Borderlands of text mapping: Experiments on Fontane's Brandenburg

Adrien Barbaresi[1, 2]

Academy Corpora, Austrian Academy of Sciences [1]
Zentrum Sprache, Berlin-Brandenburg Academy of Sciences [2]

Abstract

This article deals with the detection and projection of spatial patterns in text collections. Especially for historical corpora, researchers face a lack of general-purpose tooling. In these experiments, different maps focusing on Brandenburg at the second half of the 19th century are created based on literary works by Theodor Fontane. Using a common ground for hypothesis testing and visualization, issues related to data curation and preparation, text processing and geocoding are presented and discussed: the abstract, static nature of the results should be made up for by particular scrutiny and contextualization, by critical machine reading and by adding depth through visual cues.

1. Introduction

In digital text corpora and especially literary works, space can be analyzed as patterns to be found across texts: "Large source bases are likely to contain 'meaningful patterns,' and the ability to examine them, that is, to gain a bird's-eye view with the assistance of a computer, is tantamount to having a team of readers, even many teams of readers, at one's disposal." (Wrisley, 2017) Indeed, looking for patterns is widely considered to be a task at which distant reading and computer-based studies can be fruitful. In linguistics, a common criterion consists of frequency-based information, even in literary studies, patterns can be seen as "the strongest point of intersection between the computational strictures of text analysis and the open ended landscape of interpretative literary studies." (Ramsay, 2005) Looking for spatial patterns and displaying them can be assimilated to the recently coined concept of geocriticism, in the sense that it is not a character-centered or subjective framework. Under this assumption, the emergence of geographical patterns in texts is explicitly encouraged: "Geocriticism is a geo-centered rather than an ego-centered approach; that is, the analysis focuses on global spatial representations rather than on individual ones (a given traveler's, for example)." (Westphal,

Veröffentlicht durch die Gesellschaft für Informatik e.V. 2018 in
M. Burghardt & C. Müller-Birn (Hrsg.): INF-DH 2018 – Workshopband, 25. Sept. 2018, Berlin
Copyright © 2018 bei den Autoren. https://doi.org/10.18420/infdh2018-05

2014) In this perspective, this article deals with the detection and projection of spatial patterns in a historical text collection: based on experiments using a digitized version of works by Theodor Fontane, it focuses on extraction and visualization problems, which are exemplified by the situation of 19th century's Brandenburg. Finally, results in the form of maps are presented and commented.

2. Method

2.1 Trends in geocoding and visualization

Among the tendencies in geographic information retrieval and geocoding (Melo & Martins, 2017), 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 place names in texts (geoparsing or place names extraction) involves first the detection of words and phrases that may potentially be proper nouns and second their classification as geographic references (Nouvel, Ehrmann, & Rosset, 2015). After the identification of toponyms, a further step (geocoding or toponym resolution) resides in disambiguating and adding geographical coordinates to a place name. Geocoding mostly relies on gazetteers, i.e. geospatial dictionaries containing mostly names, locations, and metadata such as typological information, variants or dates (Hill, 2000). Toponym resolution as well as named-entity recognition can use machine learning methods, 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 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 (Alex, Byrne, Grover, & Tobin, 2015) and the necessity to complement existing resources with a precise historical gazetteer (Borin, Dannélls, & Olsson, 2014) have been highlighted. Such historical gazetteers exist, but their development is challenging (Southall, Mostern, & Berman, 2011) even for texts as late as 20th century Europe (Plini, Di Franco, & Salvatori, 2016). Existing toolboxes, such as AATOS (Tamper et al., 2017), mostly feature candidate extraction and ranking as well as entity linking. The approach used here is more light-weight, modular and adaptable, with a similar scope as CORE (Mäkelä, Lindquist, & Hyvönen, 2016) but with an overall greater focus on general-purpose usability, using texts as input, the integration of registers, and serialized map export as images.

