Journal of Open Humanities Data

Named-Entity Dataset for Medieval Latin, Middle High German and Old Norse

SPECIAL COLLECTION: COMPUTATIONAL HUMANITIES RESEARCH DATA

DATA PAPER

CLÉMENT BESNIER

WILLIAM MATTINGLY

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

ubiquity press

ABSTRACT

We present a dataset of named entities in three languages: Medieval Latin, Middle High German and Old Norse. The dataset, containing proper nouns of persons and places, was originally created to extract characters from three related medieval texts. Since the annotation is on low-resource pre-modern languages, they may be important to build named-entity recognition tools for languages with little data and high linguistic variation.

CORRESPONDING AUTHOR:
Clément Besnier
Classical Language Toolkit
clem@clementbesnier.fr

KEYWORDS:
named-entity recognition; natural language processing; Latin; Medieval Latin; Middle High German; Old Norse

TO CITE THIS ARTICLE:
Besnier, C., & Mattingly, W. (2021). Named-Entity Dataset for Medieval Latin, Middle High German and Old Norse. *Journal of Open Humanities Data*, 7: 23, pp. 1–5. DOI: *https://doi.org/10.5334/johd.36*

DOI: 10.5334/johd.36

1 OVERVIEW

REPOSITORY LOCATION

https://zenodo.org/record/4571507.

CONTEXT

The annotations were originally from a character-network analysis paper (Besnier, 2020). The aim was to compare the evolution of character sets in stories with similar backgrounds over time and space.

2 METHOD

STEPS

We first retrieved normalized (*Decem Libri Historium*/DLH, *Völsunga saga*/VÖL) or transcribed (*Nibelungenlied*/NIB) texts. We then tokenized the texts with the CLTK package (Johnson et al., 2021) and picked the tokens that start with a capital letter, since this often marks proper nouns. When such forms were also found in lowercase, we decided they were not proper nouns. For the remaining tokens and sentence-initial tokens, expert knowledge was needed to classify the results. Finally, we used translations and indexes (Anonymous, 2011) (Gregory Of Tours, 2019) to classify tokens as proper nouns or not.

SAMPLING STRATEGY

We fully annotated the three texts.

QUALITY CONTROL

We checked that our method was correct with indexes (Gregory Of Tours, 2019) (Anonymous, 2011), when available. Most mistakes occurred when word forms belonging to the same lemma were assigned different lemmata: spelling variation, for example, led us to consider *Abiti* and *Avitus* as coming from different lemmata, whereas they were both considered by the index (Krusch & Levison 1951) as belonging to *Abitus*. The peculiarity of the Latin imparisyllabic third-declension of nouns[1] made us doubt that *Agila*, on the on hand and *Agilane*, *Agilanem*, on the other hand, were from a same lemma. For VÖL and NIB, such mistakes were easily avoidable because of the rather low number of proper nouns in the texts.

3 DATASET DESCRIPTION

OBJECT NAME

The annotations are lists of lemmata of proper nouns in three texts: the *Decem Libri Historium* by Gregory of Tours written in Medieval Latin, the *Völsunga saga* written in Old Norse (ON), and the *Nibelungenlied* written in Middle High German (MHG). For each lemma, we provide a list of associated tokens.

FORMAT NAMES AND VERSIONS

CSV file. The column names are "text" (the name of the text), "category" (category of the lemma/tokens: PERSON, PLACE and GROUP), "language" (*latin*, *middle-high-german*, *old-norse*), "lemma" (the nominative singular form) and "tokens" (forms present in the text belonging to the lemma; tokens are separated by semi-colons). Current version 1.0.0.

1 In Latin third-declension nouns can take forms similar to other declensions in the nominative. In other forms, however, the number of syllables can change, immediately revealing that they in fact belong to the third declension. Declining a third-declension noun from the nominative can be done easily using known rules, but reversing the process (finding the lemma from other forms, such as the ablative or accusative) is a bit more challenging and can result in improper lemmata being used, especially if that word has a homophone in another declension.

–	DLH	NIB	VÖL
Number of tokens	123920	83961	26779
Number of PERSONS lemmata	812	67	111
Number of PERSONS tokens	1787	168	196
Number of PLACES lemmata	349	65	32
Number of PLACES tokens	990	86	37
Number of GROUPS lemmata	–	3	5
Number of GROUPS tokens	–	4	5

Statistics on the dataset and its texts.

CREATION DATES

Start: 2020-06-05. End: 2020-12-29.

LANGUAGES

Data: Latin, Middle High German and Old Norse; Metadata: English.

LICENSE

REPOSITORY NAME

https://zenodo.org/record/4571507

PUBLICATION DATE

2021-03-01

4 REUSE POTENTIAL

The dataset combines named-entity annotations of texts written in several pre-modern languages. As there are not many digital language resources for these languages, they are a good starting point to train models for named-entitiy recognition (NER). Despite the fact that Classical Latin has received much attention in Natural Language Processing (NLP), compared to other ancient and medieval languages, Medieval Latin requires different training sets because of the nominal and syntactic changes in the language during the Middle Ages. Medieval Latin has received less attention then its Classical counterpart. These datasets provide the first steps to formally incorporate Medieval Latin into Latin NLP models.

The datasets for ON and MHG allowed us to train NER neural network models for both ON and MHG and provide the necessary first steps toward wider applicability and reuse across ancient and medieval languages. When studying the datasets, we developed a workflow to automate the training of NER models for highly inflected ancient and medieval languages (Honnibal, Montani, Van Landeghem, & Boyd, 2020). Our Classical Latin model can identify PERSON (e.g. *Romanus* – a Roman man), PLACE (e.g. *Roma* – Rome), and GROUP (e.g. Romani – Romans).

