GeoEDdA: A Gold Standard Dataset for Geo-semantic Annotation of Diderot & d'Alembert's Encyclopédie

Ludovic Moncla, Denis Vigier, Katherine Mcdonough

GeoEDdA: A Gold Standard Dataset for Geo-semantic Annotation of Diderot & d'Alembert's Encyclopédie

Ludovic Moncla[1,*], Denis Vigier[2] and Katherine McDonough[3]

[1]INSA Lyon, CNRS, UCBL, LIRIS, UMR 5205, F-69621
[2]Université Lumière Lyon 2, ICAR, UMR 5142, Lyon, France
[3]Lancaster University, UK

Abstract

This paper describes the methodology for creating *GeoEDdA*, a gold standard dataset of geo-semantic annotations from entries in Diderot and d'Alembert's eighteenth-century *Encyclopédie*. Aiming to explore spatial information beyond toponyms identified with the commonly used Named Entity Recognition (NER) task, we test the newer span categorization task as an approach for retrieving complex references to places, generic spatial terms, other entities, and relations. We test an active learning method, using the Prodigy web-based tool to iteratively train a machine learning span categorization model. The resulting dataset includes labeled spans from 2,200 paragraphs. As a preliminary experiment, a custom spaCy spancat model demonstrates strong overall performance, achieving an F-score of 86.42%. Evaluations for each span category reveal strengths in recognizing spatial entities and persons (including nominal entities, named entities and nested entities).

Keywords

Geo-semantic annotations, Spatial role labeling, Gold standard dataset, Span categorization, Spatial humanities

1. Introduction

This paper presents an annotation schema and active learning method for creating a gold standard dataset of geo-semantic annotations from entries in the *Encyclopédie ou Dictionnaire raisonnée des sciences, des arts et des métiers*, a key text of the Enlightenment printed between 1751 and 1772 (in French). Geo-semantic annotation, or spatial role labeling [1], involves the identification and labeling of place-specific information and semantic classes in text. By combining geospatial information with semantic context, the *GeoEDdA* dataset aims to facilitate multi-scale and -type text analysis. This enables research that depends on evidence of the interconnection between Enlightenment ideas, historical events, people, generic spatial forms, and specific places [2].

Like projects such as *Living with Machines* and *Space Time Narratives*, we seek to diversify what we capture when we collect spatial information from historical documents. For the former,

GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, 2024, Glasgow, Scotland
*Corresponding author.
ludovic.moncla@insa-lyon.fr (L. Moncla); denis.vigier@univ-lyon2.fr (D. Vigier); k.mcdonough@lancaster.ac.uk (K. McDonough)
https://ludovicmoncla.github.io (L. Moncla)
0000-0002-1590-9546 (L. Moncla); 0009-0006-0836-0985 (D. Vigier); 0000-0001-7506-1025 (K. McDonough)

the focus was on improving recognition of formal, theoretically locatable place names at multiple scales (populated places, but also streets, buildings, or other landmarks) [3]. On the other hand, the latter project has - similar to our work here - addressed the issue of conceptualizing and annotating generic place names such as *street*, *lake* (e.g. what they term "locales", in contrast to "locations", which are usually toponyms [4]). (Generic place names may also be embedded within toponyms, of course.) Thus a critical aspect of this computational approach to annotation lies in prioritizing digital humanities concerns about information retrieval over earlier concepts of what was defined as "spatial" within the confines of Natural Language Processing (NLP) and Geographic Information Retrieval tasks such as NER [5] and Spatial Role Labeling [6]. For historical research focused on the interplay between ideas, discourse, and spatial references, the commonly used token classification task has proved more restrictive than a span categorization task.

Like many digital humanities annotation tasks, the act of labeling historical documents (whether texts, images, or other media) creates an interpretative "world" which researchers develop claims about, reducing complexity enough to think with and establish relations between specific parts of documents [7]. The rise of models in digital history has somehow rarely translated into historians (digital or not) openly talking about modeling. Here, however, we are nevertheless "opening the black box of interpretation" [8] through the formalization of modeling in activities like the ones documented in this paper. In this work, as with the examples from *Living with Machines* and *Space Time Narratives*, we move beyond NLP-based traditions of linguistic annotation to define token-level classes that reflect the needs of spatially-driven historical research.

