Latin Treebanks in Review: An Evaluation of Morphological Tagging Across Time

Marisa Hudspeth, Brendan O'Connor, Laure Thompson
University of Massachusetts, Amherst
{mhudspeth, brenocon, laurejt }@cs.umass.edu

Abstract

Existing Latin treebanks draw from Latin's long written tradition, spanning 17 centuries and a variety of cultures. Recent efforts have begun to harmonize these treebanks' annotations to better train and evaluate morphological taggers. However, the heterogeneity of these treebanks must be carefully considered to build effective and reliable data. In this work, we review existing Latin treebanks to identify the texts they draw from, identify their overlap, and document their coverage across time and genre. We additionally design automated conversions of their morphological feature annotations into the conventions of standard Latin grammar. From this, we build new time-period data splits that draw from the existing treebanks which we use to perform a broad cross-time analysis for POS and morphological feature tagging. We find that BERT-based taggers outperform existing taggers while also being more robust to cross-domain shifts.

1 Introduction

Large-scale digitized Latin archives now document cultures across many centuries in wide a variety of genres from literature to legal documents. With increasingly powerful Latin natural language processing tools (e.g. Bamman and Burns, 2020; Burns, 2023), morphological feature tagging is a promising method for Latin-based computational humanities. The goal of morphological tagging is to identify a set of morphological feature-value pairs for each token of a given sentence. These features can help researchers analyze agency, power, and other morphosyntactically-signalled phenomena which have been fruitfully investigated in English (Sap et al., 2017; Greene and Resnik, 2009) and other languages (Rashkin et al., 2017). For example, Voice (active, passive verbs) and Case (e.g., nominative, accusative ablative nouns) are useful for studying power and agency.

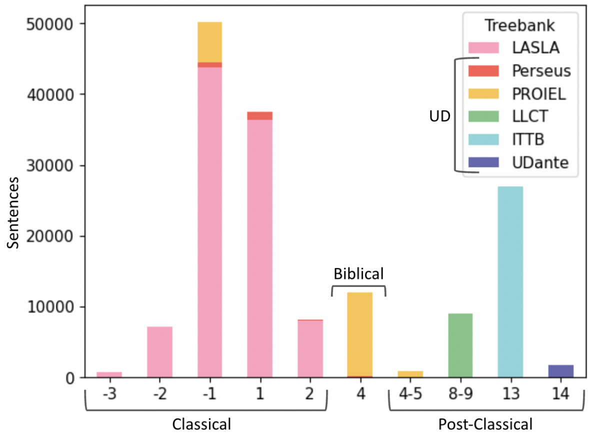

Figure 1: From our curated metadata (§2), the number of sentences per century (3rd BCE—14th CE) across the 5 UD treebanks and LASLA, shown with three broad time periods.

Although Latin taggers have relatively good performance, in our experience they often perform poorly on rarer feature values—such as passive voice–that may prove crucial for downstream analyses. Toward this end, we hope to develop a Latin morphological tagger whose accuracy is robust across time and genre by leveraging the recent development of five separate Latin Universal Dependencies (UD; de Marneffe et al., 2021) treebanks and recent efforts to harmonize their morphological tags (Gamba and Zeman, 2023a). In this work we review these harmonized treebanks[1] plus the non-UD treebank LASLA (Denooz, 2004), and conclude that more data curation is required to fully evaluate and improve morphological tagging's cross-domain accuracy.

Our contributions include: **1)** precisely documenting genre and historical context for the 544 texts within the UD treebanks as a machine-readable, cross-treebank resource that will enable

[1] Perseus (Bamman and Crane, 2011), PROIEL (Haug and Jøhndal, 2008), LLCT (Cecchini et al., 2020b), ITTB (Passarotti, 2019), and UDante (Cecchini et al., 2020a)

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 203–218
August 15, 2024 ©2024 Association for Computational Linguistics

future work to examine morphosyntactic association against these variables; **2)** harmonizing the UD and LASLA treebanks to reduce annotation differences that can affect training; **3)** proposing edits to the UD tagset that better align with standard analyses of Latin grammar to facilitate work by researchers with standard Latin training; and **4)** conducting a cross-time analysis with experimental results broken down by historical period that show the promise of our harmonization efforts and BERT-based morphological taggers.[2]

2 Latin Treebanks Revisited

2.1 Time and Genre Metadata

Detailed metadata on the texts included in the Latin UD treebanks is difficult to aggregate or lacking altogether. Information on the included works' time period, genre, author, and relative size has not been compiled in one place. Our work takes major steps to fill this gap. For all 544 texts across the five UD treebanks, we manually collected the following metadata: the source treebank, time period, century, internal treebank identifiers, cumulative and split-level sentence counts, author, and exhaustive genre labels.

Genre. Figure 2 shows the genre coverage of the UD treebanks. Previous EvaLatin campaigns (Sprugnoli et al., 2020, 2022) have implicitly defined several genres (prose, poetry, epics, and histories), which were then used to analyze cross-genre tagging accuracy on Classical era, non-UD data. We expand upon these genres by including more fine-grained labels and by covering non-Classical texts.

We annotate nine exclusive genres: short poems, epics, letters, histories, satires, speeches, legal texts, treatises, and the Bible.[6]

Time. We define the (approximate) century of each text (Figure 1 and 2). For cross-time analysis, we define three very broad time periods:

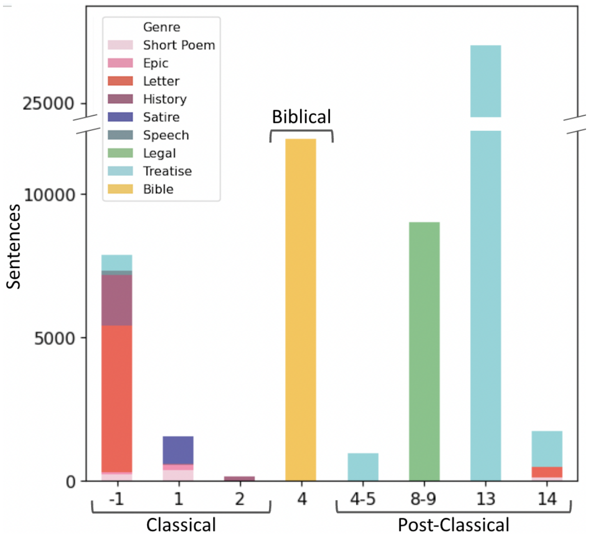

Figure 2: Number of sentences in the UD treebanks per century, colored by genre.

- **Classical** is defined as 3rd century BCE through the 2nd century CE, in line with conventional definitions of the Classical Latin language and periods (Sala and Posner, 2024) and previous Latin NLP literature (Sprugnoli et al., 2020, 2022).
- **Biblical** is defined as its own genre and time period, consisting solely of Jerome's *Vulgata* from the 4th century CE. It is significantly different from other texts given it is a translation (from much earlier material), and has relatively simpler grammar (Nunn, 1922).
- **Post-Classical** is defined as 4th century CE and later, excluding the Bible, thus including Late and Medieval Latin texts. For simplicity, we do not split it further.

