From Lemmatization to Legal Terminology: Assessing an Hybrid Pipeline on Justinian's *Digest*

Paola Marongiu, Eva Sassolini
Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche (CNR-ILC)
Via Giuseppe Moruzzi 1, Pisa
{paola.marongiu, eva.sassolini}@ilc.cnr.it

Abstract

This paper evaluates a hybrid NLP pipeline for supporting the extraction of Roman legal terminology from Justinian's Digest. Our goal is not to optimize lemmatization in isolation, but to assess whether integrating a Large Language Model (GPT-4o-mini) as a post-processing component improves lemma quality in ways that are critical for downstream glossary construction. Using LatinPipe as a baseline (F1 = 95.05), we test the integration of GPT-4o-mini under three experimental settings (zero-shot with and without prior lemma information, and few-shot prompting) against a manually annotated gold standard of 3,703 sentences and an expert-validated list of legal Latin technical terms. Results show improvement across all settings, with the best performance achieved in the few-shot configuration. Our analysis shows that the hybrid configuration produces selective improvements, significantly more likely for frequent lemmas and verbs forms, suggesting that the LLM layer primarily assists in resolving morphologically ambiguous inflected forms. Although our experimental conditions may not hold in real-world scenarios, we argue that the main contribution of this work is methodological: demonstrating how evaluation can be aligned with downstream terminological goals, rather than proposing a general-purpose solution to domain-specific lemmatization.

Keywords: Latin, lemmatization, LLM, GPT, Digest, legal text

1. Introduction

Lemmatization is a fundamental task in Natural Language Processing (NLP), which consists in assigning a word form to its canonical citation form, called lemma. Having reliable lemmatization results is crucial for performing more advanced NLP tasks, such as syntactic and semantic parsing, terminology extraction, and in general for building language resources and carrying out different types of linguistic analyses. This paper focuses in particular on the preparatory steps to the construction of a digital bilingual glossary extracted from the parallel corpus of the Justinian's *Digest*. This goal is framed as part of a broader project aimed at producing an Italian translation of the *Digest*, and extracting linguistic resources that can work as translation memories for the translators and, at the same time, provide new data for research in historical linguistics, among others. The glossary will be built using Latin as the pivot language for the term extraction. For this reason, it is essential that the lemmatization of the Latin corpus be of exceptionally high quality, especially concerning Latin legal terms.

The Latin Digest of Emperor Justinian represents a particularly demanding corpus for NLP. Compiled between 529 and 534 CE as part of the Corpus Iuris Civilis (Banchich et al., 2015; Dingledy, 2016), the Digest gathers excerpts from 1,528 writings of 39 Roman jurists, covering roughly eight centuries of legal thought—from the Twelve Tables (449 BCE) to late imperial jurists of the fourth century CE (Dingledy, 2016; Ribary and McGillivray, 2020). The text is divided into fifty books, each subdivided into titles, laws, and paragraphs, and introduced by inscriptions that identify the author and source of each excerpt (Banchich et al., 2015). As the most extensive collection of Roman legal doctrine, the Digest is invaluable not only for legal historians but also for scholars interested in Latin lexicon and morphology within a highly specialized domain. The legal register of the Digest displays terminological density, recurrent formulae, and syntactic constructions that diverge from classical Latin (CL) norms, all of which pose significant challenges for computational processing.

From a linguistic and philological perspective, developing accurate lemmatization for this corpus is a prerequisite for building comprehensive lexical resources, such as domain-specific glossaries or bilingual lexicons of legal Latin. However, manual lemmatization of such an extensive and complex corpus would be prohibitively time-consuming. Automatic or semi-automatic solutions are therefore indispensable for efficiently supporting expert validation and downstream tasks. The nature of the Digest as a Latin domain text brings specific challenges to the task. Latin is a historical language, usually defined as a low-resource

Proceedings of LT4HALA 2026 @ LREC 2026, pages 418–428.

language,[1] morphologically rich, subject to orthographic variation. Moreover, specialized corpora have an exceptional presence of domain-specific vocabulary, including lemmas that traditional NLP tools might not have seen as much during training.

Building upon recent advances in the field of Artificial Intelligence, this work explores strategies to enhance automatic lemmatization for specialized Latin corpora. By testing different approaches and experimental settings, and validating results against manually annotated samples, we aim to assess the combination of state-of-the-art NLP tools and Large Language Models (LLMs), to apply them to historical and domain-bound texts.