2.2 Crossing philology and computer science

The task at hand does not simply reside in linking text to space, it is closely related to the interpretation of texts and maps and implies to try and cross philological and computational

approaches. Even if both the methods of natural language processing and the results in the form of maps can convey an impression of scientific objectivity, the validity of mental and computerized operations described here should always be examined with respect to their potential relevance. Geospatial analysis and spatial representation may indeed be deficient or inadequate as the symbolic role and the expressive power of place names do not necessarily coincide with Western instrumental science and cartography, meaning in that particular case the registers and models used for extraction, the world geodetic system, and the chosen map projection. The static, abstract nature of this sometimes superficial reading of the texts should be made up for by putting information into context, for example using visual cues to add qualitative insights. Additionally, a particular scrutiny is required to approach the texts, most importantly concerning the proper adaptation of concepts between disciplinary boundaries and in practice concerning error analysis.

2.3 Challenges and corresponding tooling

Three common issues in geographic information extraction are specifically addressed here with a particular focus on philological soundness: detecting geographical references, disambiguating place names and developing effective user interfaces. This approach involves to gather and curate supplementary information applicable to historical texts. To this end, the aggregation of tools used here is being developed with historical texts in mind[1] and has already been used so far to map different collections ranging from the 17th to the 20th century (Barbaresi, 2016, 2017, 2018b).

The extraction of spatial entities (or geoparsing) can be considered as one of the most important part in spatial analysis (Fize, Shrivastava, & Ménard, 2017). It has been shown that a method for automatically constructing a geographic gazetteer using heterogeneous sources can mitigate coverage issues by combining information but also improve resolution and correctness by validating the datasets against another. It is for example useful to mine Wikipedia for location information (Popescu, Grefenstette, & Moëllic, 2008), which results in the constitution of minimum explicit tuples for geographic names: Entity-Name, Entity-Coordinates, and Entity-Type. In this experiment, both extraction and disambiguation are parametrized to suit the particular case at hand. Regarding toponym resolution, two different types of disambiguation methods (Buscaldi, 2011) are used so far: map-based and knowledge-based. It has been shown that an acceptable precision can be reached by including meta-information (Pouliquen et al., 2006), 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 is controlled by parameters such as distance calculations, filter level or size of the search area.

The difficulties related to German include capitalization, genitive forms (typically formed with *s*), adjectival use (frequently as modifiers), and the fact that inhabited place names are also frequently found as personal names (Volk & Clematide, 2001). These reasons make it difficult

1 The ongoing work on the toolbox is open-source: https://github.com/adbar/geokelone

to build a gazetteer and prompt for the integration of morphological information. As a general principle, it is expected that "named entity recognition in German will profit from precompiled lists as well as from learning and filtering" (*ibid.*). Especially for historical texts it is necessary to refine or complement the resources at hand: 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. The operations performed prior to the extraction comprise the bootstrapping, filtering and merging of registers; since there is no commonly adopted standard for gazetteers they have to be combined. It also includes helpers to bootstrap geographical data, as knowledge-based methods using fine-grained data improve the results (Vrandečić & Krötzsch, 2014). So far, import filters for GeoNames and structured data from Wikipedia and Wikidata are implemented, with a particular emphasis on data cleaning. Finally, the results are projected on customized maps. It is profitable to allow for adaptability of projection and design and to leave it open to the user to refine the map. It is for example possible to convey information such as typology or frequency on the maps.

3. Exploratory study

The study grounds on digitized versions of the first four volumes of Theodor Fontane's *Wanderungen durch die Mark Brandenburg*[2] (DTA, 2018), a series of travel feuilletons about Brandenburg published between 1862 and 1882 with a focus on its history and its landscape. The experimental character of Fontane's writing results in an apparent lack of unity in the *Wanderungen*, which include quotations, lists, and commentaries. The method used here can provide a visual summary of the text and its heterogeneous and possibly non-literary segments, all the more since the numerous place names structure the narration.[3] The experimental setting is derived from previous experiments and includes custom gazetteers for historical European locations. The streamlined process from text to map involves a series of decisions. Most importantly, the filtering level affects both the loading of gazetteers prior to geoparsing and the toponym recognition phase in itself. Its purpose is to allow for tighter or looser control, with either restricted options or opportunistic search. The first experiments, whose results are displayed on Figures 1 and 2, are designed to try to maximize recall with a smaller amount of filtering.