Before describing our the methodology, we clarify how we classify tokens. In token classification, the focus is on annotating individual tokens within the text such as in Named Entity Recognition (NER) [9]. Token classification assigns a class to each token, and any token is associated with at most one entity. This method allows for precise identification of specific terms or concepts, but may fall short in capturing the holistic meaning that arises from the interaction of multiple tokens. Approaches in the digital humanities that depend on token-based NER to identify spatial information in texts often miss complex expressions of spatial information , and are not designed to recognize generic place names. The former can appear as sequences ("city in France near the Atlantic Ocean"), embedded in non-spatial entities ("duchess of Brittany"), while the latter are simply not proper names ("park"). On the other hand, span categorization extends its purview to encompass such phrases or "nested" expressions [10, 11], aiming to capture contextual information. This broader approach not only accounts for the interplay between words but also considers the spatial relationships and semantic (or geo-semantic) content contained within a defined span. By addressing both token-level intricacies and broader spans, *GeoEDdA* strikes a balance between granularity and contextual richness, enabling a comprehensive exploration of the geo-semantic landscape in the *Encyclopédie*, and potentially in other historical texts. This methodological duality ensures that the dataset provides a robust foundation for research that goes beyond simple toponyms as the dominant unit of spatial analysis and embraces the flexibility of spans for spatial humanities research.

2. Annotation Schema

In order to facilitate the geo-semantic annotation process, we propose a well-defined tagset that refines the XML-TEI schema introduced in [12] and adapts it to *Encyclopédie* entries. The annotation classes reflect spatial information in the text, including named and unnamed (e.g. common/generic) entities. We also annotate non-spatial entities, thereby capturing formal and informal references to places in context alongside people, events, and objects. The tagset is as follows (with text in French followed by English translations as needed):

- *NC-Spatial*: a common noun that identifies a spatial entity (nominal spatial entity) including natural features, e.g.`ville/city`, `la rivière/river`, `royaume/kingdom`.
- *NP-Spatial*: a proper noun identifying the name of a place (spatial named entities), e.g. `France`, `Paris`, `la Chine/China`.
- *ENE-Spatial*: nested spatial entity , e.g. `ville de France/city in France`, `royaume de Naples/kingdom of Naples`, `la mer Baltique/Baltic Sea`.
- *Relation*: spatial relation, e.g. `dans/in`, `sur/on`, `à 10 lieues de/10 leagues from`.
- *Latlong*: geographic coordinates, e.g. `Long. 19. 49. lat. 43. 55. 44.`
- *NC-Person*: a common noun that identifies a person or persons (nominal spatial entity), e.g. `roi/king`, `l'empereur/emperor`, `les auteurs/authors`.
- *NP-Person*: a proper noun identifying the name of a person or persons (person named entities), e.g. `Louis XIV`, `Pline/Pliny`, `les Romains/Romans`.
- *ENE-Person*: nested person entity, e.g. `le czar Pierre/Tsar Peter`, `roi de Macédoine/king of Macedonia`.
- *NP-Misc*: a proper noun identifying spans not classified as spatial or person span, e.g. `l'Eglise/the Church`, `1702`, `Pélasgique/Pelasgian`.
- *ENE-Misc*: nested named entity not classified as spatial or person entity, e.g. `l'ordre de S. Jacques/Order of Santiago`, `la déclaration du 21 mars 1671/the declaration of March 21, 1671`.
- *Head*: name of article, e.g. `Rennes`
- *Domain-Mark*: words indicating the knowledge domain, provided by the editors, e.g. `Géographie`, `Geog.`, `en Anatomie`.

Each category within the tagset is carefully curated to capture the diverse nature of the content, ranging from nominal ("common", or generic) entities (i.e., *NC-**) and named entities (i.e., *NP-**) to nested entities (i.e., *ENE-**) and spatial relations (see examples above). As noted earlier, nominal entity mentions [13] can be used as a co-reference to a named entity, or may have the ability to refer to a unique object or place, or, alternatively, to a generic type of place. Nested spans aim to capture and structure nested entities (also known as extended named entities) [11].

By delineating the semantic content of spans such as spatial features, places, persons, and their relationships, our annotation guidelines provide clarity and consistency in the annotation process. This schema not only addresses the intricacies of spatial and person spans but also incorporates additional elements such as geographic coordinates, miscellaneous spans, domain markers, and entry names ensuring a holistic and nuanced approach to the geo-semantic

annotation of the corpus. Working with spans enables us to extend annotation to spatial references and relations that would be undetectable in an entity-driven approach. This tagset therefore represents a first step towards a spatial-historical interpretation of *Encyclopédie* content that is informed by a more thorough semantic analysis of simple and complex toponyms as well as "extra-toponymic" spatial information (e.g. the unnamed, generic words describing places, as well as spatial relations).

