Journal of Open Humanities Data

A Dataset for Toponym Resolution in Nineteenth-Century English Newspapers

DATA PAPER

MARIONA COLL ARDANUY
DAVID BEAVAN
KASPAR BEELEN
KASRA HOSSEINI
JON LAWRENCE
KATHERINE MCDONOUGH
FEDERICO NANNI
DANIEL VAN STRIEN
DANIEL C. S. WILSON

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

We present a new dataset for the task of toponym resolution in digitized historical newspapers in English. It consists of 343 annotated articles from newspapers based in four different locations in England (Manchester, Ashton-under-Lyne, Poole and Dorchester), published between 1780 and 1870. The articles have been manually annotated with mentions of places, which are linked—whenever possible—to their corresponding entry on Wikipedia. The dataset consists of 3,364 annotated toponyms, of which 2,784 have been provided with a link to Wikipedia. The dataset is published in the British Library shared research repository, and is especially of interest to researchers working on improving semantic access to historical newspaper content.

CORRESPONDING AUTHOR:

Mariona Coll Ardanuy

The Alan Turing Institute, London, UK; Queen Mary University of London, London, UK

mcollardanuy@turing.ac.uk

KEYWORDS:

benchmark; dataset; geographic information retrieval; newspapers; nineteenth-century English; toponym resolution

TO CITE THIS ARTICLE:

Coll Ardanuy, M., Beavan, D., Beelen, K., Hosseini, K., Lawrence, J., McDonough, K., Nanni, F., van Strien, D., & Wilson, D. C. S. (2022). A Dataset for Toponym Resolution in Nineteenth-Century English Newspapers. *Journal of Open Humanities Data*, 8(1), 3, pp. 1–7. DOI: *https://doi.org/10.5334/johd.56*

1 OVERVIEW

In this paper, we present a new dataset for the task of toponym resolution in digitized historical newspapers in English. Toponym resolution is a subtask of entity linking, focused on detecting and resolving mentions of places (i.e., toponyms) to their corresponding referent in a gazetteer or other knowledge base. Resolving toponyms in texts enables new forms of large-scale semantic and geographic analyses. However, most approaches to entity linking and toponym resolution are optimized to perform well with clean texts originally intended for a global audience and they do not generalize well to noisy, historical, or regional texts (Ehrmann, Romanello, Flückiger, & Clematide, 2020; Gritta, Pilehvar, Limsopatham, & Collier, 2018; Wang & Hu, 2019). Some entity linking datasets have been created to address this issue, such as Ehrmann et al. (2020) and Hamdi et al. (2021), both built from digitized historical newspaper collections.

Our dataset differs from others in its emphasis on the geographical aspect of newspaper data. The British provincial press—from which we sampled our articles—was strongly anchored in place: articles and advertisements were selected and edited with a local audience in mind. After the repeal on the 'taxes on knowledge' in the 1850s and 1860s, the provincial press proliferated; its readership expanded as did the number of titles, trumping the London-based press in size. Despite this plethora of available materials, to date historians have mostly favored the Metropolitan papers at the expense of the local press, which remains largely understudied (Beelen, Lawrence, Wilson, & Beavan, under submission; Hobbs, 2018). As shown in Lieberman, Samet, and Sankaranarayanan (2010) and Coll Ardanuy et al. (2019), the distribution of places mentioned in newspapers varies considerably depending on their intended audience (grounded in a certain place and time), hindering the resolution of ambiguous place names. Our dataset has been created to assess the robustness of entity linking and toponym resolution methods in this particularly challenging but common scenario. We hope that improved toponym resolution for these newspapers will translate into greater interest in them as research materials.

This dataset is comprised of 343 articles carefully sampled from a variety of provincial nineteenth-century newspapers based in four different locations in England. The articles have been manually annotated with mentions of places, which are linked—whenever possible—to their corresponding entry on Wikipedia. A total of 3,364 toponyms have been annotated, of which 2,784 have been linked to Wikipedia. The text of the articles is OCR-generated and has not been manually corrected. The dataset has been created with the aim of becoming a benchmark for several tasks: fuzzy string matching and toponym recognition and resolution, among others, all of which contribute to the challenging pursuit of improving semantic access to OCRed historical texts in English.

