EvaLatin 2026
-
Named Entity Recognition Shared Task Guidelines

Version 1.2

December 22, 2025

Eleonora Litta[1], Matteo Romanello[2], Valeria Boano[3],
1. CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milan, Italy
`eleonoramaria.litta[at]unicatt.it`
2. Swiss Art Research Infrastructure (SARI), University of Zurich, Switzerland
`matteo.romanello[at]uzh.ch`
3. KU Leuven, Belgium
`valeria.boano[at]kuleuven.be`

Contents

Chapter 1

Introduction

EvaLatin 2026 is the fourth edition of the evaluation campaign of Natural Language Processing (NLP) tools for the Latin language. The campaign is designed with the aim of answering two main questions:

- How can we promote the development of resources and language technologies for the Latin language?
- How can we foster collaboration among scholars working on Latin and attract researchers from different disciplines?

EvaLatin 2020 [2] was organized around 2 tasks (i. e. Lemmatization and PoS Tagging); EvaLatin 2022 [3] was organized around 3 tasks (i. e. Lemmatization, PoS Tagging, and Features Identification); EvaLatin 2024 [1] was organized around 2 tasks (i. e. Dependency Parsing and Emotion Polarity Detection). Evalatin 2026 will focus on Dependency Parsing and Named Entity Recognition. Shared data and a scorer are provided to the participants. Participants can choose to participate in either one or both tasks. The organizers rely on the honesty of all participants who might have some prior knowledge of part of the data used for evaluation not to unfairly use such knowledge.

EvaLatin 2026 is organized within the "Fourth Workshop of Language Technologies for Historical and Ancient Languages" (LT4HALA 2024), colocated at LREC-COLING 2026.[1] The workshop will be held in Palma de Mallorca, Spain, on 11th May 2026. EvaLatin 2026 is organized by the CIRCSE research centre at the Università Cattolica del Sacro Cuore in Milan.

For any update, please check the LT4HALA 2026 website: `https://circse.github.io/LT4HALA/2026/`.

1.1 Named Entity Recognition and Classification

Named Entity Recognition and Classification (NERC) is a fundamental task in Natural Language Processing that involves the automatic identification and classification of proper

[1] `https://lrec2026.info/`

names in running text into predefined categories such as Persons, Locations, Organizations, and Groups. While commercial NERC systems have achieved near-human performance on modern languages like English, largely driven by massive datasets drawn from news corpora and the web, the application of these technologies to historical languages remains an active and critical frontier. To address specific domain requirements, the Ancient Named Entities Special Interest Group has developed a Named Entity tagset tailored for Ancient Greek and Latin textual materials and their translations. This schema offers labels that are specific to ancient documentation while remaining mappable onto standard concepts employed in modern NLP datasets. In these regards, the Ancient Named Entities Special Interest Group has been working on designing a tagset and annotation strategy that could support the creation of consistent annotated datasets of Named Entities in ancient texts, minimizing annotator disagreement and ambiguity. The primary application is (not limited to) the development of Machine Learning methods. This work has been done also to ensure a basic level of interoperability and exchange across projects that involve Named Entities in Ancient Greek and Latin, by providing labels that are both specific to ancient documents and mappable onto commonly used concepts in NLP for modern datasets. From a machine learning perspective, significant challenges usually arise from the target languages high degree of inflection, flexible word order, and complexities regarding tokenization and lemmatization. These difficulties are further compounded by the potential absence of capitalization, inconsistent spelling conventions, and the heterogeneity of historical domains.

Chapter 2

Data

This NER task leverages the frameworks of the HIPE (Identifying Historical People, Places and other Entities) 2020 and 2022 campaigns (https://hipe-eval.github.io/HIPE-2022/about) and of the Ancient Named Entities Special Interest Group. The tagset has been curated specifically for Ancient Greek, Latin, and English translations of Classical works.