Latin presents certain challenges for model-based NER with two meriting particular attention. Firstly, for PERSON entities, names are complex and often multi-word tokens. They possess a *praenomen*, *nomen*, and *cognomen*.[2] Secondly, Latin is highly inflected, with all entities possessing three to six forms.[3] A model must, therefore, be able to identify names that appear in various forms.

2 Those who had received honors would often have a fourth name that detailed a triumph, i.e., *Scipio Africanus* who won the Second Punic War at the Battle of Zama.

3 Each name can be fully declined as nominative, accusative, genitive, dative, ablative, and vocative. In other words, each entity can have up to six forms (but due to overlap between these forms, i.e., dative and ablative, it is often three to four).

In the initial implementation of the ON and MHG datasets, we only identified PERSON and PLACE. We noticed that a potential disadvantage to this dataset was the absence of entities in the GROUP category.[4] For Latin, our machine learning models struggled with GROUP entities, such as "Romans". The models switched between PERSON and PLACE for such entities. By incorporating GROUP into the model, we improved the identification of PERSON. Later, we annotated the GROUP category in ON and MHG.

Unlike our manual annotations for ON and MHG, we automated the creation of a training set for Latin. To do this, we first gathered as many potential praenomen, nomen, and cognomen instances as we could for Latin PERSON entities from Wikipedia.[5] Second, we used Orbis Latinus (hosted by Columbia University) (Graesse, 1909) to collate a list of places Third, we manually compiled a list of GROUP instances from Caesar's Gallic Wars.[6] Next, we generated all potential variations of these words in their declined form via a Python script (Mattingly, 2021) that identified a noun's declension (class) and declined it accordingly. With all potential forms for each word, we created an EntityRuler in spaCy (Honnibal et al. 2020). We ran the EntityRuler over a single text: Caesar's Gallic Wars. Finally, we used this auto-generated training set to train a spaCy model.

Overall, the datasets for ON and MHG allowed us to identify patterns for automating the creation of NER training sets in highly inflected ancient and medieval languages. This demonstrates the wider applicability of these datasets. Beyond the plan we have for the data we present here, other researchers may use it to train their own models.

With NER models, complex tasks like producing prosopographies may be partly automated. Nevertheless, without correct anaphor resolution, i.e., the assignment of a personal pronoun or noun phrase to their referent, such task can only be partially performed.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Clément Besnier *orcid.org/0000-0001-5868-8723*
Classical Language Toolkit

William Mattingly *orcid.org/0000-0003-4334-5714*
Smithsonian Institution, Data Science Lab, US; United States Holocaust Memorial Museum, US; Classical Language Toolkit

REFERENCES

Anonymous. (2011). Nibelungenlied (S. Grosse, Trans.). Reclam.

Besnier, C. (2020). History to Myths: Social Network Analysis for Comparison of Stories over Time. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 1–9). Online: International Committee on Computational Linguistics. Retrieved from *https://www.aclweb.org/anthology/2020.latechclfl-1.1*

Graesse, J. G. T. (1909). *Orbis latinus*. The Electronic Text Service; Columbia University. Retrieved August 8, 2021, from *http://www.columbia.edu/acis/ets/Graesse/contents.html*

Gregory Of Tours. (2019). *Histoire des Francs* (R. Latouche, Trans.). Paris: Les Belles Lettres.

Honnibal, M., Montani, I., Van Landeghem, S., & **Boyd, A.** (2020). spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo*. DOI: *https://doi.org/10.5281/zenodo.1212303*

Johnson, K. P., Burns, P. J., Stewart, J., Cook, T., Besnier, C., & **Mattingly, W. J. B.** (2021, August). The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing: System demonstrations* (pp. 20–29). Online: Association for Computational Linguistics. Retrieved from *https://aclanthology.org/2021.acl-demo.3* DOI: *https://doi.org/10.18653/v1/2021.acl-demo.3*

4 We consider GROUP broadly defined as a collection of individuals with a proper noun.

5 Praenomina: *https://en.wikipedia.org/wiki/Praenomen* Nomina: *https://en.wikipedia.org/wiki/List_of_Roman_nomina* Cognomina: *https://en.wikipedia.org/wiki/List_of_Roman_cognomina*.

6 Caesar, Gallic Wars, The Latin Library: *http://www.thelatinlibrary.com/caesar/*.

Krusch, B., & **Levison, W.** (1951). Index. In *Gregorii episcopi Turonensis. Libri Historiarum X* (Vols. Monumenta Germaniæ Historica, Scriptores rerum Merovingicarum, pp. 540–566). Hannover. Retrieved from *https://www.dmgh.de/mghssrermerov11/index.htm#page/(540)/mode/1up*

Mattingly, W. (2021). *Latin decliner*. Retrieved January 18, 2021, from *https://github.com/wjbmattingly/latin_ner_lesson/blob/f9a9cb6db890ad66dc3d31788f89dccaf8ea7485/temp/declininglatin.py*

DOI: 10.5334/johd.36

TO CITE THIS ARTICLE:
Besnier, C., & Mattingly, W. (2021). Named-Entity Dataset for Medieval Latin, Middle High German and Old Norse. *Journal of Open Humanities Data*, 7: 23, pp. 1–5. DOI: *https://doi.org/10.5334/johd.36*

Published: 06 October 2021

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