

# LatinCy Guidelines

Patrick J. Burns

2026-02-12

# Table of contents

|  |          |
|--|----------|
| <b>Preface</b>                                     | <b>3</b> |
| Key links . . . . .                                | 3        |
| <b>Annotation Quickstart</b>                       | <b>4</b> |
| Your Task . . . . .                                | 4        |
| Lemma . . . . .                                    | 4        |
| UPOS (Part of Speech) . . . . .                    | 5        |
| Morphological Features . . . . .                   | 6        |
| Nouns, Adjectives, Pronouns, Determiners . . . . . | 6        |
| Verbs . . . . .                                    | 6        |
| Dependency Relations . . . . .                     | 7        |
| ROOT . . . . .                                     | 8        |
| nsubj . . . . .                                    | 8        |
| Named Entity Types . . . . .                       | 9        |
| PERSON . . . . .                                   | 9        |
| LOC . . . . .                                      | 9        |
| NORP . . . . .                                     | 9        |
| Quick Checklist . . . . .                          | 10       |

## Preface

“LatinCy Annotation Guidelines” is an always-a-work-in-progress describing decisions made during the annotation process for LatinCy Assets and can be used as a reference resource for annotators and other contributors to the project.

# latinCy

### Key links

Models: <https://huggingface.co/latincy>  
Universe: <https://spacy.io/universe/project/latincy>  
Preprint: <https://arxiv.org/abs/2305.04365>  
Book: <https://diyclassics.github.io/latincy-book/>

This book has been written using Quarto with support from Claude 3.7/4.0 (primarily for document outlining, structuring, and formatting). To learn more about Quarto books visit <https://quarto.org/docs/books>.

# Annotation Quickstart

This page is a quick reference for the LatinCy annotation review process. Detailed guidelines for each topic are in preparation and will be published as additional chapters.

## Your Task

You are reviewing **model output** — the LatinCy pipeline has already analyzed each text, and your job is to check and correct its predictions. Each passage is presented in a spreadsheet with the following columns:

| Column          | What it contains                   | Your action                            |
|-----------------|------------------------------------|--|
| <b>token</b>    | The word as it appears in the text | Read-only                              |
| <b>lemma</b>    | Dictionary headword                | Check and correct if needed            |
| <b>upos</b>     | Part-of-speech tag                 | Check and correct if needed            |
| <b>feats</b>    | Morphological features             | Check and correct if needed            |
| <b>deprel</b>   | Dependency relation                | Check ROOT and nsubj only<br>(for now) |
| <b>ent_type</b> | Named entity type                  | Check PERSON, LOC, NORP                |

It will take time to get used to reading the Latin in this word-by-word fashion, but after a few sentences you should begin to develop some strategies and notice patterns.

---

## Lemma

The **lemma** is the dictionary headword. It should match whatever appears in the passage vocabulary.

### Quick rules:

- **Nouns:** nominative singular — *homo, civitas, rex*

- **Adjectives:** nominative singular masculine — *bonus, omnis, felix*
- **Verbs:** first person singular present indicative — *amo, sum, fero, dico*
- **Pronouns:** citation form varies — *qui, sui, ego, hic*
- **Prepositions:** fuller form — *ab* (not *a*), *ex* (not *e*)

#### 💡 Common errors to watch for

- Adjectives lemmatized as neuter (*omne*) or plural (*omnes*) instead of masculine singular (*omnis*)
- Deponent verbs lemmatized as active
- Prepositions shortened (*a* instead of *ab*)

