Contextual morphologically-guided tokenization for Latin encoder models

Marisa Hudspeth[1] **Patrick J. Burns**[2] **Brendan O'Connor**[1]

[1]Manning College of Information & Computer Sciences, University of Massachusetts Amherst
[2]Institute for the Study of the Ancient World, New York University
{mhudspeth,brenocon}@cs.umass.edu pjb311@nyu.edu

Abstract

Tokenization is a critical component of language model pretraining, yet standard tokenization methods often prioritize information-theoretical goals like high compression and low fertility rather than linguistic goals like morphological alignment. In fact, they have been shown to be suboptimal for morphologically rich languages, where tokenization quality directly impacts downstream performance. In this work, we investigate morphologically-aware tokenization for Latin, a morphologically rich language that is medium-resource in terms of pretraining data, but high-resource in terms of curated lexical resources – a distinction that is often overlooked but critical in discussions of low-resource language modeling. We find that morphologically-guided tokenization improves overall performance on four downstream tasks. Performance gains are most pronounced for out of domain texts, highlighting our models' improved generalization ability. Our findings demonstrate the utility of linguistic resources to improve language modeling for morphologically complex languages. For low-resource languages that lack large-scale pretraining data, the development and incorporation of linguistic resources can serve as a feasible alternative to improve LM performance.[1]

1 Introduction

Tokenization is the first step in Large Language Model (LLM) pretraining pipeline, making it the foundation upon which model performance rests. A common assumption is that tokenizers should maximize compression and minimize fertility (Schmidt et al., 2024). However, recent research has challenged this view, particularly in the context of morphologically rich and lower-resource languages. Studies have shown that existing tokenization methods exhibit low morphological alignment (Jabbar, 2024; Bostrom and Durrett, 2020; Erkaya, 2022; Libovick'y and Helcl, 2024), which can negatively impact downstream performance. In this work, we investigate tokenization strategies for Latin, a morphologically rich, medium-resource language[2] with a long scholarly tradition, and moreover one with specific interest in word endings and other aspects of word formation, making it an informative test case.

We hypothesize that incorporating morphological knowledge into tokenization will improve both morphological alignment and downstream performance. While prior work has explored morphologically-aware tokenizers, they often focus on high-resource and/or morphologically simple languages (Hofmann et al., 2021; Jabbar, 2024; Bostrom and Durrett, 2020) or employ acontextual, unsupervised morphological analyzers such as Morfessor (Creutz and Lagus, 2005). Furthermore, evaluations in this area have not examined fine-grained morphological feature prediction, which should better capture whether a morphologically-aligned tokenization helps the language model's contextual embeddings capture its linguistic content.[3]

Beyond its computational implications, this question is of particular interest within Latin linguistic and philological research. The role of word endings in meaning construction has been central to Latin scholarship for centuries, making it crucial to empirically test whether morphology-informed tokenization aligns with these long-standing linguistic intuitions.

One of the unique advantages of working with Latin is the availability of expert-curated morphological and lexical resources—benefiting from long-standing philological study about the role of

[1] Code and data are available here: https://github.com/slanglab/latin-morpheme-tokenization

[2] Between 100M and 1B tokens (Chang et al., 2024)

[3] Much prior work attains a partial view of this by evaluating on POS tagging, a significantly coarser version of the problem.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5761–5775
March 24-29, 2026 ©2026 Association for Computational Linguistics

word endings in constructing meaning (Diederich, 1936; Pellegrini et al., 2021). Although Latin lacks the raw text data suitable for modern, large-scale pretraining, this abundance of linguistic resources makes Latin a useful case study to investigate whether the development of comparable resources for low-resource languages can meaningfully improve language modeling performance, especially in cases where acquiring more pretraining data is infeasible.

To test our hypothesis, we experiment with three types of tokenizers:

1. **Baseline:** Standard WordPiece (WP) and Unigram Language Model (ULM)
2. **Low-guidance approach:** Seeding morphological suffixes into the tokenizer vocabulary. This approach is lightweight, only requiring a predefined list of suffixes.
3. **High-guidance approach:** Pre-Tokenization using a morphological analyzer. Unlike prior work, we disambiguate the analyses based on POS information, making our method context-aware.

For evaluation, we pretrain Latin RoBERTa models (Liu et al., 2019) using each tokenizer and test them on **four downstream tasks**: POS and morphological feature tagging, named entity recognition (NER), word sense disambiguation (WSD), and authorship verification (AV). To our knowledge, no prior studies have evaluated their morphologically aware tokenization methods using morphological feature classification as a downstream task.

2 Related Work

2.1 Tokenization

Tokenization plays a crucial role in language model pretraining, yet its impact on morphologically rich languages remains an active area of investigation. Studies have increasingly questioned whether widely used tokenization methods such as Byte Pair Encoding (BPE) (Sennrich et al., 2016), Unigram Language Model (ULM) (Kudo, 2018), and WordPiece (Schuster and Nakajima, 2012) sufficiently capture morphological structure, and whether this matters for downstream performance.

Several studies suggest that aligning BPE and ULM tokenizers' segmentations with gold-standard morphological boundaries can enhance downstream performance on sentiment and topic classification (Hofmann et al. (2021); English), zero-shot summarization and retrieval (Jabbar (2024); English), QA, MNLI, NER (Bostrom and Durrett (2020); English and Japanese), and other classification, structured prediction, and similarity tasks (Vemula et al. (2025); English, Hindi, and Telugu).

Studies on morphologically rich languages in particular provide further evidence for the benefits of morphology-aware tokenization. Rule-based approaches have shown improvements in Romanian NLP tasks (Vasiu and Potolea, 2020), and pre-tokenization methods using either 1) a language-specific morphological analyzer (Erkaya, 2022; Nzeyimana and Niyongabo Rubungo, 2022; Vemula et al., 2025), or 2) unsupervised methods such as Morfessor (Park et al., 2021; Libovick'y and Helcl, 2024; Vemula et al., 2025) have all yielded downstream performance gains. Post-training strategies have also been effective; for instance, modifying existing BPE vocabularies improved token-level tasks in English, Dutch, and German (Bauwens and Delobelle, 2024).

While most morphologically-aware tokenization methods rely on static rules or unsupervised segmentation, some studies have experimented with adding contextual information. Yehezkel and Pinter (2023) introduced SaGe, a tokenizer whose vocabulary construction method closely resembles ULM but incorporates a SkipGram objective to refine vocabulary selection. This approach improved performance on English GLUE and NER, and Turkish Inference and NER.