This dataset has been produced as part of Living with Machines,[1] a multidisciplinary research project focused on the lived experience of industrialization in Britain during the long nineteenth century and, in particular, on the social and cultural impact of mechanization as reported in newspapers and other sources. Living with Machines is one of many projects that harness the growing volume of digitized newspaper collections for humanities research.[2] A fraction of the annotated data has been used in previous studies from Living with Machines, in particular Coll Ardanuy et al. (2019), and for fuzzy string matching in Hosseini, Nanni, and Coll Ardanuy (2020) and Coll Ardanuy et al. (2020).

2 METHOD

DATA PROCESSING

The initial source of the data was formatted as Metadata Encoding and Transmission Standard/Analyzed Layout and Text Object (METS/ALTO) files[3] and consisted of 72 newspaper titles of publications (including subsequent variant titles) from the English counties of Lancashire and Dorset. These were obtained from the genealogy company Find My Past, custodians of

1 *https://livingwithmachines.ac.uk* (last access: 2021-07-19).

2 Other notable projects are: Impresso (*https://impresso.github.io/*), NewsEye (*https://www.newseye.eu/*), Oceanic Exchanges (*https://oceanicexchanges.org/*), or ViralTexts (*https://viraltexts.org/*) (last access: 2021-08-20).

3 *https://www.loc.gov/standards/mets/* (last access: 2021-08-12).

the British Newspaper Archive, the most extensive corpus of digitised British newspapers.[4] This METS/ALTO file format contains both logical and physical layout information, along with document textual contents, expressed as Extensible Markup Language (XML).[5] It is verbose and does not lend itself directly to manipulation in natural language processing pipelines and tools. Instead, we used Extensible Stylesheet Language Transformations (XSLT)[6] to extract the plain text of each article; each article being explicitly segmented and identified in the METS logical structure map, the plain text extracted being all physical ALTO textblocks attributed to that article. This plain text is supplemented by minimal metadata extracted into in a companion file. This step is performed by `alto2txt`, which is a Python wrapper for those XSLT transformations, and is being prepared for public release via GitHub. This corpus consisted of 11,761,898 articles (as defined above). This metadata was ingested into a PostgreSQL[7] relational database for ease of querying and filtering, its relational schema mirrors directly the hierarchy of the metadata XML files.

SAMPLING

We created a subsample that consists of 343 articles published between 1780 and 1870 in local newspapers based in four different locations: Manchester and Ashton-under-Lyne (a large town and a medium-sized market town, broadly representing the industrial north of England), and Poole and Dorchester (respectively medium-sized port and market towns, representing the rural south).[8] ***Figure 1*** gives an overview of the number of annotated articles per decade and place of publication. We biased our sample toward articles that have a length between 150 and 550 words and an OCR quality confidence score greater than 0.7 (calculated as the mean of the per-word OCR confidence scores as reported in the source metadata). Most of the text is legible, even though it contains many OCR errors. See ***Table 1*** for a more detailed overview of the sample.

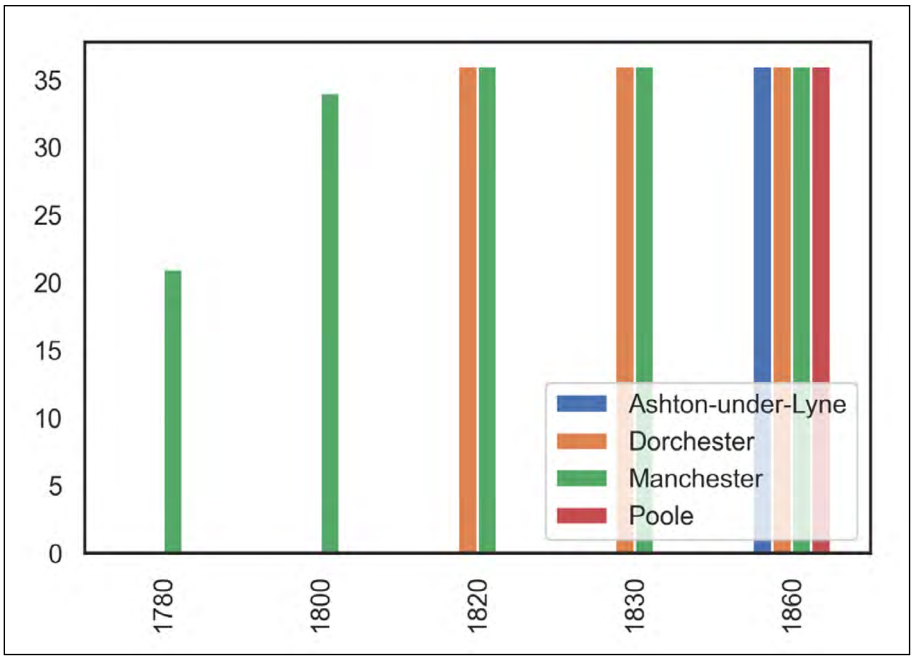

Figure 1 Number of annotated articles per decade and place of publication.