*Full lemmatization guidelines forthcoming.*

## UPOS (Part of Speech)

The UPOS column contains one of the following 17 tags:

| Tag          | Description   | Examples  |
|--------------|---|---|
| <b>ADJ</b>   | Adjective   | <i>bonus, omnis, magnus</i>                           |
| <b>ADP</b>   | Adposition (preposition)                            | <i>in, ad, de, cum, per</i>                           |
| <b>ADV</b>   | Adverb  | <i>non, iam, semper, bene</i>                         |
| <b>AUX</b>   | Auxiliary verb (for our purposes, just <i>sum</i> ) | <i>est, sunt, erat</i>                                |
| <b>CCONJ</b> | Coordinating conjunction                            | <i>et, sed, atque, vel</i>                            |
| <b>DET</b>   | Determiner  | <i>hic, ille, iste, omnis</i> (when modifying a noun) |
| <b>INTJ</b>  | Interjection  | <i>o, ecce, eheu</i>                                  |
| <b>NOUN</b>  | Noun  | <i>vir, urbs, virtus, bellum</i>                      |
| <b>NUM</b>   | Numeral   | <i>unus, duo, tres, centum</i>                        |
| <b>PART</b>  | Particle  | <i>autem, enim, igitur, quidem</i>                    |
| <b>PRON</b>  | Pronoun   | <i>qui, ego, se, quis</i>                             |
| <b>PROPN</b> | Proper noun   | <i>Caesar, Roma, Gallia</i>                           |
| <b>PUNCT</b> | Punctuation   | <i>. , ; :</i>  |
| <b>SCONJ</b> | Subordinating conjunction                           | <i>ut, cum, si, quod, quia</i>                        |
| <b>SYM</b>   | Symbol  | † (crux in charters), %                               |
| <b>VERB</b>  | Verb (including infinitives and participles)        | <i>dicit, fecit, dicere, dictus</i>                   |

| Tag      | Description                  | Examples        |
|----------|------------------------------|-----------------|
| <b>X</b> | Other (foreign words, typos) | , <i>shalom</i> |

#### **i** Key distinctions

- **PART vs. CCONJ**: Discourse particles like *autem*, *enim*, *igitur* are PART (not CCONJ). Coordinating conjunctions join clauses or phrases: *et*, *sed*, *atque*.
- **NOUN vs. PROPN**: Common nouns (*rex*, *consul*) are NOUN. Specific names (*Caesar*, *Roma*) are PROPN.

Full POS tagging guidelines forthcoming.

## Morphological Features

The **feats** column contains morphological features separated by |. The model predicts these; you should verify the key features for each word class.

### Nouns, Adjectives, Pronouns, Determiners

These should have **Gender**, **Number**, and **Case**:

| Feature | Values   |
|---------|--|
| Gender  | Fem, Masc, Neut                                |
| Number  | Sing, Plur                                     |
| Case    | Nom, Gen, Dat, Acc, Abl, Voc (also Loc rarely) |

**Example:** *virtutis* → Case=Gen|Gender=Fem|Number=Sing

### Verbs

These should have **Person**, **Number**, **Tense**, **Aspect**, **Mood**, **Voice**, and **VerbForm**:

| Feature | Values  |
|---------|---------|
| Person  | 1, 2, 3 |

| Feature  | Values   |
|----------|--|
| Number   | Sing, Plur   |
| Tense    | Pres, Past, Fut, Pqp (pluperfect)  |
| Aspect   | Perf (perfective), Imp (imperfective)  |
| Mood     | Ind, Sub (subjunctive), Imp (imperative)   |
| Voice    | Act, Pass  |
| VerbForm | Fin (finite), Inf (infinitive), Part (participle), Ger (gerund), Gdv (gerundive), Sup (supine) |

**Example:** *scripsit* → Aspect=Perf | Mood=Ind | Number=Sing | Person=3 | Tense=Past | VerbForm=Fin | Voice=Ac

### Tense and Aspect

UD uses **Tense=Past** for all past tenses. The **Aspect** feature distinguishes them:

- *scripsit* (perfect): Tense=Past | Aspect=Perf
- *scribebat* (imperfect): Tense=Past | Aspect=Imp
- *scripserat* (pluperfect): Tense=Pqp | Aspect=Perf