Prior work in cross-time tagging either used a smaller set of time periods (Sprugnoli et al., 2022) or considered each UD treebank its own time period (Gamba and Zeman, 2023a), which we argue is too approximate given our metadata findings (§2.2).

2.2 UD Treebanks

Currently, there are five UD treebanks for Latin.[7] Four of these—Perseus, PROIEL, LLCT, and ITTB—were automatically converted to UD format, while the fifth, UDante, was annotated directly in UD. Collectively, this corresponds to about 58,000 annotated sentences and 979,280 annotated tokens. As Figure 1 shows, these treebanks cover a wide range of time but far from evenly.

[2]We have publicly released our new text-level metadata, standardized morphologically tagged text from described treebanks, and conversion software on Github.

[3]Although unreleased, we determined the feature-value set by examining LatinCy's outputs.

[4]EvaLatin 2020 also has annotated data that is not directly sourced from LASLA but consists of a subset of LASLA's texts. This data is not annotated with morphological features.

[5]EvaLatin 2022 is a near-subset of LASLA because it has one non-Classical text that is not in LASLA.

[6]We also annotate additional non-exclusive genres (§A.1).

[7]In May 2024, a sixth, CIRCSE, was added; it is a subset of LASLA.

	Name	Text Data	Standard Grammar?	Data Source	Paper/Version
1	Pre-UD	4 non-UD	Mixed	Perseus, PROIEL LLCT1,ITTB	Bamman and Crane 2011; Haug and Jøhndal 2008 Korkiakangas and Passarotti 2011; Passarotti 2019
2	UD v2.8+	5 UD	Mixed	UD Site	UD v2.8-11
3	LatinCy Edits	5 UD	Yes[3]	Unreleased	Burns 2023
4	Harmonized UD	5 UD	No	Github *(acc. 1/24)*	UD v2.13; Gamba and Zeman 2023a
5	LASLA	1 non-UD	Mixed	Github *(acc. 2/24)*	Denooz 2004
6	EvaLatin 2022[4]	near-subset of 5[5]	Mixed	Github	Sprugnoli et al. 2022
7	CIRCSE	1 UD; subset of 5	Mixed	Github	UD v2.14
8	Harmonized + Standardized	5 UD + LASLA, New Splits	Yes	Github	This work

Table 1: Summary of data sources and history of Latin treebanks (for morphological tagging only).

We find that the three Post-Classical treebanks (LLCT, ITTB, and UDante) are quite distinct from each other in terms of genre and time period. LLCT consists entirely of medieval legal charters from the 8th and 9th centuries. ITTB consists of three philosophical and religious works by Thomas Aquinas from the 13th century. Finally, UDante is comprised of Dante Alighieri's 14th century Latin works, including treatises, letters, and poems.

The two remaining treebanks, Perseus and PROIEL, are more diverse. Most texts in Perseus are Classical, although 154 sentences are from Jerome's *Vulgata* (the Book of Revelation). While PROIEL also includes Classical texts, the majority (11785) of its 18411 sentences are also taken from Jerome's *Vulgata*. There is overlap between Perseus and PROIEL, as both share at least 145 sentences from the Book of Revelation.[8] Aside from Classical and Biblical texts, PROIEL also includes one 4th-5th century work, *Opus Agriculturae* by Palladius.

2.3 LASLA: Additional Classical-era treebank

LASLA is a large, non-UD treebank for Latin (Denooz, 2004). By our own count, LASLA has 134 unique texts with 95,974 sentences and about 1.8M tokens.[9] All texts are Classical. All genres included in UD are covered, in addition to plays. Unlike the UD treebanks, LASLA does not have dependency relations.

[8]See Table 8 for a breakdown of annotation agreement between these duplicate sentences.

[9]A full list of authors, works, and tokens per text is available here.

3 Harmonizing UD and LASLA Annotations

In this section, we describe steps taken to reduce the annotation differences between the Harmonized UD treebanks (Table 1 row 4) and LASLA (1 row 5). Throughout this section, we sometimes use "UD" as a shorthand for Gamba and Zeman (2023a)'s Harmonized UD treebanks.

In §3.1, we outline the annotation agreement between Harmonized UD and LASLA before any intervention on our part. Then, we describe two types of changes: harmonization (§3.2) and standardization (§3.3). During harmonization, we enforce consistency of arbitrary values to have fair training and evaluation. Standardization is more involved, where we change the grammatical system to be more Latin-specific. Both of these steps are done automatically and simultaneously through conversion scripts.

3.1 Annotation Agreement Between UD and LASLA

Author	Work	# Dups
Caesar	Gallic War	1127
Cicero	De Officiis	447
Cicero	In Catilinam	118
Ovid	Metamorphoses	0[10]
Petronius	Satyricon	407
Propertius	Elegies	183
Sallust	Bellum Catilinae	228
Tacitus	Historiae	50
Vergil	Aeneid	47

Table 2: For the nine texts shared between LASLA and UD (collectively; specifically, Perseus and PROIEL), number of duplicate sentences.

[10]Ovid's *Metamorphoses* appears in both treebanks, but they cover different books of the text.

UD and LASLA happen to have re-annotated many of the same sentences, which gives a way to analyze annotation agreement between the projects. We detect sentences that appear in both datasets (§A.2), finding 2607 such duplicates across eight Classical texts (Table 2), which may be an underestimate since our duplicate detector will miss cases where sentence segmentation or tokenization differ.

We calculate annotation agreement before and after harmonization and standardization on our reduced set of features (Table 3). Some features, such as `Degree`, `Tense`, and `VerbForm`, have low agreement due to mismatches between their possible value sets in UD and in LASLA. Other features, such as `Gender`, `Person`, and UPOS have low agreement due to remaining annotation differences.[11]

3.2 Our Harmonization Efforts

Gamba and Zeman (2023a) have already performed the bulk of the harmonization necessary for the UD treebanks. However, we are additionally attempting to harmonize LASLA with the UD treebanks.

Remaining inconsistencies we've harmonized. We have found some remaining inconsistencies, both within the UD treebanks and between UD and LASLA. Usually, neither is incorrect in their annotation, but without normalization this will cause unfair evaluation. Thus, we enforce consistent, arbitrary values in these cases. See §A.4 for specifics.

Collapsing feature values. Another issue we encountered is that some UD treebanks lack certain feature values that are present in the others. Gamba and Zeman (2023a) were aware of this issue, and chose not to harmonize these values in order to preserve as much information as possible. This is understandable, as these features may be of interest to researchers. However, for our purposes, we have collapsed certain feature values together in order to have fairer evaluation of models trained on different treebanks.

For UPOS (universal part of speech), we have collapsed INTJ into PART across all treebanks, since two UD treebanks (ITTB and LLCT) do not use INTJ, instead using the value PART. Additionally, for `Degree`, we have collapsed `Degree=Pos` into `Degree=None`, since LASLA is the only treebank to use `Pos`. The distinction between `Degree=Pos` and `Degree=None` is debated.[12] We note that Gamba and Zeman (2023a) also collapsed `Degree=Pos` and `Degree=Dim` into `Degree=None`, so this decision has precedent.

