Extraction and Visualization of Toponyms in Diachronic Text Corpora

Adrien Barbaresi 1,2 & 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.
- Austrian Academy of Sciences (Academicia Corpora) 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 1916;
- Free online access to 37 volumes, 415 issues, 922 numbers, comprising more than 22,500 pages and 6 million wordsforms;
- 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.

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. Gazetters are curated in a supervised way to account for historical differences
2. Current geographical information is used as a fallback
   - Places names with coordinates
     Geography (e.g. mittelöstliches Meer)
     curated manually
   - States (e.g. Austria-Hungary)
     curated manually
   - Regions/Landscapes (e.g. Swabia)
     supervised curation (Wikidata categories)
   - Städte (e.g. Allenstein, Tucherschach)
     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 (Leitner, 2012), a combination of several indicators is used (Pouliquen et al., 2008).

Objectives
Current geographical databases, particularly gazetters, 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.

Central Europe – opportunistic setting
Place names with coordinates
- horizon/outline: 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

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
  - disambiguation from Wikipedia
  - 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). Toponymic data extraction and gazetteer curation. Proceedings of the 8th workshop on geographic information retrieval (pp. 8–16).


Czech Academy of Sciences, Institute of Computer Science. (1994). Geonames. 1.0.3.


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


Source code 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.