2. Related work

Lemmatization of Latin has been explored through multiple paradigms, ranging from rule- and dictionary-based lexical analyzers to modern neural pipelines. Most recently, the three EvaLatin editions in 2020 (Sprugnoli et al., 2020), 2022 (Sprugnoli et al., 2022) and 2024 (Sprugnoli et al., 2024) have promoted the development of computational resources and tools for Latin NLP. EvaLatin 2020 was specifically devoted to lemmatization and PoS-tagging. The task was divided in three different subtasks, one for each dataset: CL, cross-genre (prose vs. poetry) and cross-time (authors from the 1st century BC to the 1st century CE vs. Thomas Aquinas, an author from the 13th century CE). The best tool across all sub-task was UDPipe (Straka and Straková, 2020), which scored from F1 87.69 (cross-time) to 96.74 (Classical Latin). An evolution of UDPipe, LatinPipe (Straka et al., 2024), was presented to the EvaLatin 2024 campaign, which focused on morphosyntactic analysis and sentiment analysis in Latin. LatinPipe is a graph-based dependency parser born as the evolution of its predecessors UDPipe (Straka and Straková, 2017) and UDPipe 2 (Straka, 2018). This parser leverages pre-trained language models, fine-tuned on five Latin treebanks publicly available in Universal Dependencies (Nivre et al., 2020), version 2.13, and two texts annotated in UD style (see (Straka et al., 2024) for further details):

- ITTB (Passarotti, 2019)
- LLCT (Cecchini et al., 2020a)
- PROIEL in UD 2.13 (Haug and Jøhndal, 2008) and in the UD-style harmonized version (Gamba and Zeman, 2023)
- UDante (Cecchini et al., 2020b)
- Perseus (Bamman and Crane, 2011)
- *De Latinae Linguae Reparatione* by Marcus Antonius Sabellicus (Gamba and Cecchini, 2024)
- the Latin translation of *The Spirals* of Archimedes by Jacopo da San Cassiano, part of the *Archimedes Latinus* UD-style treebank (Fantoli and de Lhoneux, 2022)

Two variants of LatinPipe ranked first and second in both EvaLatin 2024 shared tasks, demonstrating its effectiveness in Latin dependency parsing. The performance of the tools participating in the task was tested against two authors, one for prose (Tacitus' *Germania*), and one for poetry (Seneca's *Hercules Furens*).

The potential of LLMs has been tested on diverse labeling tasks, e.g. event detection and extraction (Chen et al., 2024), discourse functions in spoken corpora (Petukhova and Kochmar, 2025), caused-motion construction (Weissweiler et al., 2025). Most research has focused on modern languages, but a few studies have explored applications on historical languages too. LLMs have been tested on Named Entity Recognition in Historical Urdu (Irfan and Ali, 2025); (Volk et al., 2024) have tested GPT-4 in translation and text summarisation tasks on Late Latin texts; (Farina et al., 2025) tested open and closed-weight models in the detection of semantic features such as spatial relations for Latin motion verbs.

To our best knowledge, little to no research has been specifically devoted to assessing the performance of LLMs on lemmatization, especially for historical languages. Toporkov et al. (2025) recently tested different LLMs in the lemmatization of a selection of modern languages with different levels of morphological complexity: English, Spanish, French, German, Italian, Finnish, Icelandic, Turkish, Swedish, among others. In their work, they compared encoder-based models fine-tuned on language-specific gold data with LLMs in in-context lemmatization for multiple languages, demonstrating that while fine-tuned models still achieve superior performances, LLMs can yield competitive results through few-shot prompting without prior fine-tuning. To our knowledge, the only work on lemmatization involving LLMs in an historical language is Riemenschneider (2025), which focuses on Akkadian. At the beginning of the lemmatization pipeline, three different systems are combined and asked to provide their own lemma for each form in the text. When they do not agree on the predicted lemma, the LLM is provided with the three predictions and the passage in which the word form occurs, and is asked to provide the correct lemma.

[1] Though it should be noted that among historical languages Latin is certainly the luckiest one in terms of resources for NLP.

Recent advances in the field have expanded the possibilities offered by deploying LLMs as standalone systems, by integrating them with other computational tools and architectures to form hybrid pipelines. For instance, recent work on Latin word sense prediction combines Large Language Models with Linguistic Knowledge Graphs (LKGs) in a Graph Retrieval-Augmented Generation framework, leveraging structured semantic and contextual information from multiple resources (including Latin WordNet and Wikidata) to support word sense disambiguation. Results show that while LLMs achieve competitive performance on their own, the integration of graph-based metadata can yield measurable improvements, particularly for larger models, highlighting the potential of structured knowledge to guide LLM reasoning in low-resource historical settings (Ghizzota et al., 2026). Similarly, hybrid architectures have been successfully applied to morphologically complex and low-resource languages, such as in Jungar Tuvan, where applying an LLM at the end of the pipeline significantly improves automatic glossing over the neural baseline (Liang et al., 2026). These approaches demonstrate that integrating structured or model-based linguistic processing with LLMs can enhance performance, especially in contexts characterized by morphological complexity and limited annotated data.

As described in this section, a lot of work has been done in Latin NLP, and a few initiatives have encouraged progress in this field. In parallel, the fast evolving landascape of NLP and AI offers new means of approaching old issues. More specifically to the focus of our paper, lemmatization of highly specialized texts e.g. legal documents in our case, remains under-explored, as well as experiments on the combination of state-of-the-art tools with the most recent advances in AI in specialized lemmatization tasks. In this paper, we explore this direction with the practical goal in mind of achieving high lemmatization performances to facilitate the development of our digital glossary of legal Latin.

