HAL
archives-ouvertes.fr

Placenames analysis in historical texts: tools, risks and side effects

Adrien Barbaresi

▶ To cite this version:

Adrien Barbaresi. Placenames analysis in historical texts: tools, risks and side effects. Corpus-based Research in the Humanities, Jan 2018, Vienna, Austria. hal-01775119

HAL Id: hal-01775119
https://hal.archives-ouvertes.fr/hal-01775119

Submitted on 24 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Placenames analysis in historical texts: tools, risks and side effects

Adrien Barbaresi

Academy Corpora
Austrian Academy of Sciences
E-mail: adrien.barbaresi@oeaw.ac.at

Abstract

This article presents an approach combining linguistic analysis, geographic information retrieval and visualization in order to go from toponym extraction in historical texts to projection on customizable maps. The toolkit is released under an open source license, it features bootstrapping options, geocoding and disambiguation algorithms, as well as cartographic processing. The software setting is designed to be adaptable to various historical contexts, it can be extended by further automatically processed or user-curated gazetteers, used directly on texts or plugged-in on a larger processing pipeline. I provide an example of the issues raised by generic extraction and show the benefits of integrated knowledge-based approach, data cleaning and filtering.

1 Introduction

In Western tradition, a current of reflexion whose origin can be dated back to the 1960s has provided the theoretical foundations of the spatial turn, whose epitome is the concept of space as emergent rather than existing a priori, and composed of relations rather than structures. As a consequence, both the definition and the importance of space have been re-evaluated throughout the humanities. More recently, researchers have suggested the crossing of research objects between disciplines and the enforcement of the "spatial turn" in practice through specific methods of analysis. Even so, corpus linguistics and geographical information systems have traditionally had very little to do with each other, although both approaches can benefit from each other [13].

Distant reading practitioners employ computational techniques to mine the texts for significant patterns and make statements about them [29]. From the point of view of computational linguistics, toponyms are a particular kind of out-of-vocabulary tokens. On lexical level, they are a potential error source for natural language processing tools. On phrasal level, they are supposed to be identified by part-of-speech taggers as named entities or eventually by more fine-grained

named-entity recognition tools as placenames. The processing chains usually stop at this point, they do not provide visualizations in the geographical sense, even if the toponyms can be linked to meta-information such as type and georeference and although progresses in fulltext geocoding are tightly linked to progresses in mapping systems, mostly thanks to a technology-driven evolution [17]. On the other hand, publicly available geocoding solutions do not usually come with interfaces to linguistic methods such as disambiguation and/or annotation layers. Finally, existing cartographic software solutions are not typically built for the visualization of digital text collections.

This article summarizes issues related to historical texts and describes an effort to conveniently go from texts to maps by integrating several key steps in a modular software package: data curation and preparation, processing of linguistic corpora, geocoding, and projection on maps. The use of a toolkit creates a common ground for hypothesis testing and visualization, while at the same time being compatible with other software in terms of formats and software environment. I provide an example of the issues raised by generic extraction and show the benefits of integrated knowledge-based approach, data cleaning and filtering.

2 Previous work

Among the tendencies in geographic information retrieval and geocoding [20], the extraction and normalization of named places, itineraries, or qualitative spatial relations, as well as the extraction of locative expressions are particularly relevant to study text collections. In the field of information retrieval, named entity recognition defines a set of text mining techniques designed to discover named entities, connections and the types of relations between them. The particular task of finding placenames in texts is commonly named placenames extraction or toponym resolution. It involves first the detection of words and phrases that may potentially be proper nouns and second their classification as geographic references [21]. A further step, geocoding, resides in disambiguating and adding geographical coordinates to a placename. Geocoding mostly relies on gazetteers, i.e. geospatial dictionaries of geographic names, mostly names, locations, and metadata such as typological information, variants or dates [15]. Toponym resolution as well as named-entity recognition can use machine learning methods [18], however these are generally not ideal when tackling data not present in the training set, so that knowledge-based methods using additional fine-grained registers, for example from Wikidata [28], have already been used with encouraging results.

Especially for historical corpora, researchers face a lack of general-purpose tooling. In order to produce cartographic visualizations, both the capacity to adapt to different contexts [3] and the necessity to complement existing resources with a precise historical gazetteer [9] have been highlighted. Such historical gazetteers exist, but their development is challenging [26] even for texts as late as 20th century Europe [22]. Existing toolboxes, such as AATOS [27], mostly feature candidate

extraction and ranking as well as entity linking. HeidelPlace [24] does implement a comparable series of operations but it is currently tied to a series of engineering decisions which do not make its use on historical corpora straightforward. My approach is more light-weight, modular and adaptable, with a similar scope as CORE [19] but with an overall greater focus on usability, texts as input, integration of registers, and map export as images.

