Classificatio Sine Iactu – That Is, Zero-Shot NERC in Latin

Luisa Ripoll-Alberola[1,3], **Fernando Nicolás-Flores**[2], **Francisco Javier Muñoz Acebes**[3]
[1]Computational Humanities, Leipzig University, [2]Universidad de Alicante,
[3] Filología Digital, Universidad de Valladolid
Augustusplatz 10, 04109 Leipzig, Germany
ripoll_alberola@informatik.uni-leipzig.de, fernando.nicolas@ua.es, fjavier.munoz@uva.es

Abstract

This paper presents a zero-shot approach to Named-Entity Recognition and Classification (NERC) in Latin, applied to the EvaLatin shared task. Given the novelty and granularity of the annotation guidelines, which preclude the use of existing annotated resources, we employ the zero-shot model GLiNER2, a general information extraction system capable of CPU-efficient inference, within a cross-lingual pipeline. Latin texts are first translated into English via the Google Translate API, processed by the model, and the resulting annotations are aligned back to the original Latin using word-alignment techniques. Rule-based post-processing addresses labelling inconsistencies and low-confidence predictions. We evaluate two model variants, a large monolingual and a multilingual model, under both strict and fuzzy evaluation. The large model delivers the best results for the coarse-grained task (F1: 0.590 fuzzy), while the multilingual model outperforms it on the fine-grained task (F1: 0.432 fuzzy). Results indicate that multilingual embeddings confer an advantage for fine-grained semantic distinctions, that English embeddings introduce systematic bias in cross-lingual transfer, and that zero-shot NER represents a viable, reproducible baseline for low-resource historical languages. Fine-tuning on guideline-compliant annotated data remains a priority for future work.

Keywords: Named-Entity Recognition, Zero-Shot Learning, Latin, GLiNER.

1. Introduction

When we first read the annotation guidelines proposed by the EvaLatin coordinating team, we were impressed by their precision and ambition. The proposed schema does not merely identify broad categories of named entities; it articulates a remarkably fine-grained ontology capable of capturing subtle distinctions within historical texts. Such granularity raises a methodological challenge: how can a Named-Entity Recognition (NER) model efficiently apply this schema in a language for which annotated resources remain scarce?

Rather than normalising existing annotations to previous guidelines and fine-tuning a model, the present study explores zero-shot learning. Zero-shot models generalise to new entity types without training and achieve F1 scores of approximately 0.6 on a range of tasks in English (Alhoshan et al., 2023, Del Moral-González et al., 2025, Marchitan et al., 2025), a performance level sufficient to bootstrap silver-standard corpora. This property is of particular relevance for Latin, where philologically informed manual annotation is costly. The proposed approach therefore integrates a zero-shot model into a reusable pipeline[1].

Prior work has applied LLMs in zero-shot or few-shot settings to historical documents (Hiltmann et al., 2025, Zhang and Colavizza, 2025), including studies on classical languages (Akavarapu et al., 2025), and the reception of antiquity in modern languages (Poibeau, 2024, Ripoll-Alberola and Burghardt, 2025). Performance on CPU-executable systems remains unexplored, despite its importance for reproducibility in resource-constrained projects in the humanities. For this reason, we adopt GLiNER2 (Zaratiana et al., 2025), which extends the original GLiNER model (Zaratiana et al., 2023) to become a general information extraction system. GLiNER2 achieves performance comparable to GPT-4o in NER, operates 2.6 times faster on CPU, and is openly available[2].

The GLiNER family of architectures concatenates task specifications with the input text, computes contextualised representations jointly, and refines them through feedforward networks to produce task-specific and span embeddings. A matching score (dot-product + sigmoid) is then computed for each span–task pair, producing a probability that a given span corresponds to a given category. GLiNER2 extends this framework with a richer special token vocabulary, enabling NER, hierarchical structured extraction, and text classification within a single forward pass, whereas the original GLiNER was NER-only. For NER specifically, however, the underlying mechanism remains essentially identical. GLiNER2 also improved NER performance over the original model[3].

