Linguistic Knowledge Graphs for Sense Prediction: a case-study on Latin

Eleonora Ghizzota*, Paola Marongiu†, Pierpaolo Basile*, Stefano Ferilli*, Barbara McGillivray§
*University of Bari Aldo Moro, via Orabona 4, Bari
e.ghizzota@phd.uniba.it, {pierpaolo.basile, stefano.ferilli}@uniba.it
†Istituto di Linguistica Computazionale "A. Zampolli", CNR-ILC, Via Giuseppe Moruzzi 1, Pisa
paola.marongiu@ilc.cnr.it
§King's College London, Strand Campus, Strand, London, WC2R 2LS
barbara.mcgillivray@kcl.ac.uk

Abstract

This paper investigates the integration of the Linguistic Knowledge Graph (LKG) and Large Language Models (LLMs) for word sense prediction in Latin, a morphologically rich and low-resource historical language. Building on recent work in word sense disambiguation (WSD) and semantic change detection, we use a LKG that integrates information from a diachronic Latin corpus, a sense-annotated dataset of Latin, Latin WordNet, and Wikidata, as a structured representation of semantic and contextual relations. We present sense prediction as a binary classification task over the Latin dataset, using a Graph Retrieval-Augmented Generation approach that combines knowledge graph retrieval with LLM prompting. Two types of graph metadata are tested: author-related information (work, period, occupation) and linguistic metadata (synset and hypernyms derived from WordNet for each word sense). Experiments conducted on GPT-4o-mini, LLaMA-3.1-8B and LLaMA-3.3-70B show varying performance, with F_1 scores ranging from 0.53 to 0.77. While GPT-4o-mini achieves the best overall accuracy, LLaMA-3.3-70B benefits the most from graph-based metadata, improving its F_1 score by up to 3 points. Analysis by word type reveals that concrete and semantically shifting words are more easily disambiguated than abstract and semantically stable words. Results highlight both the promise and the challenges of combining graph-structured linguistic knowledge with LLMs for historical WSD.

Keywords: Knowledge Graphs, LLM, Latin, Sense prediction

1. Introduction

Word Sense Disambiguation (WSD) is a long-standing task in Natural Language Processing (NLP) (Navigli, 2009), concerned with automatically determining the intended sense of a word in context given an inventory of senses. It plays a central role in downstream tasks such as information retrieval, question answering, and machine translation (Chan et al., 2007; Zhong and Ng, 2012). Over the years, this task has been tackled with a variety of approaches (Bevilacqua et al., 2021), including supervised learning methods trained on sense-annotated corpora (Ng and Lee, 1996; Zhong and Ng, 2010), distributional and contextual embeddings (Basile et al., 2014; Hadiwinoto et al., 2019), and more recently Large Language Models (LLMs) (Yae et al., 2024a; Basile et al., 2025; Yae et al., 2024b).

The SemEval-2020 Task 1 introduced a benchmark for semantic change detection in various languages (Schlechtweg et al., 2020), including Latin (McGillivray et al., 2022). The task combined dictionary-based sense inventories with manual corpus annotation. In parallel, resources such as the Latin WordNet (Minozzi, 2017; Franzini et al., 2019) and the LiLa Knowledge Base of Linguistic Resources for Latin (Passarotti et al., 2020) have emerged, encoding lexical senses and relations in graph form. In this framework, McGillivray et al. (2023) presented the Linguistic Knowedge Graph (LKG), which combines Latin WordNet, WikiData, and corpus metadata into a graph-based structure for the study of sense distributions in historical data. This allows for graph-based reasoning in the exploration of semantic relationships between senses, authorship metadata, and temporal information, e.g., the semantic shift of *beatus* from 'fortunate' to 'blessed' linked to Christian authors.

These developments, together with recent advancement in Generative AI, provide an opportunity to explore KG-based approaches to WSD for Latin, where Graph Retrieval-Augmented Generation methods can be directly applied to predict the correct sense of ambiguous tokens. These developments open up new possibilities for enriching Latin WSD with structured semantic and contextual knowledge, but it is still unclear how LLMs actually make use of such information in a low-resource historical setting. In particular, we lack systematic evidence on whether graph-derived metadata ranging from authorial and temporal information to semantic relations such as synsets and hypernyms can effectively support sense prediction. Moreover, the interaction between different types of metadata, model size, and linguistic properties such as polysemy,

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10937–10952
11–16 May 2026.

part of speech, concreteness, or diachronic meaning change has not yet been explored for Latin.

1.1. Contributions

In this paper, we address the open questions described in Section 1 by evaluating Graph Retrieval-Augmented Generation methods on a Latin LKG for WSD, using the SemEval Latin dataset as a benchmark. Our approach is novel in combining graph-based inference with historical WSD, and we provide both empirical results and an analysis of how semantic relations contribute to disambiguation in a morphologically rich, low-resource language. Our study contributes to informing the design of knowledge-enhanced methods for historical and under-resourced languages.

Our contributions are threefold: (1) we operationalise an LKG-based Graph RAG framework for Latin sense prediction and provide retrieval strategies tailored to diachronic data; (2) we report a systematic evaluation across multiple LLMs and graph metadata configurations on a standard Latin benchmark; and (3) we analyse performance trends by word type, showing that concrete and semantically shifting lemmas are generally easier to disambiguate than abstract and semantically stable ones. Overall, our results clarify when and how LKG-driven retrieval can support LLM-based WSD for historical languages.

2. Related work

2.1. Latin WSD

Despite extensive research on modern languages, historical and low-resource languages remain under-explored in WSD. These languages face particular challenges: limited sense-annotated corpora, morphological richness, orthographic variation, and diachronic semantic change. Some studies have proposed the use of dictionary senses and quotations to train classifiers for WSD in Latin (Bamman and Crane, 2009; Bamman and Burns, 2020; Lendvai and Wick, 2022). To this day, the application of language models to Latin WSD task remains unexplored, and very few works have investigated this line of research in recent years. The idea of leveraging WSD to assess the capability of language models of dealing with Latin was originally proposed by Bamman and Burns (2020), which tests Latin BERT on the sense disambiguation task. Lendvai and Wick (2022) fine-tuned Latin BERT on a part of sense representations in the *Thesaurus Linguae Latinae* (Thesaurus-Kommission, 1900–)[1], the first dictionary of ancient Latin usage up to 600 AD, offering a documented overview of Latin words' history. Ghinassi et al. (2024) explored word senses annotation propagation from English to Latin using parallel corpora.

[1] https://tll.degruyter.com/about

2.2. LLM and WSD

Loureiro et al. (2021) shows that BERT-like models can effectively differentiating between various word senses, even when only a few examples are available for each. Their analysis further reveals that although language models can perform nearly perfectly on coarse-grained noun disambiguation in ideal settings where training data and resources are abundant, such conditions are rare in practical scenarios, presenting ongoing challenges. Along the lines of BERT-like approaches, Bevilacqua et al. (2021) reviews various WSD methodologies, highlighting how language models can serve both as feature extractors for contextual embeddings and as the architecture for supervised training on sense-annotated corpora. Cabiddu et al. (2023) evaluates the WSD capabilities of language models using a framework of three behavioural experiments originally designed for children. This comparative analysis emphasises the differences between human semantic acquisition and transformer-based encoding. The authors found that language models show a bias toward the most frequent sense and report a negative correlation between the training data volume and model performance.