3 Tooling

3.1 Requirements

In order to process linguistically annotated text, it is useful to be able to start from either raw text or common formats for part-of speech tags and named entities recognition. The toolkit is pluggable to existing NLP solutions or usable directly on text, although morpho-syntactical analyses are appropriate in order to narrow down the search to relevant tokens, such as phrase heads found by surface parsing [4]. Gazetteers can be curated in a semi-supervised way to account for historical differences. Knowledge-based techniques are a way to tailor the geoparsing to historical contexts. Nevertheless, bootstrapping geographical data can save a significant amount of time. The generic gazetteer GeoNames [12] and structured data from Wikipedia and Wikidata are widely known to the research community. Wikipedia's API can be used to navigate in categories and to retrieve coordinates, including for historical places or areas. Current information is usually compiled from the GeoNames database, which also often includes historical variants. Additional lexical cues like stoplists or linguistic information such as suffixes or derivation ought to be configurable, as tools trained on modern texts do not necessarily tag historical morpho-syntactic patterns as needed. To provide support for manual annotation, an additional layer can be convenient as a geocoding bypass for targeted user lists which can operate on token or lemma level (using either linguistic processing or regular expressions and wildcards).

3.2 Concrete approach

The toolbox used for the experiments below is currently being developed [5] with historical texts in mind. It has already been used so far to map different text collections ranging from the 17th to the 20th century [6].

There is no commonly adopted standard for gazetteers, they have to be combined. Consequently, my approach allows for additional input, special sorting and prioritized merging, for example to put historical variants in the foreground. Second, it includes helpers to bootstrap geographical data, as knowledge-based methods using fine-grained data improve the results [28]. So far, import filters for GeoNames and structured data from Wikipedia and Wikidata are implemented, with a particular emphasis on data cleaning. Third, an additional layer allows to bypass geocoding for targeted, easily extensible user lists which can operate on

token or lemma level (using either linguistic processing or regular expressions and wildcards).

To spare resources, the extraction is performed by a sliding window capturing single tokens as well as multi-word expressions. Two different types of disambiguation methods [10] are included so far in the toolbox: map-based and knowledge-based. It has been shown that an acceptable precision can be reached by including meta-information [23], which consists here in distance (based on a calculation relative to a contextual setting), type and importance of the entries (as known from information extracted from GeoNames or Wikipedia), as well as immediate context (e.g. the expected range and the last country seen). The process can be controlled by parameters set by the user, such as distance calculations, filter level or size of the search area.

Additionally, the toolbox integrates its own visualization component[1] which makes use of the Python module *matplotlib* and its extension *cartopy*. It is profitable to allow for adaptability of projection and design and to leave it open to the user to refine the map, in a particular emphasis on the concept of visualization.

The toolkit is bundled as Python package, currently one of the most frequently used programming languages in academia,[2] it is available under an open-source license.[3] The release includes the code, especially the functions dedicated to geographic information retrieval, which form the bases of previous studies. It is meant to ensure replicability and extendability in an open science perspective and can hopefully respond to a growing demand in this field.

3.3 Contextual settings

The streamlined process from text to map involves a series of decisions as well as a critical reading of texts and maps. As user-definable settings make results vary, experiments can lead to diverging realizations. In fact, the extraction and visualization settings have a significant influence. In order to make them easily configurable, they are all accessible in a settings file. First and foremost, the filtering level affects both the construction of gazetteers prior to geoparsing and the toponym recognition phase in itself. Its purpose is to allow for a tighter or looser control on the data, with either restricted options or opportunistic search. Second, the minimum length of tokens to consider as valid toponyms, which is a function of the frequency, can be ignored or determined in advanced. Third, the disambiguation phase can be controlled by map-based parameters, notably the reference point for distance calculations and the countries in the vicinity, which help identifying the most probable candidates. Last, the cartographic processing in itself can be configured (window size and labels). Altogether, the settings allow for an opportune handling of historical texts. The process can adapt to different texts and contexts and it can evolve to

[1] The software used in previous experiments (TileMill) is no more under active development and needed to be installed and used separately.

[2] https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages

[3] https://github.com/adbar/geokelone

reflect historical empires or regions for example, both during geoparsing (account for and disambiguate among historical names) and mapping (display historical or canonical names).