3. Data and methodology

3.1. Latin data

The legal scholars working on the translation of the *Digest* use as a reference edition for the Latin text the *editio stereotypa duodecima* edited by Mommsen and Krüger (1911), widely adopted by the research community dealing with Justinian law and its translations (Schipani, 2005, 555–557). Therefore, the text in our corpus is largely based on this version of the Digest. However, as they proceed with the translation, they sometimes modify the text of the reference edition, for various reasons, explaining the implemented changes in a footnote in the published volume. For this reason, although the text of the *Digest* is publicly available online, we process the files as we receive them from the translators.

As the translation process is still ongoing, we worked with the books that had already been translated at the time of the experiments. Therefore, we processed 36 of the 50 books in the Digest. The corpus as of now thus contains around 635k tokens, with around 37,000 sentences.

Before automatic annotation, the texts were pre-processed to lowercase characters, separate punctuation, and remove non-standard characters e.g. diacritics. Numerical indications for each paragraph were also removed, and then re-assigned in the CoNLL-U file during post-processing. Lemmatizing the Digest implies various challenges, due to the following reasons. The first reason has already been addressed in the first two sections, namely an extremely specialised lexicon which might not have been seen by the models during training. Most of the annotated data publicly available for Latin are CL or Late Latin texts, poetry or prose, but not specialised or domain-bound texts such as legal or medical literature. The second issue is that although the Digest is mostly written in Latin, it also contains Ancient Greek (AG) script, which in some cases is very extensive, e.g. the 27th book is almost entirely written in AG. Finally, the text presents some *lacunae*, due to the fact that it was not transmitted to us in its entirety. Therefore, sometimes the text contains some gaps, interrupted sentences and similar textual and philological issues which might represent a challenge for automatic annotation.

Group	Thematic scope	Books
1	General principles	I–IV
2	Property and real rights	V–XI
3	Obligations and contracts	XII–XIX
4	Family law	XX–XXVII
5	Testate succession, possession, praetorian succession, and criminal law	XXVIII–L

Table 1: Thematic groups in the Digest corpus.

3.2. Gold Standard

As mentioned in section 3, the corpus comprises a total of around 37,000 sentences, from which we built a gold standard set consisting of 3,703 sentences (10% of the total). In order to ensure both

thematic and quantitative balance, the sentences were sampled proportionally across the five major thematic groups into which the Digest is traditionally divided (see Table 1). As we only have access to books 1–36 for now, we based the extraction on groups 1 to 5. The number of sentences extracted from each group was determined according to their relative size in the corpus. Formally, if S_n denotes the number of sentences in group n and S the total number of sentences in the corpus, then the number of sentences selected from group n is given by $0.1 \times S_n$. This guarantees that each thematic group contributes to the golden dataset in direct proportion to its representation in the corpus, while keeping the total gold size fixed at 3,703 sentences. The resulting distribution is shown in Table 2.

Group	N° of sentences (S_n)	Gold size (10%)
1	5,410	541
2	5,990	599
3	6,903	690
4	7,295	729
5	11,455	1,145

Table 2: Distribution of corpus and gold sentences by thematic group.

After the pre-processing phase, the sampled sentences were lemmatised and tagged using LatinPipe. This tool was selected among others due to its recent success in the latest EvaLatin campaign, described in Section 2. We manually corrected the output of LatinPipe, focusing on lemmas and PoS, to obtain our gold standard. Our gold standard dataset, as well as the code to run our experiments, is freely available on Github (see section 8).

3.3. List of legal terms in the *Digest*

In order to start building our digital glossary, we prepared a list of candidate lemmas to be validated by our team of Roman legal scholars. In order to build the list of candidate terms, we compared the *Digest* corpus with a general-interest corpus, developed by the LASLA laboratory in Li'ege (Labora toire d'Analyse Statistique des Langues Anciennes)[2] and recently linked to the LiLa knowledge base (Fantoli et al., 2022, 2024). Specifically, for the initial word list we combined information about i) unique lemmas in the *Digest* ii) terms that appeared more frequently in the *Digest*, and iii) measured the semantic variation of the shared vocabulary between the *Digest* and the LASLA corpus. It should be noted that the comparing these two corpora does not come without risks: the LASLA corpus only contains texts from the Classical period, whereas the *Digest* dates back to the fourth century CE. However, we opted for this corpus instead of other resources for the excellent quality of its lemmatization, and ii) because the manual validation by the domain experts would anyway exclude terms emerging from the analysis because of the absence of temporal overlap. The final list, validated by the experts, contains 583 terms belonging to the legal domain, comprising 287 verbs, 264 nouns, and 32 adjectives.[3]

We first analysed the distribution of the technical terms in our gold dataset as a reference, to determine their frequency across the 3703 sentences. This step was necessary for performing in-depth evaluation of the performances of our hybrid pipeline on the technical terms specifically. In order to obtain this, we first counted both the number of occurrences for each term in the dataset, and the number of sentences in which each term occurs (noting that more than one term can occur in the same sentence). Then, we divided the list of terms into three ranges, depending on their distribution in our gold dataset: 'frequent', 'medium', 'rare'. The last category is 'absent', for technical terms that do not appear in our gold standard and therefore cannot be used for evaluation. To determine the three frequency ranges we adopted a data-driven approach: we calculated empirical quantiles based on the distribution of the technical lemmas in the corpus. Specifically, we computerd the 33rd and 67th percentiles over lemmas with at least one occurrence, and used these values as data-driven thresholds to build three balanced frequency ranges: rare lemmas have ≤ 10 occurrences, medium lemmas have 11–28 occurrences, and frequent lemmas have > 28 occurrences.

