I, RE:Claudius 256: Towards Linking Classical Latin Person Mentions to a Domain-specific Knowledge Base

Marijke Beersmans, Evelien de Graaf, Julie Nijs, Valeria Irene Boano, Alek Keersmaekers, Mark Depauw, Tim Van de Cruys, Margherita Fantoli
KU Leuven
Blijde Inkomststraat 21, 3000 Leuven, Belgium
{marijke.beersmans, evelien.degraaf, julie.nijs}@kuleuven.be
{valeria.boano, alek.keersmaekers, mark.depauw}@kuleuven.be
{tim.vandecruys, margherita.fantoli}@kuleuven.be

Abstract

This paper considers Named Entity Linking for person mentions from classical Latin texts to a domain-specific, German language knowledge base, namely *Paulys Realencyclopädie*. Following a methodology similar to Beersmans et al. (2025), adapted to classical Latin, we train a transformer-based, retrieval and ranking model (BLINK) for this task. To mitigate data scarcity, we first train the model on a newly created general-purpose dataset derived from Wikipedia. We then fine-tune it on our domain-specific dataset, which is collected from various sources and linked to our target knowledge base. Results show that while BLINK performs well on mention-entity pairs linked to entities seen during training, it performs significantly worse on mention-entity pairs linking to unseen entities. We provide a detailed error analysis, propose possible exploitation strategies for a human-in-the-loop approach, and identify directions for future improvement.

Keywords: Named Entity Linking, BLINK, Classical Latin, Paulys Realencyclopädie, domain specific knowledge base

1. Introduction

Named Entity Linking (NEL) is the task of linking mentions in a running text to their corresponding entity in a predefined knowledge base (KB). As such, the texts are enriched with entity-based information, and the mentions are disambiguated.

As an example, in the passage:

- "After their time the courses of both stars for 600 years were prophesied by **Hipparchus**, whose work embraced the calendar of the nations and the situations of places and aspects of the peoples..." (*Naturalis Historia*, book II, section 9.53, Transl. Harris Rackham)

the mention *Hipparchus* should be linked to:

- Hipparchos 18, Gr. Astronom aus Nikaia in Bithynien (Hipparchos 18, Greek astronomer from Nicaea in Bithynia). (*Paulys Realencyclopädie Wikisource*)

and not to any of the other entries in the chosen KB (here, the Paulys Realencyclopädie, Wissowa et al. 1893–1980).

This paper considers NEL for classical Latin texts. The task is in high demand within the humanities, as a prerequisite for enhancing texts through structured searching or complex analysis, such as social network analysis. Currently, almost all Entity Linking in the context of historical texts is performed manually, which is labour-intensive and expensive (Ehrmann et al., 2023). A sufficiently accurate automatic NEL model could reduce these costs and save time, at least by providing reliable suggestions for a human-in-the-loop approach. We aim to address this gap from the context of the wider NIKAW project[1], which seeks to study the circulation of Knowledge based on the mentions of person names in texts from Greco-Roman antiquity.

Most research on automatic NEL relies on large, openly available, community-curated knowledge bases that cover entities from many different types and domains, most notably Wikipedia and related projects, such as Wikidata or DBPedia (Sevgili et al., 2022). While this approach is well motivated and widely adopted, we instead rely on a domain-specific knowledge base, namely the Wikisource edition of *Paulys Realencyclopädie* (*RE*) (in German). This choice offers more comprehensive and reliable entity coverage for antiquity, but it comes at the cost of lacking pre-existing resources such as Wikipedia's hyperlink structure or Wikidata's rich property annotations. In this paper, we assess the viability of BLINK (BERT entity Linking), a knowledge-base agnostic, transformer-based approach (Wu et al., 2020) for linking entities in classical Latin texts to a German-language knowledge base. We used a methodology similar to our previous work Beersmans et al. (2025), with the workflow adapted to classical Latin and an additional training

[1] https://research.kuleuven.be/portal/en/project/3H220323

Proceedings of LT4HALA 2026 @ LREC 2026, pages 152–162.

stage based on a Wikipedia-derived dataset (see Section 4 for more information). This constitutes a challenging, cross-lingual, and domain-specific setting. We train a model tailored to this setting while still relying on a Wikipedia derived dataset in a first training stage. We provide a detailed qualitative error analysis, identify bottlenecks, and suggest potential improvements. Code and data will be provided on Github.

The paper is structured as follows: we first discuss related work in Section 2. We then introduce the Knowledge Base and the enhancements made to it in Section 3. Subsequently, we detail our approach, data and model more thoroughly in Section 4. Afterwards, we report the results and error analysis in Section 5 and discuss our findings in Section 6. Finally, we conclude with Section 7.

2. Related Work

NEL for historical corpora is a particularly challenging task. Its complexity stems both from the long-tail distribution of entities (many appear only once and therefore provide very limited contextual information for disambiguation) and from the difficulty of relying on Wikipedia or Wikidata as knowledge bases. These resources, although widely used in general NEL and Wikification pipelines (i.e. Entity Linking to Wikipedia), often lack sufficient coverage of premodern or highly specialized historical entities.

