



Evaluating UDPipe 2 and Stanza for Morphosyntactic Annotation of Late Latin: A Pilot Study on Sixth-Century Texts

DISCUSSION PAPER

ANNACHIARA CLEMENTELLI 

 jubiquity press

ABSTRACT

To date, annotated Late Latin treebanks and natural language processing (NLP) models specifically trained on this linguistic variety remain extremely limited. In this low-resource setting, the present study investigates which combination of annotation tool and model currently available for Latin provides the most reliable basis for developing a Late Latin treebank. To address this question, a manually annotated gold standard is created, comprising four sixth-century texts. The creation of this reusable dataset constitutes the first contribution of the study, as it provides a benchmark for the evaluation of existing and future NLP models on Late Latin. The evaluation focuses on part-of-speech tagging and dependency parsing, comparing two tools, UDPipe 2 and Stanza, across their available pretrained models. The results show that UDPipe 2 generally outperforms Stanza on the corpus examined, with the winning submission of EvaLatin 2024 achieving the best overall performance among the models tested. These findings suggest that UDPipe with the EvaLatin 2024 model provides a suitable starting point for a semi-automatic workflow for the creation of new Late Latin treebanks, in which automatic pre-annotation is followed by expert manual revision.

CORRESPONDING AUTHOR:

Annachiara Clementelli

Department of Italian
Language, Literature, and Arts
Around the World, University
for Foreigners of Perugia,
Perugia, Italy; Department of
Linguistics, Ghent University,
Ghent, Belgium

a.clementelli@phd.unistrapg.it

KEYWORDS:

Universal Dependencies; Late
Latin; treebanks; automatic
annotation; EvaLatin 2024

TO CITE THIS ARTICLE:

Clementelli, A. (2026).
Evaluating UDPipe 2 and
Stanza for Morphosyntactic
Annotation of Late Latin: A
Pilot Study on Sixth-Century
Texts. *Journal of Open
Humanities Data*, 12: 140,
pp. 1–12. DOI: [https://doi.
org/10.5334/johd.630](https://doi.org/10.5334/johd.630)

(1) CONTEXT AND MOTIVATION

The relevance of treebanks for historical linguistics has been widely recognized (Eckhoff et al., 2020). They allow for the automatic extraction of data, ensuring variability, quality, and speed in data retrieval (Passarotti et al., 2015, pp. 388–389), and support more detailed qualitative investigations (Bamman & Crane, 2006, p. 76). Given the breadth and chronological depth of its textual tradition, Latin constitutes a particularly suitable domain for treebank-based research. Within the Universal Dependencies framework, currently the *de facto* standard for syntactically annotated corpora (see Section (2.1)), however, the available resources for Latin are unevenly distributed across periods. Late Latin remains comparatively under-resourced, with no UD treebank specifically devoted to this period and only a limited number of texts included in broader treebanks.

Recent advances in natural language processing (NLP) offer a possible way to address this limitation. Powerful language tools developed for a wide range of languages, including historical ones, enable the automatic annotation of previously unannotated texts by performing a range of morphological and syntactic tasks. For broader overviews of automatic language processing for ancient languages, see Riemenschneider and Frank (2023) and Sommerschild et al. (2023). In this paper, I focus on supervised models, i.e. models trained on manually annotated data (see, e.g., Jurafsky & Martin, 2024). The lack of a UD treebank exclusively devoted to Late Latin also means that there are no supervised models trained specifically on, and therefore tailored to, this stage of the language.

Against this background, the present paper addresses a central methodological question: in such a low-resource setting, which combination of annotation tool and model provides the most reliable basis for developing a Late Latin treebank? To answer this question, I first compare the performance of two NLP tools for the automatic annotation of Latin within the UD framework, namely UDPipe 2 (Straka, 2018) and Stanza (Qi et al., 2020). I then examine the individual models available for the better-performing tool. Performance is evaluated against a manually annotated gold standard comprising selected portions of four sixth-century texts (for further details, see Section (3)). The analysis focuses specifically on part-of-speech (POS) and syntactic annotation, as these two levels are particularly relevant for investigating diachronic variation. Although exploratory in scope, the study lays the methodological groundwork for future research and for the large-scale automatic annotation of Late Latin texts and corpora.

The paper is structured as follows. Section (2) provides the background to the study, introducing Universal Dependencies (Section (2.1)) and presenting the UD-annotated Latin texts and treebanks (Section (2.2)), as well as the NLP tools and models considered in the analysis (Section (2.3)). Section (3) presents the evaluation dataset. Section (4) describes the methodology, detailing the annotation of the gold standard (4.1) and the evaluation of the automatic annotations (4.2). Section (5) reports and discusses the results. Finally, Section (6) outlines the implications and possible applications of the study.

(2) THEORETICAL BACKGROUND

(2.1) UNIVERSAL DEPENDENCIES

Universal Dependencies (de Marneffe et al., 2021; henceforth UD) is a framework for cross-linguistically consistent annotation of grammar, including parts of speech, morphological features, and syntactic dependencies.¹ Its general philosophy is to provide a universal inventory of categories and annotation guidelines for comparable constructions across languages, while allowing language-specific extensions to account for phenomena that are not fully captured by the shared inventory. This balance between cross-linguistic uniformity and language-specific variation makes UD suitable for both NLP and linguistic typology.