Other studies directly oppose the hypothesis that morphological alignment is beneficial. Toraman et al. (2023) found no improvements in Turkish NLP tasks when pre-tokenizing with a morphological analyzer, though they noted that errors in the analyzer itself may have influenced results. More broadly, Arnett and Bergen (2025) argued that morphological alignment is not a key factor in tokenization quality, emphasizing instead that dataset size and quality are more important. Arnett et al. (2025) even find a small negative correlation between morphological alignment and downstream task performance across 70 languages. However, this analysis only considers previously reported measures of downstream task performance when they exist–which is typically only for high-resource languages, and for large multilingual models–and perplexity when they do not (understandably, since the very nature of low resource languages means they lack evaluation data).

The existing literature provides strong, though not unanimous, evidence that morphologically-aware tokenization can improve NLP performance, particularly for morphologically rich languages. However, prior studies have largely focused on a limited set of downstream tasks—such as POS tagging and NER—which may not fully expose the benefits of morphologically-aligned tokenization. Our work extends this research by evaluating multiple tokenization strategies for Latin, including both light and high-guidance approaches. Additionally, we introduce morphological feature classification as an alternative downstream evaluation metric, hypothesizing that the fine-grained morphological feature values will reveal improvements that are less evident in coarse-grained POS tagging.

2.2 Morpheme Segmentation

Morpheme segmentation has been widely studied as an independent NLP task, distinct from its potential applications in tokenization and language model pretraining. One of the most prominent efforts in this area is the SIGMORPHON shared task on morpheme segmentation (Batsuren et al., 2022), which provided segmentation data for nine languages, including Latin. This task included both acontextual and contextual segmentation challenges; however, Latin was excluded from the contextual segmentation track. While the availability of segmentation resources for these nine languages is valuable, the dataset is too small to support pretraining efforts. Moreover, many morphological datasets, including those used in SIGMORPHON, are constructed through automatic extraction from sources such as Wiktionary, introducing data quality concerns. For example, Gorman et al. (2019) highlight extensive extraction errors in the dataset for the 2017 CoNLL-SIGMORPHON shared task for Morphological Reinflection (Cotterell et al., 2017).

Latin is unique within morpheme segmentation research due to its rich morphological tradition and availability of high-quality, expert-curated resources. Unlike many other languages, Latin benefits from centuries of linguistic study focused on word formation and morphological structure (Diederich, 1936; Pellegrini et al., 2021).

Our work builds on the precisely curated morphological resources available for Latin, incorporating linguistic knowledge at the morpheme level in a way that is not feasible for many other languages. We argue that context-aware morphological tokenization–though requiring significant language-specific effort–has the potential to bridge the gap between linguistic theory and modern NLP.

2.3 Language Modeling

Several studies have demonstrated the feasibility of pretraining transformer-based models for low-resource languages. Ogueji et al. (2021) introduced AfriBERTa, a multilingual BERT model trained on various low-resource African languages using a corpus of approximately 1GB, comparable in size to ours. Their model outperforms massively multilingual models like mBERT and XLM-R (Conneau et al., 2020) on NER and text classification. Similarly, Chang et al. (2024) systematically evaluated the relationship between model size (from tiny to small) and data size across both monolingual and multilingual GPT-2 models (Radford et al., 2019). For morphologically-rich languages in particular, prior work has shown that character-level models often outperform tradition subword models on syntactic and other lower-level tasks (Vania et al., 2018; Cao, 2023). However, this advantage also comes with a tradeoff in reduced training efficiency, and prior work has also shown that character-level models still benefit from explicit modeling of morphology (Vania et al., 2018).

Prior work in Latin language modeling has produced several pretrained, transformer-based Latin models. LaBERTa (Riemenschneider and Frank, 2023) was trained on 165M words from Corpus Corporum (Roelli, 2014),[4] a kind of super-repository of available smaller digitized Latin text repositories. Another Latin RoBERTa model (Ströbel, 2022; Liu et al., 2019) was trained on a 390M token corpus also derived from Corpus Corporum. Finally, LatinBERT (Bamman and Burns, 2020) was trained on a larger corpus (642.7M words), though a significant portion originated from noisy OCR-processed Latin texts from the Internet Archive. Its cleaner subset contained 81.6M words.

Our work differs from these prior efforts in Latin language modeling in two ways. First, our pretraining training corpus, totaling 195M words (1GB), is larger than the clean subset used in LatinBERT though smaller than the dataset used for Latin RoBERTa. Finally, we experiment with various morphologically-aware tokenization strategies, whereas existing Latin language models use base-

[4] https://mlat.uzh.ch/

line WordPiece (LatinBERT) and BPE (LaBERTa, Latin RoBERTa).

3 Background: Tokenization

Schmidt et al. (2024) conceptualize tokenization as a three-step process: 1) pretokenization, which applies an initial set of rules to define processing units–typically by segmenting on whitespace; 2) vocabulary construction or training, where subword units are learned; 3) segmentation or decoding, which determines how input text is tokenized based on the trained vocabulary. This framework helps highlight the different places where morphological guidance can be introduced, instead of viewing tokenizers as indivisible systems. We experiment with modifications to two widely-used Huggingface tokenizer implementations.

3.1 WordPiece Tokenization

BPE and WordPiece are tokenizers with similar training algorithms. While BPE is widely used, prior work finds it suboptimal (§2.1), so our experiments focus on WordPiece. For clarity, we overview both in this section.

Pretokenization for both tokenizers is typically done by splitting on whitespace and punctuation. Given a list of (pretokenized) strings and a desired final vocabulary size, an initial subword vocabulary is constructed from all unique characters. Subword types are iteratively merged until the desired vocabulary size is reached. For WordPiece, the subword bigram with the highest pointwise mutual information (PMI) (Bouma, 2009) is chosen. Its two subwords are merged into a new, single subword and added to the vocabulary. For BPE, the bigram with the highest frequency is chosen rather than highest PMI.

To tokenize new text, WordPiece performs greedy left-to-right decoding, whereas BPE applies merge rules in the order learned during training.

3.2 Unigram Language Model Tokenization

The Unigram Language Model (Kudo and Richardson, 2018) infers the most likely segmentation for a word, using the Viterbi algorithm. For learning, after initial pretokenization,[5] the model starts with a large vocabulary of all substrings in the corpus. Subwords are iteratively pruned in order to maximize the unigram likelihood of the corpus until the desired vocabulary size is reached.

