Predicate Sense Disambiguation for UMR Annotation of Latin: Challenges and Insights

Federica Gamba
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
gamba@ufal.mff.cuni.cz

Abstract

This paper explores the possibility to exploit different Pretrained Language Models (PLMs) to assist in a manual annotation task consisting in assigning the appropriate sense to verbal predicates in a Latin text. Indeed, this represents a crucial step when annotating data according to the Uniform Meaning Representation (UMR) framework, designed to annotate the semantic content of a text in a cross-linguistic perspective. We approach the study as a Word Sense Disambiguation task, with the primary goal of assessing the feasibility of leveraging available resources for Latin to streamline the labor-intensive annotation process. Our methodology revolves around the exploitation of contextual embeddings to compute token similarity, under the assumption that predicates sharing a similar sense would also share their context of occurrence. We discuss our findings, emphasizing applicability and limitations of this approach in the context of Latin, for which the limited amount of available resources poses additional challenges.

1 Introduction

Word Sense Disambiguation (WSD), i.e. the task of identifying the correct sense of a word in a specific instance or sentence, poses non-trivial challenges especially in the context of languages where resources are relatively scarce. This is the case of Latin, whose few existing resources confront the inherent complexity of the task and often resort to a binary approach revolving around the assumption that the several senses of a word can be reduced to two primary senses. This inevitably leads to resources that are overly coarse-grained. While such simplifications serve as valuable starting points for future experiments, their granularity may not universally cater to the diverse research needs.

The present work originates from the needs of a distinct project, which focuses on the annotation of Latin data according to the Uniform Meaning Representation framework (UMR) (Van Gysel et al., 2021). The text to be annotated is *De Coniuratione Catilinae* ‘Conspiracy of Catiline’ by Sallust. The UMR framework is designed to annotate the semantic content of a text, and was developed with cross-linguistic scope in mind. It is primarily based on Abstract Meaning Representation (AMR) (Banarescu et al., 2013), and aims at extending it to other languages – in particular to morphologically complex, possibly low-resource languages – in a cross-lingual and typological perspective. In AMR and UMR graphs, nodes represent semantic concepts. If word senses are available, semantic concepts are defined as word senses; participant roles associated to each predicate (e.g., ARG0, ARG1) are included in the graph if realized in the sentence. For instance, the predicate *utimur* in the sentence *Corporis servitio magis utimur* ‘Of the body we rather employ the service’ corresponds to the semantic concept `utor-03`, i.e. the sense "put into service; make work or employ for a particular purpose or for its inherent or natural purpose" to which ARG0 (first person plural, not overtly realized) and ARG1 (*servitio*) are associated. Within the whole annotation process, manual selection of the correct sense constitutes a time-consuming and demanding sub-task. We thus aim to investigate whether the existing resources allow to develop a strategy to expedite this process, by deriving annotation suggestions for unannotated predicates based on already manually annotated ones.

The paper is structured as follows. Section 2 presents an overview of related work, while Section 3 discusses Latin Vallex as the main linguistic resource that has been exploited, as well as the limitations it presents. Section 4 describes the methodology designed for the task, while its outcomes are evaluated in Section 5. Section 6 highlights some conclusive remarks and possible future research directions.

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 19–29
August 15, 2024 ©2024 Association for Computational Linguistics

2 Related Work

The exploration of WSD tasks for classical languages, and notably Latin, has recently gained attention, especially from a diachronic perspective with regard to lexical semantic change (Beelen et al., 2021; McGillivray, 2021; McGillivray et al., 2022, 2023a; Marongiu and McGillivray, 2023). However, the granularity of available resources remains a significant obstacle to successful WSD, as discussed by Navigli (2006) and McGillivray et al. (2023b). In the context of introducing the Latin BERT model, Bamman and Burns (2020) discuss a WSD task framed as a binary classification task, where only the first two major senses are selected for each headword and, thus, the sense to be predicted has to be chosen out of two possible candidates only. Building on their work, Lendvai and Wick (2022) create a new dataset based on a subset of sense representations from the *Thesaurus Linguae Latinae*,[1] and use it to fine-tune Latin BERT on a supervised WSD task. Despite achieving more robust performances, the task remains configured as binary classification, retaining only the first two sense groups for each lemma.

Pivoting a low-resource language to a high-resource one via parallel corpora has been observed to be a valid strategy to obtain WSD annotations in the under-resourced language (Pasini et al., 2021). As the issue of data scarcity applies to Latin as well, Ghinassi et al. (2024) extend such approach to historical languages, leveraging parallel corpora to pivot Latin to English. Propagating WSD annotations from English to Latin then helps tackle the challenge represented by the lack of large sense annotated corpora.