In its most recent release (2.18; May 15, 2026), UD comprises 353 treebanks, covering 193 languages.² It includes historical languages such as Ancient Greek, Latin, Old French, Akkadian, and Classical Chinese. For Latin, there are six UD treebanks, covering a wide range of genres, and amounting to 1,010,878 tokens and 60,657 sentences.

¹ <https://universaldependencies.org>, last accessed 26.09.2026.

² <https://lindat.mff.cuni.cz/repository/items/7fbbd99-ae2d-4b91-8318-d996dbe34cbc>, last accessed 26.09.2026.

Table 1 provides an overview of the Latin texts currently available in UD-annotated form, grouped by treebank and chronological period. The chronological classification follows Haverling's (1988, pp. 20–23) periodisation of Latin, which has been widely adopted in the literature (e.g. Cuzzolin & Haverling, 2009; Pinkster, 2015, 2021). Medieval Latin is included for completeness, although by this period Latin was increasingly used as a learned rather than a native language and especially in ecclesiastical, administrative, and scholarly contexts. The starting date adopted here (600 AD) is intended as a practical chronological boundary rather than a sharp linguistic *caesura*, since Medieval Latin developed in continuity with Late Latin (Banniard, 2020).

A considerable number of annotated works are available for Classical (90 BC–14 AD) and Silver Latin (14 AD–200 AD), whereas Early (240 BC–90 BC) and Late Latin (200 AD–600 AD) have comparatively received less attention. As far as Early Latin is concerned, the Pragmatic Resources in Old Indo-European Languages (PROIEL) project is currently preparing annotated versions of the complete works of Plautus and Terence.³ This gap is therefore likely to be filled in the near future. Turning to Late Latin, the only UD annotated texts currently available online are, to the best of my knowledge, Jerome's *Vulgate* (IV century AD) and Palladius' *Opus agriculturae* (IV or V century AD). The *Vulgate* is a translation from Greek and, for the Old Testament, from Hebrew. Consequently, it is not always straightforward to distinguish syntactic features arising from contact with the source languages from those that are native to Latin. Despite its undoubted importance, it is therefore difficult to determine to what extent its syntax is representative of Late Latin (see Biagetti & Brigada Villa, 2023 and the literature cited therein). Palladius' *Opus agriculturae*, by contrast, provides valuable Late Latin data. However, it represents a single author and text type. Moreover, in both cases, the texts are only one component of a broader, diachronically mixed treebank.⁴ A sufficiently large and diversified Late Latin treebank is therefore still lacking, making Late Latin a low-resource variety from an NLP perspective (see Hedderich et al., 2021 for a detailed discussion of low-resource scenarios in NLP).

PERIOD	AUTHOR	WORK AND LOCI	TOKENS	TREEBANK
Early Latin (240 BC–90 BC)				
Classical Latin (90 BC–14 AD)	Caesar	<i>Commentarii de Bello Gallico</i> (1. 1–54; 2. 1–35; 3: 1–27; 4: 1–9, 11–38; 5. 1–7, 49; 6. 1–44; 7.3, 5, 34, 77)	28,608	PROIEL
		<i>Commentarii de Bello Gallico</i> (2.1–2.3; 2.5; 2.7; 2.9; 2.14–2.18; 2.32–2.33)	352	Perseus
	Cicero	<i>In Catilinam</i> (1.1–2.11)	1,915	Perseus
		<i>Epistulae ad Atticum</i> (letters 1.1–20; 2.1–25; 3.1–27; 4.1–19, 4A and 8A; 5.1–6 and 8–21; 6.1–3 and 9; 7.1–4, 7, 10, 12–13, and 19–26; 8.1–16, 11A, 11B, 11D, 12A–C, and 15A)	42,855	PROIEL
		<i>De officiis</i> (1–39, 41–113, 116, 133–137, 139–161)	10,644	PROIEL
	Propertius	<i>Elegiae</i> (1.1–1.22)	2,801	Perseus
	Vergil	<i>Aeneid</i> (6, 1–336)	779	Perseus
	Ovid	<i>Metamorphoses</i> (1.1–1.778)	2,655	Perseus
	Augustus	<i>Res Gestae</i>	713	Perseus
	Phaedrus	<i>Fabulae</i> (1–3)	4,118	Perseus
Silver Latin (14 AD–200 AD)	Sallust	<i>Bellum Catilinae</i>	4,999	Perseus
	Seneca	<i>Hercules Furens</i>	7,714	CIRCSE
		<i>Agamemnon</i>	5,580	CIRCSE
		<i>Oedipus</i>	5,931	CIRCSE

(Contd.)

Table 1 UD annotated works available online.

³ <https://dev.syntacticus.org/proiel.html#contents>.

⁴ It should be noted that the documentation of the PROIEL project also lists the *Peregrinatio Aetheriae* among the annotated texts (<https://dev.syntacticus.org/proiel.html#contents>) However, this work is not included in the UD Latin-PROIEL treebank (https://github.com/UniversalDependencies/UD_LatinPROIEL).