This decision was based on the fact that most of the technical legal terms in our list are rare in the gold dataset, and the very frequent ones are only a few. The highly skewed distribution of technical lemmas is shown in the barplot in Figure 1. Figure 2 shows the same distribution but cut up to 200 occurrences, for better readability of the first two ranges. Having long-tail distributions of low-frequency words with few high-frequency words is widely attested in natural language texts (Baroni, 2009, 810);(Zipf, 1949). In our specialized corpus, we observe a similar behaviour for our list of technical words.

[2] https://www.lasla.uliege.be/cms/c_8508894/fr/lasla

[3] It should be noted that the list only contains single lemmas. The list of candidate multi-word expressions is currently being evaluated by the domain experts.

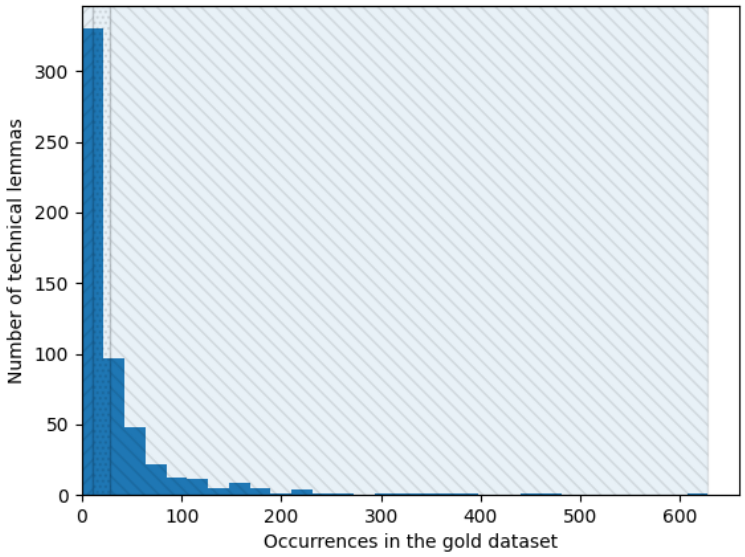

Figure 1: Frequency distribution of validated technical lemmas. Shaded regions represent quantile-based frequency bands: rare (≤10 occurrences), medium (11–28), and frequent (>28).

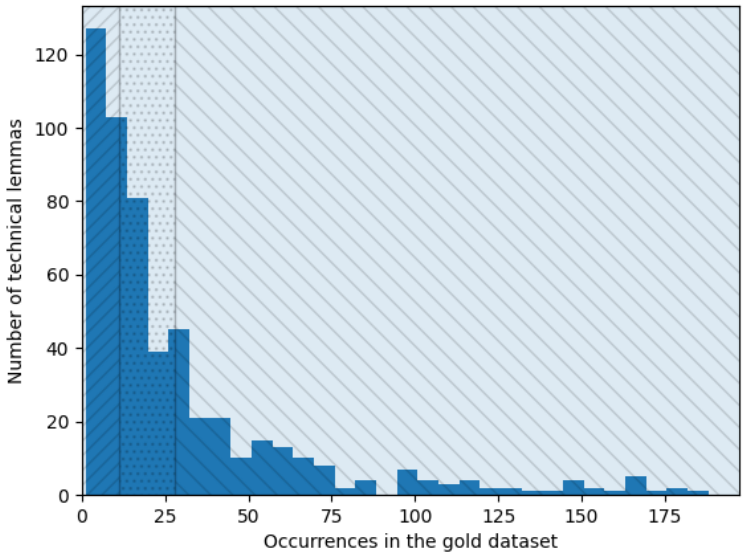

Figure 2: Frequency distribution of validated technical lemmas up to 200 occurrences.

4. Models and methodology

As mentioned in the previous sections, we obtained our baseline with LatinPipe.

The tool has a F1 score of 95.05 on the lemmatization of the Digest, which certainly leaves room for improvement. On a closer qualitative analysis of the results, we noticed that the model consistently missed lemmas like *pignus* 'pledge, security', often lemmatizing it as *pignor* and similar forms. This represents a strong limitation to the use of this tool to us, because technical words in the legal domain like *pignus* cannot be missed when pursuing the goal of extracting a specialised glossary from the corpus. Therefore, we combined the output of LatinPipe with GPT-4o-mini, to see if and how the baseline increases, and on which lemmas.