[1]All code is documented and available in: https://github.com/luisarip/classificatio_sine_iactu/

[2]https://huggingface.co/collections/fastino/gliner2-family

[3]We may refer to the model by the general name of

Proceedings of LT4HALA 2026 @ LREC 2026, pages 443–447.

Whether these advantages – computational accessibility, label flexibility, and the elimination of task-specific pretraining – translate into sufficient performance for low-resource historical languages is the question this study addresses.

GLiNER relies on pretrained bidirectional transformer encoders (DeBERTa or mDeBERTa) as its backbone. Since these embeddings cannot be replaced with Latin-specific representations without full retraining, a cross-linguistic approach is adopted, following Soffiantini (2024): the model is applied to an English translation, and identified entities are aligned back to the original Latin text.

2. Description of the system

These are the main steps of our system:

a. Translation LAT>ENG.

b. GLiNER call.

c. Rule-based post processing.

d. Alignment ENG>LAT.

e. Projection of the results back to the original TSV.

In step (a), sentence boundaries are identified using the "EndOfSentence" marker in the MISC column, and sentences are passed individually to the translation engine. Preliminary checks demonstrated that longer, entity-dense passages caused the model to miss named entities. Sentence-level inputs show consistent performance over varying sentence lengths and were therefore adopted.

For translation, the deep-translator Python package[4] was used to query the Google Translator API, with an inter-request delay to avoid rate limiting. Translation outputs were cached to avoid redundant API calls. This was considered reliable given the deterministic behaviour of the Google Translate API.

Two model variants were compared: the multilingual[5] and the monolingual large[6]. Coarse and fine-grained labels were defined in a separate JSON file, selected at runtime via the `taxonomy_name` argument. All tag names were lowercased. The GLiNER architecture accepts an optional natural-language description per tag; accordingly, the descriptions provided in the annotation guidelines were adapted for zero-shot use by removing examples and converting singular forms to plural where the category was "collective". NER is used rather than hierarchical structured extraction, as it does not perform adequately for this task.

Development on the sample data revealed two recurring issues: labelling inconsistencies among occurrences of the same entity (e.g., *Iuno* [the goddess Juno] was classified as both person and creature), and a substantial proportion of low-confidence false positives. A rule-based post-processing step was introduced: labels are normalised by majority vote, and predictions falling below a confidence threshold of 0.65 are discarded.

These results, obtained on the English text, are aligned back to the original Latin text. The alignment is performed using simalign (Jalili Sabet et al., 2020), a word alignment method which uses static and contextualised embeddings, meaning that parallel training is not required. We run simalign using the embeddings of UGARIT/grc-alignment (Yousef et al., 2022b), an XLM-RoBERTa-based alignment model trained with Ancient Greek-Latin and Ancient Greek-English pairs. These embeddings generalise well to Latin-English alignment despite the absence of training for this language pair (Yousef et al., 2022a). English-Latin alignments are computed and projected onto the original TSV file, converted to BIO format, and subjected to a final rule-based check: tokens in lowercase, assumed to be common nouns rather than named entities, are filtered out and relabelled as O. This mechanism was not always enforced. In the sample data, for instance, *populi Romani* [the Roman people] was labelled as B-COLLECTIVE/.ETHNIC, I-COLLECTIVE/.ETHNIC. The current function requires at least one uppercased token, but does not exclude cases in which additional non-entity tokens are assigned entity labels, as in *Albanique patres* [Alban fathers], were only *Albanique* bears the COLLECTIVE.ETHNIC label.

3. Results

Table 1 presents aggregate results. Per-entity results for the best-performing model in each task are also given: the large model for the coarse-grained task (Table 2) and the multilingual model for the fine-grained task (Table 3). The large model achieves higher precision for all configurations, while the multilingual model delivers higher recall.