Although the task of Wikification dates back to the early 2000s, a major shift occurred with the HIPE shared tasks in 2020 (Ehrmann et al., 2020) and 2022 (Ehrmann et al., 2022). These initiatives explicitly included NEL for historical corpora: newspapers from the 19th-20th centuries in 2020, and additionally modern commentaries to classical works in 2022. Already in the 2020 edition, some participants combined transformer models with filtering based on Wikipedia or Wikidata, such as Labusch and Neudecker (2020) and Boros et al. (2020), both discussed by Santini et al. (2026).

Among transformer-based approaches, BLINK has shown potential for NEL on Ancient Greek, though by no means solving the task (see Beersmans et al. 2025). Its key advantage lies in allowing the use of any knowledge bases, extending the task beyond standard Wikification. In this paper, we apply BLINK to a Latin corpus using a KB other than Wikipedia or Wikidata, while still exploiting the availability of the Latin Wikipedia to perform domain-specific pre-training on relevant entities. BLINK has since been extended in various directions: for example Graciotti et al. (2025a) introduced constraints to better handle non-linkable entities, achieving state-of-the-art performance. In parallel, generative models have been explored. GENRE (De Cao et al., 2020), and its multilingual adaptation mGENRE (De Cao et al., 2022), generate the target entity label directly as a sequence prediction task, rather than ranking pre-computed candidates. More recently, Graciotti et al. (2025b) evaluated BELA (Plekhanov et al., 2023), a multilingual bi-encoder based on BLINK (see Section 4.3 for an explanation of the functioning of BLINK), mGENRE, GPT-4o mini and LLaMa3-70B in zero-shot conditions for Wikidata-based NEL, with specific attention for datasets with long tails. Their experiments showed that generic Large Language Models such as GPT and LLaMa perform substantially worse on low-popularity entities. Building on these insights, Santini et al. (2026) proposed a hybrid framework that combines BELA, structured Wikidata metadata, and an instruction-tuned LLM adapted to the language of the corpus. However, this framework is explicitly tailored to Wikidata and cannot be directly applied to arbitrary KBs. Finally, to the best of our knowledge, no previous work on automated NEL has been undertaken for texts written in Latin.

In addition to neural approaches, string similarity methods (notably Levenshtein distance) remain common in NEL pipelines. Both Shen et al. (2015) and Sevgili et al. (2022), two comprehensive surveys of NEL, describe the use of string similarity for candidate generation and ranking, two core subtasks of entity linking. We therefore include a string similarity baseline in our experimental setup.

3. Knowledge Base : Paulys Realencyclopädie

Paulys Realencyclopädie der classischen Altertumswissenschaft, below *RE*, was originally published between 1893 and 1978. As an expansive reference encyclopedia on classical antiquity, it comprises around 100,000 entries contributed by prominent classical philologists. At present, a significant portion has been digitized through the German Wikisource project. De Graaf et al. (2024) have shown that this resource provides sufficiently broad entity coverage for NEL in the target domain.

At the moment of writing, the complete work is available as scanned images, accompanied by access to a manually verified *Volltext* (full text) for over 65,000 articles. A searchable index or *Register* of keywords (*Stichwörter*), including brief summaries for all entries (*Kurztext*), is also provided to support navigation. Figure 1 shows how these components are presented in the entry for the hero Abas (Abas 3).

In de Graaf (forthcoming) a separate instance of the *RE* was created on the basis of the Wikisource project. This local instance was then further preprocessed by the developers of the Trismegistos (TM) database, an expansive metadata platform fo-

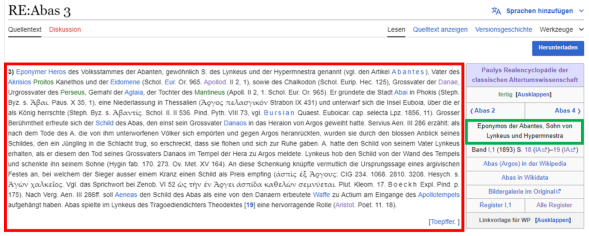

Figure 1: "Abas 3" in the Wikisource edition, highlighting *Volltext* in red and *Kurztext* in green.

cused on the Ancient World (Depauw and Gheldof, 2014). Firstly, all persons were manually identified, with the definition of “person” kept broad (including mythological creatures such as nymphs or muses, and deities). This choice ensures interoperability with NER systems and avoids confusion regarding edge cases such as “Heracles”, a hero with a demigod status. After this filtering, the KB contained a total of 54,180 entries manually identified as persons.

Secondly, for these identified people, the complete name was restored where possible by the TM team and split into its constituent elements. This is particularly important for Latin/Roman names, which often consist of three distinct parts: praenomen, nomen and cognomen. For instance, for the entry *Iulius 131*, the complete name is “C. (Gaius) Iulius Caesar”. Each of these was linked by members of the TM team to, among other things, the corresponding TM Nam ID 9067 (C./Gaius), TM Nam ID 6964 (Iulius) and TM Nam ID 9725 (Caesar) (Broux and Depauw, 2015).