PERIOD	AUTHOR	WORK AND LOCI	TOKENS	TREEBANK
	Petronius	<i>Satyricon</i> (26–78)	5,955	Perseus
	Tacitus	<i>Germania</i>	5,674	CIRCSE
		<i>Historiae</i> (1.1.21)	867	Perseus
	Suetonius	Life of Augustus (1–55)	2,066	Perseus
	Palladius	<i>Opus Agriculturae</i> (Ch. I–XXXII)	12,148	PROIEL
Late Latin (200 AD–600 AD)	Jerome	Vulgate’s Revelation	2,003	Perseus
		Vulgate New Testament (Matthew 1–28; Mark 1–16; Luke 1–24; John 1–21; Acts 1–28; Romans 1–16; 1 Corinthians 1–16; 2 Corinthians 1–13; Galatians 1–6; Ephesians 1–6; Philippians 1–4; Colossians 1–2; 1 Thessalonians 1–5; 2 Thessalonians 1–3; 1 Timothy 4 and 6; 2 Timothy 1–4; Titus 1 and 3; Philemon 1; Hebrews 1–2, 4–5, 7, 10–11, and 13; James 2–5; 1 Peter 2–3 and 5; 2 Peter 2–3; 1 John 1–2 and 4–5; 2 John 1; 3 John 1; Jude 1; Revelation 1–22)	112,454	PROIEL
Medieval Latin (600 AD–1492 AD)	521 legal charters from Tuscia (774–897 CE) ⁵		242,391	LLCT
	Thomas Aquinas and related authors	<i>Summa contra Gentiles</i> (Books 1–4)	450,480	ITTB
		Concordances of the lemma <i>forma</i>		
	Dante Alighieri	<i>De vulgari eloquentia</i>	55,286	UDante
		<i>De monarchia</i>		
		Letters		
		<i>Quaestio de aqua et terra</i>		
		Eclogues		

The lack of a dedicated treebank for Late Latin constitutes a significant gap in the literature. First, Late Latin provides important evidence for the linguistic developments that led to the Romance languages (Ledgeway, 2012). Investigating its syntax is essential not only for identifying and describing processes of linguistic change, but also for establishing their chronology. Second, supervised NLP tools generally provide multiple models, each trained on a specific treebank, with the source treebank often reflected in the model’s name. The absence of treebanks specifically designed for Late Latin has hindered the development of models tailored to the automatic annotation of this linguistic variety. Researchers seeking to annotate a Late Latin text must consequently rely on models developed for other chronological stages of Latin, despite the considerable linguistic differences between Late Latin and both earlier and later stages of the language (Gasti, 2013).

(2.3) MODELS FOR LATIN

Among the tools available for the automatic annotation of Latin within the UD framework, this study focuses on UDPipe 2 (Straka, 2018) and Stanza (Qi et al., 2020), considering all the pretrained Latin models available for each tool.

2.3.1 UdPipe 2

UDPipe 2 (Straka, 2018) is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files.⁶ It is language-agnostic and can be trained on given annotated data in CoNLL-U-formatted files.

UDPipe 2 provides trained models for almost all the treebanks included in Universal Dependencies release 2.17, published on November 15, 2025.⁷ The models available for Latin

⁵ It should be noted that, despite its name, the Late Latin Charter Treebank contains texts that, according to Haverling’s (1988) periodization, fall chronologically within Medieval rather than Late Latin.

⁶ <https://lindat.mff.cuni.cz/services/udpipe/>, last accessed 26.09.2026.

⁷ <https://universaldependencies.org/download.html>, last accessed 26.09.2026.

are listed below, while further details on the treebanks on which they are based can be found in [Table 1](#).

- **latin-itb-ud-2.17-251125**, based on the Index Thomisticus Treebank (ITTB; 450,480 tokens).⁸
- **latin-circse-ud-2.17-251125**, based on the UD Latin CIRCSE Treebank (28,287 tokens).⁹
- **latin-llct-ud-2.17-251125**, based on Late Latin Charter Treebank (LLCT; 242,391 tokens).¹⁰
- **latin-perseus-ud-2.17-251125**, based on the Perseus Treebank (28,868 tokens).¹¹
- **latin-proiel-ud-2.17-251125**, based on the PROIEL treebank (205,556 tokens).¹²
- **latin-udante-ud-2.17-251125**, based on the UDante treebank (55,286 tokens).¹³

In addition, the UDPipe LINDAT/CLARIN service provides access to LatinPipe ([Straka et al., 2024](#)), the winning submission to the EvaLatin 2024 Dependency Parsing shared task ([Sprugnoli et al., 2024](#)), hereafter referred to as EvaLatin 2024.¹⁴ The system, which is an evolution of UDPipe 2, was trained on samples of seven publicly available Latin corpora, with additional harmonization of some annotations aimed at achieving a more consistent annotation style. The training data used for the winning submission were drawn from the following resources: ITTB from UD 2.13 (391K), LLCT from UD 2.13 (194K), a UD-style harmonized version of PROIEL (178K), UDante from UD 2.13 (31K), Perseus from UD 2.13 (18K), a UD-style annotated text of *De Latinae Linguae Reparatione* by Marcus Antonius Sabellicus, written in 1489 (11K),¹⁵ and *Archimedes Latinus* UD-style treebank, based on Jacopo da San Cassiano's Latin translation of *The Spirals of Archimedes* produced around 1450 (1K).¹⁶

(2.4) STANZA

Stanza ([Qi et al., 2020](#)) is an open-source Python natural language processing toolkit supporting 66 human languages. Compared to existing widely used toolkits, Stanza features a language-agnostic fully neural pipeline for text analysis, including tokenization, multi-word token expansion, lemmatization, part of speech and morphological feature tagging, dependency parsing, and named entity recognition.¹⁷ Stanza is trained on a total of 112 datasets, including the UD treebanks and other multilingual corpora.