3. Background

3.1. Knowledge Graphs for Linguistic Analysis

Knowledge Graphs (KGs) (Hogan et al., 2021) are a key resource for encoding lexical and semantic relations. KGs can be used in linguistics to investigate diachronic semantics (McGillivray et al., 2023; Ghizzota et al., 2025), analyse etymology (De Melo (2014); Khan, 2018), examine quantitative patterns in the distribution of words and word senses across different language stages, genres, authors, and other linguistic, textual and contextual variables (Khan et al., 2018; Passarotti et al., 2020). Corpus analysis and NLP methods can help answer some of these questions, but integrating external knowledge to corpus data and metadata can give a better understanding of linguistic phenomena and the linguistic and extra-linguistic factors that drive them. Large-scale multilingual resources such as WordNet (Fellbaum, 1998) represent senses as nodes and their semantic or lexical relations as edges. Graph-based methods have exploited such resources for disambiguation (Lesk, 1986; Banerjee and Pedersen, 2002) also in hybrid or graph-based approaches (Agirre and Soroa, 2009) that generate graph representations of the text to be

disambiguated and then apply graph-based algorithms (such as PageRank) on them. The integration of NLP strategies with KGs has been advocated by Armaselu et al. (2022), who highlighted how these external resources better accommodate the diverse nature of linguistic data.

3.2. Graph Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) was introduced to overcome the limitation of LLMs parametric knowledge being bound to training data, without the option of searching for information in other external sources. RAG enables an LLM to look at trustworthy knowledge sources and collect new data that it has not seen during its training. The LLM uses the new *non-parametric* knowledge and its training data to generate better responses, thus specialising in a given domain in a cost-effective manner, without any need to re-train from scratch. Nevertheless, traditional RAG techniques face numerous limitations (Peng et al., 2024), i.e., performance deterioration with lengthy contexts (Liu et al., 2023), inability to deal with textual interconnections and global information due to RAG vector-based nature (Edge et al., 2024). Graph Retrieval-Augmented Generation (Graph RAG) (Edge et al., 2024; Hu et al., 2024; Zhang et al., 2025) tackles these concerns bringing together RAG techniques and graph-structured data, allowing RAG to take advantage of knowledge graphs and their structure. The interconnections modelled in the graph enable Graph RAG to discover knowledge related to a focal entity.

4. Data

4.1. Latin data

We used the SemEval Latin dataset (McGillivray et al., 2022), consisting of a sense-annotated portion of the LatinISE diachronic corpus of Latin (McGillivray and Kilgarriff, 2013), a 10-million-token collection of Latin texts spanning from the 5th century BCE to the present. The SemEval Latin dataset provides in-context semantic annotations for 39 lemmas: 17 lemmas known to have undergone semantic change related to Christianity or socio-cultural changes and 22 control lemmas. Each lemma is represented by 60 annotated sentences, evenly split between BCE and CE sources. Annotation followed a modified DuReL framework (Schlechtweg et al., 2020), using a four-point relatedness scale to assess semantic proximity to dictionary definitions sourced from Logeion dictionary[2] (Lewis and Short, 1879; Du Fresne Du Cange et al., 1883-1887).

[2] https://logeion.uchicago.edu/

4.2. Linguistic Knowledge Graph

The intent behind the design of the *Linguistic Knowledge Graph* (Basile et al., 2022) (LKG) is to capture different aspects of lexical resources, such as connections between words and concepts, morphological and syntactical information; furthermore, it covers diachronic aspects of language, e.g., the publication date of a literary work and the birth and death dates of the author. *In nuce*, it shapes time-sensitive linguistic knowledge with a graph database. The LKG is part of the GraphBRAIN framework (Ferilli, 2021), backed up by Neo4j, and its schema was inspired by the ontology-lexicon model LEMON (McCrae et al., 2012) and the semantic network WordNet.The GraphBRAIN framework and technology aim to apply the principles of Semantic Web to Labelled Property Graphs (LPGs) (Miller, 2013; Angles, 2018; Hogan et al., 2021) and state-of-the-art graph databases to open new exploitation possibilities, e.g., Data Mining, Knowledge Extraction and Management, Automated Reasoning. The main features of an LPG model, compared to an RDF one, are that in LPG properties (key-value pairs) can be associated to both nodes and relations; in LPG, nodes and relations are not identified via global identifiers like IRI or URI, but each of them has a unique internal ID. Essentially, the LPG data model offers a more straightforward approach, supporting richer connections thanks to the properties on edges, and more rapid traversals even in intricate networks.

The LKG originally imported the SemEval portion of Latin data (4.1). McGillivray et al. (2023) enriched the dataset by linking each annotated sense to WordNet synsets manually, and Wikidata via author metadata from the corpus. This enabled the integration of contextual metadata into the LKG, supporting diachronic semantic analysis. Ghizzota et al. (2025) applied a few refinement steps to the LKG, e.g., standardising author names (the variants 'Ovidius', 'Ouidius' and 'Ovidius Naso Publius' were aligned with 'Ovidius Naso Publius'), and disambiguating different literary works with the same title (Ovidius's 'Metamorphoses', 8 CE, and Apuleius' 'Metamorphoses', 2 CE).

For this work, some adjustments were applied to the original LKG schema before the upload on Neo4j; we are going to focus on the changes that are pivotal for this study. The updated LKG schema accommodates all the dataset enrichments presented in McGillivray et al. (2023), including the author's occupation with the new node OCCUPATION and the new directed edge OCCUPATION connecting an occupation to an author represented by the PERSON node. Figure 1 displays a sub-graph of the

OCCUPATION relation.

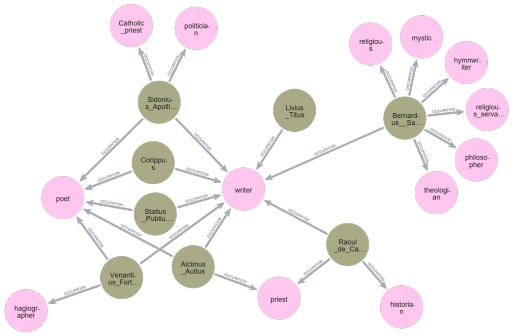

Figure 1: An example portion of LKG showing authors, PERSON nodes (dark green), linked to their occupations, OCCUPATION nodes (pink).

The directed edge HAS_EXAMPLE between a sentence (QUOTATION node) and a sense (SENSE node) in the previous version of LKG was changed into DESCRIBES, and two new properties were added: ROLE to indicate the role of the node the edge starts from, in this case EXAMPLE, and BINARY to store the binary encoding YES/NO of the DuReL scores, detailed in Section 5.1. Notice that each SENSE node stores its name in the NAME property and its gloss in the GLOSS property. Figure 2 depicts an example of this relation.

