Journal of Open Humanities Data

A Named Entity-Annotated Corpus of 19th Century Classical Commentaries

COLLECTION:
REPRESENTING THE ANCIENT WORLD THROUGH DATA

DATA PAPER

MATTEO ROMANELLO
SVEN NAJEM-MEYER
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

We release a multilingual named entity (NE) corpus of 19th century commentaries to Sophocles' *Ajax*. Selected commentaries are written in English, German and French, but are also replete with Latin and Greek quotes. Bibliographic entities were annotated along traditional named entities following our guidelines (Romanello & Najem-Meyer, 2022). The corpus contains about 300 annotated pages, 111,216 tokens and 7,334 entity mentions and was featured in the HIPE-2022 shared task. Although named entity recognition (NER) showed reassuring results, optical character recognition (OCR) mistakes and extensive use of abbreviation kept entity linking (EL) a challenging task. With such characteristics, this corpus offers an excellent way to assess the adaptability of information extraction systems to noisy, domain-specific multilingual and multiscript environments.

CORRESPONDING AUTHOR:
Matteo Romanello
Institute of Archeology and Classical Studies, University of Lausanne, Lausanne, Switzerland
matteo.romanello@unil.ch

KEYWORDS:
historical commentaries; classics; named entity recognition; entity linking; bibliographic reference extraction

TO CITE THIS ARTICLE:
Romanello, M., & Najem-Meyer, S. (2024). A Named Entity-Annotated Corpus of 19th Century Classical Commentaries. *Journal of Open Humanities Data*, 10: 1, pp. 1–7. DOI: https://doi.org/10.5334/johd.150

1 OVERVIEW

REPOSITORY LOCATION

GitHub: https://github.com/AjaxMultiCommentary/AjMC-NE-corpus, Zenodo: https://doi.org/10.5281/zenodo.8308015.

CONTEXT

Commentaries are prominent reference works in classical scholarship. Providing a text with explanatory glosses of various nature (historical, linguistic, mythological, etc.), they often contain highly domain-specific named entities (NEs) and bibliographic references. This specificity—along with the flaws of Optical Character Recognition (OCR) transcriptions and the lack of annotated data—makes this type of document extremely challenging for Natural Language Processing (NLP) tasks such as information extraction and semantic indexing.

To address this issue, we created a multilingual, named entity-annotated corpus of 19th-century commentaries on Sophocles' *Ajax*. The corpus was produced in the context of the Ajax Multi-Commentary (AjMC) project in order to support the automatic enrichment of digitised commentaries. An earlier version of this corpus (v. 0.3) was published and used in the context of the 2022 edition of the shared task HIPE – Identifying Historical People, Places and other Entities (Ehrmann, Romanello, Najem-Meyer, Doucet, & Clematide, 2022); the version of the corpus described in this paper (v. 0.4) improves data quality and includes an additional annotation layer of bibliographic references.

2 METHOD

To support the creation of this corpus we defined a set of guidelines for the annotation of domain-specific entities (Romanello & Najem-Meyer, 2022). These guidelines propose a unified approach to the annotation of both traditional and bibliographic entities in Classics publications. The corpus contains two layers of annotations (see Figure 1). The first layer captures information about bibliographic references to primary and secondary sources, following the taxonomy defined by Colavizza and Romanello (2017).

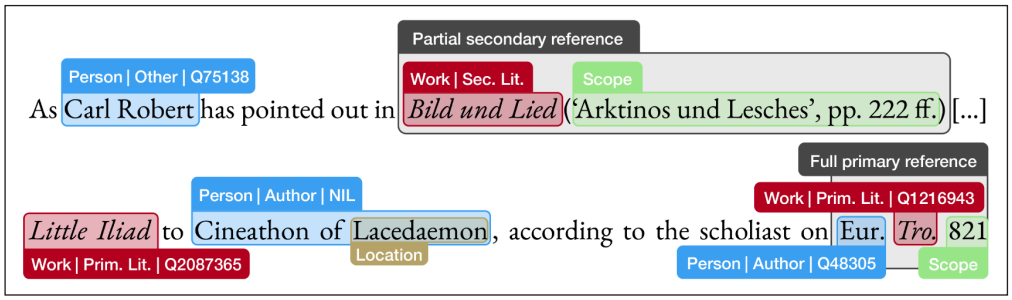

Figure 1 An example of annotated sentence.

The second layer contains both universal NEs (PERSON, LOCATION, ORGANIZATION, SCOPE) and some more domain-specific NEs (WORK, SCOPE, OBJECT). This coarse-grained tagset is complemented by a fine-grained tagset defining sub-types for certain entity types, enabling, for example, distinctions between a person being an author (PERSON.AUTHOR) and a person being a mythological character (PERSON.MYTH). Entities are linked to Wikidata (or to a `NIL` entity if no entry is available), except when they are nested or contained within a secondary bibliographic reference.[1] For entities containing OCR mistakes, a manually corrected transcription is provided, which allows for classifying mentions into OCR quality bands, or to compute the percentage of noisy mentions in the corpus (cf. Table 2).