4 *https://www.britishnewspaperarchive.co.uk/* (last access: 2021-08-12).

5 *https://www.w3.org/standards/xml/* (last access: 2021-08-12).

6 *https://www.w3.org/Style/XSL/* (last access: 2021-08-12).

7 *https://www.postgresql.org/* (last access: 2021-08-12).

8 The newspapers from which we obtained the articles to annotate are *The Dorset County Chronicle*, *The Poole and South-Western Herald*, *The Manchester Mercury and Harrops General Advertiser*, *The Manchester Mercury; and Tuesdays General Advertiser*, *The Manchester Courier and Lancashire General Advertiser*, *The Ashton Reporter*, and *The Ashton and Stalybridge Reporter*.

	ASHTON	DORCHESTER			MANCHESTER					POOLE
	1860	*1820*	*1830*	*1860*	*1780*	*1800*	*1820*	*1830*	*1860*	*1860*
Number of articles	36	36	36	36	21	34	36	36	36	36
Avg word count	300	323	313	325	311	368	378	354	312	288
Avg OCR quality: mean	0.89	0.86	0.88	0.89	0.75	0.77	0.87	0.88	0.84	0.9
Avg OCR quality: sd	0.18	0.21	0.19	0.18	0.27	0.27	0.21	0.19	0.23	0.14

Table 1 Number of annotated articles, average article word count, and average article OCR quality mean and standard deviation per decade and place of publication.

We did not perform any manual post-processing to correct the errors produced in the OCR or layout recognition steps. Therefore, the toponyms in this dataset often contain OCR errors (e.g., 'iHancfjrcter' for 'Manchester'). Additionally, our dataset is rich with name variations that are characteristic of historical data, such as spelling variations (e.g., 'Leipsic' for 'Leipzig') and other forms of name change (e.g., 'Kingstown' for 'Dún Laoghaire').

ANNOTATION

Six annotators from different disciplinary backgrounds (history, literature, data science, and linguistics) manually annotated the toponyms present in the subsample. We used the Inception annotation platform[9] (Klie, Bugert, Boullosa, de Castilho, & Gurevych, 2018). A *toponym* is a mention of a location in a text. We defined a *location* as any entity that is static and can be represented by its geographical coordinates. Toponyms were classified into the following categories: `BUILDING` (names of buildings, such as the 'British Museum'), `STREET` (streets, roads, and other odonyms, such as 'Great Russell St'), `LOC` (any other real world places regardless of type or scale, such as 'Bloomsbury', 'London', or 'Great Britain'), `ALIEN` (extraterrestrial locations, such as 'Venus'), `FICTION` (fictional or mythical places, such as 'Hell'), and `OTHER` (other types of entities with coordinates, such as events, like the 'Battle of Waterloo'). Where possible, toponyms were linked to the corresponding Wikipedia entries (from which geographic coordinates can be derived) by their URL. This would be left empty if the location had no Wikipedia entry or the annotators were uncertain as to the correct disambiguation, either because the OCR made it impossible to correctly determine the referent or due to insufficient context.[10] While the annotations were made on the OCRed text, it was possible for the annotator to consult the original page image online on the British Newspaper Archive. Annotators were encouraged to discuss difficult choices with each other, and to document their decisions in a shared document. ***Table 2*** gives an overview of the annotations for each class.

CLASS	ANNOTATIONS	UNIQUE TOPONYMS	UNIQUE WIKIPEDIA LINKS	UNLINKED TOPONYMS
LOC	2764	1348	827	133
BUILDING	354	294	83	248
STREET	240	194	32	198
OTHER	5	5	5	0
FICTION	1	1	0	1
ALIEN	0	0	0	0

Table 2 Total number of annotations, unique toponyms, unique Wikipedia links and toponyms with no link to Wikipedia, per class.

9 *https://inception-project.github.io/* (last access: 2021-08-18).