The need for automated WSD has been observed, particularly for historical languages, in light of the increasing size of corpora to annotate and of the subjectivity involved in the intuitive judgment required by sense disambiguation, even more so when native speakers cannot be exploited, as noted by Manjavacas Arevalo and Fonteyn (2022). However, efforts to expedite the annotation process do represent a more general need. For instance, in the context of expanding an event-type ontology Straková et al. (2023) try to exploit fine-tuned LLMs to generate annotation suggestions that could expedite the manual annotation process of verbs to be included in the ontology. Despite not working with a historical language – as their focus is on Czech – their remarks about the necessity of manual post-inspection and annotation of suggestions as an indispensable step can be generalized.

Furthermore, Scarlini et al. (2020) experiment with developing a semi-supervised approach[2] to obtain sense embeddings for lexical meanings within a lexical knowledge base like WordNet. Although their approach does not include Latin and thus cannot be leveraged in our work, it interestingly builds upon the semantic information already carried by contextual word embeddings.

In general – as it provides a comprehensive lexical inventory for the identification of the different word senses – WordNet is a crucial resource for WSD. The current Latin WordNet[3] (WN) (Franzini et al. 2019; Mambrini et al. 2021) is the outcome of an ongoing and substantial revision of the original LatinWordNet (Minozzi, 2010) as initiated within the MultiWordNet project (Pianta et al., 2002). In WordNet, diverse senses of a polysemic word are assigned to distinct synsets. Within the LiLa Knowledge Base (Passarotti et al., 2020), these WN synsets are mapped with valency frames of the valency lexicon Latin Vallex[4], thanks to the shared lexical entries between the two resources. As a result, the Latin Vallex contains not only valency frames but also synset definitions associated to them.

3 In between Latin Vallex and WordNet

Let us delve deeper into the examination of the linguistic resources exploited, and notably Latin Vallex.[5] Nonetheless, speaking of Vallex implies speaking of WordNet as well, as the two resources are interlinked in LiLa (Section 2).

For each lemma, Vallex contains information about the synset definition (taken from WordNet) and the valency frame associated to it. A closer look at the entries immediately reveals how some synsets are semantically close. In many cases, their strikingly similar definitions are not justified by diverging valency frames. Among the many examples, two senses of *porto*, both with frame ACT (Actor), PAT (Patient), are defined respectively as

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

[2] ARES (context-AwaRe Embeddings of Senses).

[3] `https://lila-erc.eu/lodview/data/lexicalResources/LatinWordNet/Lexicon`.

[4] `http://lila-erc.eu/lodview/data/lexicalResources/LatinVallex/Lexicon`.

[5] `https://github.com/CIRCSE/Latin_Vallex2.0`.

definition	synset_id
have on one's person	v#00047745
have with oneself; have on one's person	v#02717102

Three very similar entries are associated to *augeo*, all with the same valency frame ACT, PAT:

definition	synset_id
make strong or stronger	v#00220869
make stronger	v#00222472
make more intense, stronger, or more marked	v#00227165

The examples just mentioned represent instances of extremely high similarity of synset definitions. Although not infrequent, such cases are not the majority. *Metior* can serve as a less extreme example, yet still informative about Vallex/WN granularity; see a list of its 9 synsets, all with frame ACT, PAT:

1. measure (distances) by pacing
2. determine the measurements of something or somebody, take measurements of
3. judge tentatively or form an estimate of (quantities or time)
4. evaluate or estimate the nature, quality, ability, extent, or significance of
5. set, mark, or draw the boundaries of something
6. determine the capacity, volume, or contents of by measurement and calculation
7. travel across or pass over
8. give out as one's portion or share
9. administer or bestow, as in small portions

Although with different nuances, synsets 1-6 all revolve around the concept of *measuring*, being possibly too fine-grained for automatic detection. *Metior* does not represent an isolated occurrence, but a standard entry in Vallex/WN: in light of this consideration, it becomes apparent how Vallex itself poses additional challenges to such task of automatic synset detection.

4 Methodology

In response to the aforementioned need of deriving annotation suggestions for verbal senses, we develop a Predicate Sense Disambiguation (henceforth PSD) workflow leveraging contextual embeddings.[6] As the core of the approach, we try to assess the similarity[7] between the verbal tokens in the target text and those in the reference corpus, with the goal of disambiguating the token sense by virtue of its contextual surroundings. Reference and target corpus[8] are defined based on text paragraphs (reference: par. 1-30 + par. 41-61; target: par. 31-40). The workflow consists of the following steps:

Extracting of verbal tokens. We collect a list of all verbal tokens by extracting them from our source text, i.e., Sallust's *De Coniuratione Catilinae* annotated in the XML-based format Prague Markup Language (PML).[9] The PML files of the treebank are organized by annotation layers and linked to each other through stand-off annotation; we exploit the morphological (lemmatization and morphological tagging) and the tectogrammatical (semantic and pragmatic annotation) layers in combination. We retrieve all verbs by extracting nodes with a valency frame and the required POS.[10]
The extracted verbs are split according to the reference/target corpus partition,[11] and are then manually annotated by a single annotator.