### What to check

Focus on the features most likely to be wrong:

- **Case:** The model sometimes confuses nominative/accusative for neuter nouns, or nominative/vocative
- **Tense/Aspect:** Past perfective (*scripsit*) vs. past imperfective (*scribebat*)
- **Mood:** Indicative vs. subjunctive
- **Voice:** Active vs. passive, especially for deponent verbs

## Dependency Relations

For now, you only need to check two dependency labels:

## ROOT

The **main verb** of each sentence should be labeled ROOT. Every sentence has exactly one ROOT.

|           |        |           |
|-----------|--------|-----------|
| Poeta     | carmen | scripsit. |
| nsubj     | obj    | ROOT      |
| "The poet | a poem | wrote."   |

### Copular sentences

In copular sentences, the **predicate** (not *sum*) is the ROOT:

|         |        |      |
|---------|--------|------|
| Cicero  | consul | est. |
| nsubj   | ROOT   | cop  |
| "Cicero | consul | is." |

Here *consul* is the ROOT because it is the predicate nominal. *est* is labeled **cop** (copula).

## nsubj

The **subject** of the main verb should be labeled **nsubj** (nominal subject). Look for the nominative noun or pronoun that performs the action.

|           |           |        |
|-----------|-----------|--------|
| Rex       | exercitum | duxit. |
| nsubj     | obj       | ROOT   |
| "The king | the army  | led."  |

Not every sentence has an explicit subject — Latin often drops the pronoun (“pro-drop”):

|                  |
|------------------|
| Venit.           |
| ROOT             |
| "(He/she) came." |

*Full dependency parsing guidelines forthcoming.*



## Named Entity Types

Check whether names referring to specific people, places, or groups are correctly labeled with one of three entity types:

### PERSON

Named individuals, deities, and specific mythological groups.

| Tag                    | Not tagged                          |
|------------------------|-------------------------------------|
| <i>Caesar</i> (PERSON) | <i>consul</i> (title, not a name)   |
| <i>Venus</i> (PERSON)  | <i>dea</i> (common noun)            |
| <i>Ceres</i> (PERSON)  | <i>Cerealia</i> (derived adjective) |
| <i>Musae</i> (PERSON)  | <i>dae</i> (common noun)            |

### LOC

Locations — cities, regions, rivers, mountains, any named place.

| Tag                  | Not tagged                  |
|----------------------|-----------------------------|
| <i>Roma</i> (LOC)    | <i>urbs</i> (common noun)   |
| <i>Tiberis</i> (LOC) | <i>flumen</i> (common noun) |
| <i>Alpes</i> (LOC)   | <i>montes</i> (common noun) |

### NORP

Nationalities or religious or political groups. This includes both noun and adjective forms.

| Tag                                    | Not tagged                      |
|--|---------------------------------|
| <i>Romani</i> (NORP)                   | <i>cives</i> (common noun)      |
| <i>Troianus</i> (NORP, adjective form) | <i>milites</i> (common noun)    |
| <i>Christiani</i> (NORP)               | <i>sacerdotes</i> (common noun) |

**i** NORP is used somewhat anachronistically

In the modern NLP context it stands for “Nationalities or religious or political groups.” We apply it to ancient groups of people by analogy.

## Quick Checklist

When reviewing a passage:

1. Read through the Latin text to understand the passage
2. Scan the **lemma** column — do headwords match what you see in the vocabulary?
3. Scan the **upos** column — are content words (NOUN, VERB, ADJ) labeled correctly?
4. Scan the **feats** column — for nouns/adjectives, check gender/number/case; for verbs, check person/number/tense/aspect/mood/voice
5. Find the **ROOT** — is the main verb labeled ROOT?
6. Find the **nsubj** — is the subject of that verb labeled nsubj?
7. Scan for **names** — are people, places, and groups labeled PERSON, LOC, or NORP?
8. Correct errors in the designated cells only