[11] See appendix for agreement rates across all features (Table 9) and a comprehensive overview of the feature inventories (Table 12).

3.3 Conversion to Standard Latin Grammar

Feature	Before			After		
	% same	# same	# total	% same	# same	# total
Case	97.8	20372	20821	97.8	20372	20821
Degree	8.5	598	6998	**69.5**	598	860
Gender	74.7	14965	20034	**75.2**	14964	19911
(loose)	97.2	19481	20034	**97.8**	19477	19911
Mood	99.4	5279	5312	**97.3**	8621	8864
VerbForm	93.2	8264	8867	–	–	–
Number	97.9	25672	26211	97.9	25543	26088
Person	91.0	6089	6692	91.0	6089	6692
Tense	77.3	5228	6766	**96.7**	8184	8465
Voice	96.0	7493	7809	**96.5**	8554	8864
UPOS	93.0	34814	37425	93.0	34821	37425

Table 3: Percent and number of tokens in the duplicate LASLA and Harmonized UD sentences that have the exact same value for each feature, before and after our harmonization and standardization. Percent is out of tokens that had a non-None value in either UD or LASLA. After our changes, Mood and VerbForm are collapsed into Mood only, but we list them separately before. Percentages after our changes are **boldfaced** when there is improved agreement.

UD was developed with cross-linguistic goals in mind, offering a set of universal tags applicable to all languages. However, prior to the harmonization efforts by Gamba and Zeman (2023a), many Latin UD treebanks employed standard Latin values for certain features, reflecting a long-standing desire for a more Latin-specific tagset. Harmonization and conversion to UD has relegated these Latin-specific values to a secondary status. This poses a key challenge for evaluation, as these two annotation styles are not comparable.

Although UD provides a valuable cross-linguistic framework, we believe Latin is also useful to study on its own, within long-standing approaches to Latin linguistics (e.g. Greenough and Allen 1903). The UD treebanks remain the most complete, high-quality source of morphological annotations for Latin. To bridge the gap between UD and standard Latin linguistics, we offer an alternative version that uses more standard Latin grammar. In particular, we standardize the treebanks to follow Pre-UD Perseus's (Table 1 row 1) features: UPOS, person, number, tense, mood, voice, gender, case, and degree. This set is nearly identical to Burns

[12] See the UD documentation for Degree in Latin here.

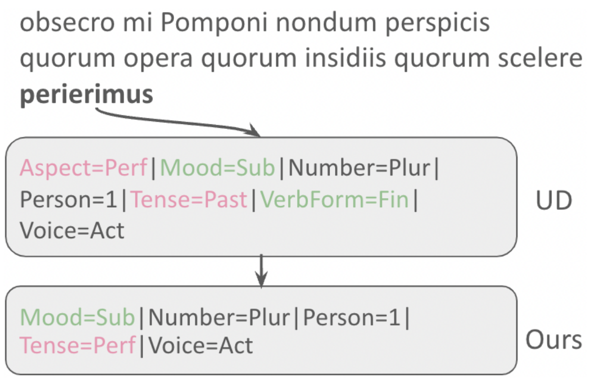

Figure 3: Example of how a token's set of morphological features changes after standardization, from Cicero's *Letters to Atticus* Book 3 Letter 9.

(2023)'s, except that LatinCy separately predicts Mood and VerbForm (which we combine). For most of these features, UD has a corresponding feature that we can easily extract. The exceptions are Tense and Mood, where we developed a more elaborate method of standardization (§A.5). For example, Latin tense traditionally has six possible values (present, imperfect, perfect, future, pluperfect, future perfect) which are standardized across pedagogical materials (Greenough and Allen, 1903; Wheelock and LeFleur, 2010). However, UD's `Tense` feature only includes four of these values (present, past, future, pluperfect), which is why we must perform a conversion.

We choose to convert to Standard Latin Grammar *before* training, rather than perform postprocessing on the predictions of a model trained on the UD tagset, for two reasons: 1) preprocessing allows for more precise conversions based on known treebank sources, addressing inconsistencies between treebanks, and 2) model predictions may combine features from various annotation schemes and be grammatically inconsistent, making postprocessing complex and potentially unreliable.

3.4 Remaining Inconsistencies

After our harmonization and standardization, most features have high annotation agreement between LASLA and UD (Table 3). Degree and UPOS, two features that already had low agreement within UD (Table 8), saw improved but not high agreement after UD-LASLA harmonization. These are likely due to fundamental differences in the annotation process which may require reannotation to fix.

We modify how our models are trained to account for the two following differences (§5.1):

- In LASLA, the `Gender` feature can take multiple values to represent possible genders based only on the word form (disregarding the context of the sentence). In the UD treebanks, `Gender` is assigned one value that depends on the sentence. This causes the low annotation agreement for `Gender` in Table 3. If we use a looser criterion—counting the annotations as the same when the UD gender value matches one of LASLA's gender values—we do see higher agreement (*(loose)* in Table 3).
- In LASLA, personal pronouns are annotated with `Person=None`, but in the UD treebanks personal pronouns have non-`None` values.[13]

We list additional differences in §A.6.

3.5 Our Custom Data Splits

Time	Train	Dev	Test
Classical (UD)	6524	201	1041
Classical (UD+LASLA)	102498	201	1041
Bible	10451	322	1021
Postclass	32661	1010	5003

Table 4: Number of sentences in our proposed train, dev, and test splits

We create new data splits to emulate EvaLatin's cross-time sub-task which evaluates models on texts of a different time period than what they are trained on. When creating train/test splits for each time period, we keep the following constraints in mind: **1)** Individual works should be within a single split. For example, Ovid's *Metamorphoses* should only appear in either the train or test set, rather than having a random sample of sentences in the train set with the rest in the test set. **2)** Make sure the test set is large enough for reasonable statistical power. We specifically choose to have a minimum of 1000 sentences in each test set. **3)** Only evaluate on UD data and not LASLA. Due to some annotation differences (see §3.4), UD treebanks have more complete information than LASLA. This is in contrast to EvaLatin campaigns which evaluate on subsets of LASLA.

To make our dev sets, we randomly sample 3% of sentences from each work in the train sets, making sure that we never sample from LASLA or any UD sentences that also appear in LASLA.

[13] This will be simple to fix in future work, since there are limited personal pronoun lemmas in Latin.

Due to these constraints, we are unable to keep the original UD test sets. Since we want test sets for each time period, we must construct Classical-specific splits. Perseus, despite being largely Classical, is too small for effective training. PROIEL contains some Classical texts but is mostly comprised of Biblical texts. We separate the Biblical content and combine the Classical texts from both treebanks to ensure a sufficiently large Classical train set. Due to the first constraint, we cannot use ITTB's original train/test splits since *Summa Contra Gentiles* appears in both the train and test set. To help meet our second constraint, we do not use UDante's and LLCT's original splits.

To achieve our third constraint, our Classical train set must include all works that appear in both LASLA and UD, shown in Table 2. We want to test two scenarios: training with and without LASLA data. In order to have enough training sentences in the UD-only scenario, we treat the letters of Cicero's *Letters to Atticus* as separate texts (i.e. that can be distributed across the Classical train and test set), even though this conflicts with our first constraint.

