Machine Learning to Geographically Enrich Understudied Sources: A Conceptual Approach

Lorella Viola[1][a] and Jaap Verheul[2][b]
[1]*Luxembourg Centre for Contemporary and Digital History (C²DH), University of Luxembourg, Belval Campus, Maison des Sciences Humaines, 11, Porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg*
[2]*Department of History and Art History, Utrecht University, Drift 6, 3512 BS, Utrecht, The Netherlands*
lorella.viola@uni.lu, j.verheul@uu.nl

Keywords: Machine Learning, Sequence Tagging, Spatial Humanities, Geographical Enrichment, Immigrant Newspapers.

Abstract: This paper discusses the added value of applying machine learning (ML) to contextually enrich digital collections. In this study, we employed ML as a method to geographically enrich historical datasets. Specifically, we used a sequence tagging tool (Riedl and Padó 2018) which implements TensorFlow to perform NER on a corpus of historical immigrant newspapers. Afterwards, the entities were extracted and geocoded. The aim was to prepare large quantities of unstructured data for a conceptual historical analysis of geographical references. The intention was to develop a method that would assist researchers working in spatial humanities, a recently emerged interdisciplinary field focused on geographic and conceptual space. Here we describe the ML methodology and the geocoding phase of the project, focussing on the advantages and challenges of this approach, particularly for humanities scholars. We also argue that, by choosing to use largely neglected sources such as immigrant newspapers (also known as ethnic newspapers), this study contributes to the debate about diversity representation and archival biases in digital practices.

1 INTRODUCTION

Advances in machine learning (ML) are allowing researchers both in computer science and the humanities to develop new tools and methods for exploring digital collections. At the same time, libraries are resorting more and more to ML methods to maximise the potential of their collections, to improve the user experience, and to discover the technical requirements necessary to facilitate the discovery and use of digital collections. One effective application of ML is enriching digital data with geographical information. Because ML uses contextual information to disambiguate entities, the method goes beyond the state of the art of place name extraction and disambiguation based on gazetteers or ensembles of NER-tools (Canale, Lisena, and Troncy 2018; Won, Murrieta-Flores, and Martins 2018; Mariona Coll Ardanuy and Sporleder 2017; Maria Coll Ardanuy 2017).

This article discusses the added value as well as the challenges of using a ML method aimed to subsequently perform *conceptual* Named Entity Recognition (cNER), a way to identify subjective and situational geographical markers and connect them to explicit geo-references to space. In doing so, cNER aligns with what has been referred to as *Spatial Turns*, a revision of past approaches to the study of space and place (Murrieta-Flores & Martins, 2019) which acknowledges that *place* and *space* are distinct entities as place is created through social experiences and can be both real and imagined.

As its evidence, the study uses a corpus of Italian American historical newspapers (*ChroniclItaly*, Viola 2018) published between 1898 and 1920. We argue that cNER opens up a way to explore the cultural context of geographical place and that, by choosing to use a largely neglected data source such as immigrant newspapers, on the whole this study contributes to the debate about language diversity representation and archival biases in digital practices.

[a] https://orcid.org/0000-0001-9994-0841
[b] https://orcid.org/0000-0001-6978-7698

2 CONTEXT OF THE STUDY

The study stands at the crossroads of migration studies, linguistics, history, and digital humanities. It developed within the context of a larger project, the *GeoNewsMiner* (GNM - Viola et al., 2019) aiming to create a visualisation tool of geo-references. The tool was intended to assist researchers working in spatial humanities, a recently emerged interdisciplinary field focussing on geographic and conceptual space, especially from a historical perspective (Bodenhamer, Corrigan, & Harris, 2010, 2015a). Essentially, spatial humanities are based on Geographic Information Systems (GIS). GIS are used to geo-reference a dataset, map it, display it, and analyse it by cross-referencing different layers of information. Such process of cross-referencing has provided researchers working in fields such as environmental history, historical demography, and economic, urban and medieval history with new perspectives leading them to challenge traditional positions and even explore new questions.