Storing annotated synsets. For each of the extracted tokens in the reference corpus, we store the synset definition that was manually assigned to it. Three cases can occur: i) Most verbs receive a synset from the Latin WN/Vallex, as linked in the LiLa Knowledge Base. For instance, *dominor* in *lubidinem dominandi* 'lust of dominion' is assigned the synset `v#02442106` "be master; reign or rule". ii) When no appropriate synset can be found in the resource, a new one is defined. The definition of the new synset can consist either of an existing WN synset which was not yet assigned to the verb, or of a new definition modeled on a dictionary entry for the verb. E.g., for *vivo* there is no entry in WN; to its occurrence in *alii alio more viventes* 'living with different customs' we assign a new frame with synset `v#02614387` "lead a certain kind

[6] Code is available at https://github.com/fjambe/PSD-Latin-UMR.

[7] Measured in terms of cosine similarity.

[8] Since we are not training any model, we decided not to call them training and test.

[9] The text is available at https://itreebank.marginalia.it/view/download.php as part of the Latin Dependency Treebank (LDT).

[10] Based on the guidelines of the Prague Dependency Treebank, whose annotation the LDT replicates, valency mainly applies to verbs, yet not exclusively. See https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/.

[11] The respective sizes of reference and target corpus are: i) tokens: 13,297 and 1,775 tokens; ii) extracted predicate tokens: 1,787 and 259. The division approximately conforms to a 9:1 ratio, while preserving the paragraph structure of the original work.

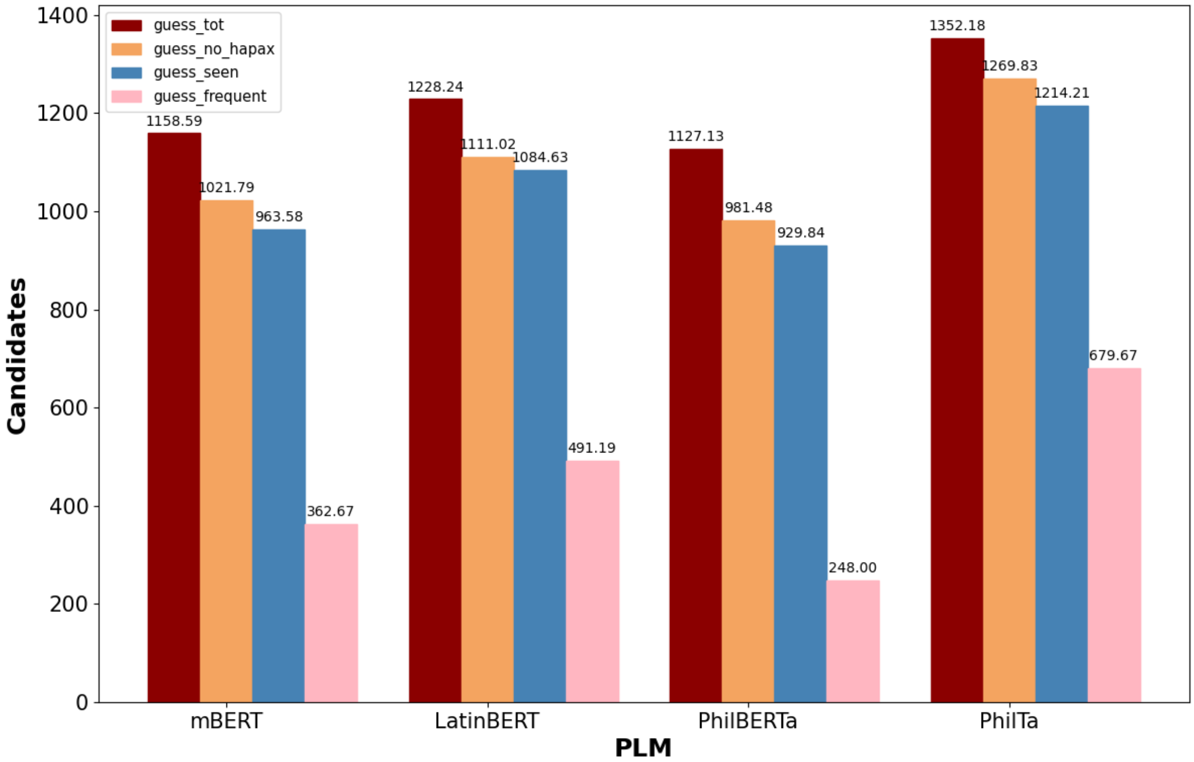

Figure 1: Comparison of different PLMs (mBERT, LatinBERT, PhilBERTa, PhilTa) with lemma constraint. For each of the four defined settings, the number of suggested candidates before retrieving one with same lemma is shown.

