Toponym Disambiguation in Historical Documents Using Network Analysis of Qualitative Relationships

Ludovic Moncla
ludovic.moncla@liris.cnrs.fr
INSA Lyon, CNRS, LIRIS UMR 5205
Lyon, France

Katherine McDonough
kmcdonough@turing.ac.uk
The Alan Turing Institute and Queen Mary, University of London
London, UK

Denis Vigier
denis.vigier@ens-lyon.fr
ENS Lyon, CNRS, ICAR UMR 5191
Lyon, France

Thierry Joliveau
thierry.joliveau@univ-st-etienne.fr
Université de Saint-Etienne, CNRS, EVS UMR 5600
Saint-Etienne, France

Alice Brenon
alice.brenon@ens-lyon.fr
ENS Lyon, CNRS, ICAR UMR 5191
Lyon, France

ABSTRACT

In this paper we use network analysis to identify qualitative "neighbors" for toponyms in an eighteenth-century French encyclopedia, but could apply to any entry-based text with annotated toponyms. This method draws on relations in a corpus of articles, which improves disambiguation at a later stage with an external resource. We suggest the network as an alternative to geospatial representation, a useful proxy when no historical gazetteer exists for the source material's period. Our first experiments have shown that this approach goes beyond a simple text analysis and is able to find relations between toponyms that are not co-occurring in the same documents. Network relations are also usefully compared with disambiguated toponyms to evaluate geographical coverage, and the ways that geographical discourse is expressed, in historical texts.

CCS CONCEPTS

• **Information systems → Geographic information systems**; **Document representation**; • **Applied computing** → *Arts and humanities.*

KEYWORDS

Toponym disambiguation, Digital Humanities, Geographic Information Retrieval

ACM Reference Format:
Ludovic Moncla, Katherine McDonough, Denis Vigier, Thierry Joliveau, and Alice Brenon. 2019. Toponym Disambiguation in Historical Documents Using Network Analysis of Qualitative Relationships. In *3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities (GeoHumanities'19), November 5, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3356991.3365471

GeoHumanities'19, November 5, 2019, Chicago, IL, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6960-2/19/11…$15.00
https://doi.org/10.1145/3356991.3365471

1 INTRODUCTION

The GÉODISCO Project uses geoparsing, cartography and linguistic analysis to identify developments in geographical discourse of historical French encyclopedias. Our corpus includes Diderot and d'Alembert's *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres* (1751-1772, EDDA), the *Encyclopaedia Universalis* (2018 digital edition), and the French Wikipedia (July 2018). We develop automatic methodologies for identifying, extracting, and analyzing geographical information in these texts. Here, we will focus our study on EDDA but the method we propose could apply to study any encyclopedia, dictionary, or entry-based text. Understanding how authors expressed geographic information in the eighteenth century illuminates the spatial dynamics of Enlightenment culture. GÉODISCO explores knowledge expressed qualitatively and quantitatively and applies qualitative and quantitative analysis to relations among data.

EDDA's 17 volumes of text contain a vast number of toponyms that have never before been extracted and studied as such. Our previous work [6] improving NER for eighteenth-century French prose means that we have been able to identify this information in the text.[1] However, because of the known challenges of toponym disambiguation and resolution, it is difficult (if not impossible or anachronistic) to translate these tagged entities into geographic locations using typical methods that depend on resources external to the historical text itself.

In this paper we use network analysis to establish an approximate location, defined by qualitative relations, for each named toponym in EDDA. Throwing a list of decontextualized toponyms at an external resource like Geonames is risky. We therefore hypothesize that defining meaningful links between places can provide essential information to improve disambiguation (and potentially replace resolution as the end goal). We establish connections between places based on the citation of "headword" toponyms (those that appear as headwords of entries) in other EDDA entries.

Using the citation method, a network reveals qualitative relations between locations (Paris is connected in the network to France), but does not define those relations. This relational data can be

[1] NER of the EDDA text is a work in progress, but uses a custom version of the Perdido geoparser. Recent results have 55.58% recall, 75.71% precision, and 64.10% F1-score.