One of the conceptual challenges digital humanities is engaging with as it evolves out of GIS mapping, is to identify the layers of meaning humans attached historically to geographical spaces. Only when humans invest meaning in such landscapes, locales and dwellings, the objective space is turned into a "place" that they can be emotionally attached to, and that can be part of collective narratives of pride, longing or loss. Whereas geographical space is "a realm without meaning […] a 'fact of life', which, like time, produces the basic coordinates for human life," the concept of place is space in which humans have invested meaning (Cresswell, 2010, p. 10; Gregory, 2014; Withers, 2009). One of the central aims of spatial humanities has become to connect these cultural constructs to an Euclidian grid of space, for instance by performing "deep mapping" (Bodenhamer, Corrigan, & Harris, 2015b, pp. 26–28, passim).

Importantly, the subjective attachment to place is expressed in the way such locations places are given proper names. Whereas GIS codes represent digital definitions of geographical space, toponyms or place names are often much more ambiguous cultural markers that represent attachment, fear, longing or other sentiments. Spatial humanities has begun to connect these subjective connotations to the toponyms that can be extracted as named entities from digitized texts, such as travel narratives, novels and newspapers (Donaldson, Gregory, & Taylor, 2017; Tally, 2011; Taylor, Donaldson, Gregory, & Butler, 2018).

Although much work has been done in the field of geographic text analysis (GTA), one of the next challenges within spatial humanities that this article addresses is how such geographical markers change over time as a result of *human movement and migration* (White, 2010, pp. 17, 18, passim). Here we attempt to tackle questions concerning the way places are mentioned by migrants, individuals who are displaced and have to make sense of their lives across contesting cultural values and traditions. The challenge is to trace how toponyms serve as *diasporic idiotopes* that migrants use to negotiate the attachment to their lost homeland and the new host society. We argue that spatial humanities can help to extract such indicators of a sense of place in large heritage collections and map the "persistence of place in a hectic world" (Pascual-de-Sans, 2004).

Drawing from the application of ML, GIS, data mining, and visualisation, the paper discusses how performing cNER on ethnic newspapers can provide researchers with new perspectives on the geographies of the past so as to open up new ways to study the socio-cognitive dimension of migration history.

3 THE TASK OF NER

In order to conceptualise the historicization of place name disambiguation, a method is needed to identify toponymic identifiers in big data collections within their proper historical context (such as place names, borders, and nations). Current Named Entity Recognizers (such as the much-used Stanford Named Entity Tagger) assume rigid designators and are historically static, as often based on contemporary word lists. This means that they cannot interpret historical changes in place names (e.g., St. Petersburg - Petrograd - Leningrad) within their proper historical context and cannot deal with culturally ambiguous geographical references (e.g., America, United States, the New World, Washington) or homonymic identifiers (e.g., Limerick – limerick) (Marrero et al, 2013; Neudecker, 2014).

The most common way to overcome these limitations is to train classifiers, employing semi-supervised ML, sometimes using crowdsourcing, or to use hand-crafted grammar-based systems that require intensive supervision by computational linguists. The required annotation labour prohibits application to big datasets of historical periodicals and other serial texts that are currently available (Ju et al., 2016).

An alternative route is to disambiguate entities on the basis of contextual information sources such as Wikipedia (Coll Ardanuy, 2017; Coll Ardanuy & Sporleder, 2017; Zhang & Iria, 2009) or to combine place name taggers with a scoring system within an ensemble (Canale, Lisena, & Troncy, 2018; Won, Murrieta-Flores, & Martins, 2018). As recent surveys confirm, however, ML algorithms based on neural networks now outperform all methods that are based on gazetteers or static databases (Yadav & Bethard, 2019). These latter methods present two major advantages in text enriching: 1) they may be based on the historical context of a historical corpus (McDonough, Moncla, & van de Camp, 2019); 2) they are able to recognize toponyms in a dynamic way as a geographical concept (Eijnatten, 2019). We propose to use the term *conceptual NER* (cNER) as a level of analysis to enrich place name analysis.