4 Risks and side effects

4.1 Examples

In order to better assess the impact of filtering and complementary registers, I present and discuss two different comparisons on close reading and on distant reading levels. Specially curated gazetteers are used, while current geographical information is used as a fallback, entries corresponding to European countries are retrieved and preprocessed.

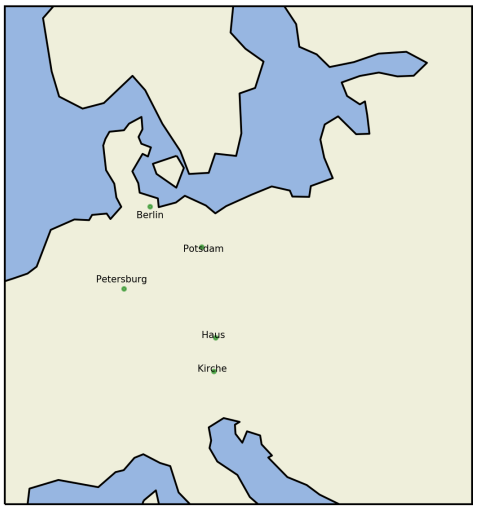

Figure 1: No filter, standard GeoNames setting

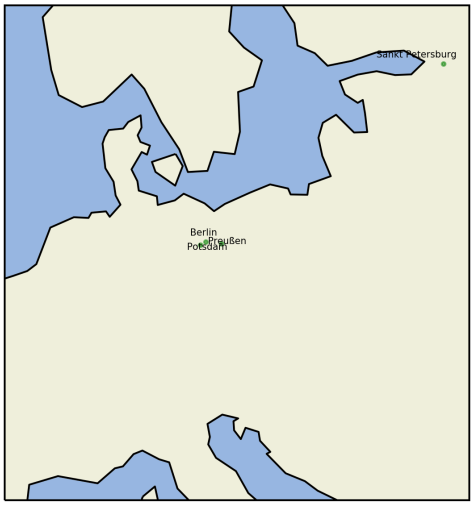

Figure 2: Cleaned data with meta-information

First, I test the coverage and the options at close level on a simple historical example. The sentence to be analyzed is from the late 19th century and features a series of proper nouns so that the experimental setting has an effect both on both form and content.[4] The standard fallback gazetteer, GeoNames, is known to be prone to coverage and data quality issues [2]. Figure 1 displays an unfiltered view using raw text and GeoNames as only gazetteer. Only one point out of five is placed correctly while two other are wrongly considered to be placenames, and one place name is missing. The most prominent error concerns the token *Berlin*, which in GeoNames corresponds to a settlement in Northern Germany without inhabitants. The capital

[4]Taken from *Der Stechlin* by Theodor Fontane: *"Ich sage Ihnen, Hauptmann, das waren Preußens beste Tage, als da bei Potsdam herum die 'russische Kirche' und das 'russische Haus' gebaut wurden, und als es immer hin und her ging zwischen Berlin und Petersburg."*

city of Germany is indeed never present as a single token in the dataset but systematically in the form of city quarters such as *Berlin-Alexanderplatz*. Figure 2 shows the impact of filtering (both knowledge-based and POS-based filtering lead here to the removal of false positives) and external resources (proper geocoding with a historical gazetteer). which lead to correct results when used in combination. This simple example illustrates how quality control and text analysis can benefit from the projection of the results on a map.

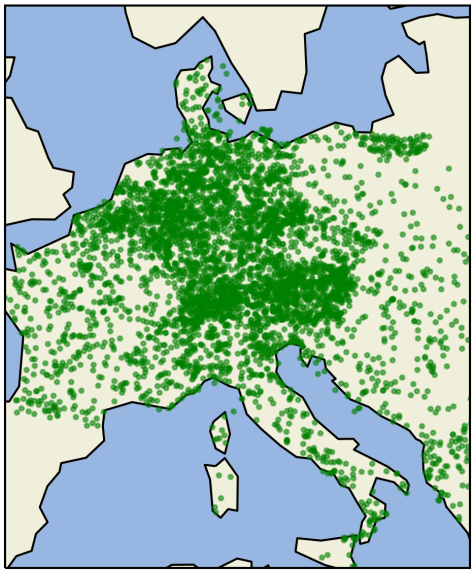

Figure 3: Minimum filtering

Figure 4: Maximum filtering

Next, the impact of filtering methods on distant reading experiments is shown. Karl Kraus (1874-1936) founded his own journal, *Die Fackel* ("The Torch"), in 1899 and published it until his death. This complex and unique work resists summary description in its whole and in detail, it has been used as a basis for distant reading experiments using placenames and collocations as entry points to provide an additional, synthetic overview of the collection [7]. The present experiments use the same text base from the digital edition of the work [8], the texts have been manually corrected as well as manually annotated with respect to the names of persons and institutions, so that most proper nouns which are not placenames can be excluded from the study. Figure 3 displays the results with a minimum filtering on a map showing most of continental Europe. Clusters can be found everywhere, not all of them being either intuitively explainable or justified with respect to the texts. In fact, the map tells more about the gazetteers used for geoparsing as about the work in itself. Current boundaries are retraceable, and numerous false positives come from plurilingual countries such as Switzerland or Belgium which are then overrepresented on the map. Figure 4 consists of a similar map featuring the results of maximum filtering level both during the construction of resources and