10 For reference, we provide the original annotation guidelines together with the dataset. However, note that the final annotations have been refined in version 2 of the dataset. These changes are described in the accompanying `README` file.

QUALITY CONTROL

To assess the quality of the annotations, we had 77 newspaper articles annotated by two people, for a total of 740 annotation pairs. We used the Inception agreement functionality to assess the inter-annotator agreement between the two sets of annotations. Using the Krippendorff's alpha (nominal) measure, we obtained an agreement of 0.87 for place name detection and classification and 0.89 for linking to Wikipedia. To further ensure the quality of our resource, after the annotation process, a curator went through all the annotations and made final decisions on which annotations to keep and which to discard, making sure the annotations were consistent throughout the dataset.

3 DATASET DESCRIPTION

OBJECT NAME

`topRes19th_v2`.

FORMAT NAMES AND VERSIONS

We are sharing the annotated files in the `WebAnno TSV` (tab-separated values) file format, version 3.2.[11] There are 343 files, one for each newspaper article. Accompanying the dataset is an additional `tsv` file that contains the metadata associated with each article: word count, OCR quality mean and standard deviation, date (and decade) of publication, place of publication, newspaper publication code and publication title, and an additional field (`annotation_batch`) in which the article is assigned to one of three batches that are similarly distributed in terms of place and decade of publication (this field was used during the sampling process, and may be useful for researchers wishing to split the dataset for experimental purposes). We have also prepared a `README` file and the original annotation guidelines in `Markdown` markup. The present paper describes version 2 of the dataset.

CREATION DATES

2019-01-01 to 2021-07-27.

LANGUAGE

Nineteenth-century English.

LICENSE

The dataset is released under open license CC-BY-NC-SA, available at *https://creativecommons.org/licenses/by-nc-sa/4.0/*.

REPOSITORY NAME

The dataset is stored in the British Library shared research repository at *https://doi.org/10.23636/r7d4-kw08*.

PUBLICATION DATE

2021-12-03.

4 REUSE POTENTIAL

The vast archive of the British Newspaper Archive and other British historical newspaper corpora will be re-used by hundreds of scholars in the coming years. Establishing benchmark datasets like this provides a foundation for others to assess the performance of methods related to the identification and location of places in historical newspapers. Although toponym density was

11 The `WebAnno TSV` format is a `CoNLL`-based file format, a format that is a widely-used in natural language processing (especially in shared tasks of the Conference on Computational Natural Language Learning). It lists one token per line, the different layers of annotation are tab-separated, and it uses blank lines to separate sentences. See *https://webanno.github.io/webanno/releases/3.4.5/docs/user-guide.html#sect_webannotsv* (last access: 2021-07-27).

always greatest for newspapers' immediate locality, all newspapers included a rich diversity of national and international place names linked to reports of trade, war, conquest and state politics. Our annotations cover the different scales of places that make up the locations of the political, economic, and everyday life reported in nineteenth-century provincial newspapers. We hope that this dataset contributes to improving methods for finding difficult-to-recognize toponyms in digitized texts and linking them to context-appropriate knowledge base records.

ACKNOWLEDGEMENTS

Newspaper data has been provided by Findmypast Limited from the British Newspaper Archive, a partnership between the British Library and Findmypast (*https://www.britishnewspaperarchive.co.uk/*).

We thank the anonymous reviewers for their careful and constructive reviews. We are grateful to Giovanni Colavizza (University of Amsterdam) and James Hetherington (University College London) for helping with the research infrastructure, to Claire Austin (British Library) for helping with data access, and to the members of Living with Machines who helped with the annotations.

FUNDING STATEMENT

This work was supported by Living with Machines (AHRC grant AH/S01179X/1) and The Alan Turing Institute (EPSRC grant EP/N510129/1). The Living with Machines project, funded by the UK Research and Innovation (UKRI) Strategic Priority Fund, is a multidisciplinary collaboration delivered by the Arts and Humanities Research Council (AHRC), with the Alan Turing Institute, the British Library and the Universities of Cambridge, East Anglia, Exeter, and Queen Mary University of London.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTION

Mariona Coll Ardanuy (conceptualization, data curation, formal analysis, project management, writing), David Beavan (resources, software, writing), Kaspar Beelen (resources, data curation, writing), Kasra Hosseini (resources, software), Jon Lawrence (conceptualization, data curation, project management), Katherine McDonough (conceptualization, data curation, writing), Federico Nanni (validation, writing), Daniel van Strien (resources, software), Daniel C.S. Wilson (conceptualization, data curation, writing).