4 METHODOLOGY AND DATASET

To establish and maintain internal cohesion whilst distinguishing themselves from others, diasporic groups adopt a collective concept of identity. Such shared identity is constructed through a variety of markers which may be religious, linguistic, performative and of other nature, for example geographical. This type of diasporic identity markers is conveyed through language to both express the bond with a 'remembered homeland' and the connection with the host society. cNER applies a pragmatic perspective to the task of NER by assessing the socio-linguistic information and historical context in which places are mentioned and discussed. This is done by combining the visualisation of place name references with linguistic, social, and historical data, both automatically -for instance by means of sentiment analysis- and non-automatically, through the qualitative analysis of sample excerpts. The overarching aim is to visualise the narratives behind distant and local places and obtain deeper insights of how such links were maintained and renegotiated over time for contemporary purposes and future generations. Here we focus on the ML (*cfr.* 4.1) and geocoding parts (*cfr.* 4.2) of the study.

4.1 Machine Learning

The collection was first tagged for entities using an advanced ML sequence tagging tool that implements Tensorflow (Riedl and Padó, 2018). The novelty of the tool lies in the fact that it combines BiLSTM and CRF and character embeddings. The two methods were tested separately against four datasets to compare both individual performances and the combination of the two. They found that combining BiLSTM with a CRF as top layer outperforms CRFs with hand-coded features consistently when enough data is available. They concluded that modern RNNs have much to recommend to researchers working in NER as they consistently yield the best performance.

Methodologically, they trained the character embeddings with pre-trained word embeddings while training the model itself. They also used character- and subword based word embeddings computed with FastText (Bojanowski et al., 2017) which, by retrieving embeddings for unknown words through the incorporation of subword information, was found to significantly alleviate issues with out-of-vocabulary words.

4.2 Geocoding

Once the tagging task was completed, locations were geocoded by using the Google API. Geocoding with Google is a two-stage process that requires Google Geocoding API[1]. First, Google Geocoding API provides users with a Place ID for each location. The Place ID uniquely identifies a place as it is stored in the Google Places database and on Google Maps. Because the language of the dataset was Italian, the language of the API was set to Italian. It was found that setting the API language as the language of the dataset improved the accuracy of the geocoding results. At the same time, however, this meant that the results were also returned in Italian. Therefore, in order to have the results returned in English, only the Place ID was extracted in this first stage. Once the Place ID was received, it was possible to use the Google Geocoding API to perform reverse geocoding, that is to obtain all the details for the location in English (e.g., geo-coordinates, administrative level).

4.3 ChroniclItaly

To demonstrate the potential of cNER, we used a corpus of Italian ethnic newspapers (i.e., *ChroniclItaly,* Viola 2018) as an example of diasporic media published in the United States between 1898 and 1920. *ChroniclItaly* is an open access collection that includes all front pages of seven Italian language newspapers published in California, Massachusetts, Pennsylvania, Indiana, Vermont, and West Virginia be-

[1] https://developers.google.com/maps/documentation/geocoding/start

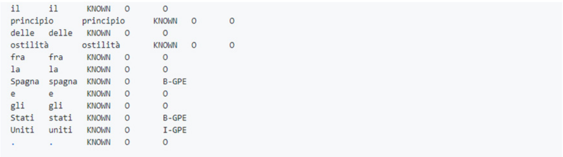
tween 1898 and 1920. The corpus, which was extracted from the *Chronicling America* newspaper collection of the Library of Congress, includes 4,810 issues and for a total of 16,624,571 words. Featuring mainstream (*prominenti*), radical (*sovversivi*), and politically independent newspapers, *ChroniclItaly* is a well-balanced resource for the study of the Italian immigrant press of the time. Moreover, because it is entirely digital, this corpus is a powerful tool for conducting text-based searches and analysis, both quantitative and qualitative. The newspapers' titles are: *L'Italia*, *Cronaca sovversiva*, *La libera parola*, *The patriot*, *La ragione*, *La rassegna*, and *La sentinella del West Virginia*.