For Latin, it has packages trained on the following treebanks: PERSEUS, PROIEL, ITTB, LLCT and UDante. I refer to the previous paragraphs for their description (Section (2.2) and Section (2.3)).

(3) DATASET DESCRIPTION

To determine which of the currently available tools and models for Latin annotation provides the most reliable basis for the development of a Late Latin treebank, a gold standard was manually annotated. It is described in [Table 2](#).

The dataset was deliberately restricted to sixth-century texts produced in the Italian peninsula in order to limit both chronological and geographical variation. To ensure evaluation on previously unseen material, texts already represented in existing Late Latin treebanks were excluded. The sample size for each text was set at 684 tokens, balancing comparability across authors with the practical constraints of manual annotation. The resulting dataset comprises 2,736 tokens drawn from works representing different text types and communicative contexts. Benedict's

⁸ https://universaldependencies.org/treebanks/la_itb/index.html, last accessed 26.09.2026.

⁹ https://github.com/UniversalDependencies/UD_Latin-CIRCSE, last accessed 26.09.2026.

¹⁰ https://universaldependencies.org/treebanks/la_llct/index.html, last accessed 26.09.2026.

¹¹ https://universaldependencies.org/treebanks/la_perseus/index.html, last accessed 26.09.2026.

¹² https://universaldependencies.org/treebanks/la_proiel/index.html, last accessed 26.09.2026.

¹³ https://universaldependencies.org/treebanks/la_udante/index.html, last accessed 26.09.2026.

¹⁴ <https://github.com/ufal/evalatin2024-latinpipe>, last accessed 26.09.2026.

¹⁵ <https://github.com/CIRCSE/Sabellicus>, last accessed 26.09.2026.

¹⁶ <https://github.com/mfantoli/ArchimedesLatinus>, last accessed 26.09.2026.

¹⁷ <https://github.com/stanfordnlp/stanza>, last accessed 26.09.2026.

Repository location	https://doi.org/10.7910/DVN/QHGETA https://github.com/annachiaraclementelli/6th-century-gold-standard-UD				
Repository name	Annotated gold standard of selections of four sixth century Late Latin texts				
Objects name	benedictus.conllu; cassiodorus.conllu; eugippius.conllu; gregorius.conllu				
Format names and versions	CoNLL-U				
Creation dates	2026-02-01–2026-07-31				
Dataset creators	Annachiara Clementelli (Università per Stranieri di Perugia & Ghent University; role: annotator)				
Language	Late Latin				
Publication date	2026-07-31				
Content	Author	Work	Period	Genre	Tokens
	Benedict	<i>Regula</i>	534 AD	Monastic rule	684
	Cassiodorus	<i>Variae</i>	527–540 AD	Administrative epistolary	684
	Eugippius	<i>Vita sancti Severini</i>	ca. 526 AD	Hagiography	684
	Pope Gregory I	<i>Regula pastoralis</i>	ca. 590 AD	Pastoral treatise	684
			Total		2,736

Table 2 Dataset description.

Regula, a guide to monastic life, is written in relatively simple Latin, displaying several features associated with Vulgar Latin. Eugippius’ *Vita sancti Severini*, by contrast, is a hagiographic work, characterized by a high register and careful stylistic elaboration. Cassiodorus’ *Variae*, a collection of official and administrative letters from the Ostrogothic court, combines a formal bureaucratic register with marked rhetorical and literary features. Finally, Pope Gregory I’s *Regula pastoralis* is a treatise on episcopal duties towards God and the faithful, characterized by a comparatively elevated style and extensive use of rhetorical figures (Gasti, 2013).

The texts were downloaded in plain-text (.txt) format from two digital resources: *Ora et Labora* for the *Regula pastoralis*, and *The Latin Library* for other three texts.¹⁸ Sentence splitting and tokenization were performed using a Python script, which is provided as supplementary file (see Supplementary file 1). Part of speech and dependency relations were manually annotated. For further details on the dataset annotation, see Section (4.1).

(4) METHOD

(4.1) DATASET ANNOTATION

The manual annotation was performed by an annotator expert in Latin and adheres strictly to the UD guidelines. To ensure consistency across the treebank, several prior decisions were made.

One of them concerned the annotation of participles, i.e., morphologically derived verbal forms that can function as adjectives (e.g., *broken window*), nouns (e.g., *the breaking of the law*), or as components of periphrastic verb forms (e.g., *has broken*; Passarotti et al., 2025, p. 103). In line with the practices adopted in the *Index Thomisticus* corpus (Busa, 1974–1980) and in the corresponding treebank (Passarotti, 2019), participles were generally assigned the POS tag VERB, reflecting the view that they are primarily verbal derivatives. Only a limited set of fully lexicalized participles—i.e., forms that have undergone nominalization or adjectivization—were assigned the POS tag NOUN (e.g., *aduentus* ‘arrival’) or ADJ (e.g., *efficiens* ‘efficient’). Their classification was based on the lexical categories recorded in authoritative Latin dictionaries, namely the *Thesaurus Linguae Latinae* and *A Latin Dictionary* (Lewis & Short, 1879), as well as on previous annotations in the UD Latin-CIRCSE treebank, consulted through TüNDRA, a web-based application for searching and exploring treebanks.¹⁹

¹⁸ <https://ora-et-labora.net>; <https://www.thelatinlibrary.com>, last accessed 26.09.2026.