AUTHOR AFFILIATIONS

Mariona Coll Ardanuy *orcid.org/0000-0001-8455-7196*
The Alan Turing Institute, London, UK; Queen Mary University of London, London, UK

David Beavan *orcid.org/0000-0002-0347-6659*
The Alan Turing Institute, London, UK

Kaspar Beelen *orcid.org/0000-0001-7331-1174*
The Alan Turing Institute, London, UK; Queen Mary University of London, London, UK

Kasra Hosseini *orcid.org/0000-0003-4396-6019*
The Alan Turing Institute, London, UK

Jon Lawrence *orcid.org/0000-0001-6561-6381*
The University of Exeter, Exeter, UK

Katherine McDonough *orcid.org/0000-0001-7506-1025*
The Alan Turing Institute, London, UK; Queen Mary University of London, London, UK

Federico Nanni *orcid.org/0000-0003-2484-4331*
The Alan Turing Institute, London, UK

Daniel van Strien *orcid.org/0000-0003-1684-6556*
The British Library, London, UK

Daniel C. S. Wilson *orcid.org/0000-0001-6886-775X*
The Alan Turing Institute, London, UK; Queen Mary University of London, London, UK

REFERENCES

Beelen, K., Lawrence, J., Wilson, D. C., & **Beavan, D.** (under submission). *Victorian Perspectives on Digital Newspapers: Addressing bias and representativeness in digital heritage collections.*

Coll Ardanuy, M., Hosseini, K., McDonough, K., Krause, A., van Strien, D., & **Nanni, F.** (2020). A deep learning approach to geographical candidate selection through toponym matching. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL): Poster papers* (pp. 385–388). DOI: *https://doi.org/10.1145/3397536.3422236*

Coll Ardanuy, M., McDonough, K., Krause, A., Wilson, D. C., Hosseini, K., & **van Strien, D.** (2019). Resolving places, past and present: toponym resolution in historical British newspapers using multiple resources. In *Proceedings of the 13th Workshop on Geographic Information Retrieval* (pp. 1–6). DOI: *https://doi.org/10.1145/3371140.3371143*

Ehrmann, M., Romanello, M., Flückiger, A., & **Clematide, S.** (2020). Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers. In *CEUR Workshop Proceedings*. DOI: *https://doi.org/10.1007/978-3-030-58219-7_21*

Gritta, M., Pilehvar, M. T., Limsopatham, N., & **Collier, N.** (2018). What's missing in geographical parsing? *Language Resources and Evaluation, 52*(2), 603–623. DOI: *https://doi.org/10.1007/s10579-017-9385-8*

Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G., & **Doucet, A.** (2021). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2328–2334). DOI: *https://doi.org/10.1145/3404835.3463255*

Hobbs, A. (2018). *A Fleet Street in Every Town: The Provincial Press in England, 1855–1900.* Cambridge: Open Book Publishers. DOI: *https://doi.org/10.11647/OBP.0152*

Hosseini, K., Nanni, F., & **Coll Ardanuy, M.** (2020). DeezyMatch: A flexible deep learning approach to fuzzy string matching. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 62–69). DOI: *https://doi.org/10.18653/v1/2020.emnlp-demos.9*

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & **Gurevych, I.** (2018,-). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 5–9). Association for Computational Linguistics. Retrieved from *https://aclanthology.org/C18-2002/*

Lieberman, M. D., Samet, H., & **Sankaranarayanan, J.** (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)* (pp. 201–212). DOI: *https://doi.org/10.1109/ICDE.2010.5447903*

Wang, J., & **Hu, Y.** (2019). Are we there yet? Evaluating state-of-the-art neural network based geoparsers using EUPEG as a benchmarking platform. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities* (pp. 1–6). DOI: *https://doi.org/10.1145/3356991.3365470*

TO CITE THIS ARTICLE:
Coll Ardanuy, M., Beavan, D., Beelen, K., Hosseini, K., Lawrence, J., McDonough, K., Nanni, F., van Strien, D., & Wilson, D. C. S. (2022). A Dataset for Toponym Resolution in Nineteenth-Century English Newspapers. *Journal of Open Humanities Data*, 8(1), 3, pp. 1–7. DOI: *https://doi.org/10.5334/johd.56*

Published: 24 January 2022