of life; live in a certain style". iii) Some tokens lack assigned synsets, as they can be treated as UMR abstract predicates;[12] for instance, the verb *sum* 'to be' can be treated e.g. as *identity-91*, *belong-91*, *have-mod[ification]-91*. We proceed to exclude such tokens from the corpus.

Computing and comparing embeddings. For each verbal token in its respective sentence, both in reference and target corpus, embeddings are computed exploiting the Flair library.[13] We then compute cosine similarity to compare embeddings, and more precisely to quantify the degree of similarity between each target token and each reference token. Similarity scores are then sorted in descending order, so that we can extract the five closest tokens (those with the highest scores — even if the scores are generally low). The synsets of these tokens are then extracted as candidate synsets.

Further constraining candidate tokens. Additionally, we retrieve all the tokens that are extracted as candidates before the first one with the same lemma as the target token[14] is found, i.e. those tokens with higher similarity score than the first one with constrained lemma. As preliminary results did not appear very promising, we decide to apply this additional lemma-based constraint on the candidate extraction. Specifically, we henceforth select as candidates only those tokens which share the lemma with the target token. The same-lemma requirement is merely an artificial constraint intended to facilitate the task, as in a real-case scenario it is possible to derive a correct synset even when the lemma differs. For instance, the synset `v#00406243` "make ready or suitable or equip in advance for a particular purpose or for some use, event, etc." is shared by *pario*, *instituo*, and *facio* among other verbs. In theory, such tokens that share synsets should be retrievable aside from whether they share the same lemma or not. Yet, the necessity of defining a simplified scenario through the imposition of a lemma constraint becomes ap-

[12] UMR features 9 types of abstract predicates, used to represent predication of properties, possession, location. They are identified by special labels serving as artificial lemmas and have their own roleset. For example, *identity-91* has an ARG1 role for the theme, and ARG2 for the equated referent.

[13] `https://flairnlp.github.io/`. We employ Transformer embeddings with default arguments; we only choose a different pooling operation to generate the final token representation from subwords – for which we select `mean`, calculating a `torch.mean` over all subword embeddings.

[14] By target token we mean the token to be annotated.

parent from the initial results of the experiments.

Output. As a result, the output file provides all retrieved information about each token: five annotation suggestions; i.e. the most plausible synsets; the number of incorrect guesses before suggesting a token with the same lemma;[15] the list of lemmas retrieved before a correct one was found.

4.1 Pretrained LMs for Embeddings

The following pretrained language models have been exploited to produce embeddings:

- mBERT (Devlin et al., 2018): multilingual BERT model (base, cased) pre-trained on 104 languages including Latin.
- Latin BERT (Bamman and Burns, 2020): pre-trained on 642.7 million words from a variety of sources spanning the Classical era to the 21st century.
- PhilBERTa (Riemenschneider and Frank, 2023): RoBERTa (Liu et al., 2019) model, pre-trained on Latin, Ancient Greek, and English, and tailored for classical philology (like PhilTa).
- PhilTa (Riemenschneider and Frank, 2023): T5 (Raffel et al., 2020) model, pre-trained on Latin, Ancient Greek, and English.

5 Evaluation

In this section we present and discuss a comparison between outputs yielded by different PLMs (Subsection 4.1), with respect to various criteria. Additionally, we manually evaluate a subset of the target corpus so as to complement the evaluation metrics with a qualitative analysis.

5.1 Quantitative Analysis

OOV. A key observation concerns out-of-vocabulary predicates, i.e. verbs that occur in the target corpus only. The amount of such verbs, for which a candidate with same lemma cannot be retrieved, is considerably high (20%). The percentage of target predicates whose lemma occurs only once in the reference corpus is quite high as well (13.7%). These figures would strongly argue against the constrained-lemma setting, when only candidates with the same lemma as the target token are retrieved. However, as mentioned before, the constraint on the lemma was deemed reasonable since preliminary results did not seem promising.

Criteria. We identify four criteria to extract some patterns from the data (see Figure 1). For all four metrics, lower scores are indicative of better performance.