¹⁹ <https://weblicht.sfs.uni-tuebingen.de/Tundra>, last accessed 26.09.2026.

The same criterion was applied to gerundives. As deverbal adjectives, when used as noun modifiers, they were assigned the POS tag `VERB` and syntactically annotated as an adnominal clause (`acl`). An exception was made for those gerundives that, already in early usage, had become fully lexicalized as adjectives (e.g., *secundus* ‘following’; see Ernout & Thomas, 1951, p. 263; Hofmann & Szantyr, 1965, p. 369).

Regarding relative pronouns, the manual annotation showed that in sixth century Late Latin they are frequently used without an overt pronominal antecedent, thus instantiating nominal ellipsis. In such cases, the current shared annotation practice does not account for the missing pronominal antecedent. Figure 1 provides an example.²⁰ However, since the antecedent is almost obligatory in Latin, its absence should be signalled in some way. A possible solution would be, as suggested by Corbetta et al. (2025), to mark these instances with a specific subtype (`:ellipsis`) to clearly identify them. Another possible solution would be to use the `orphan` relation, typically adopted in cases of head ellipsis where simple promotion would result in misleading dependency relations.²¹ Given the complexity of the issue, a more detailed investigation is left for future research.

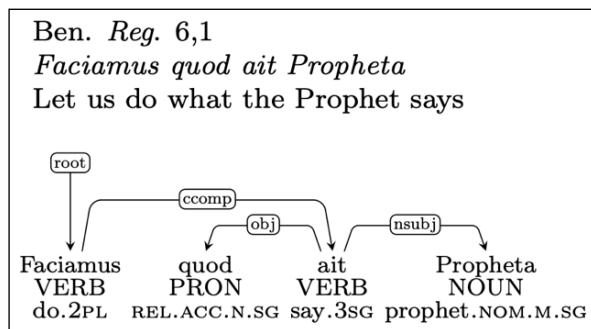


Figure 1 Example of manual POS-annotation with nominal ellipsis.

(4.2) EVALUATION

The selected tools and their models were evaluated on the same data used for the manually annotated gold standard. Each model was run separately. Only POS tagging and dependency parsing were performed, while the tokenization and sentence segmentation previously implemented in Python were retained.

For UDPipe 2, the models were accessed through the LINDAT/CLARIN online interface. Stanza was instead run in Python, with the package configuration modified according to the model being tested.

The accuracy of the automatically generated annotations was evaluated against the gold standard using the `conll18_ud_eval_with_subtypes-EvaLatin2026.py` script.²²

(5) RESULTS AND DISCUSSION

Table 3 reports the F1 score for UPOS (Universal Part of Speech), UAS (Unlabeled Attachment Score), LAS (Labeled Attachment Score) and CLAS (Content-Word Labeled Attachment Score) scores obtained by the models of UDPipe 2 and Stanza.²³ The model names refer to the UD treebanks on which the corresponding models were trained, whereas evaluation was carried out on the same textual data constituting the gold standard. As can be seen, CIRCSE and EvaLatin 2024 are reported only for UDPipe 2, since no corresponding trained models are currently

²⁰ The translation is my own.

²¹ For further details on nominal ellipsis see <https://universaldependencies.org/u/overview/specific-syntax.html#ellipsis-in-nominals>, last accessed 26.09.2026.

²² https://github.com/CIRCSE/LT4HALA/blob/master/2026/data_and_doc/conll18_ud_eval_with_subtypes-EvaLatin2026.py, last accessed 26.09.2026.

²³ The XPOS metric is the percentage language-specific POS of the word identified. Labeled Attachment Score (LAS) is the percentage of tokens assigned both the correct DEPREL and HEAD. Unlabeled Attachment Score (UAS) is the percentage of tokens assigned the correct HEAD. Content-Word Labeled Attachment Score (CLAS) is the labeled F1-score over all relations except relations for function words (aux, case, cc, clf, cop, det, mark) and for punctuation (punct). For greater details, see Buchholz and Marsi (2006).

	UDPipe 2				STANZA			
Model	UPOS	UAS	LAS	CLAS	UPOS	UAS	LAS	CLAS
Perseus	93.41	73.76	66.22	67.74	89.86	63.62	53.51	52.53
PROIEL	90.34	72.91	63.43	65.78	75.88	70.06	54.90	55.95
ITTB	92.75	72.40	64.71	66.81	90.23	69.22	60.43	61.92
LLCT	83.93	54.94	42.20	39.66	85.43	57.54	45.72	44.52
UDante	92.83	73.21	65.63	67.36	90.15	65.56	55.31	55.21
CIRCSE	92.79	73.43	65.85	67.08				
EvaLatin 2024	94.33	80.34	74.08	78.64				

Table 3 Evaluation scores of the UDPipe 2 and Stanza models, in random order.

available in Stanza. The results will first be discussed by comparing UDPipe 2 and Stanza (Section (5.1)), and then by analyzing in depth the UDPipe 2 models (Section (5.2)).