Although immigrant newspapers have often been used by migration historians also to study questions of belonging in relation to space, such socio-cognitive dimension of migration remains largely unexplored digitally. This includes the lack of use of immigrant newspapers not only as a source of data analysis but also as a starting point for creating research and analysis tools. Thus, by using *ChroniclItaly*, we also aimed to contribute to the debate about the lack of diversity, archival biases and silences in the archives in digital scholarship. Thanks to larger amounts of data, today the digital analysis of place name references in immigrant storytelling allows researchers to understand how individuals made sense of their diasporic identities within the host community and perhaps reconsider previous interpretations.

Finally, it is worth mentioning that digital scholars wishing to carry out research in languages other than English often find themselves confronted with the relative lack of appropriate computational resources, including for instance accessing already available trained models in the desired language. Thus, by both using and creating resources in Italian, the study also addresses the issue of underrepresentation of languages other than English in digital scholarship.

5 TAGGING THE CORPUS

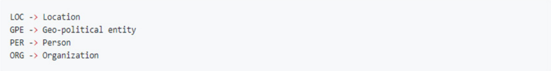
The sequence tagging model for the Italian language was trained on I-CAB (Italian Content Annotation Bank), an open access corpus annotated for entities (i.e. persons-PER, organizations-ORG, locations-LOC, and geopolitical entities-GPE), temporal expressions, and relations between entities. I-CAB contains 525 news articles taken from the Italian newspaper *L'Adige* and totals up around 180,000 words. Embeddings were computed using Italian Wikipedia and they have been trained using Fastext with 300 dimensions[2]. Once the training was complete, the output had the following format (Figure 1):



```
il      il      KNOWN   O       O
principio       principio       KNOWN   O       O
delle   delle   KNOWN   O       O
ostilità        ostilità        KNOWN   O       O
fra     fra     KNOWN   O       O
la      la      KNOWN   O       O
Spagna  spagna  KNOWN   O       B-GPE
e       e       KNOWN   O       O
gli     gli     KNOWN   O       O
Stati   stati   KNOWN   O       B-GPE
Uniti   uniti   KNOWN   O       I-GPE
.       .       KNOWN   O       O
```

Figure 1: Output of the sequence tagger for *ChroniclItaly*.

The first column is the input word, the second column specifies the pre-processed, lowercased word, the third column contains a flag, that is whether the word has been known during training (KNOWN) or not (UNKNOWN). If labels are assigned to the input file, these will appear in the third column. The last column contains the predicted tags. The no-entity tag is O. Because some entities (e.g., *Stati Uniti* "United States") have multiple words, the tagging scheme distinguishes between the beginning (tag B-...) or the inside of an entity (tag I-...). Figure 2 shows the tags:



```
LOC -> Location
GPE -> Geo-political entity
PER -> Person
ORG -> Organization
```

Figure 2: Tags of the sequence tagger.

6 ML RESULTS

The sequence tagger retrieved 1,369 unique locations (both LOC and GPE) which occurred 214,110 times throughout the whole corpus. Because each individual document was time stamped, the number of references to each location was quantified at any given time within the timeframe of *ChroniclItaly* (i.e., 1898-1920). The results of the F1 score for Italian models are shown in Table 1:

Table 1: F1 score for Italian models.

Type	Score
accuracy	98.15%
precision	83.64%
recall	82.14%
FB1	82.88

[2] https://fasttext.cc/docs/en/crawl-vectors.html

Table 2 shows the F1 score for each of the entity:

Table 2: F1 score for Italian models per entity.

Entity	Precision	Recall	FB1
GPE	83.90%	86.18%	85.02
LOC	69.70%%	44.23%	54.12
ORG	73.36%	73.08%	73.22
PER	89.78%	87.59%	88.68

The tagged version of *ChroniclItaly* is *ChroniclItaly 2.0* (Viola, 2019) and it is available as an open access resource[3].