4. Methodology

4.1. Overview

The experimental set-up of this paper is partially based on Beersmans et al. (2025), with two key differences. First, whereas the original pipeline was developed for Ancient Greek, we adapt it to link Latin texts to the *RE*. Second, we introduce an additional Wikipedia-derived pre-training phase that was not part of the original approach. We therefore train BLINK in two stages: we begin with a large, general Wikipedia-derived dataset to improve representations and mitigate data scarcity, and then continue with a smaller, heterogeneous dataset of classical Latin texts linked to the *RE* (the *RE*-dataset).

The *RE*-dataset is divided into "gold" and "silver" subsets. The gold data are manually annotated in the framework of philological projects, whereas the silver data is generated through a rule-based linking procedure that relies on various pre-existing linked identifiers. While the silver data is less time-consuming to create, it is also more error-prone than manual annotation. As such, silver data is used only during training and excluded from testing to ensure the reliability of evaluation results.

To link mentions from Latin texts to the entities of the *RE*, we additionally evaluate two scenarios. In the first, the model is trained to link exclusively to the *Kurztexts* (see Section 3), which are available for all entities (below referred to as *Kurz*). In the second, the model uses the *Volltexts* wherever they exist and falls back on the *Kurztext* otherwise (below referred to as *Voll*).

4.2. Data

This section first introduces the data used for the Wikipedia training phase (Section 4.2.1). It then describes the datasets used in the second, domain-specific phase, namely the silver data (Section 4.2.2) and the gold data (Section 4.2.3).

4.2.1. Wikipedia Data

The data for the Wikipedia-derived training phase were created based on two resources. First, Wikipedia (German and Latin versions) and second, a filtered Wikidata knowledge base, containing 30,000 persons of antiquity, created by Fantoli et al. (2026). To emulate the linguistic situation of the *RE*, only those entities that have a Wikipedia page in German were selected. Our final knowledge base contains 11,652 entities.

Based on scripts provided by De Cao et al. (2022), we derived a dataset of mention-entity pairs where the mention is a hyperlinked phrase in a Latin Wikipedia page and the entity is the corresponding page on German Wikipedia. An example can be found below:

- **Mention**: *Sed anno 1903 Sartus coronatus est Papa Pius X, et haud multum post coronationem, die* [MENTION_START] *Sanctae Caeciliae* [MENTION_END] *(patronae musicae), vulgavit Motu Proprio "Inter Sollicitudines" (cum Perosi conscriptus)* (Latin Wikipedia entry "Laurentius Perosi")
- **Entity**: *Cäcilia von Rom, Cäcilia von Rom wird in mehreren christlichen Konfessionen als Heilige, Jungfrau und Märtyrin verehrt [...])* (German Wikipedia entry "Cäcilia von Rom")

The resulting 29,421 mention-entity pairs were split into train, validation and test partitions. To ensure that certain uncommon entities were not observed during training, we deliberately reserved a small subset of mentions linked to these rare entities exclusively for the validation and test sets. An overview of the final counts per partition can be found in Table 1. The test set is only used for internal validation and results are not reported here.

Subset	Train	Dev	Test
Count	26,665	1,218	1,538
Of which unseen	n/a	122	118

Table 1: Subset counts for the Wikipedia dataset.

4.2.2. Silver Data

The starting point of the silver data pipeline was the description (*Volltext*) of a person entry in the *RE*. These descriptions sometimes cite passages in classical Latin texts, where the *RE* authors derived information about this entity. For instance, for the entity Caecilius 86 (*RE* id: 17476), the *Volltext* references *Hor. carm. II 1*, 1.

To move from the *RE Volltext* to these classical Latin source passages, we made use of various interlinked resources. One such resource is the LASLA corpus, in which all tokens have been manually linked to LiLa Lemma IDs (Fantoli et al., 2022).[2] We also rely on the Trismegistos Nam IDs assigned to each component of the restored full names for all *RE* entries, as described in Section 3. A link between LiLa:lemmas and Trismegistos Nam IDs already exists within the Trismegistos knowledge base, which facilitates this alignment. For more details, consult Beersmans et al. (2025), where a similar pipeline was used, or the provided schematic in Figure 2. Different from the Greek version of this pipeline, we did apply a rule-based procedure to reconstruct multi-token entities, where adjacent tokens assigned the same *RE* identifier are annotated as a single mention.

Since the procedure is fully automated, we evaluated the results using a random sample of 50 instances, of which 39 were correct. We found that one type of error pertained to **plural or generalized name forms**. For example, in the case where *Metellorum*, referring collectively to the Metelli, in *Cic. Planc.* 89 was incorrectly annotated as Caecilius 97 simply because he is mentioned in singular form elsewhere in the same source passage.

To avoid this, we opted to remove all plurals from the final dataset after this analysis was concluded, which was possible thanks to the LASLA morphological annotations. The final silver dataset comprises 620 mention-entity pairs.