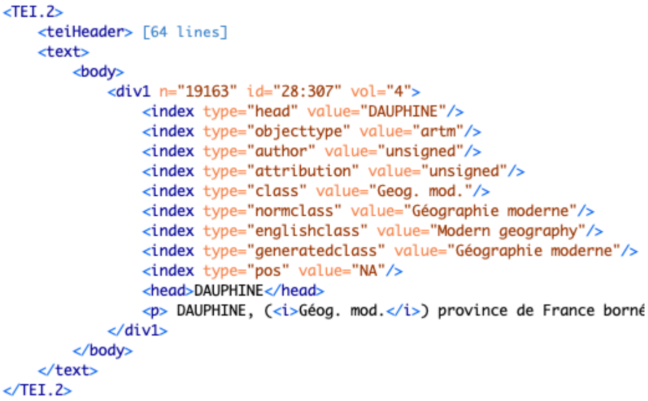
repurposed for toponym disambiguation and then resolution (e.g. using the knowledge that Paris is related to France as a way to simplify selecting the correct record in a gazetteer and establishing the location of the toponym). Even more meaningful, however, is the analysis of these relative locations as a representation of the EDDA world. Qualitative relative locations therefore replace coordinates for certain kinds of historical and linguistic analysis. This seems all the more appropriate since authors or readers may not have immediately imagined EDDA content cartographically, but rather in terms of known itineraries, political alliances, famous buildings and people, or strange and marvelous stories. Mapping EDDA toponyms could displace them from their own spatial culture: sometimes this is appropriate and other times not. Having access to both a geospatial and a non-geolocated network is useful.

Toponym disambiguation refers to the task of identifying the correct location that is referred to by a place name [4]. The result can be either geographic coordinates or knowledge base entries (as in entity linking approaches [2]). The approaches for toponym disambiguation are usually classified in three groups: map-based, knowledge-based and data-driven approaches [3]. Map-based approaches use the context of a toponym such as the location of other place names that occur in the same document in order to identify the correct location of the ambiguous toponym [11]. Data-driven approaches are based on machine learning algorithms [5, 10] while knowledge-based approaches make use of knowledge sources (gazetteers, ontologies, etc). Some of these approaches use topology-based methods for the disambiguation of toponyms and most of the methods based on knowledge graphs use Wikipedia structured data. These studies are based on the observation that spatial relatedness between locations is hidden in Wikipedia link structure. For instance, Quercini and Samet [9] explored a set of graph-based similarity measures to determine a local lexicon of a location from Wikipedia and study the spatial relatedness of concepts with location. Alencar et al. [8] proposed a method for the geographical classification of documents. They use Wikipedia as a semantic network to gather textual geographic evidence for places. Several other studies also use network analysis for toponym disambiguation. For instance, based on the hypothesis that entities share a relation if they are mentioned in the same context, Spitz et al. [12] create a network of entities' relationships from unstructured texts as opposed to methods using knowledge bases from structured data. Finally, only a few studies deal with toponym disambiguation in historical documents [1, 11].

2 BUILDING THE NETWORK FROM TEXTUAL DATA

Several studies described above consider that entities share a relation if they are mentioned in the same context, however they depend on toponym co-occurrences in the same text. We propose to build a network of EDDA articles where each node refers to the headword (i.e., article title) of a geography article and each edge represents a spatial or cultural relationship between two nodes. Our work leverages the encyclopedic structure of headwords and article classifications to construct an artificial map of all re-citations of headwords across every article.