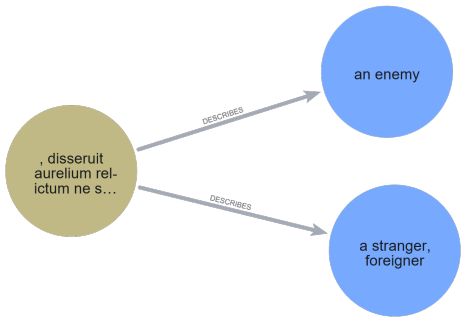

Figure 2: An example of the DESCRIBES relation: the QUOTATION node (dark yellow) is connected to two SENSE nodes (blue): the DESCRIBES edge to sense with gloss *an enemy* has property BINARY = YES, while the edge to sense with gloss *a stranger, foreigner* has property BINARY = NO.

Upon upload, the final graph on Neo4j consists of 4,276 nodes and 31,108 edges.

5. Sense prediction

The principal objective of this work is to assess the contribution of graph-structured data to the sense prediction task with Large Language Models (LLMs). To this end, we leverage a **Graph Retrieval-Augmented Generation** strategy to join together the strengths of Knowledge Graphs and Large Language Models.

At a high-level, the proposed Graph RAG methodology is straightforward: i) additional information is retrieved from the KG; ii) an LLM is fed with a single prompt containing the retrieved information and the task description. This work primarily focuses on the retrieval step, as it proposes to retrieve two different 'points of view' on the same sentence: the former leverages information about authors and literary works of the sentences most similar to the target one (5.2); the latter takes advantage of the linguistic information available in the LKG about the target sense (5.3). Ultimately, we test a combination of the two sources of information (5.4). Since the LKG has been uploaded on Neo4j, we use the Cypher query language for these retrieval steps.

The decision to include these two types of information (author and linguistic metadata) is motivated by several factors. Many of the lemmas in the dataset are known to have undergone semantic change with the advent of Christianity, acquiring meanings related to the religious sphere (e.g. *beatus* from 'happy' to 'blessed', see McGillivray et al. (2022)). We expect information about the author of the text from which a sentence is drawn, the period in which they were writing, and their role within the society and culture of the time (e.g., whether they were a religious figure or not) to help the LLM better interpret such shifts. As for linguistic metadata, the information extracted from WordNet helps contextualize the meaning of a word within a hierarchy of hypernyms. This hierarchy is supposed to clarify the type of referent involved (e.g., process/action, object, animate being) and the conceptual domain to which it belongs.

In light of the LKG described in Section 4.2, the sense prediction task is framed as predicting the value YES or NO of the BINARY property. This property belongs to the relation DESCRIBES between a QUOTATION node and a SENSE node (see 2). Therefore, this task falls under the umbrella of binary classification tasks.

5.1. Task prompt

For the purpose of this work, sense–sentence pairs were split in a stratified manner, taking into account the distribution of DuReL scores assigned to each sense. This stratification process yielded 8,989 pairs.

Since this sense prediction task is framed in a binary classification setting, the DuReL 1-to-4 scale was binary-encoded as follows:

- sense–sentence pairs with scores equal to or above 3 were labelled as YES;
- sense–sentence pairs with scores equal to or below 2 were labelled as NO.

Pairs of sense and sentence with score 0 – where the annotator could not decide – were not included, therefore the final dataset of sense–sentence pairs consists of 8,930 instances. See an example of annotated sentence in Table 1.

Ultimately, a prompt for each pair of sense and sentence was set up as below (hereafter **task prompt**):

> Given the target word form **(word)** and the sentence in input where the word is enclosed by the [TARGET] tag, and the following meaning **(gloss)**, assign a label 'yes' or 'no'.
> The label meaning is the following:
> 'yes': The sense for the target word occurrence is correct
> 'no': The sense for the target word occurrence is not correct
> Answer just with the label.
> The sentence to label is: **(sentence)**

5.2. Author metadata

The objective of this first setup is to provide the model with additional information about literary works and authors of sentences similar to the target sentence: the top-5 similar sentences where retrieved from the graph using the cosine similarity metric. Similarity was computed on the EMBEDDING property of the QUOTATION node, that represents a sentence in the graph. To this purpose, the node's VALUE property that holds the text of the sentence was embedded using OpenAI **text-embedding-3-small**, obtaining embeddings of dimension 1,536 for each sentence. The prompt specifies the literary work the sentence belongs to, the span of time in which it was written, its author and his or her occupation(s). See Appendix A.1 for an example. Figure 3 illustrates the graph path for retrieving this additional information using the QUOTATION node as entry point.

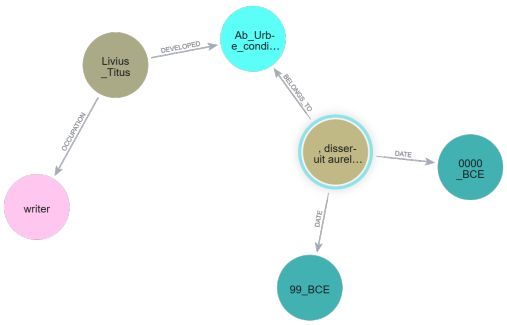

Figure 3: An example of query result for retrieving metadata about the sentence contained in the QUOTATION node (highlighted).

5.3. Linguistic metadata

This second setup aims at providing the LLM with additional linguistic information about the target sense. Notice that the senses in the graph are sourced from the Logeion online dictionary and from Latin WordNet, but sentences were annotated with Logeion senses only, thus there is no direct link between a sentence and a Latin WordNet sense. Senses from Logeion were mapped to Latin WordNet (4.2), in LKG represented via the SAME_AS link. Furthermore, LKG accommodates for representing semantic relations between senses, i.e., hyperonymy and hyponymy. In LKG, senses sourced from Latin WordNet are also linked to their hypernyms and hyponyms, if available. Therefore, when traversing the graph, by using a Logeion sense as entry point, it is possible to retrieve its corresponding Latin WordNet sense and its hypernyms and hyponyms. Figure 4 illustrates an example of graph path for retrieving linguistic information.

For this study, we decide to take into account only hypernyms, as hyponyms may introduce senses that are too specific with respect to the target sense and be misleading. A prompt example is available in Appendix A.2.