First, we compare our gold with the output of LatinPipe to spot wrongly predicted lemmas. We pass them over to GPT, with different prompt instruc-

tions (see following sections). Then, we replace the wrong lemmas in the CoNLL-U file produced with LatinPipe with the output of GPT, and perform the evaluation against our gold standard again.

4.1. Experimental settings

We tested GPT-4o-mini against the baseline obtained with LatinPipe in three different experimental settings.

Zero-shot 1. We provide the lemma as predicted by LatinPipe, and ask the model i) to evaluate if the predicted lemma is correct or not, and ii) if it is not correct, to provide the correct lemma. Below is the prompt used in this first experimental setting.

```
You are a Latin expert. I
will give you a Latin word
and the context sentence. The
lemma for this Latin word as
predicted by LatinPipe is:
'predicted lemma'
Instructions:

  • Check if the predicted
    lemma is correct
  • If it is correct give the
    same lemma in output
  • If it is wrong, correct it
  • Give me back ONLY THE
    CORRECT LEMMA, as a single
    Latin word
  • No additional comments or
    explanations

Word:
Context:
Output:
```

Zero-shot 2. We give the model the same prompt given in zero-shot 1, but we do not provide the lemma predicted by LatinPipe. This allows us to understand if providing the predicted lemma helps or hinders the accuracy of the prediction made by GPT.

Few-shot. We provide the model with a few examples of correct lemmatization. The shots consist of 30 randomly extracted sentences from the gold dataset. We give the model the word form, the associated gold lemma, and the sentence in which each word form appears. Then, we ask the model to repeat the task as formulated in zero-shot 2 i.e., without knowing the lemma predicted by LatinPipe on the remaining sentences.

Model	Setting	Precision	Recall	F1
LatinPipe (LP)		95.21	94.92	95.06
LP+GPT-4o-mini	Zero-shot 1	95.84	95.55	95.70
	Zero-shot 2	96.43	96.13	96.28
	Few-shot	96.71	96.41	**96.56**

Table 3: Results on the lemmatization task.

5. Results and discussion

Our baseline accuracy for the full sample of sentences in the dataset used for the evaluation is 95.05 (F1 score). Across all experimental settings, the combination of GPT-4o-mini with the LatinPipe output leads to an improvement of the lemmatization accuracy. Results are shown in Table 3.

When GPT is not given the lemma predicted by the system (zero-shot 2), its F1 score increases compared to when the predicted lemma is given (zero-shot 1). It seems that GPT is biased by the information given in the prompt in zero-shot 1, and performs better when it is left to freely assign the lemma to the word form.

The few shot experiment is provided with a new baseline value (F1 = 95.06), as we had to subtract from the annotation dataset the sentences selected as shots for the LLM. In this third experimental setting, GPT benefits from in-context few-shot examples and achieves its best performance (F1 = 96.49), improving the baseline by nearly 1.5 points.

On a closer qualitative analysis, we notice that the accuracy on problematic lemmas increases with the hybrid pipeline. For instance, *pignus* scores 37.3 with LatinPipe, and 88.1 with GPT (few-shot). As mentioned above, the Digest contains some AG text. LatinPipe recognizes AG script, and uses the label 'greek.expression' to fill the lemma field. GPT-4o-mini on the other hand, also recognizes AG script, but lemmatizes it as a Latin word. In particular, the conjunction καὶ 'and' is lemmatized by GPT-4o-mini as *et*, i.e. the corresponding conjunction in Latin.

5.1. In-depth analysis on technical legal terms

To better understand the actual impact of the hybrid pipeline on domain-relevant lexical material, we conducted a focused analysis on the subset of expert-validated technical lemmas. Rather than relying solely on global F1 improvements, we examined lemma-level behavior across frequency bands and part-of-speech categories, identifying where and how the hybrid configuration (LatinPipe + GPT) modifies the baseline system.

The large majority of technical lemmas (82.9%) remain stable, while 16.4% improve and only 0.7%

worsen. This confirms that the LLM layer acts as a conservative corrective component, selectively refining specific lemmas without introducing systematic regressions. The performance on lemmatization decreases for only four lemmas, suggesting that the LLM layer rarely overrides correct predictions made by LatinPipe on technical terms. In order to understand where the hybrid pipeline performs better, we conducted separate analyses on the technical terms based on their frequency range and their PoS.

Frequency range. The distribution of improvements across frequency ranges reveals a non-uniform pattern. The performance on frequent lemmas improves by 25.1%; 14.7% for medium-frequency lemmas; 6% for rare lemmas. Thus, frequent technical lemmas are approximately four times more likely to improve than rare ones. This suggests that the hybrid pipeline primarily intervenes where cumulative error impact is higher. Rather than rescuing marginal or extremely rare items, the LLM appears to stabilize lemmas that occur repeatedly in the corpus and therefore have greater influence on terminology extraction and distributional statistics. This distributional effect is particularly relevant for glossary construction, where high-frequency terms shape candidate ranking, collocational profiles, and bilingual alignment.

