# 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, UMass Amherst
[2]Institute for the Study of the Ancient World, New York University

EACL 2026
RABAT • MOROCCO
March • Mars • 2026 • مارس

## Morphologically Guided Tokenization

**Pretokenization:** define the initial "processing units" upon which the tokenizers will be trained. Tokenizers cannot merge across these units

Sentence: *fana atque domos spoliari*

### Baseline

1 Split by whitespace and/or punct → [*fana atque domos spoliari*]

### Morpheme-Based

1 Run morphological analyzer on each word

**Acontextual**

2 Use first segmentation given

**Analysis 1**
Segmentation: *spoliar -i*
POS: Noun

3 Split by morpheme
[*fan -a atque dom -os spoliar -i*]

**Contextual**

2 POS-tag sentence

| NOUN | CCONJ | NOUN | VERB |
|---|---|---|---|
| *fana* | *atque* | *domos* | *spoliari* |

3 Use segmentation for correct POS

**Analysis 2**
Segmentation: *spoli -ari*
POS: Verb

4 Split by morpheme
[*fan -a atque dom -os spoli -ari*]

**Training:** construct the vocabulary from a corpus

| | WordPiece | ULM | +MorphSeed |
|---|---|---|---|
| 1 Construct an initial subword vocab | Small vocab: all unique characters<br>[*f, a, n, t, … s, i*] | Large vocab: all substrings of pretokens<br>[*f, fa, an, na, …, spoliari*] | + morphological suffixes<br>[*…, ari, ere, ero, …*] |
| 2 Iteratively refine the vocab | Merge subwords w/ high PMI & add to vocab<br>[*f, a, …*] -> [*f, a, fa, …*] | Prune subwords to maximize unigram likelihood<br>[*f, fa, an, ~~na~~, …*] | Disallow pruning/merging of morphological suffixes |

3 Stop when desired vocab size reached

**Decoding:** at runtime, segment new text into tokens using the learned vocabulary

**WordPiece:** Greedy left-to-right
*impropero* -> [*improper, o*]

**ULM:** Viterbi algorithm to find most likely sequence of subwords
*impropero* -> [*imp, rop, ero*] Upweight morphological suffixes

The tokenization pipeline (Schmidt et al, 2024)

## Experimental Setup

### Data

- Pretraining + Tokenizers: 1.08 GB of Latin texts used to train LatinCy
- Tokenizer eval: Latin UD treebanks (test splits) augmented with gold segmentations from LEMLAT

### Base Models: 8 110M parameter RoBERTa models

- 2 tokenizer types: WordPiece (WP) and Unigram (ULM)
- 4 variations: Baseline, MorphSeed, MorphPreTokenization (Acontextual and Contextual)

### Downstream Tasks:

- POS & Morph Tagging (**Morph**)
- Named Entity Recognition (**NER**)
- Word Sense Disambiguation (**WSD**)
- Authorship Verification (**AV**)

## Tokenizer Evaluation

WordPiece is *more aligned* to gold morphological segmentations than ULM

MorphPreTok guarantees *high alignment* with the gold

| | | Sigmorphon (Actx) | | LEMLAT (Ctx) | |
|---|---|---|---|---|---|
| | | EM % | Fertility | EM % | Fertility |
| Baseline | ULM | 7.19 | 3.10 | 10.8 | 1.87 |
| MorphSeed | ULM | 7.03 | 3.09 | 12.1 | 1.89 |
| Actx | ULM | **9.89** | 3.19 | 65.1 | 2.36 |
| Ctx | ULM | 9.34 | 3.20 | 71.9 | 2.41 |
| Baseline | WP | 3.60 | 2.84 | 20.1 | 1.77 |
| MorphSeed | WP | 3.46 | 2.76 | 20.2 | 1.77 |
| Actx | WP | 8.45 | 2.67 | 75.5 | 2.13 |
| Ctx | WP | 8.28 | 2.65 | **84.3** | 2.18 |
| | | | gold=2.49 | | gold=1.94 |