3. Active learning for dataset labeling

The process of labeling data can be time-consuming, as it is usually done manually. For humanities research, annotation is often performed by experts or by students trained by experts. Tools like Recogito [14] facilitate "semi-automatic" annotation by suggesting likely labels, but for large corpora and for tasks requiring hundreds of examples for each label, machine-learning-based active learning methods can now aid with annotation. This section describes our use of such techniques to optimize the labeling process. Active learning involves an intelligent selection of data samples for annotation, emphasizing the acquisition of the most informative instances that contribute significantly to model performance. In the humanities, active learning is still a relatively new approach being explored for reducing the time required to annotate training or evaluation data from texts, but it is already showing promising results for reducing the number of annotations required to improve performance on tasks like NER [15]. By leveraging iterative model-human interaction, active learning not only enhances the efficiency of dataset labeling but also minimizes the annotation burden on human annotators (in our case, the research team).

3.1. Methodology

To execute our geo-semantic annotation process effectively, we adopted an iterative methodology using Prodigy[1], a web-based annotation tool developed by ExplosionAI that supports creating labeled data for machine learning tasks including NER and span categorization.

The initial dataset is composed of 74,000 *Encylopédie* entries provided by the ARTFL Project[2]. During annotation, data are labeled on a paragraph-by-paragraph basis rather than by whole entries. This approach enables the subsequent distribution of paragraphs from lengthy articles across training and testing datasets, ensuring a more granular and representative annotation. By annotating at the paragraph level, we enhance the flexibility in constructing datasets for training and evaluation, allowing for a finer understanding of model performance and generalization on diverse textual contexts.

Initially, a small dataset was manually annotated using Prodigy by the project team members, allowing annotators to contribute their expertise in identifying and labeling spatial and person spans according to the predefined tagset. Subsequently, a first machine learning span categorization model was trained—specifically, the spaCy[3] spancat model embedded in the Prodigy training pipeline—on this annotated subset. Although initial evaluation scores were

[1] https://prodi.gy
[2] https://artfl-project.uchicago.edu
[3] https://spacy.io

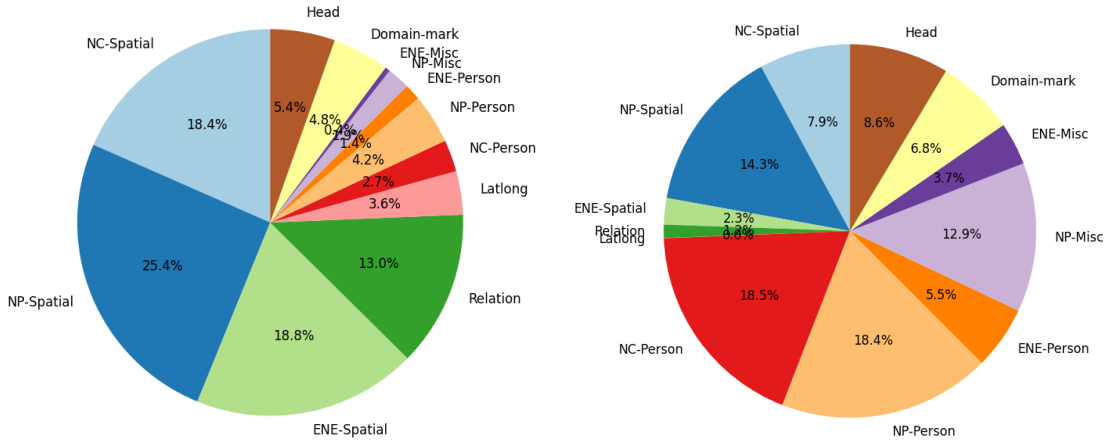
Figure 1: Example of the Prodigy interface for manual validation. Annotators can see basic metadata for the paragraph as well as the options for the spancat tags.

low, an iterative loop was established. In this loop, the trained model predicted annotations on additional paragraphs. Human annotators then interacted with these predictions through the Prodigy interface, correcting and refining the model outputs (see Figure 1). This iterative process progressively refines both the model's predictive capabilities and the overall dataset quality. By linking the strengths of human expertise with machine learning algorithms, our methodology aims to achieve an effective synergy in the geo-semantic annotation pipeline, ensuring the creation of a robust and accurate gold standard dataset. Through this iterative approach, we continuously evaluate the model's performance, correcting misclassifications, and reinforcing correct categorizations. This methodology not only contributes to building a gold standard dataset but also simultaneously trains a machine learning model, completing two goals at once.