1. `guess_tot`: average number of suggested candidates before retrieving one with the same lemma.
2. `guess_no_hapax`: average number of suggested candidates before retrieving one with same lemma, excluding *hapax legomena*.[16]
3. `guess_seen`: average number of suggested candidates before retrieving one with same lemma, considering only lemma-synset pairs which occur in the reference corpus. In other words, we try to observe what happens when evaluating only cases where there was a chance that the synset could have been guessed correctly. The results of this artificially simplified setup will be analyzed in greater depth also with respect to retrieval of synsets, by exploiting such a controlled setup to lift the lemma constraint and evaluate retrieval of synsets instead of lemmas.
4. `guess_freq`: average number of suggested candidates before retrieving one with same lemma, computed only on the 10 most frequent lemmas[17] of the whole corpus.

In light of the criteria defined, and assuming their representativeness, we observe how PhilBERTa tendentially performs best in all settings, while the worst results are achieved with PhilTa. A pattern emerges when progressively limiting the evaluation scope to ‘known’, i.e. more frequent, predicates: all four PLMs output slightly improved results, highlighting the effect of frequency on such a task. Specifically, the number of retrieved candidates before finding one with shared lemma is highest in case of overall evaluation, and it gradually decreases first when *hapax* are excluded, then when only lemma-synset pairs occurring in the reference corpus are considered, and finally when the evaluation is limited to the 10 most frequent verbs. In particular, the `guess_frequent` setting seems to impact results to a greater extent, as the number of retrieved candidates is here conspicuously lower.

[15] Of course, the fact that the lemma is shared does not guarantee that the sense is shared as well.

[16] Lemmas occurring only once, namely only in the target corpus.

[17] *Facio* ‘make’, *dico* ‘say’, *video* ‘see’, *paro* ‘prepare’, *fio* ‘become’, *do* ‘give’, *cognosco* ‘know’, *coepio* ‘begin’, *capio* ‘take’, *valeo* ‘be strong’. *Sum* ‘be’ and *habeo* ‘have’ have been discarded as they often correspond to UMR abstract concepts.

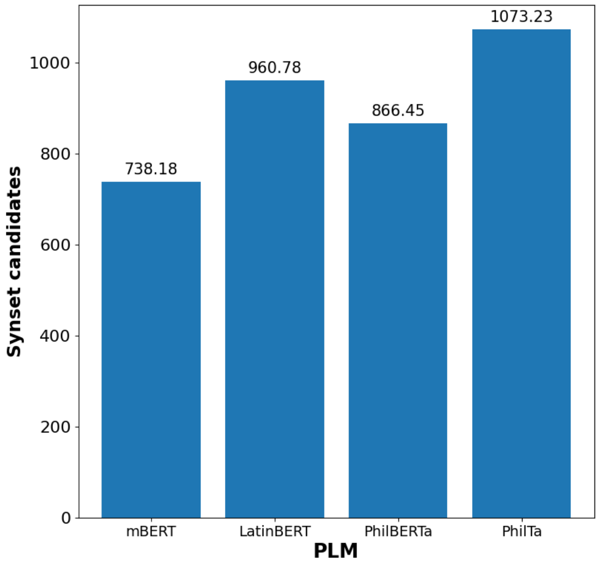

Figure 2: Evaluation of different PLMs (mBERT, LatinBERT, PhilBERTa, PhilTa) in synset retrieval. The *y* axis reports the number of candidates suggested before retrieving the correct synset, without any lemma constraint and by considering only lemma-synset pairs occurring in the reference corpus.

In addition to the evaluation settings based on lemma constraint, we then design an artificially simplified setting to analyze how PLMs behave when retrieving the correct synset without being limited by shared lemma. As mentioned when presenting the `guess_seen` evaluation criterion, in this controlled setup we focus only on lemma-synset pairs which occur in the reference corpus, excluding from the evaluation all those that do not meet this requirement. A similar setup should allow to investigate actual performances without being overly affected by data scarcity. In principle, it should be possible to retrieve tokens sharing the same synset regardless of whether they share the same lemma, as explained through the example of *pario*, *facio*, *instituo*, all sharing the synset `v#00406243` (Section 4). However, Figure 2 highlights how the number of attempts before a correct guess is still very high. The pattern is similar to what already observed when constraining on lemma, with PhilTa performing the worst. Yet, here PhilBERTa and multilingual BERT are inverted, with the latter resulting to be the model that on average needs the lowest number of attempts before a correct one.

