The UD_Latin-PROIEL as Linked Open Data: Integrating a Latin Treebank into the LiLa Knowledge Base

Lucas Dezotti[1], Marco Passarotti[2], Federica Iurescia[2], Giovanni Moretti[2]
[1] Universidade Federal da Paraíba, [2] Università Cattolica del Sacro Cuore
[1] 58051-900, João Pessoa, Brazil, [2] Largo Gemelli 1, 20123 Milan, Italy
lucas.dezotti@academico.ufpb.br, {marco.passarotti, federica.iurescia, giovanni.moretti}@unicatt.it

Abstract

This paper presents the steps taken to integrate data from the UD_Latin-PROIEL treebank into the LiLa Knowledge Base of interoperable linguistic resources for Latin. It describes how the lexical, morphological, syntactic, and citation information from the source was modeled using the Linked Open Data principles as adopted by the LiLa Knowledge Base. The process of linking tokens to the LiLa collection of Latin lemmas is detailed, addressing challenges such as ambiguities, new lemmas, and errors encountered in the source. The outcome is a syntactically annotated textual resource that is interoperable with the (meta)data of other Latin linguistic resources linked within the LiLa Knowledge Base. This integration enables new ways of analyzing linguistic information and using the content as a starting point to explore connections with other interlinked resources. A use case demonstrates this interoperability.

Keywords: Linked Open Data, Linguistic Annotation, Universal Dependencies treebank

1. Introduction

Over the past two decades, several Latin treebanks — corpora annotated with syntactic structures, morphological information, and lemmatization — have been created. Among the earliest were the Latin Dependency Treebank (LDT) (Bamman and Crane, 2006) and the Index Thomisticus Treebank (IT-TB) (Passarotti, 2019). Subsequent projects include the PROIEL Treebank (Haug and Jøhndal, 2008), the Late Latin Charter Treebank (Korkiakangas and Passarotti, 2011), UDante (Passarotti et al., 2021) and, more recently, UD_Latin-CIRCSE (Iurescia et al., 2025). Together, these resources sample Latin from a broad arc of time, stretching from Classical Antiquity to the late Middle Ages.

Treebanks have become a key resource in linguistic research and natural language processing (Nivre, 2008), enabling reproducible experiments and supporting a wide range of computational and theoretical studies (Bamman and Crane, 2011). Latin treebanks have underpinned research in morphosyntactic theory (Haug, 2015), lexicography (McGillivray and Vatri, 2015), textual similarity (Bamman and Crane, 2008), authorship style (Cecchini and Pedonese, 2022), parser development and evaluation (Burns, 2023), and language teaching (Mambrini, 2016).

The need for a common standard for the syntactic annotation of Latin texts has been recognized from the beginning, as evidenced by the shared annotation guidelines developed for the LDT and IT-TB projects (Bamman et al., 2007). A major step forward came with the Universal Dependencies (UD) project (Nivre et al., 2016), which provides a framework for cross-linguistically consistent morphosyntactic annotation and has become the de facto standard for many treebanks. Consequently, Latin treebanks originally produced under different annotation schemes have been converted, at least partially, into the UD format. Although conversion can be challenging (Cecchini et al., 2018, 2020) and often requires further refinement and harmonization to ensure data consistency (Gamba and Zeman, 2023b,a), it represents an important step toward compatibility among corpora produced by different projects.

A further step is to make Latin treebanks interoperable with other linguistic resources. This can be achieved by publishing them as Linked Open Data in the LiLa Knowledge Base (Passarotti et al., 2020), a collection of textual and lexical resources for Latin. Currently, the corpora linked to the LiLa Knowledge Base total nearly 12 million tokens; however, only two of these corpora are annotated at the syntactic level, amounting to 430,000 tokens: the IT-TB (375,000) and UDante (55,000). Despite their undeniable importance, these resources mainly represent a specific variety of Latin, namely Medieval Latin from the 13th and 14th centuries. This highlights a gap in the linked corpora with respect to diachronic representativeness.