4 Method

4.1 Morphologically-Enhanced Tokenizers

We add morphological guidance to both ULM and WordPiece tokenizer models, implemented by modifying HuggingFace's implementations of each (Wolf et al., 2020). We create and evaluate three tokenizer variations: MorphSeeding, MorphPreTokenization (acontextual), and MorphPreTokenization (contextual).

Data We train our tokenizers on our pretraining corpus.[6]

MorphSeeding We create a list of 480 morphological suffixes sourced from Lemlat, a type-level lemmatizer and morphological analyzer for Latin (Passarotti et al., 2017). For our purposes, we define morphological suffixes as all segments of a word which are not the first (root/stem) segment.

Then, we modify the ULM and WordPiece trainers to bias them to prefer segmenting with this list of suffixes. For WordPiece, all suffixes are added to the initial vocabulary with the continuing subword prefix "##" prepended. Since WordPiece's vocabulary construction is bottom-up, once added to the vocabulary a subword cannot be removed. For ULM, all suffixes are added to the initial vocabulary, and for decoding, their log-probabilities are upweighted by a fixed amount[7] in the lattice. Suffixes are not allowed to be removed from the vocabulary.

MorphPreTokenization We analyze all unique words in our corpus with Lemlat. For each word, Lemlat returns a list of possible analyses that include the word's segmentation into morphemes, as well as its lemma, declension or conjugation, part of speech, morphological features, and derivational affixes. We only utilize the segmentation and POS.

We then presegment these morphemes in our corpus, so that during tokenizer training, it will never merge Lemlat-provided morphemes. We experiment with both contextual and acontextual segmenters.

[5] Pretokenization is also done by splitting on whitespace. In the HuggingFace implementation, punctuation is not split on.

[6] The ULM tokenizers are trained on 5% of this corpus. See §8.

[7] During initial experiments, we found a weight of 0.5 to strike a good balance of encouraging these suffixes to be chosen more often when decoding, while not increasing fertility to an unreasonable degree.

For acontextual pretokenization, we simply use the segmentation in the first analysis given by Lemlat. This follows the type-level focus of previous work, either with language-specific morphological analyzers (Toraman et al., 2023; Erkaya, 2022; Nzeyimana and Niyongabo Rubungo, 2022) or the unsupervised Morfessor model (Creutz and Lagus, 2005; Libovick'y and Helcl, 2024).

But in many instances, there exists ambiguity over a word's segmentation, which can be resolved with contextual information about its grammatical role. Thus, we construct a contextual morphological segmenter by first running an off-the-shelf part-of-speech tagger on the corpus, and filtering Lemlat's output to an analysis with a matching POS tag.[8]

We tag our corpus with LatinCy (Burns, 2023). It uses the Latin UD Treebanks' tagset, which differs from Lemlat's. We create a mapping between the tag systems (Table 7), and a protocol for selecting a word's segmentation when the POS tags do not match:

- If Lemlat only gives one unique possible segmentation, use that one (occurs in 1.2% of words in UD treebanks).
- If Lemlat gives multiple possible segmentations but none match the predicted POS, do not segment the word (occurs in 0.028% of words in UD treebanks).

A word's tag usually disambiguates the segmentation, but in rare cases, one word may have multiple analyses with the same POS tag, due to multiple possible lemmas or morphological features. In these cases, we choose one segmentation based on the following criteria:

- If the candidate segmentations have the same number of subwords, choose the one with the longer suffix (i.e. out of the adjective [*adversar*, *-i*] versus infinitive verb [*advers*, *-ari*] choose the latter).
- If the candidate segmentations have a different number of subwords, choose the one with more subwords (i.e. out of participle [*inordin*, *-at*, *-o*] and imperative [*inordin*, *-ato*], choose the former).

This type of conflict occurred in 4.55% of all Lemlat analyses of unique words in the Latin treebanks.[9]

This results in four morphoglical pretokenization-based tokenizers—for each model class (ULM and WordPiece), there is both acontextual and contextual presegmentation.

We implement changes to the tokenizers to accommodate presegmentation. For ULM, the only modification is to add a new pre-tokenization step which splits on our morpheme symbol, allowing morphological suffixes to be treated as continuing subwords. Due to how the ULM vocabulary is constructed, it is possible that the suffixes will be split into multiple subwords, just like the root.

WordPiece requires modification to its trainer, not just the pretokenizer; for implementation details, see §A.2. Unlike ULM, once the suffix subword is added to the WordPiece vocabulary, it will remain unchanged, neither split or merged.

4.2 Tokenizer Evaluation

Several metrics have been proposed to assess tokenizer quality.

Renyi entropy (Zouhar et al., 2023) measures the uniformity of token frequency distributions. However, Schmidt et al. (2024) found that it correlates with Corpus Token Count, which they argue is a poor predictor of downstream performance.

A more linguistically motivated approach is morphological alignment with a gold reference segmentation, which assumes that having "meaningful" subword units improves downstream task performance. Various metrics have been introduced to quantify this.

MorphScore (Arnett and Bergen, 2025; Arnett et al., 2025) assigns a score of 1 if a tokenizer correctly segments at a specific morpheme boundary, regardless of other boundaries in the word, and 0 otherwise. Unlike other measures, it excludes words that remain unsegmented.

Suffix precision, recall, and f1 (Erkaya, 2022) evaluate how well a tokenizer captures *suffix* boundaries specifically. Subword boundary precision, recall, f1 (Bostrom and Durrett, 2020) which we adopt in this work, assess overall segmentation accuracy. In addition, we track exact matches between predicted and gold segmentations.

The reliability of these metrics depends on the quality of the gold standard segmentations. Many

[8] Interestingly, for some tasks the approach may seem circular: a predicted POS tag helps guide the LLM tokenization, and thus the eventual LLM contextual representation used to predict, for example, a POS tag. Investigating how initial tagging errors propagate would be interesting future work.