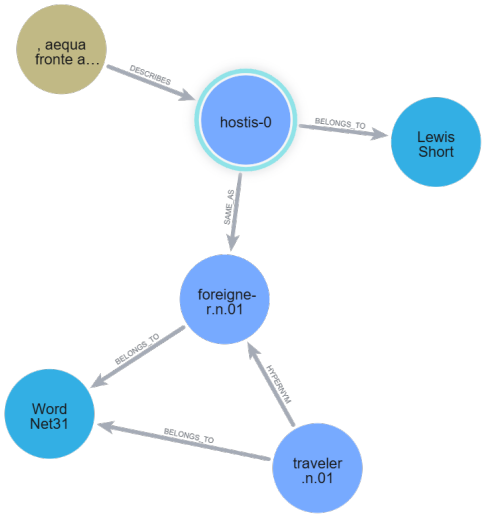

Figure 4: An example of query result for retrieving linguistic metadata: starting from SENSE *hostis-1* (highlighted) from Logeion, its corresponding sense in Latin WordNet and its respective hypernym are retrieved. Each sense is linked to a TAXONOMY (cerulean) via the BELONGS_TO relation.

5.4. Author and linguistic metadata

At last, the aforementioned author and linguistic metadata were combined together to assess whether their contribution in a joint manner enhances the performance or not. This setup leverages sentence embeddings to retrieve the top-3

Text	Senses				
	'*blessed*'	'*rich*'	'*fortunate*'	'*happy*'	'*rewarded*'
[...] Dico enim constanter grauiter sapienter fortiter. Haec etiam in eculeum coiciuntur, quo uita non adspirat beata. - Quid igitur? solane **beata** uita, quaeso, relinquitur extra ostium limenque carceris, cum constantia grauitas fortitudo sapientia reliquaeque uirtutes rapiantur ad tortorem nullumque recusent nec supplicium nec dolorem? [...]	1	1	3	3	2
	no	no	yes	yes	no

Table 1: An annotated usage of lemma *beatus* (McGillivray et al., 2022); extracted from a classical text, Cicero's '*Tusculanae disputationes*' (46 BC). The DuReL scores for each sense of *beatus* are on the first row, on the second row are the binary labels.

sentences most similar to the target one and their respective author, literary work and publication period (see Section 5.2), as well as a positive and a negative hypernym example (see Section 5.3). See Appendix A.3 for a prompt example.

The objective of joining author and linguistic metadata is twofold: on one hand, we can assess how they contribute to the task in a joint manner, on the other hand, we can compare the performance with those achieved by the LLM leveraging the metadata in a separate manner.

6. Results and evaluation

The binary encoded set for the experiment consists of 30% of the aforementioned dataset, thus 2,675 instances. These include 956 instances of class YES and 1,719 NO, making up an imbalanced dataset where class YES is 35.73% of the dataset. This is the same dataset used in Ghizzota et al. (2025), that evaluates GPT-4o-mini, LLaMA-3.3-70B and LLaMA-3.1-8B on the WSD task in a zero-shot setting. Results of this study act as the baseline of our analysis. The models were chosen in order to obtain a comprehensive overview of LLM performances on the WSD task, whether they are open-source (LLaMA) or proprietary (GPT).

For our experiment, in LKG the property BINARY of relation DESCRIBES was set to NONE for the test instances.

The results of precision, recall and F_1 for the three models in each experimental setting are shown in Table 2. Additionally, a summary of the results (F_1 score for each model) can be visualized in the two graphs in Figure 5.

Table 2 presents two separate entries for the score of each LLM alone because the input data for the experiment with author, linguistic, and joint information were slightly different. Some of the senses associated with the target words do not have a corresponding entry in WordNet. This is the case for meanings specific to Roman culture and institutions, e.g., 'virtue, personified as a deity' for the target word *virtus*. See the full list of excluded senses in Appendix A.2.2. In these cases, we did not link the sense to WordNet, therefore we had to exclude 126 corresponding passages from the experimental setting including the linguistic information. For this reason, the number of passages given to the LLM differs between author and linguistic metadata: LLM+KG$_{author}$ is tested on 2,675 instances, while LLM+KG$_{linguistic}$ and LLM+KG$_{author,linguistic}$ on 2,549.

Across the three models LLaMA 3.1 8B, GPT-4o-mini, and LLaMA 3.3 70B, the results show an improvement as model size increases for what concerns the two LLaMA models. The average F_1 rises from 0.53/0.52 in LLaMA 3.1 8B to 0.65/0.66 in LLaMA 3.3 70B, showing that scaling probably improves the model's ability to discriminate between different word senses. GPT-4o-mini has an F_1 score of 0.77/0.76, outperforming the two LLaMA models.

When author and linguistic metadata from the knowledge graph (author, period, occupation of the author, synset and hyperonyms from WordNet) are added to the prompt, performance changes unevenly across different models. LLaMA-3.3-70B is the winning model in our experiment, as its accuracy increases in all three experimental settings, managing to leverage additional information and output the correct answer without suffering information overload. On the other hand, the performance of both GPT-4o-mini and LLaMA-3.1-8B overall decreaseses.

When author metadata is added to the prompt, LLaMA-3.3-70B gains slightly (0.65 against 0.63 in F_1), whereas both GPT-4o-mini (0.68 against 0.77) and LLaMA-3.1-8B (0.42 against 0.53) decline.

Model	Setting	Precision	Recall	F_1
GPT-4o-mini	LLM	**0.7974**	**0.7634**	**0.7682**
	LLM+KG$_{author}$	0.7373	0.6755	0.6814
	LLM	**0.7906**	**0.7544**	**0.7592**
	LLM+KG$_{linguistic}$	0.7145	0.6775	0.6838
	LLM+KG$_{author,linguistic}$	0.7339	0.6163	0.6136
LLaMA-3.3-70B	LLM	0.7030	0.6301	0.6355
	LLM+KG$_{author}$	**0.7427**	**0.6505**	**0.6535**
	LLM	0.7009	0.6312	0.6361
	LLM+KG$_{linguistic}$	0.7493	**0.6559**	**0.6579**
	LLM+KG$_{author,linguistic}$	**0.7496**	0.6414	0.6411
LLaMA-3.1-8B	LLM	0.5947	**0.5271**	**0.5336**
	LLM+KG$_{author}$	**0.6061**	0.4536	0.4170
	LLM	0.5877	**0.5194**	**0.5242**
	LLM+KG$_{linguistic}$	0.6101	0.4072	0.3130
	LLM+KG$_{author,linguistic}$	**0.7345**	0.3707	0.2114

Table 2: Performance comparison of different LLMs and knowledge graph integrations. The highest scores across all models and experimental settings are underlined.

When linguistic information from our knowledge graph is provided, the models behave in different ways again. GPT-4o-mini shows similar performance as when fed with the author and text metadata, lowering its F_1 from 0.76 to 0.68. LLaMA-3.1-8B also lowers its performance accuracy, dramatically dropping from 0.52 (LLM only) to 0.31 F_1 (even lower than with author metadata). LLaMa-3.3-70B is the only model that improves its performance when given linguistic information from our knowledge graph (from 0.63 to 0.66 F_1).

Providing a combination of author and linguistic metadata does not dramatically change the trends observed in the other two experimental settings. GPT-4o-mini decreases from 0.76 to 0.61, both lower than author and linguistic metadata provided separately. LLaMA-3.1-8B goes down to 0.21. Again, LLaMA-3.3-70B is the only model that (marginally) improves its F1 score (from 0.636 to 0.641) in this setting.