We include a detailed description of which works appear in our custom train and test sets in the Appendix (Table 13).

4 Related Work: Morphological Tagging

There is a long history of work analyzing POS and morphological tagging of Latin (Eger et al., 2015, 2016; Straka and Straková, 2020). Our work follows recent trends of using transformer-based contextual representations.

Several recent papers have explored morphological tagging for Latin. As part of the 2022 EvaLatin feature identification task (Sprugnoli et al., 2022), participants trained and tested on a subset of data from the LASLA corpus that had been automatically converted to UD format (Wróbel and Nowak, 2022; Mercelis and Keersmaekers, 2022). Only a subset of UD morphological features were retained, partly to limit the task to morphological features identifiable by the word form, and partly to avoid features affected by annotation differences. Participants were then able to train on combined UD and LASLA data if they wished, but models were only evaluated on EvaLatin test sets, not UD test sets.

Nehrdich and Hellwig (2022) used LatinBERT (Bamman and Burns, 2020) to train a morphological tagger predicting the case, gender, number, tense and verbform features. Its outputs were then fed into the authors' dependency parser, outperforming prior work using UDPipe and static word embeddings (Straka et al., 2019). Their training and test data came from three UD treebanks (ITTB, PROIEL, and Perseus).

Burns (2023) developed LatinCy, a full NLP pipeline for Latin which includes morphological feature classification.[14] Notably, this pipeline was trained on all five UD treebanks with early attempts made at harmonization, using a smaller tagset than UD that is closer to standard analyses of Latin grammar (Table 1 row 3). Recently, Gamba and Zeman (2023a) performed more rigorous harmonization of morphological features across the five UD Latin treebanks (Table 1 row 4). They reported accuracy before and after harmonization, training and testing on each pair of treebanks using fast-text embeddings (Grave et al., 2018) with UDPipe (Straka et al., 2016) or Stanza (Qi et al., 2020). Harmonization was shown to improve accuracy when training and testing on two different treebanks.

Part-of-speech (POS) tagging is closely related to morphological tagging. In the 2020 EvaLatin campaign, participants trained and tested POS taggers on a subset of the LASLA corpus (Sprugnoli et al., 2020). More recently, Riemenschneider and Frank pretrained a trilingual RoBERTa (Liu et al., 2019) model on English, Ancient Greek, and Latin which surpassed the 2022 EvaLatin competitors (Table 1 row 6). Thus, the current SOTA models for Latin POS tagging are all transformer-based. Additionally, Riemenschneider and Frank's trilingual model underperformed their monolingual Ancient Greek model, suggesting a monolingual Latin model could prove even stronger, given sufficient pretraining data.

Researchers have also experimented with using GPT3.5-Turbo and GPT4 for POS tagging of 16th century Latin texts (Stüssi and Ströbel, 2024). No POS-annotated data exists for 16th century Latin, so the authors experimented with zero-shot prompting and finetuning using data from the five UD treebanks. Although the UD testsets are not entirely comparable with EvaLatin's, the accuracy of these GPT-based approaches seems low when compared to the results of EvaLatin's POS tagging shared task.

Although substantial progress has been made in Latin morphological tagging, gaps still exist. Aside

[14] Using SpaCy (Honnibal and Montani, 2017)

from Gamba and Zeman (2023a) and Burns (2023), prior work has not leveraged all five UD treebanks for training *and* evaluation. While Gamba and Zeman (2023a) measure overall tagging accuracy, more detailed analysis of specific morphological features and diachronic trends has been left to future work. Moreover, to our knowledge no recent paper has evaluated the currently available taggers on UD test data.

5 Experiments

We use three metrics in our evaluations: whole string morphological accuracy, macro F1 for individual features, and F1 for particular feature-values. See A.8 for more detailed explanations.

5.1 Our LatinBERT-based Tagger

Following other recent working finding SOTA performance with transformer-based taggers (Sprugnoli et al., 2022; Riemenschneider and Frank, 2023), we finetune a tagger on top of LatinBERT (Bamman and Burns, 2020). Similar to Riemenschneider and Frank (2023), our tagger uses a separate classification head for every morphological feature, all trained simultaneously—a simple choice which could be improved upon in future work.

When training on LASLA, we sometimes do not train a particular feature head based on a token's feature values. First, if `Gender` has multiple values we do not train the Gender prediction head. We want to keep our set of possible `Gender` values limited to the standard three (`Masc`, `Fem`, and `Neut`). Second, if the token is a personal pronoun and `Person=None`, we do not train the Person prediction head. Having a null value here is inconsistent with the rest of our data. Since we do not know the true value, we choose not to train in this instance. If either of these two cases apply to a particular token, then that token will not contribute to the loss for either the Gender or Person classifier head. Other heads are unaffected.

5.2 Comparison to Previous Taggers

In this section, we use the official UD train/test splits for comparison to previous work but converted to our harmonized and standardized tagset. We compare our BERT taggers to two sets of taggers previously evaluated on UD data: LatinCy (Table 1 row 3) and five Stanza models trained on the five Harmonized UD treebanks (Table 1 row 4). LatinCy uses a non-transformer neural architecture as part of the SpaCy pipeline, along with static floret vectors (Boyd and Warmerdam, 2022). Stanza has a Bi-LSTM architecture for its POS and morphological taggers and uses either word2vec (Zeman et al., 2018) or fasttext (Bojanowski et al., 2017) embeddings, depending on the language. For a fair comparison, we must convert between the different tagsets used by each tagger. For LatinCy, rather than retraining the SpaCy pipeline ourselves, we convert its predictions on each official UD test set to our tagset. This required little modification as LatinCy predicts a near-identical set of features and values.[15] For the Stanza models, we retrain them on our harmonized and standardized versions of each UD training set (Table 1 row 8), since Gamba and Zeman (2023b)'s models and their predictions are unreleased. Replicating Gamba and Zeman (2023b), we only train the Stanza models on each individual treebank, rather than all UD data. We also use the same Latin fasttext embeddings (Grave et al., 2018) and default training parameters.

We report whole string morphological accuracy for each UD test set in Table 5. Our BERT taggers consistently have the highest accuracy. The smallest treebanks, Perseus and UDante, see the most benefit from the BERT architecture and the out-of-domain training data.[16]

Model	Train Set(s)	perseus	proiel	llct	ittb	udante
LatinCy	All UD	.726	.740	.792	.809	.736
BERT	All UD	**.929**	.962	.969	.982	**.910**
Stanza	In-Domain UD	.787	.929	.969	.965	.819
BERT	In-Domain UD	.915	**.962**	**.977**	**.984**	.903

Table 5: Whole string accuracy of **morphological features**. Train set is either All 5 UD treebanks, or a single In-Domain UD Treebank (i.e., same as the Test column).