Notably, in the NERC coarse task, the categories benefiting most from fuzzy evaluation are LANGUAGE (which doubles in performance with the multilingual model, 0.667/1.000/0.800, and improves in 50.1% with the large model, 1.000/0.750/0.857) and WORK (which reaches 0.500/1.000/0.667 with the multilingual model under fuzzy evaluation). In the NERC fine task, LANGUAGE similarly improves under fuzzy evaluation

"GLiNER". However, we are always applying GLiNER2, the most recent version.

[4] https://github.com/nidhaloff/deep-translator

[5] https://huggingface.co/fastino/gliner2-multi-v1

[6] https://huggingface.co/fastino/gliner2-large-v1

Task	Evaluation	Model	Precision	Recall	F1 Score
Task 1: NERC Coarse	Strict	Multilingual	0.552	**0.500**	0.525
		Large	**0.575**	0.498	**0.534**
	Fuzzy	Multilingual	0.619	**0.561**	0.589
		Large	**0.635**	0.550	**0.590**
Task 2: NERC Fine	Strict	Multilingual	0.466	**0.327**	**0.384**
		Large	**0.491**	0.218	0.302
	Fuzzy	Multilingual	0.523	**0.367**	**0.432**
		Large	**0.534**	0.237	0.328

Table 1: Performance over all entities in the test sample for Task 1 (NERC Coarse) and Task 2 (NERC Fine). Results are reported using strict and fuzzy evaluation metrics. Highlighted, the best values per evaluation.

(0.800/1.000/0.889). WORK is not correctly identified by any model, unlike in the coarse task.

Performance drops between the sample (validation) data and the test set, suggesting sensitivity to textual complexity, entity density, and entity type. The *Metamorphoses*, for instance, contains a high density of mythological figures lacking a dedicated label, which may introduce ambiguity for the model. On the sample data, the large model achieved average F1 of 0.66 (COLLECTIVE 0.68, EVENT 0.67, PERSON 0.67, PLACE 0.72).

4. Discussion

The principal finding is that multilingual embeddings improve performance on the fine-grained NERC task for Latin texts, even if the input is in English. In the coarse-grained task, the large model achieves marginally higher precision at the cost of a steeper drop in recall, making the two models comparable. The multilingual model's stronger recall, more pronounced in the fine task, likely reflects residual sensitivity to Latin morphology and vocabulary, given that mDeBERTa was trained on 390 million Latin tokens (Conneau et al., 2020).

Persons and locations are the best-identified categories, which is expected, given (i) their prevalence in most NER benchmarks, and (ii) overlap between Latin toponyms and proper names and their English or Romance-language equivalents. Nevertheless, some mistakes persist: for instance, as "Troy" is a given name in English, the city of Troy is many times labelled as PERSON, even when the Latin form *Troiae* is provided as input. It can thus be concluded that English exerts a disproportionate influence on the embedding space, even in multilingual models.

LANGUAGE benefits most from fuzzy evaluation: language mentions typically appear as multi-token expressions (e.g., *Graecis litteris* [Greek letters], *lingua Britannicae* [in British language]), but the model retrieves only the capitalised adjective (*Graecis*, *Britannicae*).

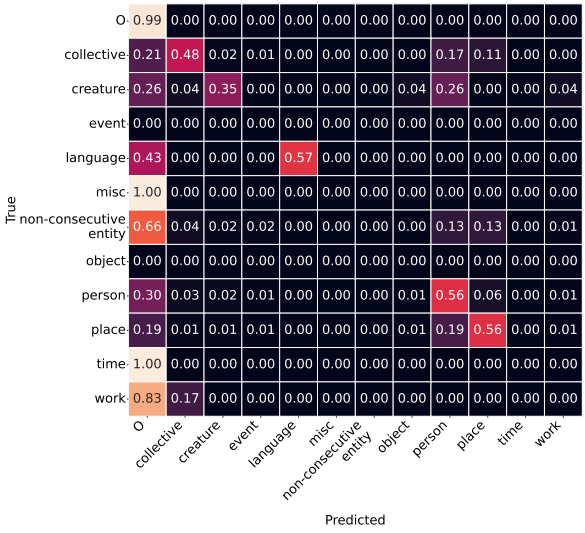

Figure 1: Normalised confusion matrix for NERC coarse task with large model.