[9] In both the UD treebanks and in LASLA (Denooz, 2004), a non-UD Latin treebank.

studies experiment with multiple languages and rely on morphological data scraped from Wiktionary. Gorman et al. (2019) highlight extensive errors in SIGMORPHON's morpheme reinflection data (Cotterell et al., 2017), demonstrating that such resources may introduce inconsistencies. This underscores the importance of carefully curating high-quality gold standards, which tends to be easier when focusing on a single language. When Latin scholar coauthors reviewed the SIGMORPHON segmentation dictionary alongside two open-source Latin morphological dictionaries (Lemlat and Whitaker's Words[10]), we judged Lemlat to be highest quality. Lemlat has also been shown to have better coverage of Latin word types and tokens than Words, and equivalent coverage to LatMor, a finite state transducer for Latin (Springmann et al., 2016). [11]

To construct an evaluation set, we extract all unique (word, POS) pairs from the five Latin UD test sets, and segment them using Lemlat. In the acontextual setting, we ignore POS and consider the first segmentation given by Lemlat as the gold segmentation. In the contextual setting, we disambiguate Lemlat's analyses using the gold UD POS, choosing the gold segmentation as described in §4.1.

We also test the morphological alignment of our tokenizers as compared to Latin SIGMORPHON 2022 segmentations, which are acontextual (Batsuren et al., 2022). This provides a fairer evaluation than testing against Lemlat segmentations, since we intentionally design our tokenizers to incorporate morphological information from Lemlat.

4.3 Model Pretraining

Data We train our tokenizers and pretrain our RoBERTa models on the same data used to train the static floret vectors (Boyd and Warmerdam, 2022) used in LatinCy, a spaCy pipeline for Latin (Burns, 2023; Honnibal and Montani, 2017). It is 1.08GB, containing 13.5M sentences and 195M whitespace-separated words. This is comparable to the pretraining data size of other Latin transformer models; for example, Riemenschneider and Frank (2023) trained LaBERTa on 167.5M words.

Models We pretrain eight base (110M parameter) RoBERTa models (Liu et al., 2019) using the HuggingFace Transformers library (Wolf et al., 2020). See §A.3 for details on hyperparameters.

4.4 Downstream Tasks

Task	Test Size	Source
Morph	5,603 sents	Hudspeth et al. (2024)
NER	3,410 sents	Beersmans et al. (2023)
WSD	533 sents (40 lemmas)	Ghinassi et al. (2024)
AV	220 text pairs per trial	Gorovaia et al. (2024)

Table 1: Scale and sources of downstream task evaluation sets.

We evaluate our models on four downstream tasks: two token-level (POS/Morphological feature classification, NER) and two sequence-level (WSD and Authorship Verification (AV)). We report overall test set sizes in Table 1, and sizes of in and out domain splits in Table 11. More detailed descriptions of each task can be found in A.5.

For each task, we report the following metrics: whole-string morphological accuracy and per-feature macro F1 for POS and morphological feature classification; BI-label F1 and per-entity micro F1 for NER; average F1 over lemmas for word sense disambiguation (WSD); and average F1 across trials for authorship verification (AV).

We do not directly compare our models to existing Latin encoder models such as LatinBERT or LaBERTa, as their training data and setup differ and are not publicly available. Our experiments aim to isolate the impact of tokenization strategies on downstream performance, so comparisons are limited to models trained under our controlled conditions. See A.7 for further discussion.

5 Results

5.1 Tokenizer Evaluation

We evaluate the morphological alignment of each tokenizer by comparing predicted segmentations to a gold-standard segmentations from Lemlat and from Sigmorphon 2022. As seen in Table 2, baseline WordPiece already exhibits relatively strong alignment with gold Lemlat segmentations, outperforming ULM in this regard (+9.4% exact match against the gold contextual segmentations). However, baseline ULM is more closely aligned to the

[10] https://latin-words.com/

[11] Lemlat does not include macrons, the accent marks that indicate vowel length in Latin (e.g., mālum 'apple' vs. malum 'evil'). While useful for phonological, morphological, or metrical analysis, macrons are an editorial choice, primarily used in poetry or educational material like textbooks and dictionaries. Most regular Latin texts do not include them, and adding them automatically can introduce errors.

Tokenizer Type	Model	Sig (Actx)		Lem (Ctx)	
	Class	EM	Fert.	EM	Fert.
Baseline	ULM	7.19	3.10	10.76	1.87
MorphSeed	ULM	7.03	3.09	12.12	1.89
MorphPreTok Actx	ULM	**9.89**	3.19	65.05	2.36
MorphPreTok Ctx	ULM	9.34	3.20	71.88	2.41
Baseline	WP	3.60	2.84	20.12	1.77
MorphSeed	WP	3.46	2.76	20.18	1.77
MorphPreTok Actx	WP	8.45	2.67	75.50	2.13
MorphPreTok Ctx	WP	8.28	2.65	**84.32**	2.18

Table 2: Tokenizers' morphological segmentation accuracy across acontextual (Sigmorphon test set) and contextual (Lemlat) settings, in terms of Exact Match rate (EM) and Fertility (Fert.). Gold fertility is 2.49 for Sigmorphon and 1.94 for Lemlat.

Model	Tok	Morph	NER	WSD	AV
Baseline	ULM	89.46	65.89	**61.79**	61.27
MorphSeed	ULM	89.51	66.54	59.34	64.23
MorphPreTok Actx	ULM	*90.98	***73.52**	60.83	65.65
MorphPreTok Ctx	ULM	*91.00	*73.07	59.47	**67.28**
Baseline	WP	89.86	66.15	58.99	65.40
MorphSeed	WP	89.82	67.72	61.65	66.01
MorphPreTok Actx	WP	*91.09	*69.47	61.08	63.38
MorphPreTok Ctx	WP	***91.18**	*72.72	59.84	64.64

Table 3: Summary of results on downstream tasks. *Starred indicates significant improvement over baseline ($p < 0.05$, paired bootstrap sampling) for Morph classification, NER, and WSD. For AV, we report the mean over three independent runs and do not perform significance testing. **Bolded** values indicate best performance per column. For Morph classification and NER, we report results using the last subword token for prediction.

Sigmorphon segmentations than baseline WordPiece. This may be due to baseline ULM's higher fertility. [12]

Introducing morphological pretokenization (MorphPreTok) significantly enhances alignment to Lemlat for both ULM and WordPiece, with exact matches exceeding 65% for all variants. Alignment to Sigmorphon also increases between 3-5%. Although the exact match rate is low compared to Lemlat, this is expected since we intentionally design our tokenizers to incorporate morphological information from Lemlat. These results suggests that explicitly incorporating morphological information during pretokenization leads to segmentations that better adhere to linguistic ground truth.

By contrast, morphological suffix seeding (MorphSeed) provides only a modest improvement for ULM (+1.4% exact match against contextual Lemlat segmentations), while having no effect on WordPiece's alignment. Alignment to Sigmorphon is roughly the same for both ULM and WordPiece. This suggests that while suffix seeding can nudge segmentation toward morphological boundaries, they are less effective than full pretokenization in enforcing linguistically coherent segmentations.