2 "Ramblings through Brandenburg" or "Walks through the March of Brandenburg"

3 For a review of studies on space by Fontane, see White (2012).

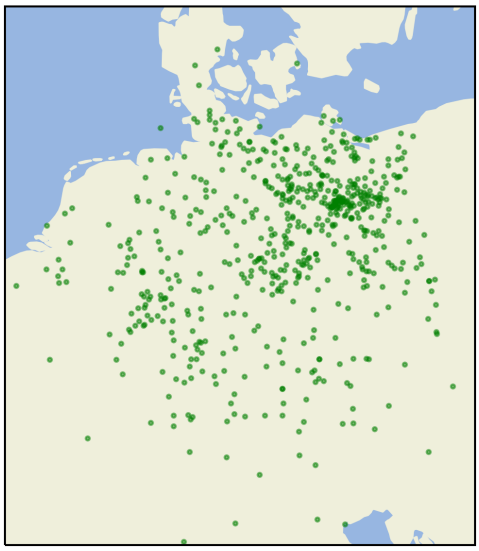

Figure 1: Raw text, medium amount of filtering, no frequency

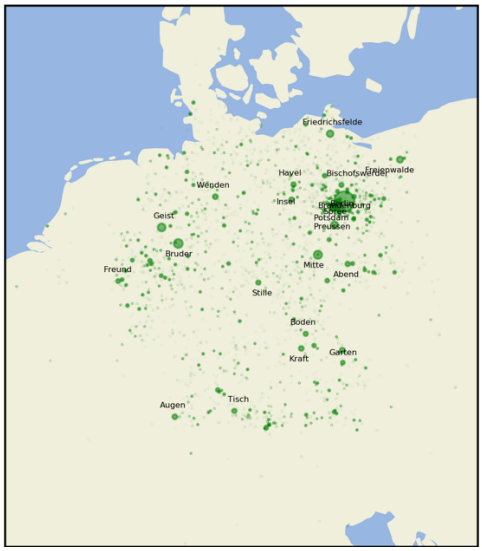

Figure 2: Raw text, minimum filtering, frequency information through size and frequency-based labeling

As can be seen from the scattering of points on Figure 1, the volumes seem to focus primarily on Brandenburg but other place names are found, so that only a loose cluster emerges around Berlin. Even so, the current borders of Germany can be distinguished, which implies that the accuracy is low, since the convergence of regional entities which was to become the German Empire and the empire as a whole had boundaries different from the current state. This fact highlights to what extent the use of current gazetteers (here GeoNames restricted to Germany) can impact the results. Figure 2 is useful for error analysis, as the filtering is reduced and a sense of frequency is given by the feature size and by the labels for the most prominent tokens. The names for former states and regions (e.g. Prussia) are linked to fixed locations in the custom gazetteer. Interestingly, the impression given by Figure 1 is corrected in the right sense by clearly focusing on Brandenburg, however, the current boundaries also become much more identifiable, notably because of the noise coming from common names matching actual locations on the map. The cluster around Berlin can also be explained by the influence of distance calculations.

In a second experiment, the texts are then tokenized with SoMaJo (Proisl & Uhrig, 2016) and POS-tagged with SoMeWeTa (Proisl, 2018) to isolate tokens which have been identified as named entities. A short stoplist is gathered manually to discard frequent proper names which could correspond to actual locations. Figure 3 depicts the results of the focus on precision with a high level of filtering and geoparsing applied on named entities only. There are less points to be seen then on Figure 1 and the sharper focus on the northwestern end of Germany appears more clearly, which is in line with the expectations concerning the contents of the works. Figure 4 features a zoomed-in view centered on Brandenburg along with a medium amount of

filtering and frequency cues. The predominance of Berlin and a few locations around it (e.g. Potsdam, Freienwalde, Rheinsberg, Küstrin, the river Havel – marked here as a point) is clear and actually corresponds to depictions in the texts and to the contrast between metropole and province in the texts (Scheiding, 2012). Overall, it can be deemed to be an adequate if somewhat biased overview of Fontane's ramblings.

