Automatic Named Entity Linking for Ancient Greek with a Domain-Specific Knowledge Base

Marijke Beersmans[1] ⓘ, Evelien de Graaf[1] ⓘ, Alek Keersmaekers[1] ⓘ, Mark Depauw[1] ⓘ,
Tim Van de Cruys[1] ⓘ, and Margherita Fantoli[1] ⓘ

[1] Faculty of Arts, KU Leuven, Leuven, Belgium

Abstract

Named Entity Linking, or disambiguating named entities by linking them to a knowledge base, is an important Natural Language Processing task, especially in the humanities. In this paper, we examine the performance of the state-of-the-art entity linking model BLINK in connecting Ancient Greek person mentions to a domain-specific German-language knowledge base. To train the model, we create both gold-standard data through manual annotation and noisier silver data through automatic extraction. We then evaluate whether incorporating the latter improves performance. Our findings suggest that, overall, the results remain suboptimal for Ancient Greek. Increasing training data, even through automatic methods, shows promise. However, as it stands, using BLINK directly would be ill-suited for Named Entity Linking in the target setting. We discuss possible causes and suggest areas for improvement.

Keywords: Named Entity Linking, Ancient Greek, Persons, BLINK, Paulys Realencyclopädie

1 Introduction

Named Entity Linking (NEL), also known as Named Entity Disambiguation (NED), is a Natural Language Processing (NLP) task that involves automatically linking named entities, often people, locations, and organizations, to entries in a predefined knowledge base (KB), such as Wikipedia or Wikidata. This process disambiguates the entities, supporting various information retrieval tasks, including Question Answering [34]. However, NEL remains challenging due to its reliance on extra-textual contextual information and the scarcity of annotated training data, which often requires expert knowledge to produce [36].

In the humanities, particularly in digital classics, entity-based indexing is in high demand. Disambiguated entity information can enhance text accessibility for researchers, streamline historians' workflows, and support the otherwise time-intensive process of manual semantic annotation [14].

This paper explores the challenges and limitations of automating NEL for individuals from antiquity on a corpus of Ancient Greek literary texts, in conjunction with a domain-specific database. As such, our experiment presents three main challenges:

- Ancient Greek is rarely included in pre-trained language models used by modern NEL systems.

- There is no Ancient Greek Wikipedia, the typical KB target for such systems. The task is therefore inherently cross-lingual.

- We rely on a domain-specific KB whose digitization is incomplete, limiting coverage and consistency. Unlike generic KBs like Wikidata, there are also fewer associated resources available, such as an internal hyperlink structure or attributed metadata.

In addition, we examine the potential for integrating automatic NEL into a human-in-the-loop verification process. This work is part of a broader project aimed at performing large-scale quantitative analysis of person mentions in Ancient Greek and Latin texts via Social Network Analysis. We focus on two main research questions:

- How accurately can a retrieval-and-ranking approach based on transformer models (viz. BLINK [37]) perform NEL on Ancient Greek texts, and what are its limitations?

- Does the inclusion of more, yet noisier data created through rule-based means enhance performance compared to relying solely on a smaller, human-annotated dataset?

The paper is structured as follows: Section 2 reviews the related work in automated NEL and its limited application in classical studies. Section 3 details three important preliminaries for this research: the Trismegistos (TM) database, the GLAUx corpus, and the Knowledge Base that was created based on the Wikisource version of Paulys Realencyclopädie. Section 4 explains our experimental set-up. Section 5 details the data creation process. Section 6 describes the BLINK model. Sections 7, 8, and 9 present results, error analysis, and performance discussion, respectively. Finally, Section 10 summarizes our findings and outlines future work. The code and data will be published on GitHub.

2 Related Work

Wikification -the process of linking entities in a text to their corresponding Wikipedia entries- has received significant attention over the past two decades. Early experiments in the 2000s relied on hand-crafted features [6; 10], while later approaches adopted deep learning techniques that use vector representations of entity mentions and KB entries to identify the most likely candidate via similarity measures [38].

More recently, pre-trained transformer models have been integrated into this pipeline and finetuned to optimize the similarity between mentions and their corresponding KB entries [9; 39], as reviewed in [33]. BLINK is a prominent example of this class of models [37]. In parallel, generative approaches have emerged, where models generate the correct entity label as a string, restricted to valid entries in the KB, as exemplified by GENRE [7].

Our experiment constitutes a case of cross-lingual entity linking, as it involves linking mentions in Ancient Greek to a German-language KB. While most research on cross-lingual entity linking has focused on linking low-resource languages to the English Wikipedia, due to the limited size or absence of other language-specific Wikipedias, models often fail to generalize to texts outside of Wikipedia, which is typically their training domain [18]. Recent multilingual approaches aim to address this by making the task language-agnostic with respect to the KB [4; 35]. Multilingual adaptations of transformer-based NEL models have been developed for this purpose [11; 31].

Despite these advances, NEL has rarely been applied to historical data. When it has, the focus has primarily been on historical newspapers, as in the HIPE 2020 shared task, which included NER and NEL for 19th- and 20th-century newspapers in French, German, and English [15; 27]. More closely aligned with our work is the HIPE 2022 task, which also included 19th-century commentaries on classical texts in these three languages [16]. More recently, Graciotti et al. [20] applied entity linking to English historical documents on musical heritage, proposing the integration of knowledge-base-derived constraints to improve the historical plausibility of predictions.

Beyond the case of 19th-century commentaries, automated entity linking has not yet been applied to texts from antiquity. However, NEL's precursor task, Named Entity Recognition (NER), has been the subject of several studies in classical languages [8; 17; 30].[1] Some annotation tools, such as Recogito, offer automated NER for places and suggest candidate entities using authority files, primarily through surface-form matching. In most cases, however, NEL in the field of classics is still performed manually. In this paper, we apply a state-of-the-art NEL model to an Ancient Greek corpus for the first time. We also introduce the first dataset for Ancient Greek NEL created through an automated approach that leverages interlinked resources.