This aligns with previous findings that pretokenization has a larger impact on the resulting vocabulary than the tokenization algorithm itself, since pretokenization places a hard constraint on the maximum token length (Velayuthan and Sarveswaran, 2025).

5.2 Downstream Performance

Overall improvement across all tasks As reported in Table 3, all tasks benefit from increased morphological guidance. For morphological feature tagging and NER, the MorphPreTok tokenization consistently improves performance for both the ULM (+1.5 morph acc, +7.6 NER BI f1) and WP (+1.3 morph acc, +6.6 NER BI f1) models. The MorphSeed method also shows minor gains for NER (+1.6 BI f1), but otherwise performs similarly to baselines.

Results are more mixed for WSD and Authorship Verification (AV). For the ULM models, WSD performance is hurt with more morphological guidance (-2.5 f1), but AV performance is improved (+6.0 avg f1). The reverse is true for the WP models (+2.7 WSD f1, -2.0 AV avg f1).

Model	Tok	Morph		NER	
		In	Out	In	Out
Baseline	ULM	93.04	77.73	85.53	32.17
MorphSeed	ULM	92.83	78.16	84.62	37.61
MorphPreTok Actx	ULM	93.81	81.61	**88.69**	**45.40**
MorphPreTok Ctx	ULM	93.72	82.09	88.01	42.31
Baseline	WP	93.09	78.94	84.22	32.94
MorphSeed	WP	93.09	78.82	85.13	35.71
MorphPreTok Actx	WP	**94.04**	81.68	87.22	36.82
MorphPreTok Ctx	WP	93.97	**82.25**	88.27	44.12

Table 4: **In-domain vs out-domain** whole-string morphological accuracy (Morph) and BI label F1 (NER)

Larger gains for out of domain texts In Table 4, we observe that increased morphological guidance resulted in the most significant gains for out-of-domain texts. The MorphPreTok variants showed the most improvement, with up to +4.4 morph accuracy and +13.2 BI f1 on out-of domain texts. For NER, in-domain performance was also improved, although there was no difference in in-domain per-

[12] For complete evaluation metrics, see Table 6.

			Per-Feature Macro F1								Per-Entity Micro F1		
Tokenizer Type	Model Class	POS Acc	Case	Degree	Gender	Mood	Number	Person	Tense	Voice	PERS	LOC	GRP
Baseline	ULM	94.78	72.31	90.47	92.41	79.74	95.94	97.27	90.27	95.04	65.59	60.84	67.62
MorphSeed	ULM	+0.02	-0.17	-0.44	+0.07	+0.25	+0.05	-0.04	+0.06	-0.05	+2.52	+0.79	-2.37
MorphPreTok Actx	ULM	+0.57	+2.04	+4.21	+1.05	+3.17	+0.76	+0.86	+3.48	+1.68	+8.88	+6.09	+2.68
MorphPreTok Ctx	ULM	+0.57	+6.04	+4.57	+1.16	+3.70	+0.80	+0.84	+3.51	+1.80	+7.01	+4.20	+1.66
Baseline	WP	94.99	73.10	91.69	92.63	82.07	96.15	97.65	91.99	95.83	66.83	56.28	63.87
MorphSeed	WP	+0.02	+4.16	+0.20	+0.05	+0.28	-0.05	+0.07	+0.14	+0.08	+0.62	+4.42	+4.30
MorphPreTok Actx	WP	+0.55	+0.55	+2.86	+0.95	+1.14	+0.69	+0.49	+2.04	+0.92	+1.04	+8.62	+7.84
MorphPreTok Ctx	WP	+0.41	+5.10	+3.30	+0.91	+1.49	+0.59	+0.55	+2.54	+1.09	+5.91	+11.44	+8.17

Table 5: Downstream POS accuracy and per-feature macro F1 scores (POS and Morphological Feature Tagging), and per-entity micro F1 scores (NER). Performance is shown for each baseline and, for morphologically-aware variants, the difference from that class's baseline.

formance for morphological feature classification.

MorphSeeding did not have a significant effect on morphological feature classification, but for NER had a +5.4 and +2.8 BI F1 gain over baseline for the ULM and WP models, respectively. These results speak to the improved generalization abilities of the morphologically guided tokenizers, especially those with morpheme-based pretokenization (MorphPreTok).

Gains for particular feature values and entities Table 5 shows the differences in performance compared to baseline tokenizers on particular morphological features and named entities. Again, MorphPreTok is helpful for all features and entities, for both ULM and WP. We see the strongest improvements (+2-6) in per-feature macro-f1 for Case, Degree, Mood, and Tense. Improvements for entities range from +1.66-11.44 in micro f1.

MorphSeed is also helpful for most features and entities, but to a lesser degree than MorphPreTok. Some features see minor regressions (ULM: Case, Degree, Person, Voice; WP: number) and the GRP entity has a more significant -2.37 micro f1 regression for ULM.

6 Discussion

The improvements observed in morphological feature tagging and NER are not unexpected. In Latin, morphological features are marked by word endings. Separating these endings from roots allows the model to treat them as informative subunits. Generalization to unseen words is made easier, as their meaning is irrelevant to the task, and their inflectional endings will have been seen during training.

For NER, separating inflectional endings helps the model generalize across different inflected forms of the same entity. Certain entity types, such as locations, also carry morphological markers. For example, the Locative case is used to express "at [place name]".

Mixed results for sentence-level tasks are also understandable. In word sense disambiguation, the goal is to infer meaning rather than grammatical form, so subwords that encode only inflectional information without independent semantic content may introduce unnecessary noise.

In theory, authorship verification could benefit from morphologically informed tokenization, if for example authors differ in their use of grammatical constructions. While morphologically guided ULM models showed slight gains (62.2 to 67.2 F1) over their baselines, they performed comparably to the baseline WordPiece model (66.8 F1). This suggests that other features, such as lexical choice, may be more informative for this task and could be overshadowed by strict morpheme-based pretokenization.

Although morpheme-based pretokenization improved token-level tasks, we found minimal differences between contextual and acontextual variants. Given the added complexity of contextual pretokenization, its application to other languages may not be justified.

7 Conclusion

We demonstrate that morphologically-guided tokenization improves downstream performance in Latin RoBERTa models, particularly for features that are strongly tied to morphological structure.