Release Materials:

- Participants Guidelines: The current document and the SIG-specific annotation protocols.
- Annotation Guidelines: a stable version of the tagset used in the Ancient Named Entities Special Interest Group annotation activities.
- Sample Data: A sample set provided to assist participants in familiarizing themselves with the annotation schema, and/or for usage in data augmentation strategies.
- Test Data: the documents participants will have to use for the resolution of the task (check the EvaLatin 2026 web page for details on the Test Data release date).

The tasks aims to improve a state of the art that is not optimal. With regard to ancient languages, NER systems currently show different degrees of harmonization, and Latin is not an exception in this respect. The diversity of the data currently available for the task is an issue we are aware of, and that needs to be addressed. This evaluation campaign aims at addressing this issue, and among the desired outcomes there are strategies to deal with it successfully.

2.1 Format

This section gives details on the format of test data for the 2026 NERC shared task. The format for EvaLatin test data is based on that developed for the HIPE 2020 and 2022 NERC shared tasks (https://hipe-eval.github.io/HIPE-2022/about).

2.1.1 HIPE format and tagging scheme

HIPE format is a simple tab-separated column textual format, similarly to that of the CoNLL format, using an IOB tagging scheme (inside-outside-beginning format).

2.1.2 File structure

Files encode annotations needed for the NER task and may contain the following lines:

- empty lines, which mark the boundaries between documents;
- comment lines, which give further information and start with the character '#' followed by a space;
- annotated lines, which contain a token followed by its tab-separated annotations.

A file contains all the documents of one dataset. Documents are separated with empty lines and are preceded with metadata comment lines (starting with the character #).

2.1.3 File contents

Each line consists of 10 columns, but only 2 columns are taken into account at evaluation time, i.e. NE-COARSE-LIT and NE-FINE-LIT as detailed in Table 1. In the HIPE format, columns have the following content (the columns that participants should consider are in bold):

1. **TOKEN**: the annotated token.
2. **NE-COARSE-LIT**: the coarse type (IOB-type) of the entity mention token, according to the literal sense.
3. NE-COARSE-METO: the coarse type (IOB-type) of the entity mention token, according to the metonymic sense.
4. **NE-FINE-LIT**: the fine-grained type (IOB-type.subtype) of the entity mention token, according to the literal sense.
5. NE-FINE-METO: the fine-grained type (IOB-type.subtype of the entity mention token, according to the metonymic sense.
6. NE-FINE-COMP: the component type of the entity mention token.
7. NE-NESTED: the coarse type of the nested entity (if any).
8. NEL-LIT: the Wikidata Q id of the literal sense, or 'NIL' if an entity cannot be linked. Rows without link annotations have value '_'.
9. NEL-METO: the Wikidata Q id of the metonymic sense, or 'NIL'.

Task	Relevant Annotation Column
NERC - Coarse	NE-COARSE-LIT
NERC - Fine	NE-FINE-LIT

Table 1: Annotation Columns Relevant to EvaLatin2026

10. **MISC**: a flag which can take the following values:

    - NoSpaceAfter: to indicate the absence of white space after the token.
    - EndOfLine: to indicate the end of a layout line.
    - EndOfSentence: to indicate the end of a sentence.
    - Partial-START:STOP: to indicate the zero-based character on-/offsets of mentions that do not cover the full token (esp. for German compounds). START and STOP follow Python's slicing semantics: “abcd" [1:3] means “bc".
    - Non-specified values are marked by the underscore character “_”.

This annotation scheme originates from the CLEF-HIPE-2020 shared task and contains detailed named entity annotation columns (reflected in the IOB file columns presented above). However, EvaLatin 2026 evaluation only focuses on a selection of annotation columns, as shown in Table 1, and uses the tagset detailed below (see Section 3.1.1).

The annotation types NE-COARSE-METO, NE-FINE-METO, NE-FINE-COMP, NE-NESTED are not considered in the EvaLatin 2026 shared tasks and evaluation scenarios.

