in Short Texts Using Structured Data from Dbpedia. In *Proceedings of the 8th workshop on geographic information retrieval* (pp. 8–16).
Juvan, M. (2015). From Spatial Turn to GIS-Mapping of Literary Cultures. *European Review*, *23*(1), 81–96.
Leetaru, K. H. (2012). Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched Wikipedia. *D-Lib Magazine*, *18*(9), 5.
Leidner, J. L., & 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., ... others (2006). Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC)* (pp. 53–58).
Sinnott, R. (1984). Virtues of the Haversine. *Sky and Telescope*, *68*(2), 158–159.
Wulfman, C. E. (2014). The Plot of the Plot: Graphs and Visualizations. *The Journal of Modern Periodical Studies*, *5*(1), 94–109.

Code on GitHub: `https://github.com/adbar/toponyms/`

TileMill: `https://www.mapbox.com/tilemill/`

Geonames: `http://www.geonames.org/`

`http://www.aac.ac.at/fackel`

Thanks to Logan Pecinovsky (BBAW) and Judith Brottrager (OEAW) for helping with gazetteer curation.

Austrian Academy of Sciences (*Academiae Corpora*)
Sonnenfelsgasse 19 – A-1010 Vienna

Berlin-Brandenburg Academy of Sciences (*Language Center*)
Jägerstr. 22/23 – D-10117 Berlin

adrien.barbaresi@oeaw.ac.at
hanno.biber@oeaw.ac.at