3 Preliminaries

The following section will detail three important resources we used to build and annotate our datasets and assess our Named Entity Linking pipeline: databases from the Trismegistos (TM, henceforth) project, the GLAUx corpus, and the knowledge base used to link the entities to: *Paulys Realencyclopedie der Classischen Altertumswissenschaft*.

3.1 TM database

TM is an expansive metadata platform that focuses on the Ancient World. Particularly relevant for this study are its TM Nam IDs and TM NamVar IDs, unique name and name variants identifiers based on onomastic gazetteers the TM team has gathered over the years. For an example of the TM information we incorporate, see Table 1.

3.2 The GLAUx corpus

The GLAUx corpus [24] contains Ancient Greek texts of various genres from the 8th century BCE to the 4th century CE (about 20 million tokens). All texts have been linguistically enriched with morphological tags, lemmas and dependency information largely automatically.

3.3 A domain-specific KB

At present, there is no dedicated KB for Ancient Greek and Latin for NEL of people [5]. There are several *separate* resources that collect people or names of entities, for example the Lexicon of Greek Personal Names (LGPN) [29], MANTO [22] for mythological people, Prosopographia Imperii Romani (PIR) [3] for persons alive during the Roman Empire, and Trismegistos (TM) [13] for a wide array of attestations of named entities. As all of these resources have specific topical, chronological, or geographic focuses, they are not suitable as a comprehensive KB for identifying *all individuals* mentioned in a large portion of Ancient Greek and Latin literature. Other annotation projects that have linked people in Ancient Greek and Latin texts have relied on project specific identifiers,[2] Wikipedia,[3] and Wikidata[4] for disambiguation.

No specific preference appears to be present in the community, but de Graaf, Depauw, and Fantoli [12] demonstrated that *Paulys Realencyclopädie der classischen Altertumswissenschaft* (*RE*) is suitable for NEL purposes.[5]

---

[1] This imbalance between NER and NEL datasets and models is common across many languages, as shown by Guellil et al. [21].

[2] E.g. the STEPBible identifiers.

[3] E.g. The Odyssey, featuring mythical people, annotated by Josh Kemp [25].

[4] E.g. Digital Harpocration.

[5] However, this study also stressed certain limitations of the *RE* as a KB, albeit affecting only a minority of cases. First and foremost, the *RE* is not fully exhaustive for people mentioned. In addition, several examples of duplication and conflation of entities in the *RE* can be found.

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

Entity	Abaskantos 3
Complete Name	Flavius Abascantus
Name Components	Flavius Abascantus
Flavius	TM Nam ID: 11846 TM NamVar ID: 30829
Abascantus	TM Nam ID: 8279 TM NamVar ID: 120213

Table 1: Complete name, Nam ID and NamVar ID for Abaskantos 3 as added to the *RE*

The *RE* (published between 1893 and 1978) is an expansive reference work on classical antiquity and it comprises around 100,000 entries contributed by leading classical philologists. A significant portion of this work has already been digitized through the German Wikisource project. As of now, the full work is available as scans, alongside access to the so-called *Volltext* (full text) of over 65,000 articles. Additionally, a searchable index or *Register* of keywords (*Stichwörter*) with brief summaries for all entries (*Kurztext*) is provided to facilitate navigation. Figure 1 illustrates the different components of the entry for the hero Abas (Abas 3).

In collaboration with the developers of the TM database, a local version of the *RE* was developed based on this Wikisource Register, which was integrated into the TM database and pre-processed semi-automatically.[6] The process focused primarily on identifying entries that describe individuals. For the purpose of linking, we employ a broad definition of a **person** as any identifiable single individual, including deities and anthropomorphic mythological figures. This decision was made to ensure interoperability of our annotation with NER systems [1; 2]. The currently available training data for NER in antiquity rarely, if ever, make a distinction between people and gods. Beside this, it would be conceptually challenging to distinguishing the two categories, for instance in cases such as Herakles and Jesus. After this filtering, our KB contains in total 54,180 entries manually identified as persons.

We are committed to maintaining our local instance in close alignment with the online version, ensuring consistency. In addition, we are constantly refining the structured attributes associated with person entities, such as relations to other entities or temporal information.

---

[6] A search tool facilitating access to the *RE* based on this work is available at `https://www.trismegistos.org/real`.
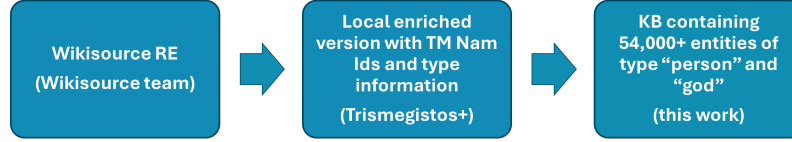
Figure 2: Overview of contributions to the *RE* Knowledge Base

4 Experimental set-up

Shen, Wang, and Han [34] defined NEL as follows: given a set of entities E in a predefined KB and a set of textual mentions M in a text corpus, the objective is to link each mention $m \in M$ to its corresponding entity $e \in E$. This definition assumes that the mention boundaries in the texts are known or defined beforehand, for example, through NER. In the present study, gold-standard mention boundaries were available to the entity linking model. The example below uses the *[Ms]* (mention start) and *[Me]* (mention end) tokens to indicate them. End-to-end entity linking, which also includes the step of identifying the entity mentions, is beyond the scope of this paper.

- **Example mention**: *...γὰρ ἡ Κύπρις πικρά [Ms] Ἀχαιός [Me] φησιν ἐν Αἴθωνι σατυρικῷ. παρ᾿ οὗ ὁ σοφὸς Εὐριπίδης λαβὼν ἔφη...*
  ('...since Cypris is a cruel goddess [...] Achaeus says in the satyric drama *Aethon*. From him the wise Euripides has borrowed the idea and has said...', Athenaeus *Deipnosophistae*, VI 270C, Transl. Charles Burton Gulick)

- **Example entity (Kurztext)**: Achaios 6, A. von Eretria, das grosse Dionysosfeste feierte, Tragiker im 5. Jh. v. Chr.
  (Achaios 6, A. from Eretria, who celebrated the Great Dionysia, tragedian in the 5th century B.C., Transl. our own)

As previously discussed, our aim is to specifically link individuals from antiquity. Our entity set E concerns the filtered people from the Wikisource version of the German-language standard work *Paulys Realencyclopädie der classischen Altertumswissenschaft* (*RE*, henceforth). This KB is comprehensive and crafted by experts, but it lacks certain features typically found in more widely used, digital born resources, such as Wikipedia's internal hyperlink structure. Since, mainly for reasons of copyright, not all the entries in the *RE* have been fully digitized, we trained different models using two different entity sets: one where we only use a short summarizing description, usually between 5 and 10 words, which is available for nearly every entity (the *Kurztext*) and one where we include the full entry text (*Volltext*) where available and the *Kurztext* otherwise (see Section 3.3 for more information). We will refer to them as the *Kurztext* model and the *Volltext* model below.