Model	Metric	perseus	proiel	llct	ittb	udante
Stanza	POS Macro F1	.072	.253	.284	.227	.122
BERT	POS Macro F1	**.066**	**.191**	**.185**	**.144**	**.101**
Stanza	Morph Acc	.058	.179	.275	.177	.077
BERT	Morph Acc	**.016**	**.069**	**.186**	**.074**	**.030**

Table 6: Average difference between in and out of domain performance, for each of the 5 UD treebank test sets (columns); this work (BERT rows) always attains a smaller difference.

[15] LatinCy lacks two possible `Tense` values, `Perf` and `FutP`, which our tagset includes. In a more generous evaluation, where `Fut` and `Imp` are considered correct predictions for gold `FutP` and `Perf`, respectively, all morphological accuracy scores in Table 5 increase by $\leq 5\%$, with maximum accuracy on the LLCT test set at 0.826.

[16] We see similar trends for UPOS; see Table 10.

Model	classical	bible	postclass
	UPOS Macro-F1		
classical-ud	**0.964**	0.937	0.864
classical-all	0.949	0.799	0.839
bible	0.868	**0.976**	0.834
postclass	0.866	0.920	**0.980**
all-ud-custom	0.961	0.975	0.976
all-both-custom	0.948	0.964	0.980
	Morph Accuracy		
classical-ud	0.946	0.936	0.905
classical-all	0.945	0.941	0.908
bible	0.914	0.956	0.885
postclass	0.916	0.931	0.973
all-ud-custom	**0.946**	0.956	0.973
all-both-custom	0.939	**0.960**	**0.974**

Table 7: Performance of our BERT-based taggers when evaluated on custom time-period test sets.

When comparing two models' performance, we calculate statistical significance via randomized permutation testing (Wasserman 2004).[17] When comparing our All-UD model to LatinCy and our in-domain models to Stanza, all comparisons were significant (p=0) for both UPOS Macro-F1 and morphological accuracy, except for LLCT UPOS Macro-F1 (p=0.13). So, in nearly all cases our BERT taggers performed significantly better at both UPOS and morphological tagging than previously released taggers, when trained on all or in-domain data.

We also find that our BERT taggers are more robust to out-of-domain data than the Stanza taggers. In Table 6, for each UD test set, we report the average difference between the in-domain test performance (training and testing on the same treebank) and out-of-domain test performance (training on a different treebank). This difference is always lower for our BERT models than for the Stanza models, suggesting that BERT has better cross-domain performance than Stanza.

5.3 Performance on Our Custom Splits

In total, we train six models including four trained on the sets described in Table 4. The other two models are `all-ud-custom`, trained on the Classical (UD Only), Bible, and Postclass train sets; and `all-both-custom`, trained on the Classical (UD+LASLA), Bible, and Postclass train sets. Since our LatinBERT taggers outperform the other taggers, we limit our focus to these BERT-based taggers. We find that it is generally unnecessary to train a period-specific model. As Table 7 shows, models trained on all time periods have only slightly reduced UPOS Macro F1 and have slightly increased morphological accuracy compared to the models trained on a single domain.

Addition of LASLA data boosts performance for some rare feature values, but decreases it for other features. Although there is only a slight difference in *overall* morphological accuracy with the addition of LASLA data, F1 of particular feature-values improves. When evaluating the `classical-ud` and `classical-all` models on the Classical test set, F1 increases from 0.907 to 0.941 for `Case=Dat` (p=0.0028), and 0.800 to 0.909 for `Mood=Ger` (p=0.0). We also found that some features' Macro F1 scores decreased with the inclusion of LASLA. This behavior is most prominent for Degree (0.96 to 0.91, p=0.0001) and UPOS (0.96 to 0.95, p=0.0). Since the duplicate sentences in LASLA and UD have low annotation agreement for Degree and UPOS (Table 3), the addition of LASLA data likely led to noisier training labels for these two features.

Most errors involve acontextual ambiguity. We randomly sample 100 tokens whose morphology was predicted incorrectly by our `all-ud-custom` model,[18] and annotated them according to six error types: illegal, lexical, genuine acontextual ambiguity, annotation differences, gold wrong, other.

Illegal. We found four illegal errors in which the model combined morphology and/or UPOS in a way that breaks rules of grammar. Three of these involved the token *quod*. For example, when the gold annotation labeled *quod* as SCONJ, the model correctly predicted SCONJ but incorrectly predicted `Gender=Neut` and `Number=Sing`, when a SCONJ should have no value for those features. In the fourth case, when the gold was PRON, the model again correctly predicted PRON but incorrectly predicted `Case=None` and `Number=None`, even though a PRON should have values for those features.

Lexical. We found eight *lexical errors* where the predicted combination of UPOS and morphological features is legal in general, but is impossible given the particular token based on lexical information. For example, let's consider the token *ista* whose gold annotation is a DET with `Case=Nom`, `Gender=Fem`, and `Num=Sing`. This word is a

[17]As detailed in §A.9, we simply report p-values based on 10,000 null simulations; thus p=0 is possible and could be more conservatively interpreted as $p < .0003$ ("rule of three": Eypasch et al. 1995).

[18]33 tokens from Classical texts, 33 from the bible, 12 from Aquinas' works, 11 from LLCT, 11 from Dante's works.

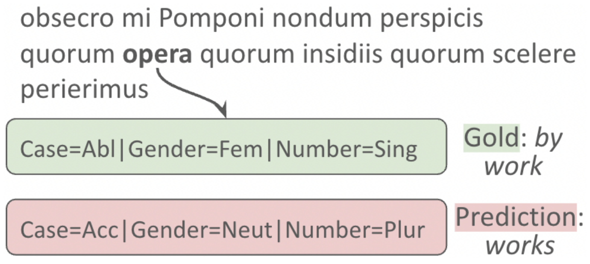

Figure 4: Example of an error in the model's prediction due to acontextual ambiguity, from Cicero's *Letters to Atticus* Book 3 Letter 9.

demonstrative adjective with 1st and 2nd declension endings, so out of context there are only a few combinations of morphological features possible: either `Case=Nom,Abl|Gender=Fem|Num=Sing` or `Case=Nom,Acc|Gender=Neut|Num=Plur`. Our tagger incorrectly predicted `Case=Acc` but correctly predicted `Gender=Fem` and `Num=Sing`. Even though its predictions for Gender and Number are correct, they do not form a valid combination of feature values for this token.

Genuine acontextual ambiguity. Most errors (67) were due to *genuine acontextual ambiguity*. This means that, out of context, the model's prediction is legal and valid given the particular token's lexical information, but in context it is incorrect. We would hope that BERT, as a contextual model, could still predict these cases correctly but it seems to struggle. Figure 4 shows an example of this error type. In other contexts, the token *opera* can be accusative plural, as the model predicted, but within this sentence it must be ablative singular. The verb *perierimus* (we have been ruined) does not take an object, so *opera* cannot be accusative. Additionally, the structure of *quorum opera* is repeated with *quorum insidiis* and *quorum scelere*. The nouns *insidiis* and *scelere* are clearly ablative, suggesting that *opera* should be the same case. This makes more sense contextually: *perierimus* (we have been ruined) *quorum opera* (by whose work).

Annotation differences. Nine errors were due to remaining annotation differences, discussed more thoroughly in §A.6.