EDDA contains 20.7 million words in 44 632 entries[2], but for this research we limit the corpus to entries classified as Geography (14 457 articles). Co-edited by Denis Diderot and Jean le Rond d'Alembert at the height of the Enlightenment, it is a work that represents the contrasts, struggles, curiosity, and imagination of the eighteenth century. EDDA is an amalgamation of authorial intentions, editorial interventions, and blatant copying from earlier works. We built the Geography subcorpus from articles with the classifications *Géographie*, *Géographie moderne*, *Géographie ancienne* and *Géographie moderne | Géographie ancienne* having respectively 5 800, 5 037, 3 218 and 167 articles.[3]



```
<TEI.2>
    <teiHeader> [64 lines]
    <text>
        <body>
            <div1 n="19163" id="28:307" vol="4">
                <index type="head" value="DAUPHINE"/>
                <index type="objecttype" value="artm"/>
                <index type="author" value="unsigned"/>
                <index type="attribution" value="unsigned"/>
                <index type="class" value="Geog. mod."/>
                <index type="normclass" value="Géographie moderne"/>
                <index type="englishclass" value="Modern geography"/>
                <index type="generatedclass" value="Géographie moderne"/>
                <index type="pos" value="NA"/>
                <head>DAUPHINE</head>
                <p> DAUPHINE, (<i>Géog. mod.</i>) province de France borné
            </div1>
        </body>
    </text>
</TEI.2>
```

Figure 1: XML-TEI content for EDDA article *Dauphine*

To identify the network nodes, we first made a list of headwords from the subcorpus. We then parse the XML-TEI documents (1 per article) (see Fig. 1) and create a node with the value of the head element. The second step normalizes headword names. Some of the names contain prepositions, punctuation marks and/or alternate names or spellings (e.g., *'Brassaw, ou Gronstat'*[4], *'Adiazzo, Adiazze ou Ajaccio'*[5]) which require attention. Where there are alternate spellings, we retain each variant in one consolidated node. It is therefore possible to match variants across the corpus.

Next we create edges representing the relationships between nodes. Edges exist when the name of one article occurs in a second article. Thus, we consider that two place names share a relationship when both place names are a geography article headword (i.e., when each place name has a corresponding node) and one of them occurs in the text of the other. Figure 2 shows the content of the article *Dauphine*[6]. The entity *Dauphine* (in green) refers to the article headword, the entities in red have a corresponding article classified as geography, while entities in blue do not (either they do not have articles or their classification is misleading). For instance, the entity *Alpes* has an article but its classification is "unclassified". Entities in red therefore all have a corresponding node in the network and they are all connected to the node *Dauphine*.

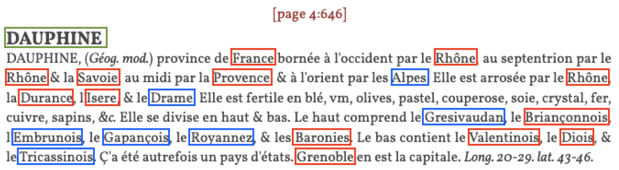
[2] https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/

[3] We also consider 25 other classifications with fewer than 10 articles, for example, *Géographie | Histoire naturelle*, *Géographie ancienne | Mythologie*, *Géographie ancienne | Géographie antique*.

[4] https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/2/3580/

[5] https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/1/810/

[6] https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/4/3260/



[page 4:646]

DAUPHINE

DAUPHINE, (*Géog. mod.*) province de France bornée à l'occident par le Rhône, au septentrion par le Rhône & la Savoie, au midi par la Provence, & à l'orient par les Alpes. Elle est arrosée par le Rhône, la Durance, l'Isere, & le Drame. Elle est fertile en blé, vm, olives, pastel, couperose, soie, crystal, fer, cuivre, sapins, &c. Elle se divise en haut & bas. Le haut comprend le Gresivaudan, le Briançonnois, l'Embrunois, le Gapançois, le Royannez, & les Baronies. Le bas contient le Valentinois, le Diois, & le Tricassinois. Ç'a été autrefois un pays d'états. Grenoble en est la capitale. *Long. 20-29. lat. 43-46.*

Figure 2: Content of *Dauphine* article. Courtesy of ARTFL

In order to create the edges of the network, we extract all place names from the content of each article. For this purpose, we use a custom version of the Perdido geoparser [7] embedding the list of EDDA geography headwords.[7] The Perdido geoparser adds geospatial semantic annotations to the XML-TEI input. It annotates spatial relations, motion expressions, and named entities (places, persons and dates). In this custom version the headword place names are tagged as such. Network edges emerge by parsing the Perdido geospatial annotation results. For each article, a new edge is created between the node for the current article and all the corresponding nodes of each headword toponym in the content.