Part-of-speech. A clearer structural pattern emerges when considering part-of-speech categories. Lemmatization of verbs improves by 20.8%; on adjectives by 12.9%; on nouns by 9.7%.

Verbs are therefore more than twice as likely to improve compared to nouns. This indicates that the integration of GPT primarily contributes to resolving morphologically complex verbal forms rather than stabilizing already well-identified technical nouns.

Table 4 shows the results of our analysis by combining PoS and frequency range information. Frequent verbs show an improvement rate of 37.8%, compared to only 5.0% for rare verbs. In contrast, nouns remain comparatively stable across different frequency ranges, with improvement rates ranging between 6.7% and 13.2%.

This pattern suggests that LatinPipe already performs robustly on domain-specific nouns, while the LLM layer contributes primarily in contexts involving verbal inflectional ambiguity.

Examples from the list of technical terms. Table 5 presents representative examples of lemmas with the highest improvements. Among verbs, notable cases include *parco* ('to spare, have mercy upon'), improved from 0.50 to 1.00; *dirimo* ('to settle, dissolve'), improved from 0.50 to 0.83; *infero* ('to bring against, to charge'), improved from 0.56 to 0.88. All these verbs are central to the *Digest*, which describes various circumstances in which the defendant or a possible defendant may be charged, tried, convicted, or spared. These improvements likely reflect enhanced contextual disambiguation of inflected verbal forms.

The case of *creditor* ('creditor'), which improves from 0.55 to 0.91 across 102 occurrences, is particularly significant due to its high frequency in the corpus and central role in Roman legal discourse. Although morphologically ambiguous in some contexts, its improved consistency directly benefits glossary-oriented analyses. However, among nouns, the most emblematic example is *pignus*, improved from 0.37 to 0.88. The Digest deals extensively with how debts can be incurred, loans granted, debts discharged, and similar circumstances, and this substantial gain seems to suggest the hybrid system's ability to recover highly relevant technical terminology.

Adjectival improvements are more moderate, *praesens* ('present'), from 0.75 to 0.94, *ratus* ('ratified, valid'), from 0.77 to 0.92. Overall, adjectival lemmas appear more stable than verbal ones, and improvements tend to be smaller. It should also be noted that these are in fact adjectival forms derived from verbs (*praesum* and *reor*), and therefore considering them adjectives is entirely debatable. The assignment of this POS is due to the fact that the legal scholars specifically required the glossary to have a specific entry for these adjectival forms.

To address a reviewer's suggestion, we also conducted an additional experiment using the type-based morphological analyzer for Latin LEMLAT (Passarotti et al., 2017). In this setup, the lemma wrongly predicted by LatinPipe was compared with the set of candidate lemmas returned by LEMLAT for each word form belonging to the list of technical lemmas. The gold lemma was used to tackle potentially ambiguous word forms and include only those belonging to lemmas included in our list. The aim was to determine potentially problematic lemmas in the list of validated terms: if neither of the two systems was able to output the correct lemma for the word form, the lemma was flagged as 'problematic' for lemmatization.

The results show that, in the vast majority of cases, the correct lemma is present among the candidates generated by LEMLAT, although not explicitly disambiguated. Any other case of mismatch was due to inconsistencies in ortography (e.g. *submoueo* vs. *summoveo* 'to remove, to banish') or lemmatization rules (e.g. *melior* 'better', comparative form from the adjective *bonus* 'good' is lemmatized as *bonus* in our gold dataset and as *melior* by LEMLAT). Only one form in our dataset is not correctly captured by either system. This form is the participle *reliquatus*, from the verb *reliquor*, which is specifically used in legal contexts and means 'to be in arrears in respect of money

PoS	Range	I / S / W	Improvement (%)	Worsening (%)
ADJ	Frequent	1 / 7 / 0	12.5	0.0
	Medium	2 / 6 / 0	25.0	0.0
	Rare	1 / 14 / 0	6.7	0.0
NOUN	Frequent	10 / 65 / 1	13.2	1.3
	Medium	8 / 74 / 0	9.8	0.0
	Rare	6 / 83 / 0	6.7	0.0
VERB	Frequent	31 / 50 / 2	37.3	2.4
	Medium	14 / 58 / 1	19.2	1.4
	Rare	4 / 76 / 0	5.0	0.0

Table 4: Lemma-level changes on technical terms by PoS and frequency range. Raw counts are reported as Improved (I) / Stable (S) / Worsened (W). Percentages indicate improvement and worsening rates within each PoS × frequency range category.