during the extraction process. The map is more easily readable and depicts an accurate centering on Vienna and its surroundings. The overall Westward tropism of the mapped locations seems to coincide with the texts. This map is thus well-suited for further analyses.

4.2 Discussion

GeoNames, arguably the most commonly used gazetteer, has to be put under scrutiny, as the entries and their classification are subject to numerous problems, mostly unevenly distributed data and sparse metadata, which impact both detection and disambiguation of placenames [1]. Nevertheless, this resource is still valuable mostly because of its coverage of language variants and thus potentially historical variants.

The status of placenames that are to be found and projected on the map also ought to be discussed. There are consubstantial ambiguities on linguistic level that complicate the search [25]: the referent ambiguity (one name used for more than one location) and the referent class ambiguity (placenames used as organization or person names) are commonly addressed by disambiguation processes, whereas reference ambiguity (more than one name for the same location) has to be dealt with during the compilation of geographical databases. In general, successful detection and disambiguation relies on a smart interplay of resources and tools at different levels. Last, the case of either imprecise, vague or vernacular names [16] is a prominently linguistic issue which can at least be addressed by manual curation and should in any case be attended to.

Concerning the maps themselves, the consensus in the research community has evolved towards a relativity in construction and uses of maps, as there is neither a ground truth nor a cartographic truth. Although the maps seem immediately interpretable, they are not an objective outcome but a construct resulting from a series of interventions. “Selection, omission, simplification, classification, the creation of hierarchies, and 'symbolization' – are all inherently rhetorical" [14]. As such, cartography is not the realization of static maps, but rather the description of emergent structures, and there is no single or best map.

Finally, the object of scientific inquiry does not simply reside in linking text to space, it is tightly linked to the interpretation of texts and maps. Even if the methodology conveys a feeling of scientific objectivity, the validity of mental and computerized operations described here should always be examined with respect to their relevance. Geospatial analysis and spatial representation may indeed be deficient or inadequate. The anthropological significance of toponyms has been emphasized by testimonies gathered on the field [11], but the symbolic role and the expressive power of placenames do not necessarily coincide with Western instrumental science and cartography, in that particular case the world geodetic system and the chosen map projection.

5 Conclusion

This article introduced theoretical and practical instruments combining philological knowledge, geographic information retrieval and visualization, in order to streamline the steps needed to go from texts to maps. Examples of the issues raised by generic extraction have been discussed, they show the advantages of a methodology centered on historical texts and subsequent data cleaning and filtering. Being able to go through all the operation in one shot is ideal to assess the risks, to spot problems in methodology or datasets, and hopefully to mitigate the side effects.

The maps play an ambiguous role in distant reading, since they have to be flexible enough to adapt to new contexts and analyses, while remaining exact and in this regard trustworthy. The information they contain and reveal cannot always be verified on a point-per-point basis, yet it can be the starting ground of scientific reasoning. In fact, text visualizations are the substrate of interpretable representations which do not follow data but rather confront them by putting them in perspective. The difference between data wrangling and research in digital humanities resides precisely in the number and diversity of conceptual and technical filters which are repeatedly applied, consciously or sometimes unknowingly. The chosen approach and its inevitable imperfections have to be brought to light, documented and criticized.

In a linguistic perspective, the tools allow for the systematization of research as well as for a critical approach to the extraction and the very concept of placenames. As quantitative and qualitative analysis can go hand in hand, digital literary studies are not mere numeric accounts. They are first and foremost a discovery process. The use of filtering, the customized gazeteers and maps, in short the human interventions as well as the technical competence to do so recreate the hermeneutic circle of the philological tradition.

References

[1] Elise Acheson, Stefano De Sabbata, and Ross S. Purves. A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320, 2017.

[2] Dirk Ahlers. Assessment of the Accuracy of Geonames Gazetteer Data. In *Proceedings of the 7th Workshop on GIR*, pages 74–81. ACM, 2013.