(5.1) A COMPARISON BETWEEN UDPipe 2 AND STANZA

When comparing the two tools, UDPipe 2 outperforms Stanza in almost all cases. The only exception is the LLCT model, for which Stanza achieves higher scores: 85.43 vs 83.93 for UPOS, 57.54 vs 54.94 for UAS, 45.72 vs 42.20 for LAS, and 44.52 vs 39.66 for CLAS. The generally stronger performance of UDPipe 2 is consistent with the findings of Corbetta et al. (2024), who reported that UDPipe 1 initially outperformed Stanza in the automatic annotation of Dante’s *Inferno*. However, after task-specific training, Stanza achieved better results than UDPipe 1.

The comparison also shows that the performance gap between the two pipelines is relatively small for UPOS tagging but becomes more pronounced for dependency parsing. Both systems generally achieve high POS-tagging accuracy, with a mean UPOS accuracy of 91.48 for UDPipe 2 (SD = 3.54) and 86.31 for Stanza (SD = 6.17). For dependency parsing, UDPipe 2 consistently obtains higher scores, with mean accuracies of 71.57 for UAS (SD = 7.81), 63.16 for LAS (SD = 9.86) and 64.72 for CLAS (SD = 11.89) compared with 65.20 (SD = 5.03), 53.97 (SD = 5.31), and 54.03 (SD = 6.32), respectively, for Stanza. It should be noted, however, that part-of-speech annotation is generally less demanding and achieves slightly higher inter-annotator agreement than dependency annotation, even when performed manually (see, for instance, Berzak et al., 2016).

Turning to the individual models, the influence of the training treebank is also noteworthy. Performance varies considerably across models, indicating that the choice and characteristics of the training data have a substantial impact on annotation quality. Given UDPipe 2’s generally stronger performance in the present evaluation, as well as the larger number of available models, the following Section (5.2) provides a more detailed analysis of its results.

(5.2) EVALUATION OF UDPipe 2 MODELS

EvaLatin 2024 achieves the best performance across all evaluation metrics, with a UPOS accuracy of 94.33, a UAS of 80.34, a LAS of 74.08, and a CLAS of 78.64. It consistently outperforms all the other models considered in the evaluation. The difference is particularly marked for the dependency parsing metrics—UAS, LAS, and CLAS—for which EvaLatin 2024 exceeds the second-best model by approximately 6 to 11 points. One possible explanation for this result is the composition of its training data: the model was trained on relatively large treebanks covering a broad chronological span and a wide range of textual genres. Notably, however, these treebanks do not include Late Latin texts, suggesting that broad chronological and genre coverage may compensate, at least in part, for the absence of training data from the target historical variety.

At the other end of the scale, LLCT obtains the lowest scores, with a UPOS accuracy of 83.93, a UAS of 54.94, a LAS of 42.20, and a CLAS of 39.66. The texts included in the LLCT treebank are chronologically relatively close to those in the present dataset, with a difference of approximately two centuries. Chronological proximity alone, therefore, does not appear to guarantee better performance. Additionally, LLCT represents a specific and relatively idiosyncratic variety, namely the legal language of Tuscia.

It is also interesting to focus specifically on the models trained on treebanks that include Late Latin texts. Turning first to the PROIEL-based model, its training data include most of the Late Latin UD-annotated works currently available, namely Jerome's Vulgate (112,454 tokens) and Palladius' *Opus agriculturae* (12,148 tokens). Despite this, the model achieves comparatively modest results: 90.34 UPOS accuracy, 72.91 UAS, 63.43 LAS, and 65.78 CLAS. Its scores are lower than those obtained by the EvaLatin, CIRCSE, UDante, and Perseus models, suggesting that the presence of Late Latin material in the training data is not, by itself, sufficient to guarantee strong performance. It should moreover be noted that Straka et al. (2024) observed, in the case of PROIEL, a low degree of accordance with the UD guidelines.

The other model trained in part on Late Latin material is Perseus. It achieves strong results (93.41 UPOS accuracy, 73.76 UAS, 66.22 LAS, 67.74 CLAS), ranking second only to EvaLatin 2024. The Perseus training data include 2,003 tokens from the Vulgate's Revelation, corresponding to approximately 6.85% of the overall training corpus (29,223 tokens). Given its relatively limited proportion, the presence of the Vulgate is unlikely to be the major factor contributing to the model's performance. Similarly to EvaLatin 2024, the Perseus treebank covers a broad chronological span and includes a wide range of textual genres, which may improve the model's ability to generalize across different linguistic varieties. In addition, the Perseus model was trained on a substantially smaller dataset than the others (see Section (2.3)). Despite this, it achieves strong performance across the evaluation metrics, further suggesting that the size of the training set alone does not determine the model's effectiveness (Nehrdich & Hellwig, 2022, p. 22).

Among the other models, CIRCSE, UDante, and ITTB achieve remarkably similar results, with differences among themselves generally remaining below one percentage point across all four metrics. Notably, the three models were trained on treebanks developed within the same research environment at CIRCSE, Università Cattolica del Sacro Cuore (Milan, Italy).²⁴ On the other hand, the treebanks represent different chronological and linguistic varieties of Latin, ranging from Silver Latin to Medieval Latin (see Section (2.3)). This suggests, once again, that the chronology of the treebank may play a secondary role.

Before concluding, it is worth noting that a detailed error analysis could shed further light on which forms are most frequently mislabelled and whether errors cluster around features specific to Late Latin. A comparison of error rates across texts could also help assess the role of chronological and genre differences. Due to space constraints, I leave this analysis for future work.