5.2 Manual Evaluation

To further investigate the performance of the models, we also conduct a manual evaluation of a subset of the results. As a sample, we extract the first 20 predicates that occur in the target text. We first assess how the models perform on this subset within the default lemma-constrained setting (`guess_tot`). We ignore the number of attempts before retrieving the correct lemma, as it is already reflected by evaluation metrics, and focus on the assignments of synsets given a shared lemma. Results are presented in Table 1, to be interpreted in the following way: `1/2` means that two synset candidates are retrieved by the model (given a constrained lemma), and the first out of the two is the correct one based on manual annotation. `1=2/2` implies that two candidates are retrieved, and that they are identical and both correct, while `0/n` means that none of the `n` retrieved candidates is correct. `1=n=5/5` corresponds to a situation where all five retrieved candidates are identical and correct.

The analysis of results shows that the models' performances do not differ substantially one from another in the defined setting. Lemmas for which none of the retrieved candidates are correct (e.g. `0/5` in the table) can be explained by the fact that the sense they have been manually annotated with never occurs in the reference corpus, either at all or in association to that specific lemma. It is e.g. the case of *credo* 'to believe' and *moveo* 'to move', despite both being quite frequent verbs. The same happens with *diffido* 'to distrust'; the sense observed in the target corpus (`v#00687926`, "regard as untrustworthy; regard with suspicion; have no faith or confidence in") never occurs in the reference corpus. In this way, even a classification that should be relatively simple -— like the binary classification of *diffido*, for which only two senses are stored in Latin WordNet – fails. In the case of *permota*, from *permoveo* 'to stir up', we can observe the similarity of definitions that was already discussed in Section 3, as the sense definitions of retrieved candidates are highly similar: "move deeply" and "disturb in mind or make uneasy or cause to be worried or alarmed" (retrieved twice).

The case of *gerere*, from *gero* 'to manage', offers interesting insights as well, since all the five retrieved candidates are assigned the same sense "direct the course of; manage or control". Such cases of candidates leading to the same sense suggestion could probably be grouped, in order to inves-

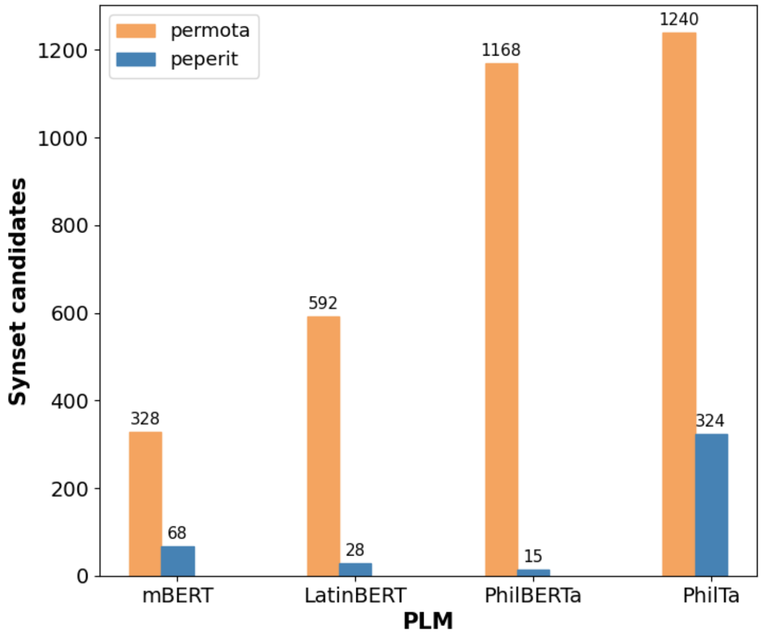

Figure 3: Evaluation of different PLMs in synset retrieval on two examples (*permota*, *peperit*).

token	lemma	hapax	**mBERT**	**Latin BERT**	**PhilBERTa**	**PhilTa**
permota	*permoveo*		1=3/3	2=3/3	2=3/3	2=3/3
pepererat	*pario*		1/2	2/2	2/2	2/2
invasit	*invado*		2/3	3/3	1/3	3/3
festinare	*festino*		1=2/2	1=2/2	1=2/2	1=2/2
trepidare	*trepido*	x				
credere	*credo*		0/5	0/5	0/5	0/5
gerere	*gero*		1=n=5/5	1=n=5/5	1=n=5/5	1=n=5/5
metiri	*metior*	x				
incesserat	*incedo*		1=2/2	1=2/2	1=2/2	1=2/2
adflictare	*afflicto*	x				
tendere	*tendo*		0/1	0/1	0/1	0/1
miserari	*miseror*		0/1	0/1	0/1	0/1
rogitare	*rogito*	x				
pavere	*paveo*	x				
adripere	*arripio*	x				
omissis	*omitto*		2/2	2/2	2/2	2/2
diffidere	*diffido*		0/1	0/1	0/1	0/1
movebat	*moveo*		0/5	0/5	0/5	0/5
parabantur	*paro*		1=3=5/5	1=2=3=5/5	1=2=3=4/5	1=2=3=5/5
interrogatus	*interrogo*		0/2	0/2	0/2	0/2