3 RELATIONAL DATA ANALYSIS

This EDDA Geography network consists of 13 819 nodes and 36 396 edges. Several well-known centrality measures exist to study the network structure (e.g., in- and out-degree centrality, closeness centrality, and betweenness centrality). Each centrality measure can indicate the influence of nodes in the network. In particular, in-degree centrality captures the number of connections that point to a node. In EDDA, the number of citations of an article headword in other articles. Betweenness centrality measures the importance of a node according to the shortest paths between each pair of nodes through the network. For example, if the nodes and edges represented a road network, EDDA nodes would represent the places which have the greatest number of short routes passing through it: these places are likely the busiest intersections. Comparing centrality measures for the headword toponyms helps GÉODISCO to disambiguate, but it also feeds back into GÉODISCO's research to locate types of geographical discourse (scientific, descriptive, relational). They allow us to compare presence and significance in the network with EDDA's actual geospatial coverage and eventually to compare EDDA to other texts' networks.

Figure 3 shows the network where label size indicates betweenness centrality. The overall network, and the top results of two centrality measures (Mer Méditerranée, France, Allemagne, Natolie, Monde, and Italie for betweenness; and France, Italie, Allemagne, Afrique, Espagne, and Naples for in-degree) are not surprising considering the state of knowledge of the world in eighteenth-century France. Among a variety of surprising findings, the use of *Afrique* deserves mention. *Afrique* has such a high in-degree ranking because this word tends to follow the name of a region or country (for example, "*AGOBEL, (Géog.) ville d'Afrique au Royaume de Maroc, dans la Province d'Ea en Barbarie*"[8]). Africa is not a country, but it is included to alert readers of an entry's position on that continent just as country mentions do for places on other continents. Additionally, while Africa has a high in-degree rank, it falls to 21st in betweenness. Instead, *Barbarie* replaces the continent for North African places and is 13th in betweenness and 44th in in-degree. *Mer méditerranée*, *France*, *Allemagne*, *Natolie*, and *Monde*, which all have very high betweenness centrality, are integral to EDDA's spatial structure. Slightly further down that list, we find Lyon (12th) together with Paris (11th), both representing key pathways in the network. Lyon falls to 183rd in in-degree centrality, however, and cannot compete with the raw citations of Paris (16th) in other articles, as shown in Fig. 4. These articles with high betweenness centrality are the places that authors refer to the most in order to direct readers' attention to understanding the geographical context of a headword. Such nodes may be the intellectual crossroads of EDDA (even if they are not the most cited), and are useful in situating historical toponyms within political, economic, or cultural spheres of influence.

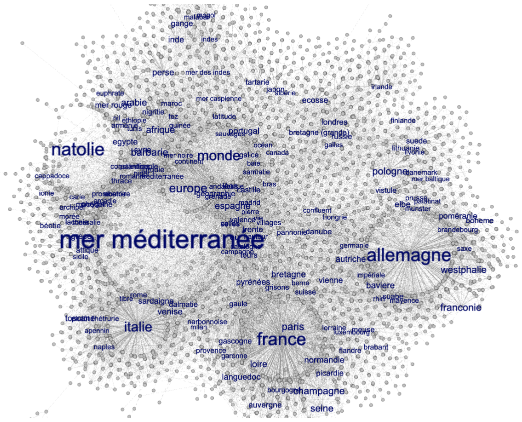


Figure 3: Visualization of EDDA headword toponyms network (label size indicates betweenness centrality)