   We use two different ways of creating data for this setup, exploiting both manual annotation (see Section 5.1) and existing linked identifiers present in the GLAUx corpus to create linked mentions (see Section 5.2). However, even using both strategies, we are still operating in a low-data setting (10,000 mentions or less).

5 Data

Given the KB's specificity, no annotated data were available for this task. Because it is time-consuming and resource-intensive, we only created a limited number of human-annotated ("gold") data for this study. We also explored ways to supplement it using automatic, rule-based annotation ("silver" data).

Inherently, the silver data were of lower quality than that of a human-annotated gold standard. As such, we decided to compare two training scenarios. In the first, we used only the gold data. In the second, both the gold and the silver data were including at training time. This allowed us to assess whether more, but slightly noisier data would improve the model's performance. In this section, we discuss the data creation process and the final dataset.

5.1 Gold data

The gold data was developed as part of a case study focusing on people co-occurring with the philosopher Plato. The texts included in the gold corpus are listed in Table 2. In these texts, all person mentions were manually linked to their corresponding entries in the *RE* by two domain experts, Evelien de Graaf and Herbert Verreth. This annotation process initially yielded 17,415 person mentions, corresponding to 2,749 unique individuals. However, the final gold dataset includes significantly fewer mentions and individuals due to several exclusion criteria. First, mentions linked to individuals not recorded in the *RE* (i.e., unlinkable or NIL mentions) were excluded. These account for 994 NIL entities and 3,686 mentions. An example of such a NIL mention is *Βακχεῖος ὁ τὴν ἐπιστολὴν φέρων* ('Baccheius, who carries this letter', Plato, *Epistulae,* 390c, Transl. our own). This individual has no entry in the RE.[7]

Second, we removed annotations that were marked as groups (e.g. 'the platonists') or as entities named after people or containing personal names (e.g. the Aeneid). While they do in fact contain a reference to an individual (Plato, Aeneas), the primary target of the mention is not the person but the group or work, and this might have a negative effect on the model's performance.

Third, we excluded nearly 5,000 mentions simply because they hadn't yet been processed as multi-token entities (e.g., "Iulius Caesar" was still annotated as two separate mentions). This was a practical decision based on the current state of processing, not due to selective filtering. Our goal is to consistently annotate all multi-token mentions, and these will be included as such in the final version.

After applying these filters, the resulting dataset comprises 8,440 mentions and 1,230 unique individuals. A breakdown of these counts by work is provided in Table 2.

---

[7] A full paper by E. de Graaf on the annotation process is forthcoming. Complexity for this task is caused primarily by historical uncertainty, ie. it is not always easy to identify an individual in the text. In cases like these, we have always followed the identification as made by the authors of the *RE* to assure consistency. An example of this is the individual Echecrates mentioned in Plato *Epistulae,* 9.358b, where scholars have argued that the description as νεανίσκον, a youth, makes identification with Echecrates from Phleius impossible [23; 28]. However, the editors of the *RE* entry Echekrates 3 do consider the identification probable (Wellmann:"schwerlich ein anderer"/ Oldfather: "wohl richtig"). Although the identification proposed by Nails and Helfer is both more recent and arguably more plausible, we retain the earlier identification with Echekrates 3 found in the *RE* to maintain consistency. The vast majority of cases of uncertainty, however, belongs to the NIL category and will therefore not be of any concern for this paper.

Author	Work	Mentions
Plato	*Epistulae*	322
Aristoteles	*Metaphysica*	224
Strabo	*Geographica*	430
Dionysius Halicarnassensis	*De Demosthenis dictione*	270
Plutarchus	*Quaestiones convivales*	1,138
Claudius Aelianus	*Varia historia*	907
Clemens Alexandrinus	*Stromata*	1,205
Origenes	*Contra Celsum*	1,723
Diogenes Laertius	*Vitae philosophorum*	2,221
	Total mentions:	8,440
	of which unique entities:	1,230

Table 2: Gold data "Plato case study": manually annotated person mentions for a case study on Plato.

5.2 Silver data

The following section details the silver (automatically linked) data, which were included as training data in the final experiments to ascertain whether more, yet noisier data would improve model performance. The silver data were created following the pipeline presented in Figure 3.

First, cited works from the *Volltexts* of *RE* entries were retrieved automatically, to then be expanded and identified using experimental algorithms created by the TM team[8]. If the relevant text was available in the GLAUx corpus, the expanded form could be linked to it. When source metadata concerning the specific text passage was provided, meaning information about which book, chapter, paragraph etc. the *RE* entry is said to appear in, it was included and used to match to the correct passage in GLAUx.

Every entry in the *RE* was linked to the full name of the person using the TM Nam IDs, as described in Section 3.3,[9] which are in turn linked to the lemma ID of the GLAUx corpus. Because of this, it was possible to retrieve the words in the passage of the GLAUx corpus which corresponded to the mention(s) of the name of the person described in the *RE* entry. If there were multiple mentions in a single passage, all were regarded as separate mentions of the entity, which is why there are more mentions (5,290) than passages found (4,961) in Table 3.