Overall, in all three experimental settings GPT-4o-mini outperforms the two LLaMA models in determining the right answer to the prompt. LLaMA 70B is the only model that improves its performance in the three experimental settings. In general, it looks like providing author and linguistic metadata combined leads to worse performances for all models, compared to providing linguistic and author information separately.

The cause of the decline in performances can be associated to the length of the prompts (see Appendix A). The often extremely lengthy text of sentences similar to the one containing the instance being disambiguated may have compromised the contribution of additional information, as it tends to be obscured within a verbose prompt. Statistical significance tests for pairs of performances with a F_1 score difference $\Delta < 0.05$ are in Appendix B.

6.1. In-depth analysis

In order to understand what type of linguistic features facilitate or hinder semantic analysis for the LLM, we evaluated the results for each word included in the experiment. In this section, we analyse results for each word considering: i) number of senses, i.e., level of polysemy; ii) Part-of Speech (noun, adjective or verb); iii) words that change their meaning from the pre-Christian era (BCE) to the Christian era (CE), based on McGillivray et al. (2022); iv) nouns that express abstract concepts compared to nouns that express concrete concepts.

Polysemy. No correlation seems to emerge between the number of senses that the word has and the performance of the model on that specific word. This holds true for both the language models alone and for their combinations with the knowledge graph. However, these results should be interpreted with caution, as the distribution of words across the different numbers of senses is highly uneven (see Table 5 in Appendix, section C). For example, in our dataset there are seven words with two senses, ten words with three senses, and twelve words with four senses. The sample contains only two–four words with five, six, or seven senses. Such imbalance prevents us from drawing any statistically significant conclusions about the level of polysemy of a word and the ability of

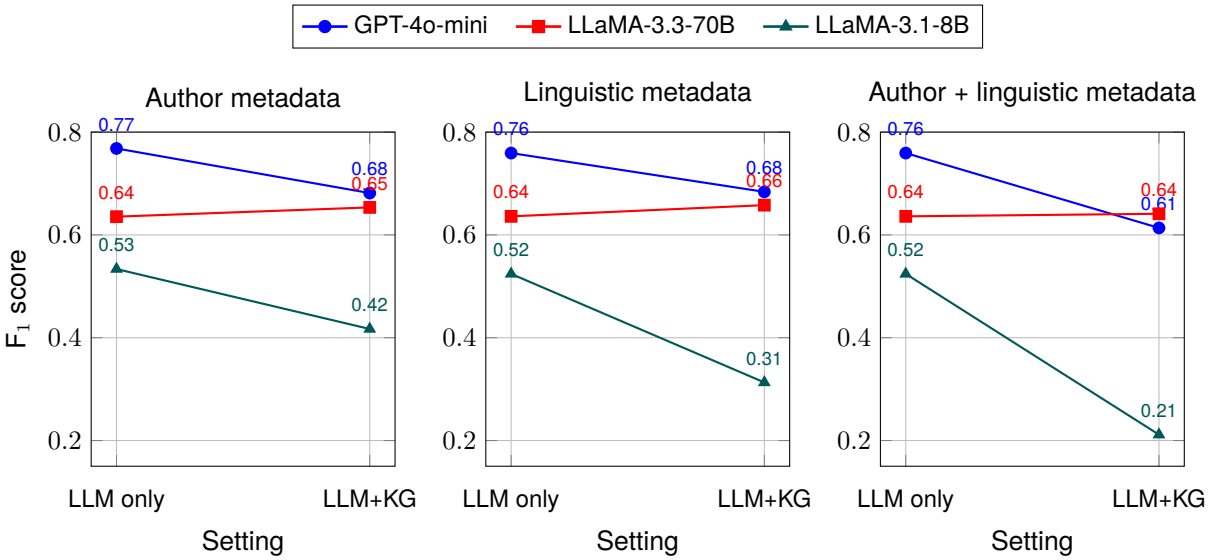

Figure 5: F_1 score comparison for different models with knowledge graph integration. Left: author metadata. Center: linguistic metadata. Right: combination of author and linguistic metadata. Numeric labels indicate F_1 values.

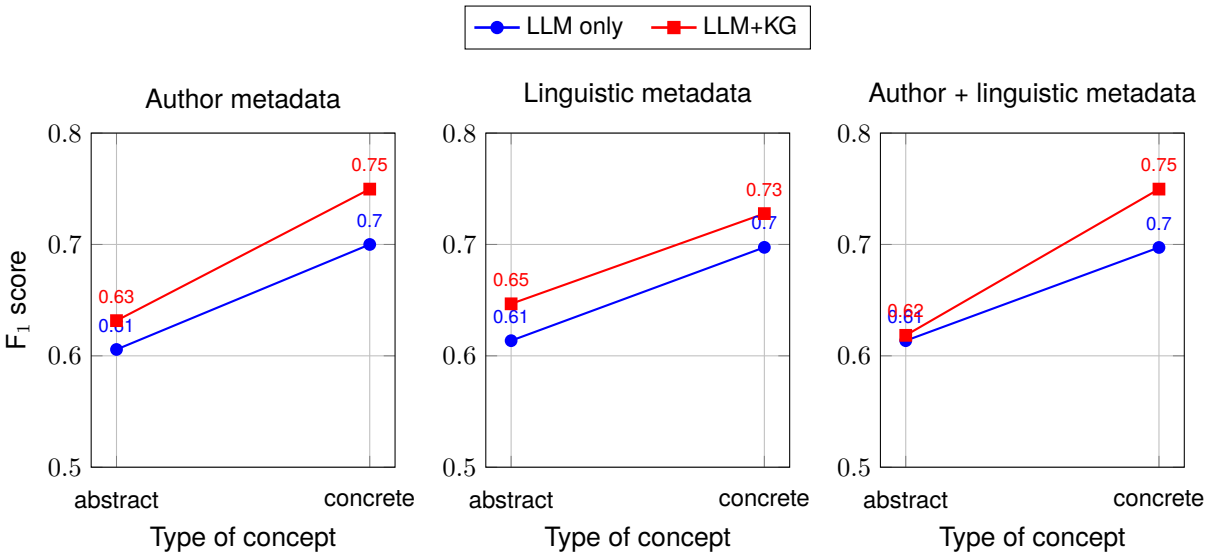

Figure 6: F_1 score comparison for LLaMA-3.3-70B in the three experimental settings, with a focus on abstract vs. concrete concepts.

the LLM to detect the right sense for that specific word. Overall, comparing the different models, both GPT-4o-mini and LLaMA-3.1-8B perform worse when combined with the KG than when used alone, whereas LLaMA-3.3-70B shows improved performance in the same setting (0.61 to 0.64).