7 VISUALISATION

To visualise the results, we chose to use a Shiny[4] app, the *GeoNewsMiner* (GNM, Viola et al 2019). This allowed us to present and analyse the data in an intuitive, interactive, and reproduceable way. Within GNM, references to place names in *Chroniclitaly* can be explored according to five different levels of aggregations:

- Time: from 1898 to 1920;
- Newspaper's title
- Type of frequency visualization: absolute, percentage
- Geographical information: include/exclude references to cities/regions
- Percentile: select the least/most mentioned places

In order to reflect the changing geo-political borders within the analysed period (1898-1920), users can additionally choose between three historical world maps displaying the different borders at three crucial points in history which intersect with the timeframe of *ChroniclItaly*: 1880, 1914, 1920. By default, GNM displays a contemporary (1994) world map. Finally, users can also share their results, download the raw data which may reflect their filters' selection, and download the visualised map as a .png file. GNM is available as an open access resource,[5] a full documentation of the project is also available on GitHub[6]. Figure 3 shows a static image of the GNM app.

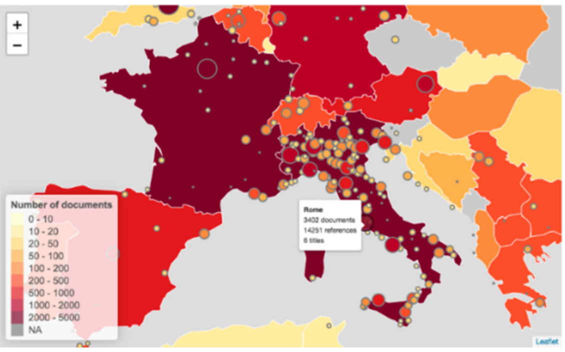


Figure 3: Static image of the GNM app.

8 DISCUSSION

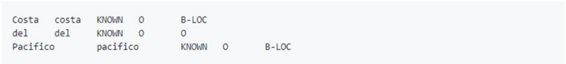
The method presented a number of technical challenges. The first remark concerns the performance of the sequence tagger. Although the overall F1 score was satisfactory, the performance for the entity LOC was rather poor. However, in *ChroniclItaly* GPE entities significantly outnumbered LOC entities (77.2% GPE vs 22.8% LOC), thus at least partially compensating for this limitation.

This first issue can be attributed to the second challenge of the study, namely OCR issues. The newspapers in *ChroniclItaly* were digitized primarily from microfilm holdings. In addition to the well-known limitations for OCR processes such as unusual text styles or very small fonts, other limitations occur when dealing with old material, including markings on the pages or a general poor condition of the original text. Such limitations also apply to the OCR-generated searchable texts in *ChroniclItaly* which therefore contain errors. However, the OCR quality was found better in the most recent texts, perhaps due to a better conservation status or better initial condition of the originals which overall improved over the course of the nineteenth century. Therefore, the quality of the OCR data can vary greatly even within the same newspaper. The OCR error limitation could however be at least partially overcome in two ways: first, it was reasonable to assume that important concept words would have been repeated several times within an article thus increasing the likelihood that OCR read them correctly in at least some of the passages. Second, the geo-coding was restricted to place names that were referred to at least more than 8 times across the whole collection as the highest number of false positives was found for occurrences <8.

Other minor issues mainly concerned multi-name locations (e.g., *Costa del Pacifico* 'Pacific Coast')

[3] https://doi.org/10.24416/UU01-4MECRO
[4] https://shiny.rstudio.com/
[5] https://utrecht-university.shinyapps.io/GeoNewsMiner/
[6] https://github.com/lorellav/GeoNewsMiner

which were tagged as individual parts, as shown in Figure 4:



```
Costa   costa   KNOWN   O       B-LOC
del     del     KNOWN   O       O
Pacifico        pacifico        KNOWN   O       B-LOC
```

Figure 4: Example of tags of the sequence tagger for multi-name locations.