Annotated datasets have sentence boundary information in the MISC column (EndOfSentence flags), including information to rebuild the original text (NoSpaceAfter flags).

2.2 Sample Data

The sample data provided contains lines from poetic (Vergil) and prose (Sallust) textual material, and comes in the format described in the previous section. Basic statistics about sample data can be found in Table 2.

Entity type	Number of mentions
collective	24
misc	4
person	32
place	26
Total	**86**

Table 2: Number of mentions per entity type.

Chapter 3

Tasks

The EvaLatin NERC Shared Task focuses on two task types, namely:

- Task 1 NERC Coarse-grained: this task includes the recognition and classification of entity mentions according to coarse-grained types, where only first level categories should be recognised.
- Task 2 - NERC Fine-grained: this task includes the recognition and classification of entity mentions according to fine-grained types, where also second level categories should be recognised.

3.1 NERC SYSTEM ANNOTATION GUIDELINES:

The types of annotation that systems are expected to produce for each task are presented in Table 2.

Table 2, above, lists the entity types to consider for each subtask

For more information about system annotation rules, please refer to the Ancient Named Entities Special Interest Group annotation guidelines.

3.1.1 Tagset

This tagset provides a set of semantic labels to classify Ancient Named Entities. The tags provided include two hierarchical levels. First-level tags represent the type of object being designated by the named entity. The usage of first-level tags is intended to be strict, to ensure interoperability. Second-level tags are designed with more consideration

Task	NE mentions with coarse types	NE mentions with fine types
NERC-Coarse	yes	no
NERC-Fine	no	yes

Table 3: Expected annotations for the NERC Shared tasks

for the semantic function of the name: they may point to the way in which a referent is designated (for example, an individual identified with an epithet rather than their own name) or the specifically identifiable subgroup to which an entity belongs (e.g. a political or racial collective). They are designed to be used, expanded, or selected for project-specific goals and contexts.

The tagset is described in the Ancient Named Entities Annotation Guidelines, to be downloaded here. Please refer to the guidelines for detailed explanations of each tag and examples that are relevant to ancient texts. Pay particular attention to issues of annotation boundaries and non-consecutive entities.

First Level Tags for NERC-Coarse

- **Person**
  Any identifiable single individual, including deities and anthropomorphic mythological figures.
  Available second-level tags: **.author**, **.ancestry**, **.epithet**, **.ethnic**, **.derivative**.

- **Place**
  A politically, culturally, or geographically defined location, including fictional spaces and structures like temples, buildings, specific urban areas (e.g., gymnasia), and houses.
  Available second-level tags: **.astronomy**, **.epithet** ('(the) Rugged' (=Ithaka)), **.derivative**.

- **Collective**
  A named group of people or other creatures with shared identifiable characteristics on social, intellectual, political, national, family, mythical, or ethnic basis.
  Available second-level tags: **.ancestry** ('sons of Priamus'), **.animal** ('cattle of Helios'), **.astronomy** ('Pleiades'), **.ethnic** ('Romana gens'), **.organization** ('Senatus'), **.epithet** ('Eumenides'), **.derivative** ('Pygmeian', 'Academic').

- **Creature**
  Mythical or real precisely identifiable non-human, non-anthropomorphic creatures.
  Available second-level tags: **.animal**, **.astronomy**.

- **Event**
  Significant named events identified by a string with a precise boundary.

- **Language**
  Languages and dialects clearly identified as such.

- **Object**
  Artifacts or groups of artifacts clearly identified with a name, such as ships, weapons, statues, columns, dedications, etc.

- **Miscellaneous**
  Entities that do not (yet) have a specific first-level tag among those provided.

- **Time**
  Any absolute date or time expression.

- **Work**
  Titles of literary or non-literary works, in any form.

Second Level Tags for NERC-Fine

Second-level tags provide further specifications for particular types of Named Entities.