4.2.3. Gold Data

The gold data was created in several contexts. In all subsets, person mentions were annotated with their respective corresponding *RE* article. Following Palladino et al. (2024), a person is defined as

[2]LiLa is a knowledge base interlinking linguistic resources, NLP tools, and corpora for Latin (Passarotti et al., 2020)

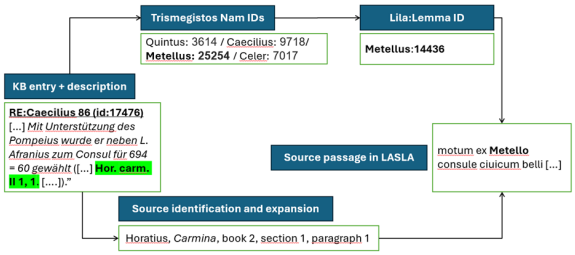

Figure 2: Schematic of the silver data creation process.

"any identifiable single individual, including deities and anthropomorphic mythological figures". The texts are taken from different digital corpora.

Ammianus' *Res Gestae* & Tacitus' *Historiae*. The first subset was created by de Graaf et al. (2024).[3] in the context of evaluating knowledge bases for NEL. It contains Tacitus *Historiae* 1 (annotated based on the LASLA version) and book XIV of the *Res Gestae* of Ammianus Marcellinus from LTA.[4]

Plato case study. The second subset consists of ancient Greek and Latin texts where the figure of Plato plays a relevant role and that are authored in diverse cultural and historical settings (de Graaf, forthcoming). It was developed in the framework of the NIKAW project.[5] The Latin texts of this corpus are annotated based on the linguistically annotated version from the *Corpus Latin Antiquité et Antiquité Tardive lemmatisé* (Clérice, 2020) (henceforth Corpus Latin).[6] The annotated texts used from this resource are:

- *Tusculanae Disputationes* by Cicero,
- *De Anima* by Tertullian,
- *Pro Se De Magia Liber* by Apuleius,
- *Divinarum Institutionum* by Lactantius.

In addition, one text from LASLA is included:

- *Ad Lucilium Epistulae Morales* by Seneca.

For *Divinarum Institutionum* and *Ad Lucilium Epistulae Morales*, only the sections that mention

[3]Data available here: https://github.com/evelien-degraaf/Named-Entity-Linking-Latin-DAAL-2024/tree/main/Gold_data.

[4]https://lta.bbaw.de/text/show/24819722_ammianus_marcellinus_res_gestae.

[5]https://research.kuleuven.be/portal/en/project/3H220323

[6]https://github.com/lascivaroma/latin-lemmatized-texts.

Plato were annotated (respectively chapters and letters).

Pliny NH. The third subset was created in the context of the MECANO doctoral network.[7] It consists of books 2-6 of Pliny the Elder's *Naturalis Historia*, taken from the Corpus Latin and manually corrected from the observed OCR-related errors.

Trismegistos/LASLA gold. A final subset consists of annotations produced by the Trismegistos team during work on LASLA/LiLa. Roughly 5,000 distinct capitalized words in the LASLA corpus were semi-automatically matched against Trismegistos variant datasets for place names (GEOVAR), divine names (GODVAR), and personal names (NAMVAR), yielding about 3,000 matches. The remaining 2,000 items were classified manually.

All 78,919 capitalized words in the corpus had been disambiguated as places (26,199), gods (8,566), or people (44,242). Places and gods were linked to TM identifiers, while only a minority of personal names (3,244) received identifiers. For these identified persons, multi-token entities were reconstructed using the same rule-based procedure described in Section 4.2.2. After removing all mentions overlapping with other annotated corpora, the final set comprised 2,982 mention–entity pairs.

4.2.4. Final *RE* Dataset

For the final dataset, all the gold subsets and the silver data were aggregated. The total counts of the mention-entity pairs for each subset, including silver, are presented in Table 2. The dataset was divided into training, validation, and test partitions. Silver entities were included exclusively in the training set. In addition, a subset of mentions linking to infrequent entities was set aside for the development and test sets to ensure that each contained at least 100 mentions linking to previously unseen entities. An overview is provided in Table 3.

Subset	# Mention-entity pairs
Trismegistos gold	2,982
Plato case study	2,388
Pliny NH	917
Silver data	620
Tacitus	619
Ammianus	202
Total	7,728

Table 2: Overview of different subset sizes

4.3. Model

BLINK (entity linking with BERT) is a transformer-based entity linking model that makes no assumptions about the nature of the KB, except that a description of each entity should be present. From a technical perspective, the model supports both knowledge-base-specific entity linking and the transfer of entity linking capabilities from one knowledge base (e.g., a large, well-resourced KB) to another that is smaller or more resource-constrained.

Subset	Train	Dev	Test
Count	6.656	537	535
Of which unseen	n/a	108	102

Table 3: Final train-validation split partitions with the numbers of mention linking to unseen entities.

BLINK uses a two-stage entity linking approach. In the first stage, both mentions and KB entity descriptions are encoded with two separate transformer models, forming a bi-encoder. The mention embedding is compared with all entity embeddings using dot-product similarity, and the top-k most similar entities (or candidates) are selected. During training, the bi-encoder is optimized to maximize the similarity between the mention and the correct entity compared to the mention and in-batch negatives (incorrect mentions found in the same batch).