More broadly, our results highlight the need for continued investment in developing high-quality linguistic resources, particularly for lower-resource and morphologically complex languages, where data availability remains a key bottleneck.

8 Limitations

Our ULM tokenizers are trained on 5% of our pretraining corpus, whereas the WordPiece tokenizers are trained on the full dataset. When ULM tokenizers were trained on the full corpus, we observed pathological behavior, including high fertility and segmentations with many single-character subwords and low morphological alignment. Training on smaller datasets, and this 5% sample, yielded much more regular results, for reasons unclear to us; future work could examine if it is an issue with the HuggingFace implementation. We decided to train ULM tokenizers on a subset of the corpus, in order to have higher-quality tokenization and a fairer comparison to the WordPiece tokenizers. All RoBERTa models were pretrained on the full corpus.

We only experiment with encoder-based models. The performance gains we observed may not scale to larger models or other architecture types.

We only pretrain and finetune a single model per tokenizer, in order to reduce computational time and cost.

Acknowledgments

We would like to thank EleutherAI for providing computational resources used for this paper. Additionally, we thank the UMass NLP group and Catherine Arnett for their support and feedback. This material is based in part upon work supported by National Science Foundation award 1845576 (CAREER). 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

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Catherine Arnett, Marisa Hudspeth, and Brendan O'Connor. 2025. Evaluating morphological alignment of tokenizers in 70 languages. In *Proceedings of the ICML 2025 Tokenization Workshop (TokShop)*.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for Classical Philology. *Preprint*, arXiv: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.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Thomas Bauwens and Pieter Delobelle. 2024. BPE-knockout: Pruning pre-existing BPE tokenisers with backwards-compatible morphological semi-supervision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5810–5832, Mexico City, Mexico. Association for Computational Linguistics.

Marijke Beersmans, Evelien de Graaf, Tim Van de Cruys, and Margherita Fantoli. 2023. Training and evaluation of named entity recognition models for classical Latin. In *Proceedings of the Ancient Language Processing Workshop*, pages 1–12, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

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.

Kris Cao. 2023. What is the best recipe for character-level encoder-only modelling? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5924–5938, Toronto, Canada. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

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, Bologna, Italy. CEUR Workshop Proceedings.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Publications in computer and information science. Report A. Helsinki University of Technology, Finland.

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.

Paul B. Diederich. 1936. Seventeen Basic Latin Endings. *Educational Research Bulletin*, 15(1):1–5.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Erencan Erkaya. 2022. A comprehensive analysis of subword tokenizers for morphologically rich languages. Master's thesis, Boğaziçi University.

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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but OK: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China. Association for Computational Linguistics.

Svetlana Gorovaia, Gleb Schmidt, and Ivan P. Yamshchikov. 2024. Sui generis: Large language models for authorship attribution and verification in Latin. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 398–412, Miami, USA. Association for Computational Linguistics.

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.

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.

Marisa Hudspeth, Brendan O'Connor, and Laure Thompson. 2024. Latin treebanks in review: An evaluation of morphological tagging across time. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 203–218, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Haris Jabbar. 2024. MorphPiece : A linguistic tokenizer for large language models. *Preprint*, arXiv:2307.07262.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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

Jindvrich Libovick'y and Jindvrich Helcl. 2024. Lexically grounded subword segmentation. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

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

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.

Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2021. The two approaches to word formation in the LiLa knowledge base of Latin resources. In *Proceedings of the third international workshop on resources and tools for derivational morphology (DeriMo 2021)*, pages 101–109. ATILF & CLLE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Philippe Roelli. 2014. The Corpus Corporum, a new open Latin text repository and tool. *Archivum Latinitatis Medii Aevi*, 72(1):289–304.

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

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Uwe Springmann, Helmut Schmid, and Dietmar Najock. 2016. LatMor: A Latin finite-state morphology encoding vowel quantity. *Open Linguistics*, 2.

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.

Phillip Benjamin Ströbel. 2022. Roberta base latin cased v1.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for Turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Clara Vania, Andreas Grivas, and Adam Lopez. 2018. What do character-level models learn about morphology? the case of dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2573–2583, Brussels, Belgium. Association for Computational Linguistics.

Mihaela Alexandra Vasiu and Rodica Potolea. 2020. Enhancing tokenization by embedding Romanian language specific morphology. *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 243–250.

Menan Velayuthan and Kengatharaiyer Sarveswaran. 2025. Egalitarian language representation in language models: It all begins with tokenizers. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5987–5996, Abu Dhabi, UAE. Association for Computational Linguistics.

Saketh Reddy Vemula, Sandipan Dandapat, Dipti Sharma, and Parameswari Krishnamurthy. 2025. Rethinking tokenization for rich morphology: The dominance of unigram over BPE and morphological alignment. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 232–252, Mumbai, India. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shaked Yehezkel and Yuval Pinter. 2023. Incorporating context into subword vocabularies. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 623–635, Dubrovnik, Croatia. Association for Computational Linguistics.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A Appendix

		Gold = Actx. Sigmorphon Segmentations				
Tokenizer Type	**Alg**	**EM**	**Recall**	**Precision**	**F1**	**Fertility**
Baseline	ULM	7.19	20.30	16.33	18.10	3.0964
MorphSeed	ULM	7.03	20.72	16.69	18.49	3.0907
MorphPreTok Actx	ULM	9.89	26.23	20.50	23.01	3.1904
MorphPreTok Ctx	ULM	9.34	25.63	19.99	22.46	3.2001
Baseline	WP	3.60	12.29	10.77	11.48	2.8434
MorphSeed	WP	3.46	11.56	10.43	10.97	2.7615
MorphPreTok Actx	WP	8.45	19.17	17.93	18.53	2.6651
MorphPreTok Ctx	WP	8.28	18.94	17.83	18.37	2.6482

		Gold = Ctx. Lemlat Segmentations				
Tokenizer Type	**Alg**	**EM**	**Recall**	**Precision**	**F1**	**Fertility**
Baseline	ULM	10.76	11.68	12.08	11.88	1.8714
MorphSeed	ULM	12.12	13.72	14.04	13.88	1.8926
MorphPreTok Actx	ULM	65.05	77.62	63.74	70.00	2.3585
MorphPreTok Ctx	ULM	71.88	84.98	68.43	75.81	2.4052
Baseline	WP	20.12	21.37	23.41	22.34	1.7688
MorphSeed	WP	20.18	21.49	23.53	22.46	1.7688
MorphPreTok Actx	WP	75.50	82.94	75.30	78.93	2.1335
MorphPreTok Ctx	WP	84.32	91.90	81.82	86.57	2.1755

Table 6: Full tokenizer evaluation metrics across acontextual (Sigmorphon, top) and contextual (Lemlat, bottom) gold segmentations. Recall, precision, and f1 are in terms of subword overlap between the predicted and gold segmentations.