- **.ancestry** (collective.ancestry, person.ancestry)
  A designation or expression that refers unambiguously to one individual or group of individuals by using a family name, patronymic, matronymic, or other indication of lineage or familial relationships.

- **.animal** (collective.animal, creature.animal)
  A type of creature or collective of creatures clearly identifiable with an animal or animal species.

- **.astronomy** (creature.astronomy, collective.astronomy, place.astronomy)
  Named stars, groups of stars, constellations, and planets.

- **.author** (person.author)
  A person clearly mentioned in relation to works they have authored. This tag may be modified or even omitted for project-specific goals.

- **.derivative** (collective.derivative, person.derivative, place.derivative)
  An adjective derived from a toponym, personal name, or group name, used to identify things that are not individuals or collectives (for individuals or collectives, see .ethnic). Only the derivative is annotated, as the common noun in the expression does not act as a rigid designator. The first-level tag depends on the name from which the adjective derives (e.g. "Iberian" will be a place, "Platonic" a person, etc.).

- **.ethnic** (collective.ethnic, person.ethnic)
  An ethnonym, demonym, or other word used to identify persons or collectives by means of their membership to a geographically or ethnically defined group. This tag is exclusively used with persons or collectives, as ethnics are mainly used in the ancient world to identify individuals via ethnic memberships (see also our rationale below). For all other uses of adjectives derived from places, use the .derivative subtag.

- **.epithet** (collective.epithet, person.epithet, place.epithet)
  A capitalized epithet used to refer unambiguously to one individual, location, or collective, including nicknames, titles, and other appellatives.

- **.organization** (collective.organization)
  Collectives identified by precise organizational structures, such as priesthoods, legions, religious, intellectual, or political groups and institutions, and so on.

Chapter 4

Evaluation

4.1 Metrics

NERC is evaluated in terms of macro and micro Precision, Recall, F1-measure. Two evaluation settings are considered: strict (exact boundary matching) and relaxed (fuzzy boundary matching). Each column is evaluated independently, according to the following metrics:

- Micro average P, R, F1 at entity level (not at token level), i.e. consideration of all true positives, false positives, true negatives and false negatives over all documents.
  - strict (exact boundary matching) and fuzzy (at least 1 token overlap).
  - separately per type and cumulative for all types.
- Document-level macro average P, R, F1 at entity level (not on token level). i.e. average of separate micro evaluation on each individual document.
  - strict and fuzzy
  - separately per type and cumulative for all types

Our definition of "macro" differs from the usual one, and macro measures are computed as aggregates on document-level instead of entity-type level. Specifically, macro measures average the corresponding micro scores across all the documents, accounting for (historical) variance in document length and not for class imbalances. Note that in the strict scenario, predicting wrong boundaries leads to severe punishment of one false negative (entity present in the gold standard but not predicted by the system) and one false positive (predicted entity by the system but not present in the gold standard). Although this may be severe, we keep this metric in line with CoNLL and refer to the fuzzy scenario if the boundaries of an entity are considered as less important.

4.1.1 Scorer

EvaLatin NERC evaluation will use the CLEF-HIPE-2020-scorer.

System performances will be computed, reported and published in terms of micro and macro P, R, and F1 for each Task. For each Task, systems will be ranked according to their F1 scores.

4.2 Evaluation period

Please check important dates on the EvaLatin2026 website. At the end of the evaluation period, participants will send their system responses via email to the task organizers, which will be evaluated using the scorer. Gold standard data will be distributed after the publication of the evaluation results.

Chapter 5

How to Participate

Participants are required to submit their runs and to provide a technical report describing their systems. At submission time, participants must declare the task type to which their submitted data belong.