In the second stage, each of the top-k entity candidates is concatenated with the mention and jointly encoded using a cross-encoder. A final linear layer assigns a score to each mention–entity pair in order to determine the best match. Although this step is slower and more computationally expensive, it yields higher accuracy because the model can attend the mention and the entity jointly.

While the original BLINK implementation is English-only (although a multilingual implementation of the bi-encoder exists, see Section 2), it can be adapted by replacing the base transformer BERT with a multilingual model. Our new base transformer is XLM-RoBERTa (Conneau et al., 2020), which contains both Latin and German.

4.3.1. Training Parameters

To choose our training hyperparameters, we began by reusing the hyperparameters from the zero-shot setting in Wu et al. (2020), since this setting involved training a model of comparable size to ours. We performed a manual search for the the learning rate and number of epochs for the bi-encoder, monitoring the in-batch accuracy using the validation set. We use a per-GPU batch size of 32, which is a standard and memory-efficient configuration for sequence length 128 and provides 31 in-batch negatives per instance. An overview of the final parameters can be found in Table 4. The hyperparameters for *Voll* and the *Kurz* are kept identical.

[7] https://mecano-dn.eu/

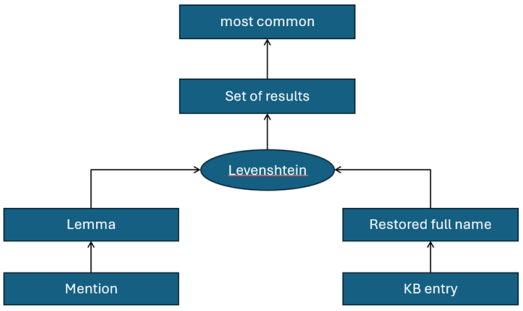

Figure 3: Schematic of the baseline approach

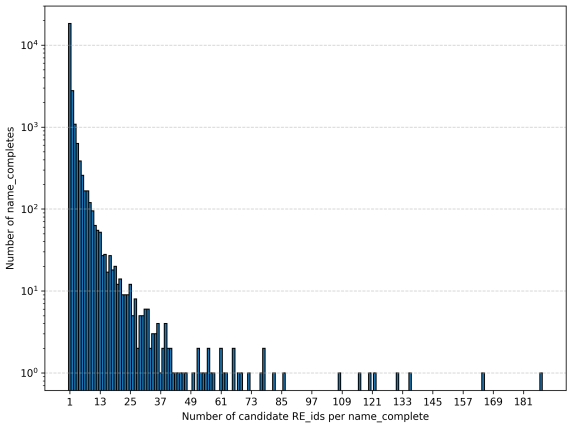

Figure 4: Number of RE_ids per complete name.

4.3.2. Baseline

We compare the performance of BLINK to a two-step approach relying first on string similarity (for the bi-encoder, or "candidate generation" step) and, second, on the frequency of the entity in the training data (for the final ranking step). For clarity, a schematic is provided in Figure 3.

The first part of the baseline implements a lexical candidate selection and identifier expansion procedure using the *RE* KB. Mention data and the *RE* data are first loaded and lowercased. From the KB data, a lookup structure is constructed that maps each available complete name to its corresponding RE_IDs. If a complete name (see Section 3) is present for an entry, it is used as the lexical representation; if not, a cleaned version of the entity's title in the *RE* (e.g. Iulius 131 would be cleaned to "Iulius") serves as a fallback.

For each mention, the baseline pipeline performs exhaustive top-1 Levenshtein matching against all candidate names. Whenever a lemma is available for the mention, the edit-distance computation is performed on the lemmatized form[8]; otherwise, the surface mention is used. For each token in the mention, the minimal Levenshtein distance to any token in the candidate name is computed, and these minimal distances are averaged to yield a similarity score. The candidate name(s) with the lowest average distance are selected, with ties retained if multiple names achieve the same best score.

In a second step, all RE_IDs associated with the selected top-1 name(s) are retrieved and aggregated. A single canonical name may correspond to multiple *RE* entries (as a straightforward example, the canonical name Thespis is associated to *RE* entries Thespis 1 (83642); Thespis 2 (83643) and Thespis 3 (83644)). This expansion can therefore result in one or multiple candidate identifiers (RE_ids) per mention (see Figure 4 for the distribution of candidate entities per name).

[8]As the data are compiled from different sources, the lemmata have differing methods of u/v normalization, similarly to the mentions.

Parameter	Wiki	*RE*
Bi-encoder		
LR	2×10^{-5}	2×10^{-6}
# Epochs	10	5
Batch size	32	32
Cross-encoder		
LR	1×10^{-5}	1×10^{-5}
# Epochs	2	2
Batch size	1	1
Top-k	64	64
Context length	128	128

Table 4: Hyperparameters across training phases.

The entire group of entities retrieved by this setup is compared to the output of the bi-encoder. Although the Levenshtein pipeline is allowed to retrieve an undefined amount of RE_IDs and the bi-encoder retrieves a fixed amount, the comparison provides insights into the effectiveness of the surface matching for the task. The Levenshtein pipeline retrieves 17 RE_IDs on average, with outliers up to 509.