Evaluation using seqeval (Nakayama, 2018) produces the normalised confusion matrix in Figure 1. Multitoken entities are predominantly categorised as one-token entities, thereby explaining the high prevalence of O among entities. Two problems recur in the classification of person entities. First, many mythological figures such as Iuno are sometimes misclassified as CREATURE, bypassing the anthropomorphic reading. This points to a potential gap in the tagset: a PERSON.MYTH subclass would guide the model in such cases, and serve annotation tasks focused on mythological rather than historical persons. Second, COLLECTIVE.ETHNIC, PLACE.DERIVATIVE and PERSON.ETHNIC are highly specific categories, all three of which are derived adjectives with toponymical roots. While they offer annotators the opportunity to capture nuances, they also present under-

Entity Type	P	R	F1	TP	FP	FN	F1 Fuzzy
ALL	0.575	0.498	0.534	1327	982	1338	0.590
COLLECTIVE	0.555*	0.467*	0.507*	197	158	225	0.533
CREATURE	0.320*	0.348*	0.333*	8	17	15	0.333
LANGUAGE	0.667	0.500	0.571	2	1	2	0.857
PERSON	0.633	0.515	0.568	262	152	247	0.628
PLACE	0.604	0.504	0.549	858	563	846	0.614

Table 2: Performance (strict evaluation) for the NERC Coarse task using the best-performing model *gliner2-large-v1*. Entity types EVENT, MISC, OBJECT, TIME and WORK were omitted because precision, recall, and F1 were zero. Entity types marked with an asterisk (*) were better identified by the multilingual model: COLLECTIVE (0.729/0.645/0.684) and CREATURE (0.389/0.609/0.475). The last column reports the F1 score obtained using fuzzy evaluation.

Entity Type	P	R	F1	TP	FP	FN	F1 Fuzzy
ALL	0.466	0.327	0.384	872	999	1793	0.432
COLLECTIVE.ETHNIC	0.167	0.105	0.129	41	204	350	0.142
CREATURE	0.500	0.294	0.370	5	5	12	0.370
LANGUAGE	0.400*	0.500	0.444*	2	3	2	0.889
PERSON	0.652*	0.365*	0.468*	133	71	231	0.518
PERSON.AUTHOR	0.147*	0.109*	0.125*	5	29	41	0.200
PLACE	0.605	0.417	0.494	686	447	961	0.554

Table 3: Performance (strict evaluation) for the NERC Fine task using the best-performing model *gliner2-multi-v1*. The categories COLLECTIVE, COLLECTIVE.ANCESTRY, COLLECTIVE.DERIVATIVE, CREATURE.ANIMAL, CREATURE.ASTRONOMY, EVENT, MISC, OBJECT, PERSON.ANCESTRY, PERSON.DERIVATIVE, PERSON.EPITHET, PERSON.ETHNIC, PLACE.ASTRONOMY, PLACE.DERIVATIVE, TIME and WORK were omitted because precision, recall and F1 were zero. Entity types marked with an asterisk (*) were better identified by the large model: LANGUAGE (0.667/0.500/0.571), PERSON (0.662/0.393/0.493) and PERSON.AUTHOR (0.400/0.304/0.346). The last column reports the F1 score obtained using fuzzy evaluation.

standing challenges for LLMs. The translation step is crucial in the differentiation of these three fields, see an example below:

> *Progeniem sed enim Troiano a sanguine duci audierat* (Verg. A. 1.19)
> Troiano → place.derivative
>
> *but anxiously she heard that of the Trojan blood there was a breed then rising* (trans. Williams 1910)
>
> *But he had heard of the race of the Trojans, led by blood* (Google Translate)
> Trojans → collective