The *GeoEDdA* gold standard dataset is composed of 2,200 randomly selected paragraphs from 2,001 entries. All paragraphs were written in French (mid eighteenth-century) and are distributed among the *Encyclopédie* knowledge domains as shown on Table 1. Knowledge domains were assigned to each paragraph using a BERT-based supervised text classification model trained on *Encyclopédie* entries (these represent a simplified, composite set of domains compared to the original domains noted in the entries) [16].

Table 1
Distribution of the annotated paragraphs among a simplified set of *Encyclopédie* knowledge domains.

Knowledge domain	Paragraphs	Knowledge domain	Paragraphs
Geography (*Géographie*)	1,096	Literature (*Belles-lettres*)	65
History (*Histoire*)	259	Military (*Militaire*)	62
Law (*Droit Jurisprudence*)	113	Commerce	48
Physics (*Physique*)	92	Fine arts (*Beaux-arts*)	44
Professions (*Métiers*	92	Agriculture	36
Medicine (*Médecine*)	88	Hunting (*Chasse*)	31
Philosophy (*Philosophie*)	69	Religion	23
Natural history (*Histoire naturelle*)	65	Music (*Musique*)	17



(a) in paragraphs classified as Géographie

(b) in paragraphs not classified as Géographie

Figure 2: Spans distribution

Figure 2 shows the spans distribution within paragraphs classified under Géographie or not. 80% of spans in paragraphs classified under Géographie refer to spatial spans (i.e., **-Spatial*, *Relation*, *Latlong*) (see Figure 2a) against only 25% in other paragraphs (see Figure 2b). This goes to 25% and 42% for person spans in paragraphs classified under Géographie, and paragraphs not classified as Géographie, respectively.

Table 2
Distribution of spans (tokens and paragraphs) across the datasets.

Span	Train	Validation	Test	Span	Train	Validation	Test
NC-Spatial	3,252	358	355	NP-Person	1,599	170	150
NP-Spatial	4,707	464	519	ENE-Person	492	49	57
ENE-Spatial	3,043	326	334	NP-Misc	948	108	96
Relation	2,093	219	226	ENE-Misc	255	31	22
Latlong	553	66	72	Head	1,261	142	153
NC-Person	1,378	132	133	Domain-Mark	1,069	122	133
Paragraphs	1,800	200	200	Spans	20,650	2,187	2,250
Tokens	132,398	14,959	13,881				

With the aim of utilizing this dataset for training and evaluating span categorization algorithms, a train/val/test split was performed. The validation and test sets each consist of 200 paragraphs: 100 classified under Géographie and 100 from another knowledge domain. The datasets can be downloaded from the HuggingFace Hub[4] and Zenodo[5] [17] and also from

[4]https://huggingface.co/datasets/GEODE/GeoEDdA
[5]https://zenodo.org/records/10530177

the Github project repository[6]. They are available in the JSONL format provided by Prodigy and the binary spaCy format (ready to use with the spaCy train pipeline). Table 2 shows the distribution of each span class and the total number of paragraphs, tokens, and spans across the datasets. As already observed in Figure 2, Table 2 shows that Spatial spans (*NC-Spatial*, *NP-Spatial*, *ENE-Spatial*, *Relation* and *Latlong*) are over represented in comparison to Person or Miscellaneous spans. This highlights the importance of geographical information even on paragraphs not classified as `Géographie`, but we also note the near absence non-Spatial spans within paragraphs classified as `Géographie`, e.g. *NC-Spatial*, *NP-Spatial*, *ENE-Spatial*, *Relation* and *Latlong* cover more than 75% of all labeled spans. Witnessing this particularly strong spatial signature of `Géographie` paragraphs - even for this small annotated dataset - points to the importance of distinguishing between types of content at different levels: word-, paragraph-, and article-level spatial information is not equally distributed across the *Encyclopédie*.