Finally, for the second step of the baseline, we select the most common entity from this set of retrieved entities in the training data. Naturally, this comparison is mostly relevant for the mention-entity pairs where the entity was seen during training. However, we also include in the evaluation of the second step the cases in which the entity is unseen but the Levenshtein pipeline only retrieved one entity, and hence we simply assess whether it is correct or not.

5. Results

5.1. General Overview

The following section details the results of the models after training on the RE. We only report the results on our final KB and not the intermediate results on the Wikipedia test set. The bi-encoder is evaluated on recall@64, whether or not the correct entity is present in the retrieved set of 64 most likely entities. 64 is one of the standard values tested by Wu et al. (2020) and also used in the zero-shot setting. The cross-encoder is then evaluated on its correct predictions (using accuracy) for those instances where the bi-encoder was able to retrieve the correct entity. The overall accuracy, where both bi-encoder retrieval errors and cross-encoder errors are taking into account, is also reported. The column "in training" refers to mention-entity pairs in the test set where entities were seen during training (m=433), and "unseen" refers to mention-entity pairs where the entity has not been seen during training (m=102). We additionally report whether the correct entity was present in the final top 5.

An overview can be found in Table 5. Overall, the Levenshtein baseline is competitive as a candidate generation step compared to the bi-encoder, even outperforming it in the "unseen" setting. Furthermore, the *Voll* outperforms the *Kurz* and both models outperform the most-common baseline in all settings. As expected, performance drops notably when the models are evaluated on unseen entities in both variants. Interestingly, when evaluated on mentions linking to entities unseen in the training data, *Kurz* and *Voll* score comparably. When evaluating whether the correct prediction is among the top 5 entities ranked by the cross-encoder, results improve significantly, up to +20% for unseen entities. We discuss this point in Section 6.

Model	Total	In training	Unseen
Baseline			
Lev	79.44	79.21	80.39
Most common	64.67	75.06	20.58
Kurz			
Bi recall@64	85.42	87.53	76.47
Cross acc.	90.37	94.72	69.23
Total acc.	77.20	82.91	52.94
In top 5	84.30	87.07	72.54
Voll			
Bi recall@64	87.66	91.22	72.55
Cross acc.	93.17	96.96	72.97
Total acc.	81.68	88.45	52.94
In top 5	86.54	90.76	68.63

Table 5: RE dataset results for the total test set, mention linking to entities seen during training and those linking to unseen entities, evaluated on bi-encoder recall@64 (bi recall@64), cross-encoder accuracy (Cross acc.) and total accuracy (Total acc.).

5.2. Error Analysis

Among the 535 entities in the test set, *Kurz* produced 122 incorrect predictions and *Voll* produced 98. Of these errors, 76 were shared: in 31 cases both models made the same incorrect prediction, while in the remaining 45 their incorrect outputs differed. The observations below are based on a closer inspection of 50 incorrect predictions per model.

Firstly, the analysis highlights several aspects of the datasets that could have influenced the models. We observed several issues, including the **inconsistent representation of the phoneme /v/** as ‹v› or ‹u› across editions in the test set. For example, both the form *uespasiano* (Tacitus) and *vespasiano* (Pliny NH) occur. Additional errors originate in the training data due to imperfect OCR, such as in Plato case study *omdius* for *ovidius*, *maxumus* for *maximus*, *heropbilus* for *herophilus* and, in NH_Plinius, *vlixis* for *ulixis*.[9]

In addition, we observed **inconsistent annotations across datasets**. These issues occur primarily when multiple entries in the KB refer to the same individual. Examples include separately recorded epithets, aliases, *Verweise* (redirects from one *RE* entry to the other), and Latin and Greek name variants. Such duplications can produce false negatives: *ulixes*, for instance, the Latin name for the hero Odysseus, may be linked to either Ulixes or Odysseus. Epithets for Jupiter illustrate the problem as well: *elicius* in Pliny NH is annotated as Iuppiter, yet *Voll* predicts Elicius, which is counted as an error despite being semantically correct. In the case of *uespasiano*, both models return the *Verweis* in the *RE*, "Vespasianus → Flavius 206" rather than the target entity Flavius 206. Not all predictions are defensible, however: for *elissa*, an alias of queen Dido, *Kurz* predicts Elix 1 and *Voll* predicts Elis 5, both unrelated. Other issues arise when a person is referred to only by alias which is not independently represented in the KB. This is the case for *magne* for Pompeius 31 (Pompeius Magnus), where *Kurz* predicts Magnes 1 and *Voll* predicts Mago 15.

The remaining errors show that both models struggle to disambiguate individuals who share the same name (homonyms). As an example, in the test set, there are 32 instances where a lemma refers to more than one historical individual. These

[9] This represents a case of OCR error missed during the correction process of the Pliny NH data. The error has been corrected in a more recent version of the dataset.

32 occurrences involve six different lemmas.[10] For these lemmas, we evaluated whether the model predictions were correct. Out of the 32 homonymous cases, the Voll model misclassified 18, while the Kurz model misclassified 22. Importantly, this count reflects only homonymy within the test set; other lemmas may be homonymous in the training or development splits.