Table 1: Manual assessment of PLMs' performances (with lemma constraint).

tigate whether additional and different senses are retrieved after the main one; then, retrieved sense suggestions could possibly be weighted by the number of times they are proposed. However, in this specific case in the reference corpus we can find ten occurrences of the verb *gero*, all assigned that same sense. The effect of frequency can be observed with *gero* in the number of total guesses before a token with the same lemma is retrieved: 459 for PhilBERTa and 304 for mBERT, considerably lower than the average number (Figure 1).
Hapax legomena, marked as such in Table 1, have been set aside also in the manual evaluation, as the lemma-constrained setting inevitably prevents the retrieval of any candidate.
Overall, what emerges from Table 1 is that no PLM consistently outperforms the others, with all models exhibiting similar performance within the defined setting.[18]

Within the proposed manual assessment, we also evaluate the sub-task of synset retrieval. Let us take again the token *permota*[19] as an example. mBERT and PhilBERTa, the two models that have proved to perform better, take respectively 328 and 1168 guesses before retrieving the correct synset. Their performances differ substantially here, with mBERT outperforming PhilBERTa by much. Nonetheless, the synset definitions of the first 5 out of the 328 candidates suggested by mBERT are sufficient to highlight the absence of a clear, reliable rationale in such retrieval, as they appear uncorrelated: "give a certain impression or gave a certain outward aspect", "enter or assume a certain state or condition", "from a critical opinion of", habitually do something (used only in past tense)", "have with oneself; have on one's person".

Moreover, deriving discernible patterns from the outputs of PLMs presents considerable challenges (see Figure 3). In the case of *permota*, beside performances by mBERT and PhilBERTa, we observe the number of guesses by PhilTa and Latin BERT amounting to 592 and 1240 respectively – not totally consistently with the pattern observed e.g. in Figure 1. However, if we take into account the second token of the target corpus, i.e. *peperat* from *pario* with the meaning of "cause to happen, occur or exist", the number of suggestions before retrieving the correct sense does not mirror what has been observed so far (PhilBERTa: 15 suggested candidates; Latin BERT: 28; mBERT: 68; PhilTa: 324). Once again, it is hard to interpret why specific senses associated to candidate suggestions are retrieved. For instance, mBERT retrieves the following: 1) "be willing to concede", 2) "spur on", 3)"impose a penalty on; inflict punishment on", 4) "confess to a punishable or reprehensible deed, usually under pressure", 5) "take or capture by force". PhilBERTa, i.e. the model with lowest retrieval score in this specific case, outputs these candidates: 1) "make a solicitation or entreaty for something; request urgently or persistently", 2) "order, request, or command to come", 3) "get to know or become aware of, usually accidentally", 4) "assign a specified (usually proper name) proper name to", 5) "decide with authority". Not only their similarity to the actually assigned one ("cause to happen, occur or exist") is irrelevant, but the two sets of candidates do not look mutually similar in any way.

6 Conclusions

The complexity of the task has been apparent from the beginning, and is confirmed by observations from related studies. Bamman and Burns (2020) already discuss comparable challenges, emphasizing the inherent difficulty of the WSD task and the lack of suitable resources for Latin – an observation also echoed by Keersmaekers et al. (2023). In light of such complexity, our study was never truly conceived as a solution to a specific task, but rather as a qualitative assessment of the available resources as well as of the results they can lead to. Therefore, our main objective revolved around a thorough examination of the task, its objectives, and challenges, with the intention of critically analyzing and identifying realistic possibilities within the constraints of the available resources. One of the key questions concerned whether we can actually exploit available resources: in particular, can Latin Vallex represent a suitable resource for PSD? At its present stage, its exploitation for PSD does not appear to be feasible; its fine-grained granularity definitely presents challenges for this specific task. Nevertheless, adopting a binary classification approach, as suggested by previous works (Bamman and Burns, 2020; Lendvai and Wick, 2022), may not offer a satisfactory solution either. As an

[18] It is important to note that these results may be influenced by the limited sample size.

[19] Occurring in the sentence *Quibus rebus permota civitas atque inmutata urbis facies erat* (Sall., *De Coniuratione Catiline* XXXI), translated as "By such proceedings as these the citizens were struck with alarm" in Perseus, at https://www.perseus.tufts.edu/.

illustrative example, the verb *postulo* demonstrates the need for at least three distinct frames, even under a coarse-grained granularity: i) 'to ask, demand, require' (ACT, ADDR, PAT); ii) 'arraign before a court, to prosecute, accuse' [juridical] (ACT, PAT, REG); iii) 'to contain, measure' [of things] (ACT, PAT). Currently, Latin Vallex/WN provides nine frames for *postulo*. The granularity of Latin Vallex and the simplicity of a binary classification demand a thoughtful exploration of alternative strategies to address such challenges. A possibility could be represented by sense clustering, as described e.g. by Navigli (2006) and Martelli et al. (2022).