The approach made it possible to go from the entry of the *RE* discussing the Ancient Greek tragedian Achaios (Achaios 6), who, based on the *RE Volltext* entry, is mentioned in the sixth book, section 270C, of Athenaeus *Deipnosophistae*, to the corresponding passage in the GLAUx corpus; we could then annotate the token Ἀχαιός (Achaiòs) as a mention of Achaios 6 of the *RE*.[10]

Completing the full pipeline was nevertheless only possible in a minority of cases, due to errors in the retrieval phase and lack of standardization between the divisions of the texts used in the *RE* and in the GLAUx corpus: this significantly reduced the final number of silver annotations. A short error analysis of 50 randomly selected entity mentions from the final set revealed that this approach establishes correct links from a Greek text to the KB in 41 out of 50 cases. Mistakes occurred most

---

[8] The rule-based algorithms are part of the internal workflow of the TM database and are not currently published. An article by M. Depauw, B. Thijs, F. Pietowski and L. Benrais is in preparation. We were able to use preliminary version through an API.

[9] The database TM Names collects all names attested for the Ancient Graeco-Roman world and links them to all the attested variants. Because of this, even if a text registers a different variant of the name than the one chosen by the *RE*, the matching is still possible. See `https://www.trismegistos.org/ref/about_naw.php` for more information.

[10] We excluded the cases in which the specific passage only gave information about the book in which the person was mentioned, because it is a very broad unit that could potentially introduce errors in the pipeline.

often due to mentions of entities with the same name appearing in the same passage. Table 3 describes the final numbers for the silver data creation.
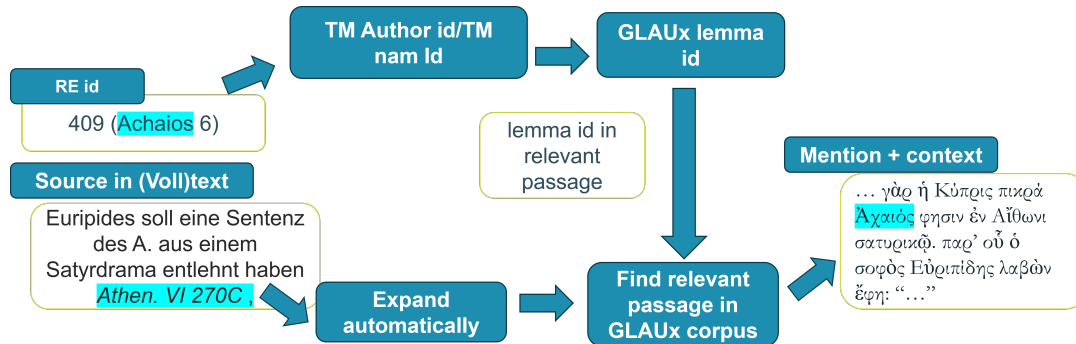

Figure 3: Schematic of the silver (automatically linked) data creation pipeline.

	# of (potential) links
expanded sources from the *RE* entries	68,241
Of which in GLAUx corpus	20,706
Of which nam ID and lemma ID could be linked	13,880
Of which passage found (*books* excluded)	4,961
Total mentions	5,290
Of which unique entities	2,104

Table 3: Overview of the different components of the automated pipeline for data creation and the number of results they yielded.

5.3 Final dataset

The models were trained according to two different data scenarios; one where only the gold data was used and one where the silver data was additionally included. Since some entities appeared very frequently in the gold dataset, due to it being created for a narrowly focused case study, we decided to cap their occurrences at 50 to prevent them from skewing the overall distribution.[11] Finally, we split the mentions into train, validation, and test sets, ensuring that there was no overlap in entities between them. This way, the model would have to link mentions exclusively to unseen entities during test time, and we could fully assess the generalizing capabilities of the model. Final mention counts for each set can be found in Table 4.

	train		**validation**		**test**	
	gold + silver	gold	gold + silver	gold	gold + silver	gold
Mention count	9,066	4,851	1,143	597	1,010	567
Of which unique entities	2,552	984	319	123	319	123

Table 4: Number of mentions in the train, validation and test set for both data quality scenario's

---

[11] Hence the different counts in this part compared to those reported in Section 5.1.

6   Model: BLINK

In this section, we introduce the model we used for this task, its training procedure, and hyperparameters.

Given the reliance on a domain-specific KB with only textual descriptions of the entities (i.e. no structured data like, for instance, Wikidata), we chose to test the BLINK system [37]. BLINK (entity linking with BERT) is a conceptually simple entity linking approach that nonetheless performs well compared to other, more complex and/or resource-intensive systems [33]. The model is transformer-based and makes no assumptions about the nature of the KB, except that a description of each entity should be present.

BLINK uses a two-stage entity linking approach. In the first stage, both the mentions and the entity descriptions in the KB are encoded using two separate BERT encoders, commonly referred to as a bi-encoder or dual encoder strategy. The mention encoding is then compared to all the entity encodings using dot product similarity. Subsequently, the top k most similar entities are retrieved for the second step. At training time, the bi-encoder is optimized to maximize the similarity of the mention and the correct entity compared to the mention and in-batch negatives.

During the second stage, the mention is concatenated to each of the top k most similar entities. Those concatenations are then jointly encoded (cross-encoder). They are subsequently fed through an added linear layer, which outputs a score for each mention-entity combination, making the final decision. The second step is slower and more resource-intensive, but it produces more accurate results by allowing the system to attend jointly between the mention and the entity.

Bi-encoder performance is evaluated using *recall@topk* (whether or not the correct entity is present in the retrieved set). For all our experiments, the top k was fixed to 64, following the original paper. Other hyperparameters can be found in Table 5. We follow the hyperparameters used by Wu et al. [37] in their zero-shot setting, not because we operate in a zero-shot setting ourselves, but because their setup involved training a transformer model comparable in size to ours. In the *Volltext* scenario, it was necessary to extend the bi-encoder training to ten epochs. Training for only five epochs resulted in suboptimal model convergence.

BLINK was originally tested in an English-only setting, although it needs to be mentioned that a multilingual implementation of the bi-encoder exists [31]. The full system can be adapted to a multilingual setting using multilingual transformers.