ANNOTATION PROCEDURE AND SAMPLING STRATEGY

The pages to annotate were sampled from the introductions and glosses (i.e. the main commentary sections) of five 19th-century commentaries on the *Ajax* available in the public

1 Nested entities are a marginal phenomenon in our corpus (see Table 2) and mostly occur in person names; for example, a mention of "Ion of Chios" will be linked only to the Wikidata entry of the person and not of the place of origin (Chios). As to entities contained within secondary bibliographic references, they are not linked individually as it is rather the reference as a whole that should be disambiguated.

domain. Commentaries by Schneidewin (1853), Tournier (1866), Campbell (1881), Wecklein (1894) and Jebb (1896) were all chosen because of their importance and their availability. An earlier latin commentary by Lobeck (1835) was excluded from sampling as Latin is an otherwise under-represented language in the AjMC commentary corpus. OCR was performed with Tesseract[2] and annotation was done by two independent annotators using INCEpTION (Klie, Bugert, Boullosa, de Castilho, & Gurevych, 2018). Pages that were considered too short (fewer than 100 tokens) or problematic (e.g. presence of errors in the page reading order, missing text due to erroneous layout recognition, etc.) were discarded by the annotators. In order to ensure consistently annotated data, ambiguities arising during annotation were shared and discussed with two curators. Before starting the annotation campaign, annotators and curators annotated a small reference corpus (approx. 3,900 tokens), in order both to familiarise themselves with the annotation guidelines and to test their robustness and clarity. Finally, all produced annotations were reviewed by the two curators in order to ensure consistency across annotators and to correct annotation mistakes.

QUALITY CONTROL

We performed double annotation on a sub-set of approximately 2,000 tokens per language (22 commentary pages and 6,400 tokens in total) in order to calculate the inter-annotator agreement (IAA) rates by using Krippendorf's α (see Table 1).[3] Overall, both named entity recognition (NER) and entity linking (EL) show good agreement between annotators, with an average IAA rate of 0.81 and 0.88 respectively. Some difficulties of the annotation task which led to disagreements are, for example, the erroneous inclusion of end-of-sentence punctuation as part of the entity mention and, for EL, the presence of homonymous entities in Wikidata, leading to the wrong entity being selected (e.g. Sophocles the Classical playwright vs. the Hellenistic tragic poet named Sophocles).

SUB-CORPUS	NERC	EL
English	0.83	0.95
French	0.74	0.87
German	0.85	0.81
Avg.	0.81	0.88

Table 1 IAA rates computed on a double-annotated sample of the corpus.

LANG.	FOLD	DOCS	TOKENS	MENTIONS				
				ALL	FINE	NESTED	%NOISY	%NIL
de	Train	76	22,695	1,738	1,738	11	13.81	0.92
	Dev	14	4,701	403	408	2	11.41	0.49
	Test	16	4,845	382	382	0	10.99	1.83
Total		106	32,241	2,528	2,523	13	13.00	0.99
en	Train	60	30,932	1,823	1,823	4	10.97	1.65
	Dev	14	6,506	416	416	0	16.83	1.68
	Test	13	6,052	348	348	0	10.34	2.59
Total		87	43,490	2,587	2,587	4	11.83	1.78
fr	Train	72	24,669	1,621	1,621	9	30.72	0.99
	Dev	17	5,425	391	391	0	36.32	2.56
	Test	15	5,391	360	207	0	27.50	1.45
Total		104	35,487	2,219	2,372	9	31.16	1.31
Grand Total		**297**	**111,216**	**7,334**	**7,334**	**26**	**18.66**	**1.36**

Table 2 Basic statistics for the `AjMC NE corpus` (version 0.4).

2 Tesseract was chosen based on a benchmark of several OCR models (Romanello, Najem-Meyer, & Robertson, 2021).

3 To compute IAA rates we used the Python library INCEpTALYTICS v. 0.1.1 (Hamacher & Zesch, 2022).

3 DATASET DESCRIPTION

OBJECT NAME

`AjMC-NE-corpus`.

FORMAT NAMES AND VERSIONS

CoNLL-like HIPE TSV format[4]

CREATION DATES

November 2021 to March 2022.

DATASET CREATORS

Carla Amaya (UNIL, annotation), Kevin Duc (UNIL, annotation), Sven Najem-Meyer (EPFL, data curation), Matteo Romanello (UNIL, data curation & supervision).