The third challenge concerned the geocoding phase. In addition to the geo-coordinates, Google provides further details, such as the tag *type[]* indicating why those specific geo-coordinates have been attributed to a location. Understanding the *type[]* of a location is therefore very important, especially when working with historical data. The Google Places database stores places based on a contemporary world map, however the locations in a historical dataset may have changed name or may no longer exist. Moreover, categories such as country, city, region, municipality, etc. which Google uses to determine the location *type[]* are highly dependent on the location itself and consequently, certain categories may not apply or they may change from country to country. In geocoding the places in *ChroniclItaly 2.0*, we encountered cases in which the type and level given by Google required a manual edit. The full list of remarks and manual edits is available in the GNM GitHub repository.

Despite the described challenges, we argue in favour of ML methods to enrich digital collections. One of the biggest advantages of using ML for tagging, for instance, is that it is not based on gazetteers, therefore place name extraction and disambiguation is more reliable. Another advantage is that models can be continuously improved by increasing the amount and quality of data the algorithms learn from, so that they can make faster and more accurate predictions. Finally, ML methods offer the huge benefit of being completely unsupervised thus bearing great potential of assistance also for not highly computationally trained researchers, such as humanities scholars. Because a solid training in ML was not necessary for using the sequence tagger, indeed, we found that the most challenging part of the project was the geocoding phase. Therefore, ML bears great potential for the humanities: with modest investment of time, it may be possible to enrich automatically large amounts of data, saving scholars precious time and resources that can be dedicated to investigating new avenues of data analysis.

9 CONCLUSIONS

This paper discussed the added value of applying ML to contextually enrich digital collections. In this study, we employed ML as a method to geographically enrich a historical dataset. Specifically, we used a deep learning architecture for NER tasks (Riedl and Padó 2018) which implements TensorFlow to perform NER on a corpus of historical immigrant newspapers (*ChroniclItaly*, Viola 2018). The aim was to prepare large quantities of unstructured data for a conceptual historical analysis of geographical references, which we called *conceptual* Named Entity Recognition, cNER. Triangulating the quantitative information provided by ML with data visualisation and a qualitative (i.e., sentiment), socio-historical and linguistic analysis, cNER enables us to trace and understand the changing cultural constructions that are attached to place names, as they are derived from the historical context. This further pragmatic level of analysis will help us to establish how imagined "place" is defined over time in relation to changing realities of space. We argue that overall the cNER method does better justice to the historical ambiguities that are embedded in the texts themselves than extraction on the basis of gazetteers or static external information allow us to achieve.

Despite a number of limitations which were mainly encountered during the geocoding phase, we found that the method has much to recommend particularly to humanities scholars who are more and more confronted with the challenge of exploring collections larger than before and in a digital format. Finally, we also argued that, by choosing to use largely neglected sources such as Italian immigrant newspapers, this study contributed to the debate about language diversity representation and archival biases in digital practices.

REFERENCES

Ardanuy, Maria Coll. (2017). *Entity-Centric Text Mining for Historical Documents*. Georg-August-Universitat Gottingen, Göttingen.

Ardanuy, Mariona Coll, & Sporleder, C. (2017). Toponym disambiguation in historical documents using semantic and geographic features. *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017*, 175–180. https://doi.org/10.1145/3078081.3078099

Bodenhamer, D. J., Corrigan, J., & Harris, T. M. (Eds.). (2010). *The spatial humanities: GIS and the future of humanities scholarship*. Bloomington, Ind.: Indiana Univ. Press.

Bodenhamer, D. J., Corrigan, J., & Harris, T. M. (Eds.). (2015a). *Deep maps and spatial narratives*. Bloomington: Indiana University Press.

Bojanowski, P. Grave, E., Joulin, A. and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5:135–146.

Canale, L., Lisena, P., & Troncy, R. (2018). A Novel Ensemble Method for Named Entity Recognition and Disambiguation Based on Neural Network. In D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, … E. Simperl (Eds.), *The Semantic Web – ISWC 2018* (Vol. 11136, pp. 91–107). https://doi.org/10.1007/978-3-030-00671-6_6

Cresswell, T. (2010). *Place: A short introduction* (Repr.). Malden, Mass.: Blackwell.