Part-of-Speech. We encounter a similar distributional imbalance. Most of the words in the dataset are nouns (28 in total), while adjectives and verbs are fewer than ten each. For nouns, GPT-4o-mini and LLaMA-3.1-8B both perform worse when coupled with the KG, although the combination GPT-4o-mini + KG still yields slightly better results than LLaMA-3.1-8B + KG. In contrast, LLaMA-3.3-70B benefits from the inclusion of information from the KG, in all three settings: author metadata, linguistic information, and the two of them together. This improvement is likely linked to the fact that most words that changed meaning between the BCE and CE periods are nouns. These include e.g. *humanitas* (from 'human nature, humanity' to 'humanity, philanthropy' and 'mankind'), *virtus* (from 'manliness' to 'Christian virtues'), or *sacramentum* (from 'civil suit; military oath; oath; secret; mystery' to 'a sacrament' in the Christian sense). We should note that within the sample of so-called stable words, there are, of course, polysemous terms that may

have inevitably undergone semantic change. An example is *hostis*, meaning both 'foreigner' and 'enemy'. However, these semantic shifts are not the result of any changes in the external world, and are therefore considered stable in the experiment.

Diachronic change. When comparing changed and stable words, all models consistently perform better on changed words across all experimental settings. In particular, LLaMA-3.3-70B appears to take advantage of additional contextual information alone, suggesting that especially linguistic information supports the model in identifying when the meaning of a words has evolved, possibly in connection with cultural or societal shifts such as the advent of Christianity or new power structures. Interestingly, linguistic information seems even more beneficial than author metadata, though this warrants further investigation.

Abstract vs. concrete. As for abstract versus concrete nouns, our analysis shows that all models achieve better results on the disambiguation of concrete concepts compared to abstract ones. Two words in the sample, i.e., *consilium* and *senatus*, were labeled as both abstract and concrete, since they encompass both a concrete and an abstract dimension (e.g. for *consilium* 'the persons who deliberate, a council', and 'wisdom, judgment, sense'). Due to their limited number, these cases were excluded from the analysis. In this setting as well, GPT-4o-mini and LLaMA-3.1-8B achieve better results without KG information, while LLaMA-3.3-70B benefits from both types of metadata (author and linguistic) available in the KG. The difference between these two kinds of information does not appear to be substantial, as shown in Figure 6, with the performances remaining roughly similar in the three experimental settings.

7. Discussion and conclusion

Across the three models tested, the baseline results, i.e., those obtained without the integration of metadata from our KG in the prompt, reveal a clear upward trend in F_1 score, with LLaMA-3.1.-8B being the worst model and GPT-4o-mini being the best performing one. The interpretation of these results in terms of model size is partially hindered by the fact that the number of parameters of GPT-4o-mini are not known. However, a large number of parameters appears to help the model correctly identify the right answer, as suggested by the notable difference in performance between LLaMA 8B and LLaMA 70B. Additionally, it is reasonable to expect GPT-4o-mini to have seen more Latin data than the two LLaMA models, as its baseline performances are much better in all experimental settings (with and without knowledge graph), also in line with recent work (Farina et al., 2025).

The introduction of additional information in the prompt through our knowledge graph affects the models differently. When author metadata is added, only LLaMA-3.3-70B shows a slight gain, and a similar pattern emerges when linguistic information is incorporated, and when linguistic and author information are combined.

The superior performance of LLaMA-3.3-70B in handling graph knowledge seems to come from its larger parameter size and consequently its stronger integration strategies for external information. Its large size allows it to balance previously acquired knowledge with newly introduced information and therefore to benefit from additional data rather than being misled by it. Most probably, both LLaMA models were not exposed to large amounts of Latin data during training. However, LLaMA-3.3-70B is capable of leveraging additional information given in the prompt, improving accuracy. LLaMA-3.1-8B, on the other hand, has a limited number of parameters, and consequently a limited representational capacity. Due to its structure, it does not seem to correctly integrate the information given in the prompt within its prior knowledge. The prompt probably ends up being too noisy and misleading for the model to correctly predict the right answer.

Some tendencies also emerge from our word-level in-depth analysis. Words that designate concrete concepts are more easily interpreted by all three models, though with different levels of accuracy, compared to abstract concepts. Similarly, words that acquired a new meaning or underwent semantic change with Christianity are easier to interpret in context for all models, compared to semantically stable words.

Some conclusions and directions for future work can be drawn from our results. Firstly, among open models, a large parameter size is to be preferred when performing Graph RAG on complex tasks, such as sense prediction in Latin. More specifically, LLaMA-3.3-70B proves to be receptive to Graph RAG strategies, compared to a smaller size model of the same family. Secondly, prompt and Graph RAG strategies need to be further investigated and implemented in order to make sure that new information is not misleading, and is successfully used by the model to predict the correct answer. Lastly, semantic features linked to the concept indicated by the target word (e.g. abstract/concrete; new meaning/stable word), more than linguistic features such as level of polysemy or PoS, seem to make the task easier for the LLM.

Acknowledgements

Eleonora Ghizzota acknowledges the Ph.D. fellowship within the framework of the Italian Ministerial Decree (D.M.) n. 630, date of D.M.: 24.04.2024 - under the National Recovery and Resilience Plan, Mission 4, Component 2, Investment 3.3 - Ph.D. Project 'Development and application of Generative Artificial Intelligence models based on the Symbiotic AI paradigm to support the Public Administration', co-supported by company InnovaPuglia S.p.A. (CUP B91I24000240007).
Paola Marongiu acknowledges the support of COST Action CA23147 GOBLIN - Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs, supported by COST (European Cooperation in Science and Technology, https://www.cost.eu)
Pierpaolo Basile acknowledges the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.
Barbara McGillivray acknowledges the support of the UKRI under the Horizon Europe Guarantee (grant number UKRI947) for the project COALA (Computational Corpus Annotation for Quantitative Analysis of Latin Lexical Semantics) successfully evaluated by the ERC.

8. Authors' contributions

Eleonora Ghizzota worked on the adjustments to the LKG with Stefano Ferilli and Pierpaolo Basile, conceptualisation of the methodology and development of the software (Neo4j LKG import, Graph RAG experiments), wrote sections 3, 4, 5, A and B, and contributed to sections 1, 2, 6.
Paola Marongiu wrote sections 6, 7, and C, and contributed to sections 1.1 and 5.
Pierpaolo Basile was responsible for the conceptualisation of both the RAG methodology and the experimental design and provided overall supervision throughout manuscript preparation, including critical review and editing.
Barbara McGillivray wrote sections 1, 2, 4.1, provided overall supervision throughout manuscript preparation, including critical review and editing, and contributed to the design of the study and the analysis of results.

9. Bibliographical References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece. Association for Computational Linguistics.

Renzo Angles. 2018. The property graph database model. In *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management, May 21-25, 2018*, Cali, Colombia. CEUR Workshop Proceedings.

Florentina Armaselu, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Andrius Utka, Giedrė Valūnaitė Oleškevičienė, and Marieke van Erp. 2022. LL(O)D and NLP perspectives on semantic change for humanities research. *Semantic Web*, 13(6):1051–1080.