4. Training a Span Categorization Model

4.1. Custom spaCy spancat model

As a first experiment, a machine learning model was trained and evaluated using the proposed Gold Standard dataset. Developed within the spaCy natural language processing library, this model is tailored to our unique dataset and annotation schema requirements. It goes beyond traditional NER tasks by specifically categorizing spans (including longer phrases and nested entities), allowing for a more nuanced understanding of the relationships between entities within the text. Subsequently, we employed the spaCy span categorizer pipeline, comprising a suggester function that proposes candidate spans, potentially with overlaps, and a labeler model that predicts zero or more labels for each candidate. The model can be downloaded and installed directly from the HuggingFace Hub[7] and executed using the spaCy Python library.

4.2. Evaluation

To assess the performance of the geo-semantic span categorization model, we evaluated it using the test set described above. The overall model performance (on the Test set) demonstrates strong precision, recall, and F-score values, attaining 93.98%, 79.82%, and 86.33%, respectively. The model's performance by span category is presented in Table 3. Notably, the model exhibits high precision, recall, and F-score for crucial categories such as `NC-Spatial`, `NP-Spatial`, and `ENE-Spatial`. The model also excels in accurately categorizing `Head` and `Domain-mark` entities, achieving high precision and recall values. However, certain categories like `Latlong` and `ENE-Misc` are not recognized at all. This can be explained by the low number of examples (see Table 2) and the high diversity of forms (or values) of these two categories of spans. Only 255 `ENE-Misc` spans are present in the training set compared to 3,043 for `ENE-Spatial` for instance. Furthermore, the `Latlong` spans consist of numerical values that might not be as effectively represented as words in the embeddings [18].

[6] https://github.com/GEODE-project/ner-spancat-edda
[7] https://huggingface.co/GEODE/fr_spacy_custom_spancat_edda

Table 3
Model performance by span (Test set)

Tag	Precision	Recall	F-score	Tag	Precision	Recall	F-score
NC-Spatial	96.50	93.24	94.84	NP-Person	92.47	90.00	91.22
NP-Spatial	92.74	95.95	94.32	ENE-Person	92.16	82.46	87.04
ENE-Spatial	91.67	95.51	93.55	NP-Misc	93.24	71.88	81.18
Relation	97.33	64.60	77.66	ENE-Misc	0.00	0.00	0.00
Latlong	0.00	0.00	0.00	Head	97.37	24.18	38.74
NC-Person	93.07	70.68	80.34	Domain-mark	99.19	91.73	95.31

5. Conclusion

In this paper, we presented the creation of the *GeoEDdA* gold standard dataset for geo-semantic annotation of the *Encyclopédie*, including a review of the annotation schema and the active learning approach to labeling spans in the text. *GeoEDdA* includes over 20,000 annotated spans across 2,200 paragraphs. As an initial experiment, we trained and evaluated a spaCy span categorization model, yielding insights into its strengths and areas for improvement. This evaluation provides a comprehensive understanding of the model's effectiveness in handling specific entities. It also lays a foundation for future refinements and broader applications in geo-semantic annotation across a wider variety of types of spatial and non-spatial information than traditional NER.

As further work, we intend to expand the *GeoEDdA* dataset by annotating additional data, aiming to enhance the model's performance across all span categories, with particular focus on improving geographical coordinates and miscellaneous spans categorization (e.g. new deposits will be added to the Zenodo record). Next, we plan to compare span categorization algorithms, further advancing our understanding and refining the capabilities of next generation, post-NER geo-semantic annotation models. The final step is using inferred results that depend on this dataset as training data for historical research. This will allow us to examine spatial language across all of the *Encyclopédie* alongside additional information about, for example, knowledge domains. Ultimately, this documentation of our thinking about spatial information reflects a commitment to open historical research and promotes a more flexible approach to recognizing spatial language.

Acknowledgments

The authors, all members of the GEODE team, are grateful to the ASLAN project (ANR-10-LABX-0081) of the Université de Lyon, for its financial support within the French program "Investments for the Future" operated by the National Research Agency (ANR).

References

[1] J. Zhou, W. Xu, End-to-end learning of semantic role labeling using recurrent neural networks, in: Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on NLP, Beijing, China, 2015, pp. 1127–1137. doi:10.3115/v1/P15-1109.