Additionally, an important limitation of the study arises from the decision not to fine-tune PLMs, whose performances would most probably be enhanced through fine-tuning. However, fine-tuning requires training data, and the annotated dataset currently at our disposal is of limited size. The quantitative results, as illustrated in Figure 1, clearly highlight the substantial impact of the limited amount of available data on results. Therefore, what can be also inferred from the present study is the need for a larger reference corpus, to be obtained by enlarging the existing dataset with additional data.

An envisioned extension to the presented workflow involves the computation of sentence embeddings for definitions. Without constraining either on same lemma or on same synset, and thus handling even OOV cases, cosine similarity could be leveraged to identify the most probable synset by comparing all the synset definitions associated to the target token against the synset definition of the extracted candidates, to find the most similar one(s). In other words, embeddings for the synset definition of retrieved candidates could be generated, as well as for the list of synset definitions as available in Vallex/WN for the lemma under scrutiny. We could then select candidate synset definitions by computing cosine similarity between all synsets associated in Vallex/WN to the target lemma and synsets of the extracted candidate tokens in the reference corpus, in order to be able to deal not only with synsets shared by verbs with different lemma, but also with synsets that do not occur in the reference corpus. However, we expect the issues encountered so far (to name one, the dataset size) to pose similar challenges even in this further-defined setting.

Acknowledgments

This work has been supported by the Charles University, project GAUK No. 104924. It has also been partially supported by the SVV project number 260 698 and by the grant *Language Understanding: from Syntax to Discourse* (No. 20-16819X) of the Czech Science Foundation.

The project has been using data and tools provided by the *LINDAT/CLARIAH-CZ Research Infrastructure* (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Viviana Ventura, Federica Zampedri, et al. 2019. Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Bari, Italy.

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.

Alek Keersmaekers, Wouter Mercelis, and Toon Van Hal. 2023. Word Sense Disambiguation for Ancient Greek: Sourcing a training corpus through translation alignment. In *Proceedings of the Ancient Language Processing Workshop*, pages 148–159.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. Interlinking valency frames and wordnet synsets in the LiLa knowledge base of linguistic resources for Latin. In *Further with Knowledge Graphs*, pages 16–28. IOS Press.

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 123–134, Taipei, Taiwan. Association for Computational Linguistics.

Paola Marongiu and Barbara McGillivray. 2023. Preliminary guidelines for manual annotation of word senses in Latin and ancient Greek corpora.

Federico Martelli, Marco Maru, Cesare Campagnano, Roberto Navigli, Paola Velardi, Rafael-J. Ureña-Ruiz, Francesca Frontini, Valeria Quochi, Jelena Kallas, Kristina Koppel, Margit Langemets, Jesse de Does, Rob Tempelaars, Carole Tiberius, Rute Costa, Ana Salgado, Sanni Nimb, Sussi Olsen, Simon Krek, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, and Tina Munda. 2022. *D3.8 Lexical-semantic analytics for NLP*. ELEXIS - European Lexicographic Infrastructure.

Barbara McGillivray. 2021. Latin lexical semantic annotation. King's College London. DOI: `10.18742/16974823.v1`.

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

Barbara McGillivray, Iacopo Ghinassi, and Paola Marongiu. 2023b. The challenges of sense granularity in word sense disambiguation for Latin. In *La memoria digitale: forme del testo e organizzazione della conoscenza. Atti del XII Convegno Annuale AIUCD*, pages 279–283, Siena, Italy. AIUCD.

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. 2010. The Latin WordNet project. In *Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik*, pages 707–716, Innsbruck. Institut für Sprachen und Literaturen der Universität Innsbruck.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia. Association for Computational Linguistics.

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13648–13656.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa knowledge base of linguistic resources for Latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Frederick Riemenschneider and Anette Frank. 2023. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539, Online. Association for Computational Linguistics.

Jana Straková, Eva Fučíková, Jan Hajič, and Zdeňka Urešová. 2023. Extending an event-type ontology:

Adding verbs and classes using fine-tuned LLMs suggestions. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 85–95, Toronto, Canada. Association for Computational Linguistics.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. 2021. Designing a Uniform Meaning Representation for Natural Language Processing. *KI-Künstliche Intelligenz*, 35(3-4):343–360.