PoS	Lemma	LP Acc.	LP+GPT Acc.	Δ	Occ.	Range
VERB	parco	0.50	1.00	+0.50	2	rare
	creditor	0.55	0.91	+0.36	102	frequent
	dirimo	0.50	0.83	+0.33	6	rare
	infero	0.56	0.88	+0.31	16	medium
NOUN	pignus	0.37	0.88	+0.51	59	frequent
	tribunal	0.25	0.75	+0.50	4	rare
	suspectus	0.83	1.00	+0.17	6	rare
	emptum	0.77	0.92	+0.15	13	medium
ADJ	praesens	0.75	0.94	+0.19	17	medium
	ratus	0.77	0.92	+0.15	13	medium
	absens	0.91	1.00	+0.09	11	rare
	bonus	0.87	0.88	+0.01	67	frequent

Table 5: Top four improved lemmas per PoS (LatinPipe → LatinPipe + GPT). Δ indicates the increase in lemma-level accuracy.

owed'.

These results highlight two main points. On the one hand, LEMLAT provides broad coverage of Latin legal technical vocabulary, at least with respect to the material attested in the Digest. Indeed, all forms except one are associated with the correct lemma. On the other hand, however, LEMLAT is not a lemmatizer in itself, but a morphological analyzer. Crucially, it does not perform contextual disambiguation: for each form, it outputs all possible morphological analyses and, consequently, all possible lemmas, which differ from one another. Therefore it does not constitute a direct alternative to the LLM in our specific experimental context. This observation further supports our choice of a hybrid pipeline in which ambiguity is resolved through context-sensitive models rather than through type-based analyses.

6. Conclusions and future work

Currently, LLMs do not represent valid replacements for state-of-the-art tools in lemmatization tasks. This limitation is due to several factors. Lemmatization is not a simple binary or multiple-choice classification task; rather, it involves a complex process that is computationally and economically costly, given the large number of input and output tokens and the need for high-performance GPUs. Moreover, the generation of CoNLL-U formatted data seems to be an excessively complex task, failing particularly when applied to datasets exceeding ten sentences.

Given this as a starting point, our evaluation was intentionally conceived as a goal-aligned assessment rather than as a generic improvement of lemmatization accuracy. Instead of optimizing for global performance, we restricted the analysis to a curated list of expert-validated technical lemmas that are directly relevant to glossary construction. This methodological choice reflects the downstream objective of the project: building a reliable glossary of Roman legal terminology. From this perspective, the hybrid pipeline shows desirable properties. It is largely conservative, introduces minimal regressions, and selec-

tively improves morphologically complex or distributionally central technical lemmas. The analysis demonstrates that improvements are not randomly distributed, but show patterns according to frequency range and PoS dimensions. In particular, the system disproportionately benefits frequent verbal lemmas, which are more prone to inflectional ambiguity.

However, this evaluation scenario should not be interpreted as a fully realistic setting, for the following reasons. Firstly, the analysis assumes access to a validated list of domain-specific lemmas. Such curated resources are rarely available in real-world scenarios, especially in low-resource historical domains, and for specialized languages like the Roman legal lexicon. The ability to focus evaluation on a predefined terminology set presupposes expert knowledge and manual curation, which may not scale to other corpora or domains. Moreover, the hybrid correction strategy implicitly assumes computational resources sufficient to post-process and evaluate lemma-level discrepancies. In large-scale or real-time applications, correcting or re-evaluating each lemma individually through an LLM layer may be computationally too heavy. Finally, the improvements observed here are selective rather than systemic. The hybrid pipeline does not substantially alter the overall lemmatization performances; most lemmas remain stable. While this stability is methodologically reassuring, it also means that the hybrid approach does not eliminate the need for human validation in large-scale terminology extraction workflows.

Our comparison with LEMLAT shows that, while it provides exhaustive coverage of morphologically plausible analyses of legal Latin terms, it does not perform contextual disambiguation. As a consequence, its integration into the lemmatization pipeline in place of the LLM would shift the task from lemma prediction to candidate selection, which would still require either manual intervention or an additional disambiguation component. In this perspective, LEMLAT can be effectively used as a consistency check, or as a candidate generator for future external validation via an LLM or a human annotator.

To conclude, the present results should not be interpreted as evidence that LLM-based post-processing provides a general solution to domain-specific lemmatization. Rather, they indicate that, under controlled conditions and with clearly defined terminology targets, hybrid pipelines can selectively improve structurally relevant portions of the lexicon. By shifting the focus from global accuracy to terminology-relevant behavior, we obtain a more nuanced understanding of where hybrid pipelines provide added value, and where they do not.

7. Author contributions

PM: Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Validation; Visualization; Writing – original draft (all sections); Writing – review & editing.
ES: Supervision; Funding Acquisition; Writing – review and editing.
Contributions follow the Contributor Role Taxonomy: https://doi.org/10.5281/zenodo.18421448

8. Resources

We release all resources (gold standard and code) in the GitHub repository: https://github.com/paoma370/Justinian-Digest-project.git.

9. Acknowledgements

This work stems from the project PRIN 2022 PNRR P20224NJLK – SH2 – TESTO “Translating, Encoding, Sharing The Origins”. From the Littera Florentina to an open-access Italian translation of Justinian’s Digest”. Piano Nazionale Ripresa e Resilienza (PNRR) M4C2 – Investimento 1.1 “Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)- funded by the European Union – NextGeneration EU” – CUP B53D23032480001.
The authors also wish to thank Michele Pedone (PI of the TESTO project) and Alessandro Grillone (Pisa University) for lending their expertise to validate the list of legal Latin terms, which was foundational for carrying out this work.