We used a multilingual base transformer instead of BERT: UGARIT-grc alignment [40] (below UGARIT). This is an XLM-RoBERTa model that contains German in its pre-training data and was then further trained using various tasks on Ancient Greek texts. While other multilingual encoder-only transformers that contain Ancient Greek exist (e.g. PhilBerta [32]), this is the only one that has seen both Ancient Greek and German during its training stages. To improve representations, we added another stage of MLM training on the *RE* texts and on the GLAUx corpus, similar to Logeswaran et al. [26], before finetuning the models for NEL.

6.1   Baseline

The bi-encoder approach was tested against the following baseline: its recall@top64 was compared to a method that backlinked the mention's lemma ID to its Nam ID, retrieving all associated *RE* entities. In other words, the baseline tries to retrieve all entities in the RE that match the target name in the text, allowing us to assess whether the bi-encoder could outperform a simple rule-based match. Linking via the TM Nam ID instead of via the surface form, allowed us to bypass the difference in scripts between Ancient Greek and German (TM Nam IDs are linked to surface forms both in latinized and Ancient Greek script). In cases where the resulting list contained more than 64 entities, a random subset of 64 was selected. For multi-token mentions, all lemmas and Nam IDs of the individual tokens were considered.

Bi-encoder hyperparameters	
context length	128
epochs	5 (10)
learning_rate	2e-5
batch size	32
top k	64
Cross-encoder hyperparameters	
context length	256
epochs	2
learning rate	1e-5
batch size	1

Table 5: Bi-encoder and cross-encoder hyperparameters of the BLINK model

7   Results

		gold test set (567 entities)		
		rc@64	cross acc	overall acc
gold	Kurz (5 epochs)	51 %	45 %	23 %
	Voll (10 epochs)	58 %	30 %	17 %
gold + silver	Kurz (5 epochs)	71 %	50 %	35 %
	Voll (10 epochs)	64 %	35 %	23 %
baseline		56 %	N/A	N/A

Table 6: Results of the BLINK model (bi-encoder recall@64, cross-encoder accuracy and total accuracy) in the *Kurztext*/*Volltext* scenarios on the gold-standard test set

Table 6 shows the results of the BLINK models for the gold test set. For completeness sake, the results on the test set including the silver test data are reported in Appendix A. We do not base our analysis on them because the silver test data was automatically created and therefore not entirely reliable for testing purposes, as discussed in Section 5. For the bi-encoder, we report the recall when retrieving the top 64 entities (rc@64), for the cross-encoder we report the accuracy on the entities it received from the bi-encoder. Finally, we report the overall accuracy.

The results are significantly worse those than benchmarked performance on unseen mentions for English data (76.8 % total accuracy, as reported in Wu et al. [37]). Some performance drops are expected given the lower-data scenario and the multilingual setting. However, even in the best performing setting the model reaches only 35.44% overall accuracy. Adding the lower-quality silver data does improve performance considerably, especially bi-encoder recall. Surprisingly, the *Kurztext*-only model outperforms the *Volltext*-where-available model in all data scenarios, both in bi-encoder recall (+7% points) and cross-encoder accuracy (+15% points), which now outperforms the baseline more convincingly.[12] This could indicate that the model fails to utilize the additional contextual information effectively. The potential reasons for this are discussed in Section 8.

---

[12] Although the poor performance of the baseline can partially be explained by the fact that the TM Nam IDs have not been added for the gods in the *RE*.

8 Error Analysis

In this section, we explore the predictions of the BLINK models in more detail, focusing exclusively on the models trained on the full dataset as they demonstrated superior performance. We discuss general trends and evaluate the results for a human-in-the-loop workflow.

Firstly, since most *RE* entries are composed of a base name and a numerical identifier (The entry for Augustus, for example, is Iulius 132), we can utilize the number of entries sharing a base name (below *RE*-basenames) as a proxy of the ubiquity of the name and, by extension, how challenging the disambiguation task is for a given entity.

Building on this, we can first observe that both when using the *Kurztext* and the *Volltext*-where-available model, the correctly predicted entities are the less ambiguous and less challenging ones. The average number of entities with the same *RE*-basename in the gold test dataset is 15.84, while for the correctly predicted entities it is 4.64 and 7.82 for the *Kurztext* and the *Volltext* models, respectively. Within a human-in-the-loop framework, this insight can be leveraged to prioritize mentions for manual verification, specifically those associated with entities that exhibit a high number of homonyms. However, this does not imply that mentions linked to entities with fewer homonyms can be considered reliably correct. Given the overall low accuracy of the models, it remains uncertain whether even the surface name of the matched entity is accurate.

Additionally, 171 mentions in the test set link to entities that only have a single *RE*-basename. It is mainly on this subset of unambiguous names that the *Kurztext*-only model outperforms the *Volltext* model. If these mentions are not taken into consideration, the performance gap is reduced, as seen in Table 7. This could indicate that the *Kurztext* model's main strength is being a better surface matcher between the Latinized *RE*-entries and the Ancient Greek mentions, and that the extra contextual information detracts from this, perhaps because it introduces noise or patterns that the model cannot consistently leverage in this lower-resource multilingual setting.

	1 *RE*-basename	>1 *RE*-basename
Acc Voll	33 %	18 %
Acc Kurz	64 %	23 %
Support	171	396

Table 7: BLINK accuracy comparison across *RE*-basename subsets

Second, we examine whether the model's confidence (certainty), as indicated by the cross-encoder output score, correlates with prediction accuracy. This analysis is motivated by potential human-in-the-loop applications, where high-confidence predictions could be accepted automatically while low-confidence ones are flagged for manual review. Figure 4 and 5, show the prediction accuracy (blue line) at different confidence thresholds, together with the number of mentions where the prediction has a confidence of this threshold or higher (red line).

From this, it is apparent that while the accuracy improves the more confident the model is (at least for the *Kurztext* model), high confidence does not guarantee a correct prediction. For the *Volltext* model, the increase in accuracy is minimal. As for *Kurztext* model, the accuracy remains at 50 % even if we limit ourselves to mentions where the model's prediction has a score of 0.99 or higher. These findings suggest that, in its current form, BLINK may not be directly suitable for human-in-the-loop entity linking in this setting.