Our proposal for disambiguation using the network consists to identify the most related place of a place name with unknown location. Our hypothesis is that the quantitative citation network reveals qualitative relations. For this purpose, we compute an ego-centered network (i.e., neighborhood graph, with a radius of 2) from the node of a place name with unknown location. We compute the betweenness centrality measure of this ego-network and the node with the highest value is selected as the most related.[9]

For experiments with and evaluation of our proposal, we use 100 articles randomly selected from the Geography subcorpus. The proposed method identifies for each of the 100 headwords the most related place name from the network. A manual review of the results shows that 83/100 responses are correct (e.g. the "most related" place in the ego network is actually a place that helps to disambiguate a place, whether defined by a relation of distance or of culture. In most cases, for a city, the method returns the name of the country to which it belongs (contained in) and for country, the name of

[7] We remove from the headwords list those that are too ambiguous such as common nouns (e.g., montagne). Demonyms (e.g., *Basques (les)*), massively present as headwords in the geography subcorpus, occur as nodes, but they do not necessarily link to citations of demonyms in the text because those occurrences may not have been tagged as place entities.

[8] https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/1/1173/

[9] The implementation is done with the NetworkX python library.

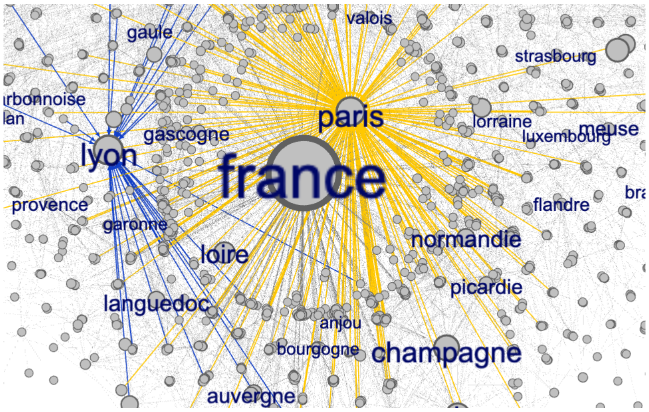


Figure 4: Zoom of *lyon* and *paris* nodes. Blue edges refer to inward directed edges to *lyon* and yellow edges inward to *paris* (node and font size indicate betweenness centrality)

a neighboring location (e.g., country, region, city, or a relation of proximity). For instance, for the country *Pérou* the method returns the name *Popayan* which is in present-day a city in the south of Colombia. In 18% of correct answers (15 over 83), the returned name is not present in the content of the article, such as *Natolie* for *Isaurie* and *Inde* for *Azmer*. This means that in these cases the network plays a unique role that extends beyond the analysis of the article in question, drawing on inter-article relations to identify the most important connection. The relations in the network help to disambiguate and identify an approximate location or cultural link without any external resources. Additionally, among the 17 false positive results, 3 refer to demonyms and 6 concern Italian cities for which the answers are *Venise* instead of Italy or more precise locations, such as *Calabrie* for *Stilo*, *Toscane* for *Nepi* and *Naples* for *Capitanate*. In this case, Venise is not always a "true" false positive, since these articles may refer to Venice as a major intellectual center of the Enlightenment, or to the Gulf of Venice as a bordering geographical feature.

4 DISCUSSION AND CONCLUSION

In this study we have proposed to use network analysis to identify an approximate location or cultural relation for place names in EDDA. This experimental study has shown that connecting entries of an encyclopedia can help with toponym disambiguation. This is a first step in disambiguation which aims to build geographic information retrieval solutions for historical documents without using anachronistic external resources and in addition to standard natural language understanding methods.

A remaining issue related to toponym ambiguity is that when multiple articles have the same name (e.g., *Londres* has 2 entries and *Aix* 3), it is impossible to say in connections which one is being referred to. It would be possible to create a new node for each entry (for instance adding an index on the name) and then to apply text analysis methods to disambiguate and identify for each occurrence (from the content of other articles) which entry it refers to.