## Results: Overall

Consistent improvement on **syntax-level tasks**

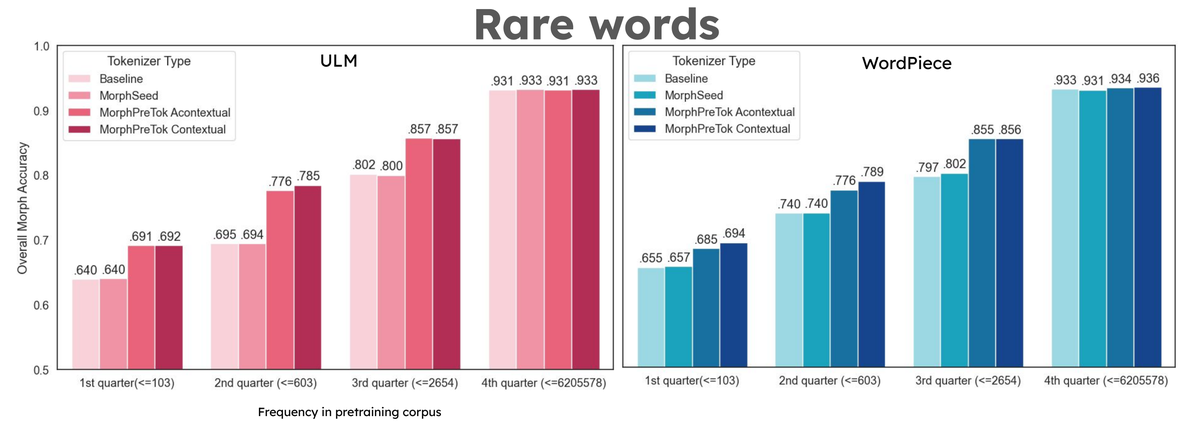
MorphPreTok shows strongest gains

*signficant compared to baseline

| | | Morph | NER | WSD | AV |
|---|---|---|---|---|---|
| Baseline | ULM | 89.46 | 65.89 | **61.79** | 61.27 |
| MorphSeed | ULM | 89.51 | 66.54 | 59.34 | 64.23 |
| Actx | ULM | *90.98 | ***73.52** | 60.83 | 65.65 |
| 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 |
| Actx | WP | *91.09 | *69.47 | 61.08 | 63.38 |
| Ctx | WP | ***91.18** | *72.72 | 59.84 | 64.64 |

## Results: Generalization

### Rare words




### Rare features and entities

| | | Acc | | Per-Feat Macro-F1 | | | | | | | | Per-Entity Micro-F1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | POS | Morph | Case | Degr | Gend | Mood | Num | Pers | Tense | Voice | PERS | LOC | GRP |
| Baseline | ULM | 94.78 | 89.46 | 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 | 0.06 | -0.17 | -0.44 | 0.07 | 0.25 | 0.05 | -0.04 | 0.06 | -0.05 | 2.52 | 0.79 | -2.37 |
| Actx | ULM | 0.57 | 1.52 | 2.04 | 4.21 | 1.05 | 3.17 | 0.76 | 0.86 | 3.48 | 1.68 | 8.88 | 6.09 | 2.68 |
| Ctx | ULM | 0.57 | 1.54 | 6.04 | 4.57 | 1.16 | 3.70 | 0.80 | 0.84 | 3.51 | 1.80 | 7.01 | 4.2 | 1.66 |
| Baseline | WP | 94.99 | 89.86 | 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 | -0.04 | 4.16 | 0.20 | 0.05 | 0.28 | -0.05 | 0.07 | 0.14 | 0.08 | 0.62 | 4.42 | 4.3 |
| Actx | WP | 0.55 | 1.23 | 0.55 | 2.86 | 0.95 | 1.14 | 0.69 | 0.49 | 2.04 | 0.92 | 1.04 | 8.62 | 7.84 |
| Ctx | WP | 0.41 | 1.32 | 5.10 | 3.30 | 0.91 | 1.49 | 0.59 | 0.55 | 2.54 | 1.09 | 5.91 | 11.44 | 8.17 |

### Out-of-domain texts

| | | Morph | | NER | |
|---|---|---|---|---|---|
| | | **in** | **out** | **in** | **out** |
| Baseline | ULM | 93.04 | 77.73 | 85.53 | 32.17 |
| Seed | ULM | -0.21 | +0.43 | -0.91 | +5.44 |
| Actx | ULM | +0.77 | +3.88 | +3.16 | +13.23 |
| Ctx | ULM | +0.68 | +4.36 | +2.48 | +10.14 |
| Baseline | WP | 93.09 | 78.94 | 84.22 | 32.94 |
| Seed | WP | +0.00 | -0.12 | +0.91 | +2.77 |
| Actx | WP | +0.95 | +2.73 | +3.00 | +3.88 |
| Ctx | WP | +0.88 | +3.31 | +4.05 | +11.18 |

Morphologically-guided tokenization allows the model to learn the **meaning** of word endings, instead of relying on **memorization** of frequent words

University of Massachusetts Amherst | NYU | ISAW