David Bamman and Patrick J Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International joint conference on artificial intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc.

Francesco Cabiddu, Mitja Nikolaus, and Abdellah Fourtassi. 2023. Comparing children and large language models in word sense disambiguation: Insights and challenges. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

Stefano Ferilli. 2021. Integration strategy and tool between formal ontology and graph database technology. *Electronics*, 10(21):2616.

Eleonora Ghizzota, Pierpaolo Basile, Lucia Siciliani, and Giovanni Semeraro. 2025. The meaning of beatus: Disambiguating latin with contemporary ai models.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F.

Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Springer.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. 2024. Grag: Graph retrieval-augmented generation.

Anas Fahad Khan. 2018. Towards the representation of etymological data on the semantic web. *Information*, 9(12):304.

Fahad Khan, Javier E Díaz-Vera, and Monica Monachini. 2018. Representing meaning change. *Formal Representation and the Digital Humanities*, page 59.

Piroska Lendvai and Claudia Wick. 2022. Fine-tuning latin bert for word sense disambiguation on the thesaurus linguae latinae. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 37–41.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Justin J Miller. 2013. Graph database applications and concepts with neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324, pages 141–147.

Marco Passarotti, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, and Francesco Mambrini. 2020. The lila knowledge base of linguistic resources for latin. In *Proceedings of LREC*.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph retrieval-augmented generation: A survey.

Thesaurusbüro München Internationale Thesaurus-Kommission, editor. 1900–. *Thesaurus linguae latinae*. Mouton de Gruyter, Berlin.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, Yi Chang, and Xiao Huang. 2025. A survey of graph retrieval-augmented generation for customized large language models.

10. Language Resource References

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology. In *ArXiv*, volume abs/2009.10053.

David Bamman and Gregory Crane. 2009. Computational linguistics and classical lexicography. *Digital Humanities Quarterly*, 3.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, page 136–145, Berlin, Heidelberg. Springer-Verlag.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Pierpaolo Basile, Pierluigi Cassotti, Stefano Ferilli, and Barbara McGillivray. 2022. A new time-sensitive model of linguistic knowledge for graph databases. In *Proceedings of the 1st Workshop on Artificial Intelligence for Cultural Heritage co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022)*, page 69. CEUR Workshop Proceedings.

Pierpaolo Basile, Lucia Siciliani, Elio Musacchio, and Giovanni Semeraro. 2025. Exploring the word sense disambiguation capabilities of large language models.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint*

Conference on Artificial Intelligence, IJCAI-21, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic. Association for Computational Linguistics.

Gerard De Melo. 2014. Etymological wordnet: Tracing the history of words. In *LREC 2014*, pages 1148–1154.

Charles Du Fresne Du Cange, G. A. Louis Henschel, P. Carpentier, Johann Christoph Adelung, and Léopold Favre. 1883-1887. *Glossarium mediæet infimælatinitatis*. L. Favre, Niort.

Andrea Farina, Andrea Ballatore, and Barbara McGillivray. 2025. Mapping meaning in latin with large language models: A multi-task evaluation of preverbed motion verbs and spatial relation detection in llms. In *CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy*.

Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Viviana Ventura, and Federica Zampedri. 2019. *Nunc Est Aestimandum*: Towards an evaluation of the latin wordnet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Accademia University Press.

Iacopo Ghinassi, Simone Tedeschi, Paola Marongiu, Roberto Navigli, and Barbara McGillivray. 2024. Language pivoting from parallel corpora for word sense disambiguation of historical languages: A case study on Latin. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10073–10084, Torino, Italia. ELRA and ICCL.

Eleonora Ghizzota, Pierpaolo Basile, Claudia d'Amato, and Nicola Fanizzi. 2025. Enhancing linguistic resources for diachronic analysis via linked data. *GWC 2025*, page 192.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Piroska Lendvai and Claudia Wick. 2022. Finetuning Latin BERT for word sense disambiguation on the thesaurus linguae latinae. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 37–41, Taipei, Taiwan. Association for Computational Linguistics.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, page 24–26.

Charlton T. Lewis and Charles Short. 1879. *A Latin Dictionary, Founded on Andrews' edition of Freund's Latin dictionary revised, enlarged, and in great part rewritten by Charlton T. Lewis, Ph.D. and Charles Short*. Clarendon Press, Oxford.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46:701–719.

Barbara McGillivray, Pierluigi Cassotti, Davide Di Pierro, Paola Marongiu, Fahad Khan, Stefano Ferilli, and Pierpaolo Basile. 2023. Graph databases for diachronic language data modelling. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 86–96.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In *New Methods in Historical Corpus Linguistics*, pages 247–257, Tübingen. Narr.

Barbara McGillivray, Daria Kondakova, Annie Burman, Francesca Dell'Oro, Helena Bermúdez Sabel, Paola Marongiu, and Manuel Márquez Cruz. 2022. A new corpus annotation framework for latin diachronic lexical semantics. *Journal of Latin Linguistics*, 21(1):47–105.

Stefano Minozzi. 2017. Latin wordnet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. *Strumenti digitali e collaborativi per le Scienze dell'Antichita*, (14):123–134.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate

word sense: an exemplar-based approach. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, page 40–47, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1–23. International Committee for Computational Linguistics.

Jung H. Yae, Nolan C. Skelly, Neil C. Ranly, and Phillip M. LaCasse. 2024a. Leveraging large language models for word sense disambiguation. *Neural Comput. Appl.*, 37(6):4093–4110.

Jung H. Yae, Nolan C. Skelly, Neil C. Ranly, and Phillip M. LaCasse. 2024b. Leveraging large language models for word sense disambiguation. *Neural Comput. Appl.*, 37(6):4093–4110.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: a wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, page 78–83, USA. Association for Computational Linguistics.

Zhi Zhong and Hwee Tou Ng. 2012. Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–282, Jeju Island, Korea. Association for Computational Linguistics.

11. Resources

We release the source code, the testing data, the experiment outputs and figures in a GitHub repository `https://github.com/Midorilly/latin-sense-prediction`.

Appendix

A. Prompts

Tabel 3 illustrates the number of test instances and the average length of prompts for each experimental setting. See Sections A.1, A.2, A.3 for the prompt structure.

Experimental setting	Test instances	Prompt length
LLM+KG$_{author}$	2,675	4,104.95
LLM+KG$_{linguistic}$	2,549	4,253.82
LLM+KG$_{author,linguistic}$	2,549	6,731.48

Table 3: Number of test instances and average prompt length for each experimental setting.

A.1. Prompt with author metadata

SYSTEM PROMPT:
You are a linguist proficient in Latin and you are tackling a Latin word sense disambiguation task.

USER PROMPT:
Consider the following n sentences and their metadata as examples for the task:

Sentence **(sentence)** belongs to work **(literary work title)** written in **(year(s))**. Its author is **(author name)** and was **(author occupation)**.
⋮
Sentence n

TASK PROMPT

A.1.1. Example

SYSTEM PROMPT:
You are a linguist proficient in Latin and you are tackling a Latin word sense disambiguation task.