9 Discussion

The results presented above provide a clearer understanding of the challenges involved in applying a state-of-the-art NLP model to domain-specific Ancient Greek entity linking. Several factors likely contribute to the observed suboptimal performance.
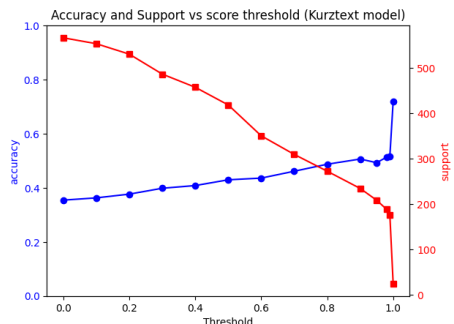
Figure 4: Accuracy and remaining number of mentions across various score thresholds, BLINK Kurztext model
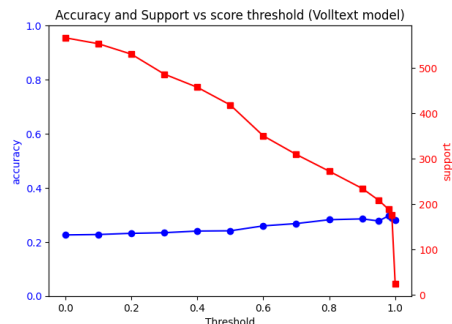

Figure 5: Accuracy and remaining number of mentions across various score thresholds, BLINK Volltext model

One limitation is the relatively small size of the dataset, particularly in terms of the number of unique entities, which may restrict the model's ability to generalize, especially if all entities are to remain unseen at test time. The inclusion of additional data, even though it is noisy, appeared to improve performance, suggesting that data sparsity is a bottleneck.

Another issue is the lack of surface similarity between mentions and candidate entities, largely due to differences in language and script, since the *RE* contains a Latinized version of the name and the mention is in Ancient Greek. Since surface similarity appears to be important for disambiguation, it might be worth looking into ways of enhancing it, such as transliteration or, in this specific case, including (a) Latinized version(s) based on the TM Namvar if the associated lemma ID of the mention in GLAUx has one.

Finally, many mentions in the corpus are inherently ambiguous. The *RE* frequently includes individuals who share the same name, family lineage, and professional role, such as political or military titles, and who are often discussed in close textual proximity. This may confuse a model that relies on contextual disambiguation. For instance, the *RE* entries *Aratos 2* and *Aratos 3* refer to a father and a son who not only share their name but also hold the same office in the Acheaen league. Given a mention such as οἱ μὲν οὖν Ἀχαιοὶ συνεληλυθότες εἰς Αἴγιον ἐκεῖ τὸν Ἄρατον ἐκάλουν., (Accordingly, the Achaeans came together at Aegium and invited Aratus thither, Plutarch *Aratus*, 42.1, Transl. Bernadotte Perrin) the model may struggle to distinguish between these two entities, because this context does not provide sufficiently distinctive features. This suggests that richer forms of context, such as temporal metadata, may be necessary to resolve these ambiguities reliably.

10   Conclusion and Future Work

In this paper, we explored the use of the transformer-based retrieval-and-ranking model BLINK to link Ancient Greek named entities to a German-language KB. Our focus was on data creation strategies, including both manual annotation and automatic generation, and we investigated whether the inclusion of noisier, automatically generated data could improve model performance. We found that BLINK's performance was insufficient for reliable entity linking in the target domain.

Future work could explore several directions to address this. First, incorporating domain-specific constraints, such as historical consistency checks, as was done by Graciotti et al. [20], may help guide the model toward more plausible entity candidates.

Second, we could address the atypicality of the language combination (Greek-German) either by automatically translating the *RE* to English or by using language distillation techniques to aug-

ment transformers that contain more specialized knowledge of Ancient Greek (such as PhilBerta) with knowledge of German [19].

Third, we are also experimenting with the generative approach mGENRE [11], testing various hyperparameter combinations and exploring strategies to enhance its performance; however, it currently struggles to converge well in this domain.

Finally, recent techniques in contextual augmentation using large language models, as in Vollmers et al. [36] may help improve candidate generation and disambiguation in low-resource settings such as Ancient Greek.

References

[1] Beersmans, Marijke, Graaf, Evelien de, Van de Cruys, Tim, and Fantoli, Margherita. "Training and Evaluation of Named Entity Recognition Models for Classical Latin". In: *Proceedings of the Ancient Language Processing Workshop*, ed. by Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sept. 2023, pp. 1–12. URL: `https://aclanthology.org/2023.alp-1.1/`.

[2] Beersmans, Marijke, Keersmaekers, Alek, Graaf, Evelien de, Van de Cruys, Tim, Depauw, Mark, and Fantoli, Margherita. ""Gotta catch 'em all!": Retrieving people in Ancient Greek texts combining transformer models and domain knowledge". In: *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, ed. by John Pavlopoulos, Thea Sommerschield, Yannis Assael, Shai Gordin, Kyunghyun Cho, Marco Passarotti, Rachele Sprugnoli, Yudong Liu, Bin Li, and Adam Anderson. Hybrid in Bangkok, Thailand and online: Association for Computational Linguistics, Aug. 2024, pp. 152–164. DOI: `10.18653/v1/2024.ml4al-1.16`.

[3] Berlin-Brandenburg Academy of Sciences and Humanities. "Prosopographia Imperii Romani (PIR)". `https://pir.bbaw.de/`. Accessed: 2025-07-16. 2021.

[4] Botha, Jan A., Shan, Zifei, and Gillick, Daniel. "Entity Linking in 100 Languages". en. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 7833–7845. DOI: `10.18653/v1/2020.emnlp-main.630`. URL: `https://www.aclweb.org/anthology/2020.emnlp-main.630` (visited on 07/03/2025).

[5] Broux, Yanne. "Cooperation of Databases". en. In: *Digital Classics Online Bd. 10* (May 2025), 10–36 Seiten. DOI: `10.11588/DCO.2024.10.103974`. (Visited on 06/27/2025).