Gold wrong. Nine errors were caused by incorrect gold annotations. These include missing `Case` value for nouns, and incorrect UPOS and morphological features.

Words segmented by the tokenizer have a higher error rate. Because of the presence of lexical errors in our model's predictions, we investigated whether the LatinBERT tokenizer segments words in a morphologically-aware manner. We find that the majority (81%) of words in our three custom test sets correspond to a single subtoken for the tokenizer. For these word tokens, our `all-ud-custom` model achieves 98.3% accuracy on UPOS and 97.2% accuracy on all morphological features. In the case that word tokens are split into *multiple* subtokens, performance degrades; Accuracy drops slightly for UPOS to 97.5% and more dramatically for morphological features to 93.6%. Since most words are not segmented and those that are have worse performance, we hypothesize that the model is not able to learn Latin's inflections, which could hypothetically aid in the tagging of rarer words. The relationship between token frequency, word segmentation, and downstream performance is a promising direction for analysis in future work. This aligns with previous findings for English that transformer models with WordPiece tokenizers have lower generalization ability than those with morphologically-aware tokenization (Hofmann et al., 2021).

6 Conclusion and Future Work

In this work, we consider the diverse time periods represented in the Latin treebanks when training and evaluating morphological taggers. We hope the genre metadata we provide can be used for future cross-genre analysis of Latin, similar to the cross-time analysis we present in this paper.

We also believe further improvements can be made through the harmonization of remaining annotation differences (§3.4) and more informed modeling choices. Specifically, we hypothesize that (1) conditioning morphological feature prediction on UPOS, or vice versa; (2) enforcing grammatical constraints through modeling, rather than only through training data; and (3) constructing a morphologically-aware tokenizer may all lead to improved performance.

Acknowledgments

We thank Gregory Crane, David Smith, Geoff Bakewell, the UMass NLP group, and anonymous reviewers for feedback. This work was supported by NSF CAREER 1845576. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A Contextual Language Model for Classical Philology. *ArXiv*, abs/2009.10053.

David Bamman and Gregory Crane. 2011. The Ancient Greek and Latin Dependency Treebanks. In *Language Technology for Cultural Heritage*, pages 79–98, Berlin, Heidelberg. Springer Berlin Heidelberg.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Adriane Boyd and Vincent D. Warmerdam. 2022. floret: lightweight, robust word vectors.

Patrick J. Burns. 2023. LatinCy: Synthetic Trained Pipelines for Latin NLP. *Preprint*, arXiv:2305.04365.

Flavio Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020a. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin works.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020b. 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, Marseille, France. European Language Resources Association.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Joseph Denooz. 2004. Opera Latina: une base de données sur internet. *Euphrosyne*, 32:79–88.

Steffen Eger, Rüdiger Gleim, and Alexander Mehler. 2016. Lemmatization and Morphological Tagging in German and Latin: A Comparison and a Survey of the State-of-the-art. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1507–1513, Portorož, Slovenia. European Language Resources Association (ELRA).

Steffen Eger, Tim vor der Brück, and Alexander Mehler. 2015. Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 105–113, Beijing, China. Association for Computational Linguistics.

Ernst Eypasch, Rolf Lefering, CK Kum, and Hans Troidl. 1995. Probability of adverse events that have not yet occurred: a statistical reminder. *BMJ*, 311(7005):619–620.

Federica Gamba and Daniel Zeman. 2023a. Latin Morphology through the Centuries: Ensuring Consistency for Better Language Processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stephan Greene and Philip Resnik. 2009. More than Words: Syntactic Packaging and Implicit Sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado. Association for Computational Linguistics.

James B. Greenough and Joseph Henry Allen. 1903. *Allen and Greenough's New Latin Grammar for Schools and Colleges*. Ginn Company.

Dag Trygve Truslew Haug and Marius L. Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European BibleTranslations.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

Timo Korkiakangas and Marco Carlo Passarotti. 2011. Challenges in Annotating Medieval Latin Charters. *J. Lang. Technol. Comput. Linguistics*, 26:103–114.

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.

Wouter Mercelis and Alek Keersmaekers. 2022. An ELECTRA Model for Latin Token Tagging Tasks. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 189–192, Marseille, France. European Language Resources Association.

Sebastian Nehrdich and Oliver Hellwig. 2022. Accurate Dependency Parsing and Tagging of Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 20–25, Marseille, France. European Language Resources Association.

H. P. V. Nunn. 1922. *An introduction to ecclesiastical Latin*, page x–xi. Cambridge University Press.

Marco Passarotti. 2019. *The Project of the Index Thomisticus Treebank*, pages 299–320. De Gruyter Saur, Berlin, Boston.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Hannah Rashkin, Eric Bell, Yejin Choi, and Svitlana Volkova. 2017. Multilingual Connotation Frames: A Case Study on Social Media for Targeted Sentiment Analysis and Forecast. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 459–464, Vancouver, Canada. Association for Computational Linguistics.

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.

Marius Sala and Rebecca Posner. 2024. Latin language. In *Encyclopedia Britannica*.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation Frames of Power and Agency in Modern Films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

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).

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

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 Jan Hajič. 2019. Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing. *Preprint*, arXiv:1908.07448.

Elina Stüssi and Phillip Ströbel. 2024. Part-of-Speech Tagging of 16th-Century Latin with GPT. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 196–206, St. Julians, Malta. Association for Computational Linguistics.

Larry Wasserman. 2004. *All of Statistics*. Springer Science & Business Media.

F.M. Wheelock and R.A. LeFleur. 2010. *Wheelock's Latin: The Classic Introductory Latin Course, Based on Ancient Authors*. Elsie Giddings Helman Memorial. HarperCollins.

Krzysztof Wróbel and Krzysztof Nowak. 2022. Transformer-based Part-of-Speech Tagging and Lemmatization for Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 193–197, Marseille, France. European Language Resources Association.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 UD Genres

We mark the following 12 genres: narrative, poem, short poem, letter, epic, history, satire, speech, treatise, Christian, Bible, legal.

These genres are not exclusive, so each text will have at least one, but possibly more genres marked.

In Figure 2, for simplicity we showed a subset of genres which are mutually exclusive. This also ensures the number of sentences shown in the figure exactly matches the number of sentences that exist in the UD treebanks. The additional genres that we left out of the figure are broader, covering multiple sub-genres. Specifically, *narratives* includes some (not all) texts from every genre except for legal texts and speeches. *Poems* includes epics and short poems, the two of which are mutually exclusive. *Christian* includes the Bible itself, as well as the religious treatises of Thomas Aquinas.

A.2 Finding Duplicate Sentences in LASLA and UD Treebanks

In order to detect *duplicate sentences* between the treebanks, we first normalize the orthographic variation across the UD treebanks and LASLA. We used CLTK's (Johnson et al., 2021) JV replacer on the harmonized UD treebanks, since LASLA's texts do not use the letters 'j' or 'v'. We also remove any punctuation present in the UD treebanks, as LASLA does not have punctuation.

We search for duplicate sentences by finding sentence pairs with exact character or token overlap at the beginning or end of each sentence.