To help close this gap, this paper describes the steps taken to link the UD_Latin-PROIEL treebank to the LiLa Knowledge Base. It focuses on how the lexical, morphological, and syntactic annotations, as well as citation information, are modelled using the Linked Open Data principles as adopted in LiLa. The remainder of the paper is organized as follows: (a) an overview of the LiLa KB structure,

Proceedings of LT4HALA 2026 @ LREC 2026, pages 353–360.

especially the model used to publish annotated textual resources as Linked Open Data; (b) the main characteristics of the UD_Latin-PROIEL treebank; (c) the process of modelling and linking tokens to the LiLa collection of Latin lemmas, including issues related to ambiguity, new lemmas, and errors in the source; and (d) two use case scenarios illustrating the research possibilities enabled by the resulting linked resource.

2. Textual Resources in the LiLa Knowledge Base

LiLa is a Linked Open Data-based Knowledge Base of linguistic resources for Latin. It promotes interoperability by adopting widely used ontologies for representing linguistic information, together with Semantic Web and Linked Data standards. In particular, OLiA is used to model linguistic annotation (Chiarcos and Sukhareva, 2015), Ontolex-Lemon to represent lexical data (McCrae et al., 2017), and POWLA to model corpus data (Chiarcos, 2012). LiLa relies on the Resource Description Framework (RDF) as its data model, representing information as triples (McBride, 2004).

Its core is the Lemma Bank, a collection of approximately 170,000 Latin headwords and 200,000 written representations.[1] In the LiLa ontology, the class Lemma is subsumed under the class Form, as defined in the Ontolex ontology (Cimiano et al., 2016). A lemma is defined as an inflected form chosen as the citation or canonical form of a lexical item. Each lemma is assigned an exclusive part of speech and an inflectional type. Some lemmas are additionally assigned a morphological base – i.e., the lexical morpheme of a word that is neither a prefix nor a suffix (Passarotti et al., 2020) –, a property that is essential for disambiguating homographic lemmas such as `occido` (*ob* + *caedo* 'to strike down') and `occido` (*ob* + *cado* 'to fall down').

Because lemmatization is a shared annotation layer across both lexical resources and corpora, the Lemma Bank serves as a central hub for integrating distributed resources in LiLa. Interoperability is achieved by linking lexical entries and corpus tokens to their corresponding lemmas in the Lemma Bank.

As described in Mambrini et al. (2022), text corpora in LiLa are modelled using POWLA, which provides classes and properties to represent document stratifications and subdivisions, and organizes different types of annotation into separate layers. A corpus is represented as an instance of POWLA's `Corpus` class. Its internal subdivisions — i.e., the works it comprises — are instances of `Document` and are linked to the corpus via `hasSubDocument`.

For each document, four annotation layers store information on tokenization, sentence division, morphology, and citation hierarchy. As regards morphological features, they are represented as properties predicated of the annotation units (i.e., tokens), using POWLA's `hasBody` to list the features and `hasTarget` to connect them to the corresponding token. In turn, dependency relations are represented as instances of the local class `DependencyRelation` (defined as a subclass of POWLA's `Relation`), and two dedicated properties connect each relation to its head and dependent nodes (the properties `hasHead` and `hasDep`, as specializations of POWLA's `hasSource` and `hasTarget`, respectively). Finally, the citation-hierarchy layer provides a way to track token positions in the text that are especially relevant for research in historical linguistics and digital humanities.

3. The PROIEL Latin treebank

The Pragmatic Resources of Old Indo-European Languages (PROIEL) project (Haug and Jøhndal, 2008) created a multilingual treebank to investigate the grammatical means by which information structure is expressed in five languages: Greek, Latin, Armenian, Gothic, and Church Slavonic. Initially, the project adopted an annotation tagset derived from the LDT, enriched with additional fine-grained distinctions (Haug, 2010). In 2017, the data were automatically converted to the UD scheme, and the Latin portion was released as "UD_Latin-PROIEL".