[3] Beatrice Alex, Kate Byrne, Claire Grover, and Richard Tobin. Adapting the Edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing*, 9(1):15–35, 2015.

[4] Adrien Barbaresi. A one-pass valency-oriented chunker for German. In *Proceedings of the 6th Language & Technology Conference*, pages 157–161, 2013.

[5] Adrien Barbaresi. Towards a Toolbox to Map Historical Text Collections. In *Proceedings of the 11th Workshop on Geographic Information Retrieval*, GIR'17. ACM, 2017.

[6] Adrien Barbaresi. A constellation and a rhizome: two studies on toponyms in literary texts. In Bubenhofer Noah and Kupietz Marc, editors, *Visual Linguistics*. Heidelberg University Publishing, Heidelberg, 2018. To appear.

[7] Adrien Barbaresi. Toponyms as Entry Points into a Digital Edition: Mapping Die Fackel. *Open Information Science*, 2018. To appear.

[8] Hanno Biber, Evelyn Breiteneder, Heinrich Kabas, and Karlheinz Mörth. Austrian Academy Corpus AAC–FACKEL. Die Fackel. Herausgeber: Karl Kraus, Wien 1899-1936, Edition No 1, Online Version: http://www.aac.ac.at/fackel.

[9] Lars Borin, Dana Dannélls, and Leif-Jöran Olsson. Geographic visualization of place names in Swedish literary texts. *Literary and Linguistic Computing*, 29(3):400–404, 2014.

[10] Davide Buscaldi. Approaches to disambiguating toponyms. *Sigspatial Special*, 3(2):16–19, 2011.

[11] S. Feld. Waterfalls of song: An acoustemology of place resounding in Bosavi, Papua New Guinea. In S. Feld and K.H. Basso, editors, *Senses of place*, pages 91–135. School of American Research Press, 1996.

[12] Unxos GmbH. Geonames, 2017. http://www.geonames.org.

[13] Ian N. Gregory and Andrew Hardie. Visual GISting: Bringing together corpus linguistics and Geographical Information Systems. *Literary and Linguistic Computing*, 26(3):297–314, 2011.

[14] John Brian Harley. Deconstructing the map. *Cartographica: The international journal for geographic information and geovisualization*, 26(2):1–20, 1989.

[15] Linda Hill. Core elements of digital gazetteers: placenames, categories, and footprints. *Research and advanced technology for digital libraries*, pages 280–290, 2000.

[16] Christopher B. Jones, Ross S. Purves, Paul D. Clough, and Hideo Joho. Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, 22(10):1045–1065, 2008.

[17] Marko Juvan. From spatial turn to GIS-mapping of literary cultures. *European Review*, 23(1):81–96, 2015.

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

[19] Eetu Mäkelä, Thea Lindquist, and Eero Hyvönen. CORE – a Contextual Reader Based on Linked Data. *Digital Humanities 2016*, pages 267–269, 2016.

[20] Fernando Melo and Bruno Martins. Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, 21(1):3–38, 2017.

[21] Damien Nouvel, Maud Ehrmann, and Sophie Rosset. *Les entités nommées pour le traitement automatique des langues*. ISTE editions, 2015.

[22] Paolo Plini, Sabina Di Franco, and Rosamaria Salvatori. One name one place? Dealing with toponyms in WWI. *GeoJournal*, pages 1–13, 2016.

[23] Bruno Pouliquen et al. Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of LREC*, pages 53–58. ELRA, 2006.

[24] Ludwig Richter, Johanna Geiß, Andreas Spitz, and Michael Gertz. HeidelPlace: An Extensible Framework for Geoparsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 85–90, 2017.

[25] David A. Smith and Gideon S. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of Geographic References*, pages 45–49. Association for Computational Linguistics, 2003.

[26] Humphrey Southall, Ruth Mostern, and Merrick Lex Berman. On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5(2):127–145, 2011.

[27] Minna Tamper, Petri Leskinen, Esko Ikkala, Arttu Oksanen, Eetu Mäkelä, Erkki Heino, Jouni Tuominen, Mikko Koho, and Eero Hyvönen. AATOS – a Configurable Tool for Automatic Annotation. In *International Conference on Language, Data and Knowledge*, pages 276–289. Springer, 2017.

[28] Denny Vrandečić and Markus Krötzsch. Wikidata: a Free Collaborative Knowledge Base. *Communications of the ACM*, 57(10):78–85, 2014.

[29] Clifford E. Wulfman. The Plot of the Plot: Graphs and Visualizations. *The Journal of Modern Periodical Studies*, 5(1):94–109, 2014.