A.1 Disambiguating Segmentations with POS Tags

When attempting to match a UD POS tag to a Lemlat POS tag, the Lemlat tags are checked in the order they appear in Table 7.

A.2 Tokenizer Implementation Details

Training Hyperparameters For all tokenizers, we fix the vocabulary size at 30k. For ULM, we set the shrinking factor to the default HuggingFace value, 0.75.

UD POS Tag	Lemlat POS Tags
NOUN	Noun, Adjective
PROPN	Noun, Adjective
VERB	Verb
ADJ	Adjective, Noun
PRON	Pronoun, Noun, Invariable
ADV	Invariable
ADP	Preposition, Invariable
CCONJ	Conjunction, Invariable
SCONJ	Conjunction, Invariable
PART	Interjection, Invariable
INTJ	Interjection, Invariable
DET	Pronoun, Adjective
X	Invariable, Other
AUX	Verb
PUNCT	Invariable
NUM	Noun, Adjective, Invariable

Table 7: Mapping from Universal Dependencies (UD) POS Tags to Lemlat POS Tags

MorphPreTokenization For ULM, we use a sequence of the default `Metaspace()` pretokenizer, followed by a `CharDelimiterSplit(delimiter="@")`. The `Metaspace` pretokenizer replaces whitespace with a special underscore-like symbol, then splits on this character and prepends it to the next word. Functionally, this means that the first subword of a word is differentiated from continuing subwords with this symbol. Then, the `CharDelimiterSplit` pretokenizer will split on our morpheme symbol, allowing morphological suffixes to be treated as continuing subwords.

WordPiece requires more modification than ULM, since the continuing subword symbol "##" is added at train time rather during pretokenization. First, we use a sequence of pretokenizers: `WhitespaceSplit()` followed by `Split(pattern = "@", behavior = "merged_with_next")`. Then, we add a `morph_delimiter="@"` argument to the `WordPieceTrainer`. During training, if a subword is encountered that starts with the `morph_delimiter`, the delimiter is replaced with the continuing subword symbol "##" and the new subword is added to the vocabulary.

A.3 Pretraining Details

For all 8 models, pretraining took around 4460 GPU hours on A40s and H200s.

Hyperparameter	Value
Epochs	10
Effective batch size	256
Learning rate	2×10^{-5}
Weight decay	0.01
Warmup steps	1000
Max position embeddings	513
Number of attention heads	12
Number of hidden layers	12
Hidden size	768
MLM probability	0.15
Architecture	`RobertaPreLayerNorm`

Table 8: Pretraining hyperparameters for all 8 models. Training follows a HuggingFace `RobertaPreLayerNorm` architecture with masked language modeling (MLM) objective.

A.4 Finetuning Details

A.4.1 Hyperparameters

Hyperparameter	POS/Morph	NER	WSD
Batch size	8	16	8
Epochs	15	10	20
Dropout	0.1	–	0.25
Weight decay	0.01	0.01	–
Learning rate	5×10^{-5}	2×10^{-5}	5×10^{-5}

Table 9: Fine-tuning hyperparameters for downstream tasks.

Hyperparameters are taken from prior work in Latin POS/Morphological feature tagging (Hudspeth et al., 2024), NER (Beersmans et al., 2023), WSD (Ghinassi et al., 2024).

For WSD, we specicially train on Ghinassi et al. (2024)'s SemEval + Pers$_{inter}$ train set.

To determine which checkpoint is best-performing, we use the validation whole-string morphological accuracy for POS/Morphological feature tagging, and the validation F1 for NER. WSD always saves the model from the last epoch.

The AV task does not require finetuning. We re-implemented the baseline method described by Gorovaia et al. (2024).

A.4.2 Subword Aggregation Strategies

As reported in Table 10, we generally saw the best performance on the token level tasks when predicting from the last subword. For Latin, we did not find any prior literature using the last subword, but we did find prior work using the first subword embedding (Riemenschneider and Frank, 2023) and the mean-pooled word embedding (Bamman and Burns, 2020).

		Morph Acc		
Model	**Tok**	**First**	**Last**	**Mean**
Baseline	ULM	76.88	77.73	75.18
MorphSeed	ULM	77.44	78.16	75.64
MorphPreTok Actx	ULM	81.76	81.61	78.58
MorphPreTok Ctx	ULM	81.70	82.09	78.51
Baseline	WP	77.19	78.94	77.36
MorphSeed	WP	77.40	78.82	77.33
MorphPreTok Actx	WP	81.11	81.68	78.96
MorphPreTok Ctx	WP	81.41	**82.25**	79.54
Per-col AVG		79.36	**80.16**	77.64

		NER BI F1		
Model	**Tok**	**First**	**Last**	**Mean**
Baseline	ULM	28.91	32.17	29.56
MorphSeed	ULM	30.15	37.61	31.04
MorphPreTok Actx	ULM	44.47	45.40	**46.40**
MorphPreTok Ctx	ULM	42.70	42.31	43.99
Baseline	WP	30.22	32.94	30.96
MorphSeed	WP	34.76	35.71	33.25
MorphPreTok Actx	WP	36.81	36.82	41.11
MorphPreTok Ctx	WP	36.01	44.12	39.33
Per-col AVG		35.50	**38.39**	36.96

Table 10: Out-domain performance across **subword aggregation strategies** (first subword, last subword, mean pooling) for Morph and NER tasks.

A.5 Downstream Task Descriptions

Task	In-domain	Out-domain	Unit
Morph	4,245	1,358	sents
NER	975	2,435	sents

Table 11: Sizes of in and out domain test sets for POS/Morphological Feature tagging and NER.

A.5.1 POS and Morphological Feature Tagging

Data We use the official train/test splits of five Latin Universal Dependencies (UD) Treebanks (de Marneffe et al., 2021),[13] harmonized to have more consistent morphological features (Gamba and Zeman, 2023) and standardized to use Latin-specific morphological features (Hudspeth et al., 2024).

When comparing in versus out domain performance, we consider the Perseus and UDante test sets as the out-domain. They comprise the smallest portion of the finetuning data, and they are stylistically distinct from the other treebanks: Perseus is primarily Classical era histories, poems, epics, satires; and UDante Medieval treatises, letters, poems (Hudspeth et al., 2024).