[2] K. McDonough, L. Moncla, M. Van de Camp, Named entity recognition goes to old regime france: geographic text analysis for early modern french corpora, International Journal of Geographical Information Science 33 (2019) 2498–2522. doi:10.1080/13658816.2019.1620235.

[3] M. C. Ardanuy, F. Nanni, K. Beelen, L. Hare, The past is a foreign place: Improving toponym linking for historical newspapers, in: Proceedings of the Computational Humanities Research Conference 2023, Paris, France, December 6-8, 2023, volume 3558 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 368–390.

[4] I. Ezeani, P. Rayson, I. Gregory, E. Haris, A. Cohn, J. Stell, T. Cole, J. Taylor, D. Bodenhamer, N. Devadasan, E. Steiner, Z. Frank, J. Olson, Towards an extensible framework for understanding spatial narratives, in: Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities, ACM, Hamburg, Germany, 2023, p. 1–10. doi:10.1145/3615887.3627761.

[5] C. B. Jones, R. S. Purves, Geographical information retrieval, International Journal of Geographical Information Science 22 (2008) 219–228. doi:10.1080/13658810701626343.

[6] P. Kordjamshidi, M. Van Otterlo, M.-F. Moens, Spatial role labeling: Towards extraction of spatial relations from natural language, ACM Transactions on Speech and Language Processing (TSLP) 8 (2011) 1–36.

[7] D. Gerstorfer, E. Gius, J. Jacke, Working on and with Categories for Text Analysis: Challenges and Findings from and for Digital Humanities Practices, Digital Humanities Quarterly 17 (2023).

[8] S. Schwandt, Opening the black box of interpretation: Digitla history practices as models of knowledge, History and Theory 61 (2022) 77–85. doi:10.1111/hith.12281.

[9] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named entity recognition and classification in historical documents: A survey, ACM Computing Surveys 56 (2023) 1–47. doi:10.1145/3604931.

[10] M. Gaio, L. Moncla, Extended named entity recognition using finite-state transducers: An application to place names, in: Proceedings of the 9th international conference on advanced geographic information systems, applications, and services (GEOProcessing 2017), Nice, France, 2017.

[11] Y. Wang, H. Tong, Z. Zhu, Y. Li, Nested named entity recognition: a survey, ACM Transactions on Knowledge Discovery from Data (TKDD) 16 (2022) 1–29. doi:10.1145/3522593.

[12] L. Moncla, M. Gaio, A multi-layer markup language for geospatial semantic annotations, in: Proceedings of the 9th Workshop on Geographic Information Retrieval, Paris, France, 2015, pp. 1–10. doi:10.1145/2837689.2837700.

[13] A. Medad, M. Gaio, L. Moncla, S. Mustière, Y. Le Nir, Comparing supervised learning algorithms for spatial nominal entity recognition, in: Proceedings of the 23rd AGILE

Conference on Geographic Information Science, Chania, Greece, 2020, pp. 1–18. doi:10.5194/agile-giss-1-15-2020.

[14] R. Simon, E. Barker, L. Isaksen, P. De Soto Cañamares, Linked data annotation without the pointy brackets: Introducing recogito 2, Journal of Map & Geography Libraries 13 (2017) 111–132. doi:10.1080/15420353.2017.1307303.

[15] A. Erdmann, D. J. Wrisley, B. Allen, C. Brown, S. Cohen-Bodénès, M. Elsner, Y. Feng, B. Joseph, B. Joyeux-Prunel, M.-C. de Marneffe, Practical, efficient, and customizable active learning for named entity recognition in the digital humanities, in: Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, ACL, Minneapolis, Minnesota, 2019, pp. 2223–2234. doi:10.18653/v1/N19-1231.

[16] A. Brenon, L. Moncla, K. McDonough, Classifying encyclopedia articles: Comparing machine and deep learning methods and exploring their predictions, Data & Knowledge Engineering 142 (2022) 102098. doi:10.1016/j.datak.2022.102098.

[17] L. Moncla, D. Vigier, K. McDonough, GeoEDdA: A Gold Standard Dataset for Named Entity Recognition and Span Categorization Annotations of Diderot & d'Alembert's Encyclopédie, 2024. doi:10.5281/zenodo.10530178.

[18] D. Sundararaman, S. Si, V. Subramanian, G. Wang, D. Hazarika, L. Carin, Methods for numeracy-preserving word embeddings, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2020, pp. 4742–4753. doi:10.18653/v1/2020.emnlp-main.384.