[6] Bunescu, Razvan and Paşca, Marius. "Using Encyclopedic Knowledge for Named entity Disambiguation". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*, ed. by Diana McCarthy and Shuly Wintner. Trento, Italy: Association for Computational Linguistics, Apr. 2006, pp. 9–16. URL: `https://aclanthology.org/E06-1002/`.

[7] Cao, Nicola De, Izacard, Gautier, Riedel, Sebastian, and Petroni, Fabio. "Autoregressive Entity Retrieval". In: *International Conference on Learning Representations*. 2021. URL: `https://openreview.net/forum?id=5k8F6UU39V`.

[8] Chastang, Pierre, Aguilar, Sergio Torres, and Tannier, Xavier. "A Named Entity Recognition Model for Medieval Latin Charters". In: *Digital Humanities Quarterly 15*, no. 4 (2021).

[9] Chen, Shuang, Wang, Jinpeng, Jiang, Feng, and Lin, Chin-Yew. "Improving Entity Linking by Modeling Latent Entity Type Information". In: *Proceedings of the AAAI Conference on Artificial Intelligence 34*, no. 05 (Apr. 2020), pp. 7529–7537. ISSN: 2374-3468, 2159-5399. DOI: `10.1609/aaai.v34i05.6251`. URL: `https://ojs.aaai.org/index.php/AAAI/article/view/6251`.

[10]   Cucerzan, Silviu. "Large-Scale Named Entity Disambiguation Based on Wikipedia Data". In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, ed. by Jason Eisner. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 708–716. URL: `https://aclanthology.org/D07-1074/`.

[11]   De Cao, Nicola, Wu, Ledell, Popat, Kashyap, Artetxe, Mikel, Goyal, Naman, Plekhanov, Mikhail, Zettlemoyer, Luke, Cancedda, Nicola, Riedel, Sebastian, and Petroni, Fabio. "Multilingual Autoregressive Entity Linking". In: *Transactions of the Association for Computational Linguistics 10* (2022), ed. by Brian Roark and Ani Nenkova. Place: Cambridge, MA Publisher: MIT Press, pp. 274–290. DOI: `10.1162/tacl_a_00460`. URL: `https://aclanthology.org/2022.tacl-1.16/`.

[12]   de Graaf, Evelien, Depauw, Mark, and Fantoli, Margherita. ""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*, ed. by Colin Swaelens, Maxime Deforche, Ilse De Vos, and Els Lefever. Ghent, Belgium: Language & Translation Technology Team (LT3), 2024, pp. 1–19. ISBN: 978-90-78848-12-7.

[13]   Depauw, Mark and Gheldof, Tom. "Trismegistos: An Interdisciplinary Platform for Ancient World Texts and Related Information". en. In: *Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops*, ed. by Łukasz Bolikowski, Vittore Casarosa, Paula Goodale, Nikos Houssos, Paolo Manghi, and Jochen Schirrwagen. Communications in Computer and Information Science. Cham: Springer International Publishing, 2014, pp. 40–52. ISBN: 978-3-319-08425-1. DOI: `10.1007/978-3-319-08425-1_5`.

[14]   Ehrmann, Maud, Hamdi, Ahmed, Pontes, Elvys Linhares, Romanello, Matteo, and Doucet, Antoine. "Named Entity Recognition and Classification in Historical Documents: A Survey". In: *ACM Comput. Surv. 56*, no. 2 (2023), pp. 1–47. DOI: `10.1145/3604931`.

[15]   Ehrmann, Maud, Romanello, Matteo, Flückiger, Alex, and Clematide, Simon. "Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers". en. In: *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, ed. by Linda Cappellato, Carsten Eickhoff, Nicola Ferro, and Aurélie Névéol. Vol. 2696. CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR, Sept. 2020. URL: `https://ceur-ws.org/Vol-2696/#paper_255`.

[16]   Ehrmann, Maud, Romanello, Matteo, Najem-Meyer, Sven, Doucet, Antoine, and Clematide, Simon. "Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, ed. by Alberto Barrón-Cedeño, Giovanni Da San Martino, Mirko Degli Esposti, Fabrizio Sebastiani, Craig Macdonald, Gabriella Pasi, Allan Hanbury, Martin Potthast, Guglielmo Faggioli, and Nicola Ferro. Vol. 13390. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 423–446. ISBN: 978-3-031-13642-9. DOI: `10.1007/978-3-031-13643-6_26`. URL: `https://link.springer.com/10.1007/978-3-031-13643-6_26`.

[17]   Erdmann, Alexander, Brown, Christopher, Joseph, Brian, Janse, Mark, Ajaka, Petra, and Elsner, Micha. "Challenges and Solutions for Latin Named Entity Recognition". en. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. 2016, pp. 85–93.

[18] Fu, Xingyu, Shi, Weijia, Yu, Xiaodong, Zhao, Zian, and Roth, Dan. "Design Challenges in Low-resource Cross-lingual Entity Linking". en. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 6418–6432. DOI: `10.18653/v1/2020.emnlp-main.521`. URL: `https://www.aclweb.org/anthology/2020.emnlp-main.521` (visited on 07/03/2025).

[19] Gao, Tianyu, Yao, Xingcheng, and Chen, Danqi. "SimCSE: Simple Contrastive Learning of Sentence Embeddings". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. DOI: `10.18653/v1/2021.emnlp-main.552`.

[20] Graciotti, Arianna, Lazzari, Nicolas, Presutti, Valentina, and Tripodi, Rocco. "Musical heritage historical entity linking". en. In: *Artificial Intelligence Review 58,* no. 5 (Feb. 2025), p. 140. ISSN: 1573-7462. DOI: `10.1007/s10462-024-11102-9`.

[21] Guellil, Imane, Garcia-Dominguez, Antonio, Lewis, Peter R., Hussain, Shakeel, and Smith, Geoffrey. "Entity linking for English and other languages: a survey". en. In: *Knowledge and Information Systems* (Apr. 2024). ISSN: 0219-3116. DOI: `10.1007/s10115-023-02059-2`. URL: `https://doi.org/10.1007/s10115-023-02059-2` (visited on 05/09/2024).

[22] Hawes, Greta and Smith, Scott. "MANTO". `https://www.manto-myth.org/manto/`. Accessed: 2025-07-16. 2020.