Donaldson, C., Gregory, I. N., & Taylor, J. E. (2017). Locating the beautiful, picturesque, sublime and majestic: Spatially analysing the application of aesthetic terminology in descriptions of the English Lake District. *Journal of Historical Geography*, *56*, 43–60. https://doi.org/10.1016/j.jhg.2017.01.006

Eijnatten, J. V. (2019). Something about the Weather. Using Digital Methods to Mine Geographical Conceptions of Europe in Twentieth-Century Dutch Newspapers. *BMGN - Low Countries Historical Review*, *134*(1), 28–61. https://doi.org/10.18352/bmgn-lchr.10655

Gregory, I. N. (2014). Further Reading: From Historical GIS to Spatial Humanities: An Evolving Literature. In I. N. Gregory & A. Geddes (Eds.), *Toward spatial humanities: Historical GIS and spatial history* (pp. 186–202). Bloomington, Ind.: Indiana Univ. Press.

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., & McKenzie, G. (2016). Things and Strings: Improving Place Name Disambiguation from Short Texts by Combining Entity Co-Occurrence with Topic Modeling. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge Engineering and Knowledge Management* (Vol. 10024, pp. 353–367). https://doi.org/10.1007/978-3-319-49004-5_23

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, *35*(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004

McDonough, K., Moncla, L., & van de Camp, M. (2019). Named entity recognition goes to old regime France: Geographic text analysis for early modern French corpora. *International Journal of Geographical Information Science*, *33*(12), 2498–2522. https://doi.org/10.1080/13658816.2019.1620235

Murrieta-Flores, P., & Martins, B. (2019). The geospatial humanities: Past, present and future. *International Journal of Geographical Information Science*, *33*(12), 2424–2429.

Neudecker, C. (2014, March 3). Named Entity Recognition for digitised newspapers – Europeana Newspapers. Retrieved 10 November 2019, from http://www. europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/

Pascual-de-Sans, A. (2004). Sense of place and migration histories Idiotopy and idiotope. *Area*, *36*(4), 348–357. https://doi.org/10.1111/j.0004-0894.2004.00236.

Riedl, M. and Padó, S. 2018. A Named Entity Recognition Shootout for German. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Short Papers), pages 120–125. Melbourne, Australia, July 15 - 20, 2018

Tally, R. T. (Ed.). (2011). *Geocritical explorations: Space, place, and mapping in literary and cultural studies*. New York: Palgrave Macmillan.

Taylor, J., Donaldson, C. E., Gregory, I. N., & Butler, J. O. (2018). Mapping Digitally, Mapping Deep: Exploring Digital Literary Geographies. *Literary Geographies*, *4*(1), 10–19.

Viola, L. (2018). *ChroniclItaly: A corpus of Italian American newspapers from 1898 to 1920. Utrecht University*. Retrieved from https://public.yoda.uu.nl/i-lab/UU01/T4YMOW.html

Viola, L. (2019). *ChroniclItaly 2.0. A corpus of Italian American newspapers annotated for entities, 1898-1920 (Version 2.0)*. Retrieved from https://doi.org/10.24416/UU01-4MECRO

Viola, L., De Bruin, J., van Eijden, K., & Verheul, J. (2019). *The GeoNewsMiner (GNM): An interactive spatial humanities tool to visualize geographical references in historical newspapers (v1.0.0)*. Retrieved from https://github.com/lorellav/GeoNewsMiner

White, R. (2010). Spatial History Project. Retrieved 8 November 2019, from https://web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29

Withers, C. W. J. (2009). Place and the 'Spatial Turn' in Geography and in History. *Journal of the History of Ideas*, *70*(4), 637–658. https://doi.org/10.1353/jhi.0.0054

Won, M., Murrieta-Flores, P., & Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, *5*. https://doi.org/10.3389/fdigh.2018. 00002

Yadav, V., & Bethard, S. (2019). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *ArXiv:1910.11470 [Cs]*. Retrieved from http://arxiv.org/abs/1910.11470

Zhang, Z., & Iria, J. (2009). A novel approach to automatic gazetteer generation using Wikipedia. *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, 1–9. Retrieved from http://dl.acm.org/citation.cfm?id=1699765.1699766