USER PROMPT:
Consider the following n sentences and their metadata as examples for the task:

Sentence **', magna pars ablegati. quam multitudinem consul alter romae praetorque alias ex aliis fingendo moras retinebat. et primo quidem ignari ludificationis minime inviti domos revisebant; postquam neque reverti ad signa primos nec ferme alium quam qui in campania hibernassent praecipueque ex his seditionis auctores mitti viderunt, primum admiratio, deinde haud dubius timor incessit animos consilia sua emanasse: iam quaestiones, iam indicia, iam occulta singulorum supplicia impotensque et crudele consulum ac patrum in se regnum passuros. haec qui in castris erant occultis sermonibus serunt,**

nervos coniurationis electos arte consulis cernentes. [...]' belongs to work **Ab Urbe condita** written in **99 BCE**, **0000 BCE**. Its author is **Livius Titus** and was a **writer**.
⋮
Sentence n

TASK PROMPT

A.2. Prompt with linguistic metadata

SYSTEM PROMPT:
You are a linguist proficient in Latin and you are tackling a Latin word sense disambiguation task.

USER PROMPT:
Consider the following information and examples about the target sense **(sense gloss)**: it has the same meaning as **(Latin WordNet sense gloss)** and its hypernym is **(hypernym gloss)**.
The following sentence is a positive example of the target sense and was labelled with 'yes': **(positive example)**.
The following sentence is a negative example of the target sense and was labelled with 'no': **(negative example)**.

TASK PROMPT

A.2.1. Example

SYSTEM PROMPT:
You are a linguist proficient in Latin and you are tackling a Latin word sense disambiguation task.

USER PROMPT:
Consider the following information and examples about the target sense **'evil intent, wrongdoing'**: it has the same meaning as **a cunning or deceitful action or device: he played a trick on me; he pulled a fast one and got away with it** and its hypernym is **any clever (deceptive) maneuver; he would stoop to any device to win a point**.
The following sentence is a positive example of the target sense and was labelled with 'yes': **[...]a eris, ecce, color, tum cum sine nubibus aer nec tepidus pluvias concitat auster aquas; ecce, tibi similis, quae quondam phrixon et hellen diceris inois eripuisse dolis ; hic undas imitatur, habet quoque nomen ab undis: crediderim nymphas hac ego veste tegi. [...]**.
The following sentence is a negative example of the target sense and was labelled with 'no': **[...] tegminibus telisque super[sigeaque praeter] eriperet reditus, alter vulcania ferro vulnera protectus depellere navibus instat.) hos erat aeacides vultu laetatus honores, dardaniaeque alter fuso quod sanguine campis hector lustravit devicto corpore troiam. rursus acerba fremunt, paris hunc quod letat et huius arma dolis ithaci virtus quod concidit icta. [...]**.

TASK PROMPT

A.2.2. Excluded senses

Due to the absence of their hypernyms in the graph, the following senses and related sentences were excluded from the experiment with linguistic metadata:

- *virtus*: 'related to Christian virtue', 'related to virtue, personified as a deity';
- *templum*: 'a space marked out, an open place for observation';
- *sacramentum*: 'the military oath of allegiance';
- *consul*: 'municipal official', 'an epithet of Jupiter', 'consul'.

A.3. Prompt with linguistic and author metadata

SYSTEM PROMPT:
You are a linguist proficient in Latin and you are tackling a Latin word sense disambiguation task.

USER PROMPT:
Consider the following information and examples about the target sense **(sense gloss)** and the metadata of the examples:

It has the same meaning as **(Latin WordNet sense gloss)** and its hypernym is **(hypernym gloss)**.
The following sentence is a positive example of the target sense and was labelled with 'yes': **(positive example)**.
The following sentence is a negative example of the target sense and was labelled with 'no': **(negative example)**.

The following sentences are the most similar to the target sentence:

Sentence **(sentence)** belongs to work **(literary work title)** written in **(year(s))**. Its author is **(author name)** and was **(author occupation)**.
⋮
Sentence n

TASK PROMPT

B. Statistical significance test

We compute the p-value with the Wilcoxon signed-rank test for pairs of performances with a F_1 score difference $\Delta < 0.05$. Table 2 shows that the slight performance increment achieved in the LLM+KG$_{linguistic}$ setting w.r.t. LLM only, 0.6579 versus 0.6361, is statistical significant for $p < 0.05$.

p-value	**LLaMA-3.3-70B**
LLM+KG$_{author}$	0.0612
LLM+KG$_{linguistic}$	**0.0214**
LLM+KG$_{author,linguistic}$	0.1178

Table 4: Wilcoxon signed-rank test of LLaMA-3.3-70B.

C. Data

The lemmas used for the experiments are described in Table 5 with reference to the following categories: part-of-speech, level of polysemy (i.e., number of senses), abstract vs. concrete (only for nouns), and changed vs. stable.

Dimension	Category	Lemmas	N
PoS	ADJ	acerbus, beatus, dubius, fidelis, necessarius, sanctus, simplex	7
	NOUN	ancilla, civitas, cohors, consilium, consul, dolus, dux, honor, hostis, humanitas, imperator, jus, nepos, nobilitas, poena, pontifex, potestas, regnum, sacramentum, salus, sapientia, scriptura, senatus, sensus, templum, titulus, virtus, voluntas	28
	VERB	adsumo, credo, itero, licet	4
Polysemy	2 senses	adsumo, ancilla, hostis, itero, necessarius, senatus, simplex	7
	3 senses	acerbus, consilium, dubius, dux, fidelis, imperator, poena, pontifex, salus, templum	10
	4 senses	civitas, cohors, consul, dolus, humanitas, licet, nepos, nobilitas, regnum, sanctus, sapientia, scriptura	12
	5 senses	beatus, honor, titulus, voluntas	4
	6 senses	jus, potestas, sacramentum, virtus	4
	7 senses	credo, sensus	2
Ontological Class (nouns only)	Abstract	civitas, dolus, honor, humanitas, jus, nobilitas, poena, potestas, regnum, sacramentum, salus, sapientia, sensus, titulus, virtus, voluntas	16
	Concrete	ancilla, cohors, consul, dux, hostis, imperator, nepos, pontifex, scriptura, templum	10
	Both (collective nouns)	consilium, senatus	2
Semantic Change	Changed	beatus, civitas, cohors, consul, credo, dolus, dux, fidelis, humanitas, imperator, itero, pontifex, potestas, sacramentum, sanctus, scriptura, virtus	17
	Stable	acerbus, adsumo, ancilla, consilium, dubius, honor, hostis, jus, licet, necessarius, nepos, nobilitas, poena, regnum, salus, sapientia, senatus, sensus, simplex, templum, titulus, voluntas	22

Table 5: Summary of lemmas in the study across morpho-syntactic, semantic, ontological and diachronic dimensions.