(6) IMPLICATIONS/APPLICATIONS

This study has highlighted an important methodological gap: the very limited availability of annotated Late Latin treebanks and NLP models specifically trained on this linguistic variety.

In such a low-resource setting, the first contribution of the paper is the creation of a manually annotated gold-standard, consisting of selected portions of four sixth-century texts. This dataset may have several further applications. First, it provides a benchmark for the evaluation of existing and future NLP tools and models of Late Latin. Second, it facilitates linguistic studies of sixth-century Latin (e.g., pronominal ellipsis; see Section (4.1)). In addition, if expanded and enriched, the dataset could be combined with the Late Latin subsets currently available in existing treebanks to train a model exclusively on Late Latin data. Such an experiment could help disentangle the effect of chronological proximity from those of training-data size and genre diversity.

By focusing on part-of-speech and syntactic annotation, the research shows that UDPipe 2 generally performs better than Stanza on the data examined here. Among the models tested, the EvaLatin 2024 model achieves the highest overall accuracy. This finding has a direct practical implication: this model could be used as the initial stage of a semi-automatic workflow for the creation of new Late Latin treebanks. Automatic pre-annotation would reduce the amount of manual work required at the beginning of the process, while expert revision would remain

²⁴ <https://centridiricerca.unicatt.it/circse/en.html>, last accessed 26.09.2026.

essential to correct errors, resolve ambiguous cases, and ensure the consistency and reliability of the final resource.²⁵

More generally, the paper suggests that, even in the absence of models trained specifically on a variety of language, existing models trained on a broad range of textual genres can provide a useful starting point for corpus annotation. Such models are clearly preferable to having no suitable model at all, but previous literature (see, among others, Khan et al., 2013) has shown the importance and benefits of using chronologically in-domain training data. This study therefore points to key research priority: the development of larger and more representative Late Latin treebanks, which would provide a stronger empirical basis for training models specifically tailored to this stage of the language.

AI DECLARATION

During the preparation of this work, the author used ChatGPT (OpenAI) in order to: Paraphrase and reword and Grammar and spelling check. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplementary File 1.** Python script used for sentence splitting and tokenization. DOI: <https://doi.org/10.5334/johd.630.s1>

ACKNOWLEDGEMENTS

The author gratefully acknowledges the CIRCSE research community for its welcoming and stimulating environment. Special thanks go to Marco Passarotti for his role in fostering it and to Federica Iurescia for her invaluable guidance and support. The author also wishes to thank the anonymous reviewers for their careful reading and constructive comments, which helped improve the manuscript.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATIONS

Annachiara Clementelli  orcid.org/0009-0005-4198-0900

Department of Italian Language, Literature, and Arts Around the World, University for Foreigners of Perugia, Perugia, Italy; Department of Linguistics, Ghent University, Ghent, Belgium

REFERENCES

- Bamman, D., & Crane, G. (2006). The Design and Use of a Latin Dependency Treebank. In: J. Hajič & J. Nivre (eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories* (pp. 67–78). Prague: ÚFAL MFF UK.
- Banniard, M. (2020). *Viva voce: Comunicazione scritta e comunicazione orale nell'Occidente latino dal IV al IX secolo* (L. Cristante & F. Romanini, Eds.). Trieste: Edizioni Università di Trieste.
- Berzak, Y., Huang, Y., Barbu, A., Korhonen, A., & Katz, B. (2016). Anchoring and agreement in syntactic annotations. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2215–2224). Austin: Association for Computational Linguistics. <https://doi.org/10.18653/v1/D16-1239>
- Biagetti, E., & Brigada Villa, L. (2023, August 29–September 1). Verb–object pairs in the Latin, Gothic, Armenian, and Old Church Slavonic translations of the Gospels: A treebank-based quantitative study

²⁵ At the time of writing, the model developed for the EvaLatin 2026 dependency parsing task is not yet available among the models distributed for UDPipe 2. The task and its results were presented at LT4HALA 2026, held in conjunction with LREC 2026 in May 2026 (Iurescia et al., 2026). Once available, the EvaLatin 2026 model will also need to be evaluated, in order to determine whether, and to what extent, it improves on the performance observed here.