LANGUAGE

Primarily French, German and English, and to a lesser extent Ancient Greek and Latin.

LICENSE

CC BY Attribution 4.0 International.

REPOSITORY NAME

GitHub and Zenodo.

PUBLICATION DATE

The current version of the corpus (v. 0.4) was published on 2023-09-01; version 0.3 was published on 2022-05-20 as part of the HIPE-2022 data (v. 2.1).

CORPUS STATISTICS

With about 300 annotated pages, 111,218 tokens and 7,482 entity mentions (see Table 2), this corpus is relatively small compared to other corpora (Ehrmann, Hamdi, Pontes, Romanello, & Doucet, 2023). Nested entities constitute a marginal phenomenon, more frequent in German and French than English. OCR noise affects on average 18.66% of entity mentions, but the French commentary has a rate of noisy entities almost three times higher than English and German, indicating a much lower OCR quality. As to the number of fine-grained mentions (see Table 3), we observe that certain entity types—namely OBJECT, LOC and DATE—are heavily under-represented. Wikidata provides excellent coverage for EL in this corpus as 98.55% of mentions have a corresponding entry in the knowledge base.

USAGE OF THE CORPUS IN THE HIPE-2022 SHARED TASK

The dataset was featured in two of the challenges into which the HIPE-2022 shared task (Ehrmann et al., 2022) was organised: the *Multilingual Classical Commentary Challenge* and the *Global Adaptation Challenge*.[5] While the former focused on commentaries, the latter aimed at testing the ability of participating systems to handle both commentaries and newspapers written in at least two languages.

NER results were reassuring. Despite considerable OCR noise, the best systems reached an overall F1-score of 85.4% for English, 84.2% for French and 93.4% for German on the coarse tagset.

4 See https://doi.org/10.5281/zenodo.3677171 for further details.

5 For more details on the shared task's organisation please refer to the participation guidelines, https://doi.org/10.5281/zenodo.6045662.

COARSE TAG SET	FINE TAG SET	NB. MENT.	LINKING	EXAMPLES
PERS	PERS.AUTHOR	1,212	yes	"Sophocles", "Euripid."
	PERS.EDITOR	153		"Triclinius", "Schndw."
	PERS.MYTH	933		"Tekmessa", "Ajax"
	PERS.OTHER	237		"Musgrave", "Perikles"
Total (PERS)		2,535		
WORK	WORK.PRIMLIT	1,566	yes	"O. T.", "Iliad"
	WORK.SECLIT	82		"Lexicon Sophocleum", "L. and S."
	WORK.FRAGM	12		"Frag. adesp.", "fragm."
	WORK.JOURNAL	1		"the Cambridge Journal of Philology"
	WORK.OTHER	3		
Total (WORK)		1,664		
OBJECT	OBJECT.MANUSCR	25	no	"Laurentianus A"
LOC	–	109	yes	"Athènes", "Salamisinsel"
DATE	–	26	no	"um 770 v. Chr.", "A.D. 1618"
SCOPE	–	2,975	no	"1340 f.", "1083"

Table 3 The tagset of annotated entities in the `AjMC NE corpus` (version 0.4); for each entity type, the total number of mentions in the corpus and some selected examples are provided.

Much work remains to be done, however, in the EL task; here, the best F1-score for the linking of pre-extracted mentions is 38.1% for English, 47% for French and 50.3% for German.

Two other factors compound with OCR noise make EL on this corpus an extremely challenging task. First, the style of commentary writing favours conciseness and uses abbreviations abundantly (approx. 47% of all entity mentions are abbreviated). Secondly, abbreviations rely heavily on context: in a commentary on a tragedy by Sophocles, the commentator will refer to the tragedy *Philoctetes* as simply *Ph.* (instead of *Philoct.*), thus making such abbreviations hard to resolve for EL systems.

4 REUSE POTENTIAL

Despite being the first named entity-annotated corpus—to the best of our knowledge—to contain classical commentaries, it has certain limitations. Firstly, given the research context it originates from, the selection of commentaries is limited to Attic tragedy; to make this corpus more generic, commentaries to works of Ancient Greek prose, as well as to works of Latin prose and poetry should be considered too. Secondly, some entity types are heavily under-represented (see Table 3), a limitation which could be mitigated by applying data augmentation or meta-learning approaches to training.

The main anticipated use of this corpus is to train and evaluate domain-specific models for NER and EL on historical documents.[6] In fact, the sampling strategy adopted—both in the selection of commentaries and of page sections to annotate (introduction and glosses)—has led to a dataset which is not complete nor representative enough to study higher level characteristics of the commentary genre (e.g. by means of textual analysis).