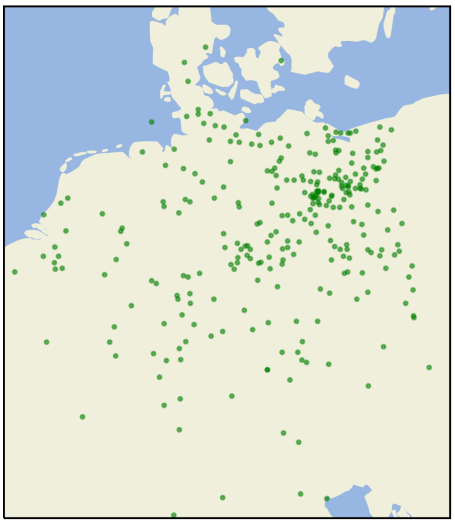

Figure 3: POS-tagged text, maximum filtering

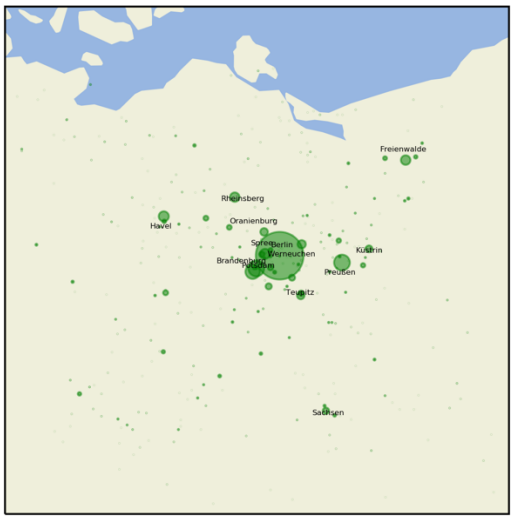

Figure 4: POS-tagged text, medium filtering, frequency cues

4. Discussion

The status of place names that are to be found and projected on the map ought to be discussed since there are consubstantial ambiguities on linguistic level that complicate the search (Smith & Mann, 2003): the referent ambiguity (one name used for more than one location) and the referent class ambiguity (place names 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 (Jones, Purves, Clough, & Joho, 2008) is a prominently linguistic issue which can at least be addressed by manual curation and should in any case be attended to in order to meet the expectations of philological research. Additionally, from a literary standpoint, it seems necessary to take the changes of perspective into account as well as the literary devices which are used to guide the reader's gaze.

The artifical character of research results and especially maps in this context also ought to be mentioned. The "selection, omission, simplification, classification [...] and symbolization"

steps of mapping as shown here are all "inherently rhetorical" (Harley, 1989), they are both the foundation and the product of scientific reasoning and have to be taken with a grain of salt. One must bear in mind that "computers are lousy readers", and "our current digital tools walk a delicate line between analytical power and accessibility" (Wilkens, 2011). As the products of machine learning and machine reading can diverge from the expectations, there is an estrangement in distant reading experiments which one has to face and to overcome in order to make proper use of the tools.

5. Conclusion

This article introduced theoretical and practical instruments combining philological knowledge, geographic information retrieval and visualization. A common ground for hypothesis testing and visualization has been presented, with the particular example of the detection and projection of spatial patterns in a historical text collection. Maps focusing on Brandenburg at the second half of the 19th century have been compared, they exemplify issues related to the unsupervised processing of historical resources and they show the advantages of an appropriate methodology and subsequent data cleaning and filtering. Rapidly spotting problems in methodology or datasets can hopefully help mitigating the side effects of large-scale analysis and distant reading. Text visualizations are indeed the substrate of interpretable representations which do not follow data but rather confront them by putting them in perspective and trying to overcome the superficiality of computational reading, be it through a flattening constellation or by dealing with the rhizomatic character of literary texts (Barbaresi, 2018a). 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.

As quantitative and qualitative analysis can go hand in hand, digital literary studies are not mere numeric accounts, they are first and foremost a exploratory process. The parametrization of tools, the linguistic annotation of both texts and gazetteers as well as an appropriate level of filtering and map customizing can lead to a representative overview of ramblings through Brandenburg's March, whereas the borderlands of text mapping require a careful examination and scrutiny to let patterns emerge, in a constant loop of questioning and feedback from datasets and methods.

References

Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the Edinburgh geoparser for historical georeferencing. *International Journal of Humanities and Arts Computing*, *9*(1), 15–35.

Barbaresi, A. (2016). Visualisierung von Ortsnamen im Deutschen Textarchiv. In *DHd 2016* (pp. 264–267). Digital Humanities im deutschprachigen Raum e.V.