Within the duplicate sentences, we identify *duplicate tokens* by searching for the longest overlapping, contiguous subsequence of tokens of each sentence. We search for exact token matches. Our reported number of duplicate tokens is an underestimate, since there are sometimes token mismatches within sentences that are genuine duplicates. For example, one sentence may have a numeral where the other has the word form of the number.

A.3 Annotation Agreement

In Table 8, we show the annotation agreement between duplicate sentences in Perseus and PROIEL after our standardization and harmonization. Notably, these are both (Harmonized) UD treebanks (Table 1 row 4), and some annotation differences still remain, although agreement is generally still higher than between UD and LASLA.

Feature	% same	# same	Total
Case	96.1	794	826
Degree	50.0	5	10
Gender	94.7	767	810
Mood	93.1	349	375
Number	97.7	1097	1123
Person	100.0	347	347
Tense	95.4	356	373
Voice	98.4	369	375
UPOS	97.6	1538	1576

Table 8: Percent and number of tokens in the duplicate Perseus and PROIEL sentences that have the **exact same value** for each feature, after our harmonization and conversion to Standard Latin grammar

Feature	% same	# same	Total
AdpType	76.8	2115	2753
AdvType	0.0	0	357
Aspect	97.1	8608	8864
Case	97.8	20372	20821
Compound	0.0	0	1
ConjType	0.0	0	5
Degree	8.5	598	6998
Foreign	0.0	0	2
Gender	74.7	14965	20034
Gender_loose	97.2	19481	20034
InflClass	0.0	0	27580
InflClass[nominal]	0.0	0	3394
Mood	99.4	5279	5312
Number	97.9	25672	26211
Number[psor]	100.0	281	281
NumForm	0.0	0	268
NumType	71.2	497	698
PartType	6.2	4	65
Person	91.0	6089	6692
Person[psor]	96.3	501	520
Polarity	35.1	267	760
Poss	96.3	501	520
PronType	78.2	4952	6333
Reflex	91.7	584	637
Tense	77.3	5228	6766
Variant	0.0	0	43
VerbForm	93.2	8264	8867
Voice	96.0	7493	7809
UPOS	93.0	34814	37425

Table 9: Percent and number of tokens in the duplicate LASLA and Harmonized UD sentences that have the **exact same value** for each feature, before any harmonization or standardization by us.

In Table 9, we show the annotation agreement between Harmonized UD and LASLA, before our harmonization and standardization. This table also shows the union of UD and LASLA's feature sets. There are many features we did not consider which could benefit from harmonization.

Finally, Table 12 is a venn diagram showing all possible features and values in UD and LASLA, before our harmonization and standardization.

A.4 Remaining Inconsistencies We've Harmonized

Here is a full list of arbitrary values we've enforced for certain grammatical constructions. For each of these items, there are arguably multiple correct ways to annotate—and the Latin treebanks were annotating these differently.

- Gerunds, Infinitives, and Supines should have `Number=None`
- Gerunds, Gerundives, and Supines should have `Tense=None`
- If UPOS is AUX, then `Voice=Act`. This almost entirely applies to forms of *sum*.
- Gerunds should have `Voice=Act`, and Gerundives should have `Voice=Pass`.
- Supines have `Voice=Act`, unless used in a construction with *iri*, in which case `Voice=Pass`.
- Gerunds, Infinitives, and Supines should have `Gender=None`.

A.5 Standardizing Tense and Mood

Tense We use the `TraditionalTense` field of the harmonized treebanks (Gamba and Zeman, 2023a), rather than the UD approach to tense. Altogether, the UD Latin treebanks include four tenses (present, past, future, pluperfect) and four aspects (imperfective, perfective, prospective, inchoative). When tense and aspect are considered together, they can represent the seven traditional Latin tenses. However, this is less intuitive for Classicists or those whose goal is to study only Latin. We chose to revert back to the traditional tenses. We were able to use the `TraditionalTense` field for most tags, but to differentiate between future and future perfect it is also necessary to look at `Aspect`. Additionally, we found that infinitives did not have a `TraditionalTense`, so we looked to the `Aspect` feature value to determine the tense of infinitives.

For LASLA, since it does not have a `TraditionalTense` field, we look at both `Tense` and `Aspect` feature values to determine tense.

Our final set of tenses is: present, imperfect, perfect, pluperfect, future, and future perfect.

Mood Similar to tense, the non-finite moods are represented by a combination of the `Mood` and `VerbForm` fields in Gamba and Zeman (2023a)'s harmonized treebanks, with references to Latin-specific constructions being moved to the `TraditionalMood` field. Strictly speaking, non-finite verbs do not have mood, but traditional Latin grammars still classify the different non-finite verbforms as "mood."[19] Again, we opt to use the traditional terminology and follow the same tagset as the Perseus treebank. For finite verbs, this includes indicative, subjunctive, imperative; and for non-finite verbs, infinitive, participle, gerund, gerundive, and supine.

For LASLA, we are able to take the mood directly from the `Mood` feature for finite verbs, and from `VerbForm` feature for non-finite verbs. This is because LASLA uses the Latin-specific `Ger,Gdv,Sup` values for `VerbForm`, unlike the harmonized UD treebanks.

A.6 Remaining Inconsistencies We're Unable to Harmonize

We are aware of the following differences, but leave their harmonization to future work:

- The pre-UD Perseus treebank (Table 1 row 1) has an additional `Voice` value for deponent verbs. After Gamba and Zeman (2023a)'s harmonization, deponent verbs in Perseus always have `Voice=Act`, but deponent verbs in every other UD treebank have `Voice=Pass`. We would like a system more similar to pre-UD Perseus with an additional `Voice=Dep` value.
- ITTB is the only treebank that sometimes marks *esse*, the infinitive of *sum*, as NOUN with `Mood=None`.

The following annotation differences were found to cause 9% of sampled errors in our BERT tagger's morphological predictions:

[19] This is explained in the EvaLatin 2022 guidelines: https://github.com/CIRCSE/LT4HALA/blob/master/2022/data_and_doc/EvaLatin_2022_guidelines_v1.pdf

Model	Train Set(s)	per-seus	pro-iel	llct	ittb	uda-nte
LatinCy	All UD	.729	.800	.800	.786	.737
BERT	All UD	**.872**	.974	.982	.980	**.855**
Stanza	In-Domain UD	.809	.967	.982	.977	.841
BERT	In-Domain UD	.867	**.977**	**.984**	**.986**	.880

Table 10: Macro F1 of **UPOS**. Train set is either All 5 UD treebanks, or a single In-Domain UD Treebank (i.e., same as the Test column).

- Whether to have `Case=None` for undeclined nouns.
- Whether deponent verbs should be labeled as `Voice=Act` or `Voice=Pass`.
- Whether infinitives should have a value for `Case`.
- Whether infinitives can have their UPOS be NOUN, `Mood=None`, and `Tense=None`.
- Whether the pronoun *sui* should always have `Number=None`.

A.7 Finetuning Details

We use the same hyperparameters that Bamman and Burns (2020) used to finetune a POS tagger: Adam optimizer with learning rate 5×10^{-5}, early stopping patience of 10 epochs, batch size 32, dropout rate 0.25. We keep the model with the lowest validation loss across all epochs.