- on the influence of Greek word order [Paper presentation]. 56th Annual Meeting of the Societas Linguistica Europaea, Athens.
- Buchholz, S., & Marsi, E.** (2006). CoNLL-X shared task on multilingual dependency parsing. In L. Màrquez & D. Klein (Eds.), *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)* (pp. 149–164). New York City: Association for Computational Linguistics. <https://doi.org/10.3115/1596276.1596305>
- Busa, R.** (1974–1980). *Index Thomisticus*. Stuttgart-Bad Cannstatt: Frommann-Holzboog.
- Corbetta, C., Moretti, G., & Passarotti, M.** (2025, January 29–30). How to ... Ellipsis: a Proposal from Old Italian for the Annotation of Elliptical Structures in UD [Paper presentation]. *UniDive 3rd General Meeting*, Hungarian Research Centre for Linguistics, Budapest. <https://doi.org/10.63317/32zvso57vq2k>
- Corbetta, C., Passarotti, M., & Moretti, G.** (2024). The rise and fall of dependency parsing in Dante Alighieri's *Divine Comedy*. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024* (pp. 50–56). Torino: ELRA and ICCL.
- Cuzzolin, P., & Haverling, G.** (2009). Syntax, sociolinguistics, and literary genres. In P. Baldi & P. Cuzzolin (Eds.), *New perspectives on historical Latin syntax: Volume 1 Syntax of the Sentence* (pp. 19–64). Berlin/New York: De Gruyter Mouton. <https://doi.org/10.1515/9783110205626.19>
- de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D.** (2021). Universal Dependencies. *Computational Linguistics*, 47(2), 255–308. https://doi.org/10.1162/coli_a.00402
- Eckhoff, H. M., Luraghi, S., & Passarotti, M.** (Eds.). (2020). *Diachronic treebanks for historical linguistics*. Amsterdam: John Benjamins. <https://doi.org/10.1075/bct.113>
- Ernout, A., & Thomas, F.** (1951). *Syntaxe latine*. Paris: Klincksieck.
- Gasti, F.** (2013). *Profilo storico della letteratura tardolatina*. Pavia: Pavia University Press.
- Haverling, G.** (1988). *Studies on Symmachus' language and style*. Göteborg: Acta Universitatis Gothoburgensis.
- Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D.** (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545–2568). Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.201>
- Hofmann, J. B., & Szantyr, A.** (1965). *Lateinische Syntax und Stilistik*. München: Beck.
- Iurescia, F., Passarotti, M., & Sprugnoli, R.** (2026). Overview of the dependency parsing task at EvaLatin 2026. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026) @ LREC 2026* (pp. 211–218). Palma: European Language Resources Association. <https://doi.org/10.63317/2j3y8jv8ktyd>
- Jurafsky, D., & Martin, J. H.** (2024). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed. draft). Retrieved from <https://web.stanford.edu/~jurafsky/slp3/> (last accessed: 21 September 2026).
- Khan, M., Dickinson, M., & Kübler, S.** (2013). Towards domain adaptation for parsing web data. In R. Mitkov, G. Angelova, & K. Bontcheva (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* (pp. 357–364). Hissar: INCOMA Ltd. Shoumen.
- Ledgeway, A.** (2012). *From Latin to Romance: Morphosyntactic typology and change*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199584376.001.0001>
- Lewis, C. T., & Short, C.** (1879). *A Latin dictionary: Founded on Andrews' edition of Freund's Latin dictionary*. Oxford: Clarendon Press.
- Nehrdich, S., & Hellwig, O.** (2022). Accurate dependency parsing and tagging of Latin. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages* (pp. 20–25). Marseille: European Language Resources Association.
- Passarotti, M.** (2019). The Project of the Index Thomisticus Treebank. In M. Berti (Ed.), *Digital Classical Philology: Ancient Greek and Latin in the Digital Revolution* (pp. 299–320). Berlin/Boston: De Gruyter Saur. DOI: <https://doi.org/10.1515/9783110599572-017>
- Passarotti, M., Iurescia, F., & Ruffolo, P.** (2025). Harmonizing divergent lemmatization and part-of-speech tagging practices for Latin participles through the LiLa Knowledge Base. In S. Peng & I. Rehbein (Eds.), *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)* (pp. 103–114). Vienna: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.law-1.8>
- Passarotti, M., McGillivray, B., & Bamman, D.** (2015). A treebank-based study on Latin word order. In G. Haverling (Ed.), *Latin linguistics in the early 21st century: Acts of the 16th International Colloquium on Latin Linguistics, Uppsala, June 6th–11th, 2011* (pp. 338–350). Uppsala: Acta Universitatis Upsaliensis.
- Pinkster, H.** (2015). *The Oxford Latin syntax: Volume 1, The simple clause*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199283613.001.0001>

- Pinkster, H. (2021). *The Oxford Latin syntax: Volume 2, The complex sentence and discourse*. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780199230563.001.0001>
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-demos.14>
- Riemenschneider, F., & Frank, A. (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)* (pp. 15181–15199). Toronto, Canada: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-long.846>
- Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., & de Freitas, N. (2023). Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3), 703–747. https://doi.org/10.1162/coli_a_00481
- Sprugnoli, R., Iurescia, F., & Passarotti, M. (2024). Overview of the Evalatin 2024 evaluation campaign. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024* (pp. 190–197). Torino: ELRA and ICCL. <https://doi.org/10.63317/2ip45gum7dfg>
- Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In D. Zeman & J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 197–207). Brussels: Association for Computational Linguistics. <https://doi.org/10.18653/v1/K18-2020>
- Straka, M., Straková, J., & Gamba, F. (2024). ÚFAL LatinPipe at Evalatin 2024: Morphosyntactic analysis of Latin. In R. Sprugnoli & M. Passarotti (Eds.), *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024* (pp. 207–214). Torino: ELRA and ICCL. <https://doi.org/10.63317/5ogjqhfdzgra>
- Thesaurus linguae Latinae*. (1900–). Leipzig: B. G. Teubner. Retrieved from <https://thesaurus.badw.de/tll-digital/tll-open-access.html> (last accessed: 21 September 2026).

TO CITE THIS ARTICLE:

Clementelli, A. (2026). Evaluating UDPipe 2 and Stanza for Morphosyntactic Annotation of Late Latin: A Pilot Study on Sixth-Century Texts. *Journal of Open Humanities Data*, 12: 140, pp. 1–12. DOI: <https://doi.org/10.5334/johd.630>

Submitted: 07 August 2026

Accepted: 22 September 2026

Published: 26 October 2026

COPYRIGHT:

© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <https://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.