Barbaresi, A. (2017). Towards a Toolbox to Map Historical Text Collections. In *Proceedings of the 11th Workshop on Geographic Information Retrieval*. ACM. https://doi.org/10.1145/3155902.3155905

Barbaresi, A. (2018a). A constellation and a rhizome: two studies on toponyms in literary texts. In B. Noah & K. Marc (Eds.), *Visual Linguistics*. Heidelberg: Heidelberg University Publishing.

Barbaresi, A. (2018b). Toponyms as Entry Points into a Digital Edition: Mapping Die Fackel. *Open Information Science*.

Borin, L., Dannélls, D., & Olsson, L.-J. (2014). Geographic visualization of place names in Swedish literary texts. *Literary and Linguistic Computing*, *29*(3), 400–404.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *Sigspatial Special*, *3*(2), 16–19.

DTA (2018). Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Berlin-Brandenburg Academy of Sciences (ed.) http://www.deutschestextarchiv.de/

Fize, J., Shrivastava, G., & Ménard, P. A. (2017). Geodict: an integrated gazetteer. In *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017)*.

Harley, J. B. (1989). Deconstructing the map. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *26*(2), 1–20.

Hill, L. (2000). Core elements of digital gazetteers: place names, categories, and footprints. *Research and Advanced Technology for Digital Libraries*, 280–290.

Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, *22*(10), 1045–1065.

Mäkelä, E., Lindquist, T., & Hyvönen, E. (2016). CORE – a Contextual Reader Based on Linked Data. In *Digital Humanities 2016* (pp. 267–269). ADHO.

Melo, F., & Martins, B. (2017). Automated Geocoding of Textual Documents: A Survey of Current Approaches. *Transactions in GIS*, *21*(1), 3–38.

Nouvel, D., Ehrmann, M., & Rosset, S. (2015). *Les entités nommées pour le traitement automatique des langues*. ISTE editions.

Plini, P., Di Franco, S., & Salvatori, R. (2016). One name one place? Dealing with toponyms in WWI. *GeoJournal*, 1–13.

Popescu, A., Grefenstette, G., & Moëllic, P. A. (2008). Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 85–93). ACM.

Pouliquen, B. (2006). Geocoding multilingual texts: Recognition, disambiguation and visualisation. In *Proceedings of LREC* (pp. 53–58). ELRA.

Proisl, T. (2018). SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In *Proceedings of LREC*. ELRA.

Proisl, T., & Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop* (pp. 57–62). Association for Computational Linguistics.

Ramsay, S. (2005). In Praise of Pattern. *TEXT Technology: The Journal of Computer Text Processing*, *92*(2), 177–190.

Scheiding, K. (2012). *Raumordnungen bei Theodor Fontane*. Marburg: Tectum Verlag.

Smith, D. A., & Mann, G. S. (2003). Bootstrapping toponym classifiers. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of Geographic References* (pp. 45–49). Association for Computational Linguistics.

Southall, H., Mostern, R., & Berman, M. L. (2011). On historical gazetteers. *International Journal of Humanities and Arts Computing*, *5*(2), 127–145.

Tamper, M., Leskinen, P., Ikkala, E., Oksanen, A., Mäkelä, E., Heino, E., … Hyvönen, E. (2017). AATOS – a Configurable Tool for Automatic Annotation. In *International Conference on Language, Data and Knowledge* (pp. 276–289). Springer.

Volk, M., & Clematide, S. (2001). Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition. In *Proceedings of NLDB* (pp. 153–163).

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a Free Collaborative Knowledge Base. *Communications of the ACM*, *57*(10), 78–85.

Westphal, B. (2014). Foreword. In R. T. J. Tally (Ed.), *Geocritical Explorations : Space, Place, and Mapping in Literary and Cultural Studies* (pp. ix–xv). Palgrave Macmillan.

White, M. J. (2012). *Space in Theodor Fontane's Works: Theme and Poetic Function* (Vol. 38). Modern Humanities Research Association.

Wilkens, M. (2011). *Contemporary Fiction by the Numbers*. https://web.archive.org/web/20180208085407/http://post45.research.yale.edu/2011/03/contemporary-fiction-by-the-numbers/

Wrisley, D. J. (2017). Locating Medieval French, or Why We Collect and Visualize the Geographic Information of Texts. *Speculum*, *92*(1), 145–169.