When analysing the fine-grained task, a broad pattern is observed: the model strongly favours first-level tags, which are more frequent in the data, and rarely commits to second-level distinctions. As a result, the model has an acceptable performance but completely misses its purpose. PERSON.AUTHOR, PERSON.DERIVATIVE, PLACE.DERIVATIVE, CREATURE.ANIMAL, and CREATURE.ASTRONOMY are largely collapsed into their parent categories or labelled as O. Among subcategories, only COLLECTIVE.ETHNIC and PERSON.AUTHOR generate true positives, and PERSON is the only category with true positives in more than two tags, given that all collectives are labelled as COLLECTIVE.ETHNIC. Therefore, the model retains coarse-grained knowledge but does not generalise to finer distinctions.

5. Conclusion

Although performance ranged approximately between 0.4 and 0.6 over entities and models, zero-shot NER with GLiNER may suffice for simpler Latin texts. Multilingual models are a better option, even in a cross-linguistic setting. The primary limitation is the impossibility of applying Latin-specific embeddings; conversely, the ability to process Latin texts without any annotated data represents a direction worth further investigation. All code has been

made available on GitHub to ensure reproducibility in Latin texts of choice. Future work should address three directions: replicating these experiments with higher-quality Latin–English translation; adapting the annotation guidelines for better understanding of the transformer model; and fine-tuning the GLiNER model on additional annotated data following the guidelines.

6. Acknowledgements

Luisa Ripoll-Alberola acknowledges funding from the European Union's Horizon Europe Research and Innovation Programme through the Marie Sklodowska-Curie Actions Doctoral Network MECANO, Grant Agreement No. 101120349.

7. Bibliographical References

V.S.D.S.Mahesh Akavarapu, Hrishikesh Terdalkar, Pramit Bhattacharyya, Shubhangi Agarwal, Dr. Vishakha Deulgaonkar, Chaitali Dangarikar, Pralay Manna, and Arnab Bhattacharya. 2025. A case study of cross-lingual zero-shot generalization for classical languages in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 2745–2761, Vienna, Austria. Association for Computational Linguistics.

Waad Alhoshan, Alessio Ferrari, and Liping Zhao. 2023. Zero-shot learning for requirements classification: An exploratory study.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Rodrigo Del Moral-González, Helena Gómez-Adorno, and Orlando Ramos-Flores. 2025. Comparative analysis of generative llms for labeling entities in clinical notes. *Genomics Informatics*, 23(1):3.

Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, Wiebke Sczeponik, and Anica Skibba. 2025. Ner4all or context is all you need: Using llms for low-effort, high-performance ner on historical texts. a humanities informed approach.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Teodor-George Marchitan, Claudiu Creanga, and Liviu P. Dinu. 2025. Few-shot text-based emotion detection.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Thierry Poibeau. 2024. Annotating references to mythological entities in french literature. (arXiv:2412.18270). ArXiv:2412.18270 [cs].

Luisa Ripoll-Alberola and Manuel Burghardt. 2025. Llm-based approaches to canonical reference extraction in academic texts: Initial results. *AII-EEKE 2025 Joint Workshop of the 5th AI + Informetrics (AII) and the 6th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE), in 20th International Conference on Scientometrics Informetrics (ISSI)*.

Laura Soffiantini. 2024. Cross-Linguistic Annotation Transfer in Geoparsing Experiments with Classical Texts. *DH Benelux Journal*, (6):155–168.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022a. An automatic model and gold standard for translation alignment of Ancient Greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5894–5905, Marseille, France. European Language Resources Association.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022b. Automatic translation alignment for ancient greek and latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. Gliner2: An efficient multi-task information extraction system with schema-driven interface.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer.

Shibingfeng Zhang and Giovanni Colavizza. 2025. Named entity recognition of historical text via large language model.