A.8 Metrics

Whole-String Morphological Accuracy Following the convention of Gamba and Zeman (2023a) and Sprugnoli et al. (2022), we consider the model's prediction correct when every morphological feature is correctly predicted. We construct a morphological feature string from the predicted feature set, making sure to sort the features alphabetically. Then, we can test whether the predicted morphological string is an exact match to the gold string. Although this is a strict criteria, it indicates whether the model understands how all the morphological features fit together.

Macro F1 for Individual Features For UPOS and each individual morphological feature, we report Macro F1 in order to see how the model performs on rare feature values. If we define F as a particular feature and $\mathcal{V}_F = \{v_1, ..., v_n\}$ as the set of possible values that F can take, then macro F1 is defined as $\frac{1}{n}\sum_{i=1}^{n} \text{F1}(v_i)$. Note that $v =$ `None` is a possible value for every morphological feature, and is included in our calculation.

Feature	classical	bible	postclass
Case	0.946	**0.953**	0.948
Degree	0.977	**0.987**	0.965
Gender	0.968	0.977	**0.982**
Mood	0.859	0.938	**0.982**
Number	0.987	0.988	0.992
Person	0.994	0.993	0.992
Tense	0.955	0.977	0.954
Voice	0.969	0.973	0.990

Table 11: Macro f1 of each individual feature for the `all-ud-custom` model. Note that macro f1 for `Mood` on the Classical test set seems low (0.859) because the model never predicts `Mood=Sup` (supine). Excluding that value, its macro f1 is 0.967.

A.9 Randomized Permutation Testing

Within a null simulation, for each test set sentence we shuffle the two models' predictions, and store the absolute difference in the performance metric calculated from the entire shuffled test set. We finally report the p-value as the fraction of 10,000 simulated absolute differences that are larger than the observed absolute difference. p=0 simply means the observed difference is larger than in all simulations; it could be more conservatively interpreted as $p < .0003$ (Eypasch et al., 1995) due to Monte Carlo error.

Feature Union	UD Only Values	Value Intersection	LASLA Only Values
Abbr		Yes	
AdpType	Post	Prep	
AdvType	Loc, Tim		
Aspect	Inch	Imp, Perf, Prosp	
Case		Loc, Acc, Abl, Voc, Nom, Dat, Gen	
Compound	Yes		
ConjType			Cmpr
Degree	Dim	Abs, Cmp	Pos
Foreign		Yes	
Form	Emp		
Gender		Fem, Neut, Masc	Fem,Neut, Fem,Masc,Neut, Fem,Masc, Masc,Neut
InflClass		LatPron, LatI, LatAnom, IndEurU, IndEurO, LatI2, IndEurI, LatA, IndEurE, IndEurA, LatX, Ind, IndEurX, LatE	IndEurA,IndEurO, IndEurInd
InflClass[nominal]	IndEurX	IndEurI, IndEurO, Ind, IndEurA	IndEurA,IndEurO, IndEurU
Mood		Ind, Sub, Imp	
NameType	Lit, Ast, Oth, Met, Giv, Nat, Let, Rel, Cal, Com, Sur, Geo		
Number		Plur, Sing	Plural
Number[psor]		Plur, Sing	
NumForm	Reference, Word	Roman	
NumType		Card, Dist, Mult, Ord	
NumValue	2		
PartType		Int, Emp	
Person		2, 3, 1	
Person[psor]		2, 3, 1	
Polarity			Neg
Poss		Yes	
PronType	Ind, Rel, Art, Rcp	Tot, Neg, Con, Prs, Rel, Int, Dem, Ind	Emp
Proper	Yes		
Reflex		Yes	
Tense		Past, Pqp, Fut, Pres	
Typo	Yes		
UPOS	PUNCT	SCONJ, ADP, ADJ, AUX, VERB, X, NUM, _, PART, INTJ, ADV, NOUN, DET, CCONJ, PROPN, PRON	
Variant			Greek
VerbForm	Conv, Vnoun	Fin, Inf, Part	Ger, Gdv, Sup
VerbType	Mod		
Voice		Pass, Act	

Table 12: Feature and Values Comparison between UD and LASLA. Note that Perseus and PROIEL (the only UD treebanks that overlap with LASLA) lack some feature values that the other UD treebanks have, but this shows the union of all UD features.

Classical (UD Only)		Bible		Post Classical	
Work	# Sents	Work	# Sents	Work	# Sents
BellumGallicum	1445	jerome_vulgata-Mark	1257	aquinas_summa-contra-gentiles	23687
DeOfficiis	557	jerome_vulgata-1-John	12	dante_de-vulgari-eloquentia	419
InCatilinam	137	jerome_vulgata-2-John	3	dante_letters	376
Metamorphoseon	183	jerome_vulgata-3-John	4	dante_questio-de-aqua-et-terra	133
PetroniusSatiricon	547	jerome_vulgata-John	1765	dante_eclogues	111
PropertiusElegiae	224	jerome_vulgata-Luke	2044	llct_39	165
Catilina	336	jerome_vulgata-Galatians	189	llct_79	670
TacHistoriae	64	jerome_vulgata-Titus	39	palladius_opus-agriculturae	955
Aeneis	68	jerome_vulgata-1-Thessalonians	97	llct_36	276
cicero_letters-to-atticus-1	703	jerome_vulgata-James	7	llct_80	571
cicero_letters-to-atticus-2	800	jerome_vulgata-Acts	1490	llct_72	271
cicero_letters-to-atticus-4	703	jerome_vulgata-Hebrews	13	llct_83	812
cicero_letters-to-atticus-5	688	jerome_vulgata-Colossians	29	llct_73	518
cicero_letters-to-atticus-6	270	jerome_vulgata-revelation	763	llct_40	324
		jerome_vulgata-1-Corinthians	736	llct_84	771
		jerome_vulgata-2-Peter	2	llct_86	826
		jerome_vulgata-Matthew	1978	llct_75	462
		jerome_vulgata-2-Corinthians	345	llct_38	276
				llct_74	404
				llct_81	288
				llct_77	333
				llct_76	216
				llct_85	807
Total	6725	**Total**	10773	**Total**	33671
phaedrus_fabulae	389	jerome_vulgata-1-Peter	5	aquinas_forma	3290
augustus_res-gestae	38	jerome_vulgata-1-Timothy	4	dante_monarchia	682
suetonius_life-of-augustus	109	jerome_vulgata-2-Thessalonians	37	llct_37	170
cicero_letters-to-atticus-3	420	jerome_vulgata-2-Timothy	47	llct_78	389
cicero_letters-to-atticus-7	85	jerome_vulgata-Ephesians	100	llct_82	472
		jerome_vulgata-Jude	22		
		jerome_vulgata-Philemon	25		
		jerome_vulgata-Philippians	97		
		jerome_vulgata-Romans	684		
Total	1041	**Total**	1021	**Total**	5003

Table 13: Number of UD sentences in our custom train (top) and test (bottom) splits. Works that appear only in LASLA are not listed, as there are too many. See LASLA's website for a full list.