Network results should also be weighted by their context, (e.g. the very common structure of "petite ville de France" should not have the same strength in the network as a mention of "France" that exists outside this phrase. In future work we will include the immediate lexical environment as an attribute of network nodes.

Place names that appear in the content of articles, but have no entry in EDDA, do not have nodes in the network yet. This implies that we do not know if those names are very frequent in the content of articles. Thus, the next step of our work is to include all place names identified by the geoparser as nodes in the network and not only those with an entry. Following the same idea, another perspective is to add to the network entries of all classifications (not limited to Geography). This might lead to visualizations of relations between articles of the EDDA not only focusing on geographic relations. Beyond the problem of toponym disambiguation, such an approach can help us build a knowledge graph.

ACKNOWLEDGMENTS

This research is supported by an IDEXLYON project of the University of Lyon within the framework of the Investments for the Future Program (ANR-16-IDEX-0005). We thank ARTFL for sharing EDDA data.

REFERENCES

[1] Mariona Coll Ardanuy and Caroline Sporleder. 2017. Toponym Disambiguation in Historical Documents Using Semantic and Geographic Features. In *Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2017)*. ACM, New York, NY, USA, 175–180. https://doi.org/10.1145/3078081.3078099

[2] Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly* 7 (July 2016), 60 – 80. https://doi.org/10.7250/csimq.2016-7.04

[3] Davide Buscaldi. 2011. Approaches to Disambiguating Toponyms. *SIGSPATIAL Special* 3, 2 (July 2011), 16–19. https://doi.org/10.1145/2047296.2047300

[4] Jochen L Leidner. 2008. *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names.* Universal-Publishers.

[5] Bruno Martins, Ivo Anastácio, and Pável Calado. 2010. *A Machine Learning Approach for Resolving Place References in Text.* Springer Berlin Heidelberg, Berlin, Heidelberg, 221–236. https://doi.org/10.1007/978-3-642-12326-9_12

[6] Katherine McDonough, Ludovic Moncla, and Matje van de Camp. 2019. Named entity recognition goes to old regime France: geographic text analysis for early modern French corpora. *International Journal of Geographical Information Science* 0, 0 (2019), 1–25. https://doi.org/10.1080/13658816.2019.1620235

[7] Ludovic Moncla, Walter Renteria-Agualimpia, Javier Nogueras-Iso, and Mauro Gaio. 2014. Geocoding for Texts with Fine-grain Toponyms: An Experiment on a Geoparsed Hiking Descriptions Corpus. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14)*. ACM, New York, NY, USA, 183–192. https://doi.org/10.1145/2666310.2666386

[8] Rafael Odon de Alencar, Clodoveu Augusto Davis, Jr., and Marcos André Gonçalves. 2010. Geographical Classification of Documents Using Evidence from Wikipedia. In *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR '10)*. ACM, New York, NY, USA, Article 12, 8 pages. https://doi.org/10.1145/1722080.1722096

[9] Gianluca Quercini and Hanan Samet. 2014. Uncovering the Spatial Relatedness in Wikipedia. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '14)*. ACM, New York, NY, USA, 153–162. https://doi.org/10.1145/2666310.2666398

[10] João Santos, Ivo Anastácio, and Bruno Martins. 2015. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal* 80, 3 (01 Jun 2015), 375–392. https://doi.org/10.1007/s10708-014-9553-y

[11] David A. Smith and Gregory Crane. 2001. Disambiguating Geographic Names in a Historical Digital Library. In *Research and Advanced Technology for Digital Libraries*, Panos Constantopoulos and Ingeborg T. Sølvberg (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 127–136.

[12] Andreas Spitz, Johanna Geiß, and Michael Gertz. 2016. So Far Away and Yet So Close: Augmenting Toponym Disambiguation and Similarity with Text-based Networks. In *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data (GeoRich '16)*. ACM, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/2948649.2948651