Instead, due to the very nature of the data, this corpus is particularly suitable for testing the adaptability of NER systems to noisy, multilingual and multiscript texts. The density of abbreviated entity mentions also makes this corpus an excellent testbed for evaluating the ability of EL systems to deal with domain-specific—and oftentimes cryptic—abbreviations. Furthermore, the bibliographic reference annotation layer contains a particularly rich set of primary source references. As recent initiatives demonstrate,[7] bibliographic reference extraction

6 Schweter, März, Schmid, and Çano (2022) released three language-specific hmBERT models fine tuned on this dataset for the task of (coarse-grained) NER. The models are available via the HuggingFace model hub: https://huggingface.co/hmbert.

7 See for example the workshop *New approaches for extracting heterogeneous reference data* (Frankfurt, 2023), https://mpilhlt.github.io/reference-extraction/.

in the humanities is far from a solved problem, and the scarcity of available data for this task is a real issue (Colavizza & Romanello, 2019). Hopefully, this dataset will help alleviating this issue in the domain of Classics.

Finally, based on the encouraging results obtained on noisy and highly abbreviated texts such as commentaries, it is reasonable to expect that NER models trained on this corpus will perform fairly well on commentaries, books and journal articles which are either born digital or have better OCR quality.[8]

FUNDING INFORMATION

This research was funded by the Swiss National Science Foundation under the Ambizione scheme (Grant number: PZ00P1_186033).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Matteo Romanello: Conceptualization, Data Curation, Funding Acquisition, Methodology, Supervision, Writing – Original Draft, Writing – Review & Editing. Sven Najem-Meyer: Methodology, Data Curation, Software, Writing – Review & Editing.

AUTHOR AFFILIATIONS

Matteo Romanello orcid.org/0000-0002-7406-6286
Institute of Archeology and Classical Studies, University of Lausanne, Lausanne, Switzerland
Sven Najem-Meyer orcid.org/0000-0002-3661-4579
Digital Humanities Laboratory, Swiss Federal Institute of Technology Lausanne, Lausanne, Switzerland

REFERENCES

Colavizza, G., & **Romanello, M.** (2017, November). Annotated References in the Historiography on Venice: 19th–21st centuries. *Journal of Open Humanities Data*, *3*, 2. DOI: https://doi.org/10.5334/johd.9

Colavizza, G., & **Romanello, M.** (2019). Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead. *Journal of European Periodical Studies*, 4. DOI: https://doi.org/10.21825/jeps.v4i1.10120

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & **Doucet, A.** (2023, June). Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*. DOI: https://doi.org/10.1145/3604931

Ehrmann, M., Romanello, M., Najem-Meyer, S., Doucet, A., & **Clematide, S.** (2022). Extended overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In G. Faggioli, N. Ferro, A. Hanbury & M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum* (Vol. 3180). CEUR-WS. DOI: https://doi.org/10.1007/978-3-031-13643-6_26

Hamacher, M., & **Zesch, T.** (2022, September). *INCEpTALYTICS – An easy-to-use API for analyzing INCEpTION annotation projects*. Retrieved from https://github.com/catalpa-cl/inceptalytics. DOI: https://doi.org/10.5281/zenodo.5654690

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). Retrieved from https://aclanthology.org/C18-2002

Romanello, M., & **Najem-Meyer, S.** (2022, March). *Guidelines for the annotation of named entities in the domain of classics* [Documentation]. DOI: https://doi.org/10.5281/zenodo.6368101

Romanello, M., Najem-Meyer, S., & **Robertson, B.** (2021, September). Optical Character Recognition of 19th Century Classical Commentaries: The Current State of Affairs. In *The 6th International Workshop*

8 Examples of openly accessible resources to which these models could be applied include the commentaries from the Dickinson College commentaries (https://dcc.dickinson.edu/), the online publications of Harvard's Center for Hellenic Studies (https://chs.harvard.edu/browse-online-publications/), and the (French) journal articles from Persée (https://persee.fr/).

on Historical Document Imaging and Processing (HIP '21). Lausanne: Association for Computing Machinery. DOI: https://doi.org/10.1145/3476887.3476911

Schweter, S., März, L., Schmid, K., & **Çano, E.** (2022, September). hmBERT: Historical Multilingual Language Models for Named Entity Recognition. In G. Faggioli, N. Ferro, A. Hanbury & M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum* (Vol. 3180, pp. 1109–1129). Bologna, Italy: CEUR. Retrieved from https://ceur-ws.org/Vol-3180/paper-87.pdf

TO CITE THIS ARTICLE:
Romanello, M., & Najem-Meyer, S. (2024). A Named Entity-Annotated Corpus of 19th Century Classical Commentaries. *Journal of Open Humanities Data*, 10: 1, pp. 1–7. DOI: https://doi.org/10.5334/johd.150

Submitted: 01 September 2023
Accepted: 26 October 2023
Published: 02 January 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.