Modeling We use a separate classification head for each morphological feature, the same architecture as Riemenschneider and Frank (2023)'s finetuned Greek model. See §A.4 for details on hyperparameters. Additionally, we experiment with three subword aggregation strategies, as we noticed different methods being used in prior work: predicting from the first subword (as in Riemenschneider and Frank (2023)), predicting from the last subword, and predicting from the averaged word embedding (as in Bamman and Burns (2020)'s POS tagging task for LatinBERT).

Metrics We report whole-string morphological accuracy, following the convention of Gamba and Zeman (2023) and Sprugnoli et al. (2022). This metric considers the model's prediction correct when every morphological feature is correctly predicted, indicating whether the model understands how all the morphological features fit together.

Additionally, we report per-feature macro-F1 scores in order to emphasize the performance on rare feature values.

These metrics are aggregated across the test sets, weighted by the number of labels (words) in each test treebank.

A.5.2 Named Entity Recognition (NER)

Data We finetune our models on the Herodotos Project's[14] manually annotated NER dataset (Erdmann et al., 2016, 2019), using Beersmans et al. (2023)'s in-domain/out-domain splits. The in-domain consists of prose texts (Caesar's *Bellum Gallicum* and *Bellum Civile*, Pliny the Younger's *Epistulae*, and Pliny the Elder's *Naturalis Historia*), and the out-domain a single poetry text, Ovid's *Ars Amatoria*.

Modeling Following Beersmans et al. (2023)'s finetuning of LatinBERT, we treat this as a single-head token classification task. Similarly to our morphological feature classification task, we test three subword aggregation strategies: for each word, predicting on the first subword, the last subword, or an average embedding of that word's subwords.

Metrics We primarily report the overall micro f1 (accuracy) for BI labels, excluding the O label, consistent with Beersmans et al. (2023).

[13] Perseus (Bamman and Crane, 2011), PROIEL (Haug and Jøhndal, 2008), LLCT (Cecchini et al., 2020a), ITTB (Passarotti, 2019), and UDante (Cecchini et al., 2020b): https://universaldependencies.org/la/

[14] https://u.osu.edu/herodotos/

A.5.3 Word Sense Disambiguation (WSD)

Data We employ Ghinassi et al. (2024)'s train/test split of SemEval's 2020 shared task on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020). In addition to the SemEval training data, we also train on Ghinassi et al. (2024)'s Pers$_{inter}$ silver data, created by propagating word senses from English. Unlike previous Latin WSD experiments (Bamman and Burns, 2020; Lendvai and Wick, 2022), this dataset includes more than two senses per lemma, derives senses from multiple dictionaries, and includes texts from a wider time range.

Modeling We replicate Ghinassi et al. (2024), who finetuned LatinBERT using a single classification head on top of a mean-pooled representation of the target word's subword embeddings.

Metric We report the macro average F1 across the 40 lemmas in the test set. For each lemma, the F1 is the weighted average of each sense.

A.5.4 Authorship Verification

Data The dataset is a subset of the Patristic Sermon Textual Archive (PaSTA), curated by Gorovaia et al. (2024). They selected 22 authors who preached during the 3rd to 7th centuries, sampling 15 positive and 15 negative text pairs per author. In total, there are 660 text pairs.

Method We replicate Gorovaia et al. (2024)'s baseline method, which generates sentence embeddings from the base model using mean-pooling, tunes a cosine similarity threshold on two-thirds of the data, and tests the threshold on the remaining one-third of the data. Unlike the other downstream tasks, this method does not require finetuning.

Metrics We divide the data into three disjoint portions and conduct three trials, each using a different portion as the test set, and report the average F1 score across trials. Thus, the size of the test set will always be 220 text pairs.

A.6 Effect of Word Frequency on Morphological Classification Accuracy

Performance on morphological feature classification is unchanged for the words seen most frequently in the pretraining corpus, but rarer words see a boost from the MorphPreTok tokenizers, indicating better generalization.

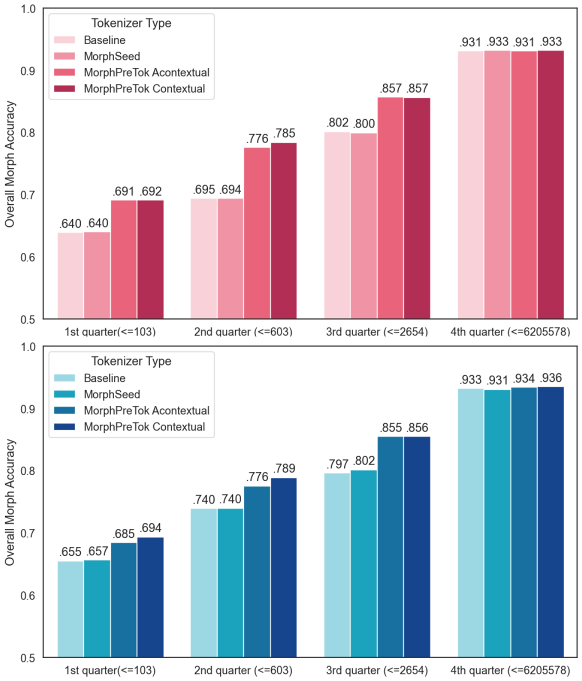

Figure 1: Word frequency in the pretraining corpus versus whole-string morphological accuracy, for ULM (top) and WordPiece (bottom).

A.7 Performance of Existing Models on Downstream Tasks

Model	Tok	Morph	NER	WSD	AV
LatinBERT	WP	**96.07**	81.19*	**63.34**	**72.60**
LaBERTa	BPE	88.86	77.60/**85.21**	52.60	70.70

Table 12: Summary of results on downstream tasks for existing Latin encoder models. For Morph and NER, reported results use mean pooled word representations for prediction. For LaBERTa NER, we show results on lowercased/cased data, since LaBERTa's tokenizer is cased but LatinBERT's is not. LatinBERT's NER* BI f1 is computed based on the reported in and out domain BI f1 in Beersmans et al. (2023) (we did not finetune it ourselves). LatinBERT's WSD f1 is slightly different than reported in Ghinassi et al. (2024) because we discovered a bug in their calculation of the f1 score.

LatinBERT dominates on most downstream tasks, likely because its pretraining corpus (643M words) is nearly 4x larger than LaBERTa's (165M) or the corpus used in this work (§4.3).