[23] Helfer, Ariel. *Plato's "Letters": The Political Challenges of the Philosophic Life*. Ithaca: Cornell University Press, 2023.

[24] Keersmaekers, Alek. "The GLAUx corpus: methodological issues in designing a long-term, diverse, multi-layered corpus of Ancient Greek". In: *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 39–50. DOI: `10.18653/v1/2021.lchange-1.6`.

[25] Kemp, Josh. "Beyond Translation: Building Better Greek Scholars". Pelagios, 2021. URL: `https://medium.com/pelagios/beyond-translation-building-better-greek-scholars-561ab331a1bc` (visited on 02/22/2024).

[26] Logeswaran, Lajanugen, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, Devlin, Jacob, and Lee, Honglak. "Zero-Shot Entity Linking by Reading Entity Descriptions". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3449–3460. DOI: `10.18653/v1/P19-1335`.

[27] Menzel, Sina, Schnaitter, Hannes, Zinck, Josefine, Petras, Vivien, Neudecker, Clemens, Labusch, Kai, Leitner, Elena, and Rehm, Georg. "Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten". In: *Qualität in der Inhaltserschließung,* ed. by Michael Franke-Maier, Anna Kasprzik, Andreas Ledl, and Hans Schürmann. De Gruyter, Sept. 2021, pp. 229–258. ISBN: 978-3-11-069159-7. DOI: `10.1515/9783110691597-012`. URL: `https://www.degruyter.com/document/doi/10.1515/9783110691597-012/html` (visited on 07/03/2025).

[28] Nails, Debra. *The people of Plato: a prosopography of Plato and other Socratics*. Indianapolis: Hackett, 2002. ISBN: 978-0-87220-564-2.

[29] Oxford University. "Lexicon of Greek Personal Names". `https://www.lgpn.ox.ac.uk/`. Accessed: 2025-07-16. 1996.

[30] Palladino, Chiara and Yousef, Tariq. "Development of Robust NER Models and Named Entity Tagsets for Ancient Greek". In: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, ed. by Rachele Sprugnoli and Marco Passarotti. Torino, Italia: ELRA and ICCL, May 2024, pp. 89–97. URL: `https://aclanthology.org/2024.lt4hala-1.11`.

[31] Plekhanov, Mikhail, Kassner, Nora, Popat, Kashyap, Martin, Louis, Merello, Simone, Kozlovskii, Borislav, Dreyer, Frédéric A., and Cancedda, Nicola. "Multilingual End to End Entity Linking". In: , no. arXiv:2306.08896 (June 2023). arXiv:2306.08896 [cs]. DOI: `10.48550/arXiv.2306.08896`. URL: `http://arxiv.org/abs/2306.08896`.

[32] Riemenschneider, Frederick and Frank, Anette. "Exploring Large Language Models for Classical Philology". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 15181–15199. DOI: `10.18653/v1/2023.acl-long.846`. URL: `https://aclanthology.org/2023.acl-long.846`.

[33] Sevgili, Özge, Shelmanov, Artem, Arkhipov, Mikhail, Panchenko, Alexander, and Biemann, Chris. "Neural entity linking: A survey of models based on deep learning". EN. In: *Semantic Web 13*, no. 3 (Apr. 2022), pp. 527–570. ISSN: 1570-0844. DOI: `10.3233/SW-222986`.

[34] Shen, Wei, Wang, Jianyong, and Han, Jiawei. "Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions". In: *IEEE Transactions on Knowledge and Data Engineering 27*, no. 2 (Feb. 2015), pp. 443–460. ISSN: 1558-2191. DOI: `10.1109/TKDE.2014.2327028`.

[35] Tsai, Chen-Tse, Upadhyay, Shyam, and Roth, Dan. *Multilingual Entity Linking*. en. Synthesis Lectures on Human Language Technologies. Cham: Springer Nature Switzerland, 2024. DOI: `10.1007/978-3-031-74901-8`. (Visited on 07/03/2025).

[36] Vollmers, Daniel, Zahera, Hamada, Moussallem, Diego, and Ngonga Ngomo, Axel-Cyrille. "Contextual Augmentation for Entity Linking using Large Language Models". In: *Proceedings of the 31st International Conference on Computational Linguistics*, ed. by Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 8535–8545. URL: `https://aclanthology.org/2025.coling-main.570/`.

[37] Wu, Ledell, Petroni, Fabio, Josifoski, Martin, Riedel, Sebastian, and Zettlemoyer, Luke. "Scalable Zero-shot Entity Linking with Dense Entity Retrieval". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 6397–6407. DOI: `10.18653/v1/2020.emnlp-main.519`.

[38] Yamada, Ikuya, Shindo, Hiroyuki, Takeda, Hideaki, and Takefuji, Yoshiyasu. "Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, ed. by Stefan Riezler and Yoav Goldberg. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 250–259. DOI: `10.18653/v1/K16-1025`.

[39] Yamada, Ikuya, Washio, Koki, Shindo, Hiroyuki, and Matsumoto, Yuji. "Global Entity Disambiguation with BERT". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3264–3271. DOI: 10.18653/v1/2022.naacl-main.238. URL: https://aclanthology.org/2022.naacl-main.238/.

[40] Yousef, Tariq, Palladino, Chiara, Shamsian, Farnoosh, Ferreira, Anise d'Orange, and Ferreira dos Reis, Michel. "An automatic model and Gold Standard for translation alignment of Ancient Greek". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 5894–5905. URL: https://aclanthology.org/2022.lrec-1.634.

A   Appendix A: results on the silver + gold test sets

		silver + gold test set		
		rc@64	cross acc	overall acc
gold	Kurz (5 epochs)	47 %	32 %	16 %
	Voll (10 epochs)	57 %	21 %	12 %
gold + silver	Kurz (5 epochs)	71 %	43 %	31 %
	Voll (10 epochs)	68 %	38 %	26 %
baseline		68 %	N/A	N/A

Table 8: Results of the BLINK model (bi-encoder recall@64, cross-encoder accuracy and total accuracy) in the different Kurztext/Volltext scenarios, silver test set added