5.1 Submitting Runs

Each participant can submit a maximum of 2 runs for each task. Participants are allowed to use any approach (e.g. from traditional machine learning algorithms to Large Language Models) and any resource (annotated and non-annotated data, embeddings): all approaches and resources are expected to be described in the system's report. Once the system has produced the results for the task over the test set, participants have to follow these instructions to complete their submissions:

- Files must be in UTF-8, tsv encoded (.tsv extension), with annotations in the same format as in the sample data.
- name the runs with the following filename format: `task_docName_teamName_runID.tsv`.
  For example: `ner-coarse_XXX_unicatt_1.tsv` would be the first run of a team called *unicatt* for the NER Coarse Grained task on the `XXX.tsv` document.
  `ner-fine_XXX_unicatt_2.tsv` would be the second run of a team called *unicatt* for the NER Fine Grained task for the `XXX.tsv` document.
- send the file to the following email address: `eleonoramaria.litta[AT]unicatt.it`, using the subject "EvaLatin Submission: task - teamName", where the "task" is either *NERC-Coarse* or *NERC-Fine*.

5.2 Writing the Technical Report

Technical reports will be included in the proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) as short papers and will be published along the LREC-COLING 2026 proceedings. Reports must be submitted through the START platform (https://softconf.com/lrec2026/LT4HALA2026/). All the reports must meet the following requirements:

- they must be written in English;
- they must be formatted according to the LREC-COLING 2026 conference style[1];
- the maximum length is 4 pages (excluding references);
- they should contain (at least) the following sections: description of the system, results, discussion, references.

Reports will receive a light review: we will check for the correctness of the format, the exactness of results and ranking, and overall exposition. If needed, we will contact the authors asking for corrections.

[1] https://lrec2026.info/authors-kit/

Appendix A

Selection of Resources for Latin

- Lemma embeddings: https://embeddings.lila-erc.eu/
- Latin texts and embeddings: http://www.cs.cmu.edu/~dbamman/latin.html
- Latin BERT: https://github.com/dbamman/latin-bert
- Word embeddings: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1989
- CLTK: http://cltk.org/
- UD Latin PROIEL: https://github.com/UniversalDependencies/UD_Latin-PROIEL
- UDLatin ITTB: https://github.com/UniversalDependencies/UD_Latin-ITTB
- UD Latin Perseus: https://github.com/UniversalDependencies/UD_Latin-Perseus
- UD Latin LLCT: https://github.com/UniversalDependencies/UD_Latin-LLCT
- UD UDante: https://github.com/UniversalDependencies/UD_Latin-UDante
- Latin texts: https://github.com/PerseusDL
- Collatinus: https://outils.biblissima.fr/en/collatinus/index.php
- LEMLAT v.3: https://github.com/CIRCSE/LEMLAT3
- Treetagger: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
- Glossaria: https://glossaria.eu/outils/lemmatisation/#page-content
- Word Formation Latin (WFL) lexicon: http://wfl.marginalia.it/
- Lemmatized corpus with morphological analysis "Corpus Latin antiquité et antiquité tardive lemmatisé": https://doi.org/10.5281/zenodo.4337145

- Latin Lasla Model (Thibault Clérice) for lemmatization and morphological analysis: `https://doi.org/10.5281/zenodo.3773327`
- LatinCy: `https://spacy.io/universe/project/latincy`
- LatinAffectus v4: `https://github.com/CIRCSE/LT4HALA/blob/master/2024/LatinAffectusv4.tsv`
- Deucalion: `https://dh.chartes.psl.eu/deucalion/latin`
- UDPipe Models: `https://ufal.mff.cuni.cz/udpipe/1/models`

Bibliography

[1] Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. Overview of the EvaLatin 2024 evaluation campaign. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia, May 2024. ELRA and ICCL.

[2] Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of* LT4HALA *2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France, May 2020. European Language Resources Association (ELRA). Retrievable at `https://www.aclweb.org/anthology/2020.lt4hala-1.16`.

[3] Rachele Sprugnoli, Marco Passarotti, Cecchini Flavio Massimiliano, Margherita Fantoli, and Giovanni Moretti. Overview of the evalatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), Language Resources and Evaluation Conference (LREC 2022)*, pages 183–188, 2022.