10. Bibliographical References

Thomas M Banchich, John Marenbon, and Charles J Reid Jr. 2015. The revival of roman law and canon law. In *A Treatise of Legal Philosophy and General Jurisprudence: Volume 6: A History of the Philosophy of Law from the Ancient Greeks to the Scholastics*, pages 251–265. Springer.

Marco Baroni. 2009. Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus linguistics: An international handbook. Volume 2*, pages 803–821. Mouton de Gruyter, Berlin.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17772–17780.

Frederick W Dingledy. 2016. The corpus juris civilis: a guide to its history and use. *Legal Reference Services Quarterly*, 35(4):231–255.

Margherita Fantoli, Marco Passarotti, Dominique Longrée, and Concepción Cabrillana. 2024. Lemmas in dialogue: Linking the lasla corpus to the lila knowledge base. *Recent Trends and Findings in Latin Linguistics: Volume I: Syntax, Semantics and Pragmatics. Volume II: Semantics and Lexicography. Discourse and Dialogue*, page 297.

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the LASLA corpus in the LiLa knowledge base of interoperable linguistic resources for Latin. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34, Marseille, France. European Language Resources Association.

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 Italian Conference on Computational Linguistics. Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025) Cagliari, Italy, September 24-26, 2025.*

Eleonora Ghizzota, Paola Marongiu, Pierpaolo Basile, Stefano Ferilli, and Barbara McGillivray. 2026. Linguistic knowledge graphs for sense prediction: a case-study on latin. In *Proceedings of the 15th Language Resources and Evaluation Conference*. Anthology of Computers and the Humanities.

Saniya Irfan and Syed Juned Ali. 2025. Qallm: An llm-based ner dataset curation, annotation and evaluation in historical urdu elegies. In *Computational Humanities Research 2025*, pages 921–936. Anthology of Computers and the Humanities.

Siyu Liang, Talant Mawkanuli, and Gina-Anne Levow. 2026. Hybrid neural-LLM pipeline for morphological glossing in endangered language documentation: A case study of jungar tuvan. In *Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics*, pages 16–30, Rabat, Morocco. Association for Computational Linguistics.

Theodor Mommsen and P. Krüger, editors. 1911. *Corpus iuris civilis: editio stereotypa duodecima. Institutiones*, volume 1. Weidmannos, Berlin.

Kseniia Petukhova and Ekaterina Kochmar. 2025. A fully automated pipeline for conversational discourse annotation: Tree scheme generation and labeling with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15829–15852.

Marton Ribary and Barbara McGillivray. 2020. A corpus approach to roman law based on justinian's digest. In *Informatics*, volume 7, page 44. MDPI.

Frederick Riemenschneider. 2025. Beyond base predictors: Using LLMs to resolve ambiguities in Akkadian lemmatization. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 226–231, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Sandro Schipani. 2005. *Iustiniani Augusti Digesta seu Pandectae. Digesti o Pandette dell'Imperatore Giustiniano. Testo e traduzione*, volume 1. Giuffrè Editore.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197, Torino, Italia. ELRA and ICCL.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the EvaLatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 183–188, Marseille, France. European Language Resources Association.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Olia Toporkov, Alan Akbik, and Rodrigo Agerri. 2025. Lemma dilemma: On lemma generation without domain-or language-specific training data. *arXiv preprint arXiv:2510.07434*.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language*

Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024, pages 122–128, Torino, Italia. ELRA and ICCL.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2025. Hybrid human-llm corpus construction and llm evaluation for the caused-motion construction. *Northern European Journal of Language Technology*, 11(1):27–57.

George Kingsley Zipf. 1949. *The Principle of Least Effort*. Addison-Wesley.

11. Language Resource References

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language technology for cultural heritage: Selected papers from the LaTeCH Workshop Series*, pages 79–98. Springer.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020a. A new latin treebank for universal dependencies: Charters between ancient latin and romance languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 933–942.

Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020b. Udante: First steps towards the universal dependencies treebank of dante's latin works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 66–72.

Margherita Fantoli and Miryam de Lhoneux. 2022. Linguistic annotation of neo-latin mathematical texts: a pilot-study to improve the automatic parsing of the archimedes latinus. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 129–134.

Federica Gamba and Flavio Massimiliano Cecchini. 2024. De latinae linguae reparatione treebank. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Federica Gamba and Daniel Zeman. 2023. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34. Prague.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Marco Passarotti. 2019. The project of the index thomisticus. *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution*, 10:299.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The lemlat 3.0 package for morphological analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31, Gothenburg. Linköping University Electronic Press.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2020. UDPipe at EvaLatin 2020: Contextualized embeddings and treebank embeddings. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 124–129, Marseille, France. European Language Resources Association (ELRA).

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.