We can observe three main patterns when models are confronted with homonymity.

1. **Defaulting to one entity**: The lemma *Cinna* refers in the test set to both Helvius 12 (C. Helvius Cinna) and Cornelius 106 (L. Cornelius Cinna): the two models predict Cornelius 106 for all but one instance, where *Kurz* incorrectly selects Helvius 12.

   The case of Caesar is equally illustrative. Correct predictions would include Constantius 5, Iulius 131, Sulpicius 63, and Claudius 256. In seven of nine occurrences of *caesar*, the two models default to Constantius 5; although for *claudius caesar* (correct: Claudius 256), both predict Claudius 92, despite the correct entity being more prevalent in the training data.

2. **Confusion between multiple candidates**: In case of *Claudius*, where three individuals are possible (Claudius 115, 239, and 256). Only for *claudius maximus* (plato_case_study) both models correctly return Claudius 239. Elsewhere, predictions diverge and are uniformly wrong: Claudianus 9, Claudius 328 (*Voll*) or Claudius 109 (*Kurz*) and Claudius 92.

3. **Correct disambiguation (rare)**: *Cato* is the sole example where one model succeeds for multiple individuals: *Voll* correctly links both mentions of Porcius 16 and the single mention of Porcius 9.

Notably, for the examples highlighted in this paragraph, these outcomes cannot be explained simply by frequency in the training data alone, since the models do not consistently choose the most common individual.

The 46 errors made by *Kurz* but not by *Voll* indicate that the *Voll* model handles variation in mentions' surface forms more effectively. These cases suggest that the bi-encoder in *Voll* is less sensitive to variation: *Kurz* correctly predicts Alexandros 10 only for the token *alexandri*, while for the other eight forms (*alexander*, *alexandri*, and *alexandro*) Alexandros 10 does not even appear in *Kurz*'s top-64 candidates. *Voll*'s predictions of these entities, instead, are all correct. A similar improvement appears for the entity Amphitryon, in the form of *amphitruo*, which is correctly predicted by *Voll*. On the other hand, *Kurz* correctly predicts only once the form *amphitruone*, while it never links correctly the form *amphitruo*.

In the 22 errors made by *Voll* but not by *Kurz*, we found no consistent patterns.

6. Discussion

From our results and error analysis, we can investigate whether our trained BLINK is viable for humanities research. While model performance is competitive for entities seen during training, for unseen entities, roughly a fifth of the test set, both *Kurz* and *Voll* were correct in only about half the cases, making them hard to employ without further verification.

The error analysis revealed that KB phenomena such as duplicated entries, redirects, aliases, and parallel Greek/Latin forms, produce systematic false negatives and errors. This indicates that KB selection is crucial for the task: inconsistencies within the KB propagate into both the training and test sets, inevitably affecting model performance and undermining the soundness of the evaluation. The Wikisource *RE* partially suffers, from this standpoint, from its incompleteness with regards to the *Volltext*, due to being a work-in-progress subject to copyright laws, and from the fact that it mirrors the structure of a printed resource without a rigorous data model behind it. However, Wikidata and Wikipedia, being community-based, notably suffer from similar issues (e.g., duplicate entries) and currently offer less extensive coverage (Fantoli et al., 2026). In the future, we plan to compare the results obtained using the *RE* to those with Wikidata.

Moreover, while both *Kurz* and *Voll* models struggle with homonym disambiguation, *Voll* handles variation in surface forms more robustly than *Kurz*, especially for inflected mentions. This may indicate that, when available in the *Volltext*, contextual information is leveraged to disambiguate less straight-forward surface mentions.

We find that string-based top-1 Levenshtein is a competitive method for candidate generation on unseen entities. However, in rare cases it returns an unmanageably large number of candidates (see Section 4.3.2), which limits its direct usability. This can be mitigated by discarding Levenshtein suggestions that correspond to full names mapped to many RE_ids. With such filtering, its outputs can effectively supplement the BLINK bi-encoder's initial selection by adding Levenshtein matches that aren't already in the retrieved set. Finally, the findings are promising for a human-in-the-loop approach. The top 5 suggestions from BLINK have accuracies of 0.9 and 0.7 for seen and unseen entities, respectively. This, together with the first sugges-

[10] Caesar (8), Cato (3), Cinna (6), Claudius (6), Lucilius (2), and Maximus (7)

tion for unseen entities being right in 50% of cases, indicates that it could reduce the verification time.

Secondly, Figure 4 shows that many names correspond to only a small number of RE_IDs. This suggests that a system could reliably propose the correct match either by selecting the RE_ID with a very high surface-matching score when no homonyms are present, or by presenting the few possible candidates. Such an approach could serve as a useful aid for human verification.

7. Conclusions

In this paper, we explored the viability of a transformer-based BLINK model for cross-lingual, domain-specific NEL in a highly specialised humanities context.

The results showed that while BLINK scored competitively for data that was seen during training, generalising to unseen entities remains a challenge. The analysis highlights that issues related to KB structure and annotation choices limit performance. More broadly, the study shows that contextual richness in the KB and data quality and homogeneity are decisive factors in this domain. We also found that both *Kurz* and *Voll* models struggled to disambiguate individuals with the same name. However, the top 5 scores and the performance of the Levenshtein-based baseline, point toward the feasibility of human-in-the-loop workflows, where automated suggestions can reduce verification time without requiring full automation.

Future improvements could include incorporating Levenshtein retrieval. Additionally, we could experiment with adding real-world rule constraints such as the time-based constraints Graciotti et al. (2025a) added, to make sure no temporally unlikely entities are added. Next, KB enrichment could be helpful, for exemple exploiting the already existing interlinks between Wikidata and the RE to resolve ambiguity and add more textual data in the *Kurz* setting. Lastly, we could explore more advanced modeling strategies, such as hard negative mining (Gillick et al., 2019), to improve bi-encoder performance.

8. Funding statement

Part of this work has been carried out under 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. Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

9. Bibliographical References

Marijke Beersmans, Evelien De Graaf, Alek Keersmaekers, Mark Depauw, Tim Van De Cruys, and Margherita Fantoli. 2025. Automatic Named Entity Linking for Ancient Greek with a Domain-Specific Knowledge Base. *Anthology of Computers and the Humanities*, 3:540–556.

Emanuela Boros, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Ahmed Hamdi, José G. Moreno, Nicolas Sidère, and Antoine Doucet. 2020. Robust Named Entity Recognition and Linking on Historical Multilingual Documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, Paper 171. CEUR-WS Working Notes.

Yanne Broux and Mark Depauw. 2015. Developing onomastic gazetteers and prosopographies for the ancient world through named entity recognition and graph visualization: Some examples from trismegistos people. In *Proceedings of the International Conference on Social Informatics (SocInfo)*, volume 8852, pages 304–313, Barcelona. Springer.

Thibault Clérice. 2020. Corpus latin antiquité et antiquité tardive lemmatisé.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive Entity Retrieval. Version Number: 3.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290.

Evelien de Graaf. forthcoming. Challenges in disambiguating ancient individuals with Paulys Realencyclopädie. In *Semantic Annotation for the Ancient World 2025*.

Evelien de Graaf, Mark Depauw, and Margherita Fantoli. 2024. "Nescio Carneades iste qui fuerit": Evaluation of Knowledge Bases for Named Entity Linking for Latin Texts. In *Proceedings of The First Workshop on Data-driven Approaches to Ancient Languages*, pages 1–19. Language & Translation Technology Team (LT3).

Mark Depauw and Tom Gheldof. 2014. Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information. In *Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops*, Communications in Computer and Information Science, pages 40–52, Cham. Springer International Publishing.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2):1–47.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of clef hipe 2020: Named entity recognition and linking on historical newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310, Cham. Springer International Publishing.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Overview of hipe-2022: Named entity recognition and linking in multilingual historical documents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 423–446, Cham. Springer International Publishing.

Margherita Fantoli, Valeria Irene Boano, Evelien de Graaf, and Camillo Carlo Pellizzari di San Girolamo. 2026. Wikidata as a knowledge base for people of the greco-roman world. *Journal of Open Humanities Data*, 12(1).

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34. European Language Resources Association.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, page 528–537, Hong Kong, China. Association for Computational Linguistics.

Arianna Graciotti, Nicolas Lazzari, Valentina Presutti, and Rocco Tripodi. 2025a. Musical heritage historical entity linking. *Artificial Intelligence Review*, 58(5):140.

Arianna Graciotti, Leonardo Piano, Nicolas Lazzari, Enrico Daga, Rocco Tripodi, Valentina Presutti, and Livio Pompianu. 2025b. KE-MHISTO: Towards a multilingual historical knowledge extraction benchmark for addressing the long-tail problem. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20316–20339, Vienna, Austria. Association for Computational Linguistics.

Kai Labusch and Clemens Neudecker. 2020. Named entity disambiguation and linking historic newspaper ocr with bert. In *CLEF 2020 Working Notes*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.

Chiara Palladino, Margherita Fantoli, Evelien de Graaf, Monica Berti, Matteo Romanello, Tariq Yousef, Marijke Beersmans, Tom Gheldof, Laura Soffiantini, and Eleonora Litta Modignani Picozzi. 2024. Experience and challenges with named entities – workshop at DHBenelux 2024.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58(1).

Mikhail Plekhanov, Nora Kassner, Kashyap Popat, Louis Martin, Simone Merello, Borislav Kozlovskii, Frédéric A. Dreyer, and Nicola Cancedda. 2023. Multilingual End to End Entity Linking.

Cristian Santini, Marieke Van Erp, and Mehwish Alam. 2026. It's All About the Confidence: An Unsupervised Approach for Multilingual Historical Entity Linking using Large Language Models.

Ozge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3).

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27:443–460.

Georg Wissowa, Wilhelm Kroll, Karl Mittelhaus, Konrat Ziegler, and Hans Gärtner, editors. 1893–1980. *Paulys Realencyclopädie der classischen Altertumswissenschaft*. Metzler, Stuttgart. Multivolume encyclopaedia.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 6397–6407, Online. Association for Computational Linguistics.