The latest release (r2.17) (Haug, 2025) contains most of the Vulgate New Testament translations, as well as selections from Caesar's *Commentarii de Bello Gallico*, Cicero's *Epistulae ad Atticum*, Palladius' *Opus Agriculturae*, and the first book of Cicero's *De officiis*, totaling approximately 205,500 tokens. Table 1 reports the number of tokens per work and the proportion of each work covered by the treebank.[2]

This release serves as the source for the Linked Open Data version presented in this paper.

[1] In LiLa, written representations (`https://www.w3.org/ns/lemon/ontolex#writtenRep`) account for spelling variations of a lemma. A lemma may have one or more such representations.

[2] The percentages are based on the word counts of the full texts available at *The Latin Library* website (`https://www.thelatinlibrary.com/`) and at a snapshot of the *Forum Romanum* website preserved in the Internet Archive's Wayback Machine (`https://web.archive.org/web/20230530004518/http://www.forumromanum.org/literature/palladius/agr.html`).

Table 1: Texts treebanked in UD_Latin-PROIEL.

document	tokens	% of full work
New Testament	109517	87.4%
De bello Gallico	27386	53.4%
De officiis	11375	33.3%
Epistulae ad Atticum	45308	36.3%
Opus agriculturae	11980	29.9%

Table 2: Results of the lemma matching process.

unique lemmas	n	%
Total	8,663	100.0%
1:1 single matches	7,844	90.5%
1:N ambiguous matches	519	6.0%
1:0 no matches (total)	300	3.5%
1:0 with a single candidate	102	1.2%
1:0 with multiple candidates	196	2.3%
1:0 with no candidates	2	<0.1%

4. Modeling and Linking Treebank Data

The PROIEL Latin treebank data are modeled as Linked Open Data through an automatic conversion of the CoNLL-U files from the aforementioned release. The data are extracted as-is from the source, with a single typographical correction.[3] However, the CoNLL-U file must be enriched with additional information on the citation hierarchy and, most importantly, with lemma URIs — since linking token URIs to lemma URIs via the `lila:hasLemma` property is what connects the corpus to the LiLa Knowledge Base.

The token linking process matches the lemmas annotated in the corpus to their written representation(s) recorded in the LiLa Lemma Bank. Due to lemmatization problems in the source,[4] a manual inspection is required before automatic matching can proceed, in order to either link these tokens or mark them as ineligible for linking. Of the approximately 860 affected tokens, most remain unlinked. This includes monetary values such as $\overline{\text{DCCC}}$ (800,000) and 492 Greek words (mostly from Cicero's *Letters to Atticus*). Conversely, some tokens are manually linked to the relevant Lemma Bank entries, including calendar expressions (full or abbreviated) such as 'Decembribus', 'Febr.' (i.e. *Februariis*), and 'a.d.III' (i.e. *ante diem tertius*), as well as monetary expressions such as 'HS' (*sestertius*).

To improve matching accuracy and reduce ambiguity and false positives, lemmas are extracted from the source corpus, then normalized and enriched with part-of-speech tags and (where applicable) gender labels. Normalization includes lowercasing and the substitution of *j* with *i* and *v* with *u*. Part-of-speech and gender labels are derived from the morphological features annotated in the source and mapped to the LiLa tagset. This procedure produces a composite string consisting of the lemma and its relevant features (e.g. `res_NOUN_f` and `dico_VERB`).

Matching is performed programmatically using a progressive, three-step approach. First, composite strings are matched. Second, unmatched items are reprocessed using only the lemma string. Finally, for the remaining unmatched lemmas, the edit distance to the lemma collection in LiLa is computed (with a threshold of 2) to generate a set of candidate links. The results are classified into three categories: single matches (1:1), ambiguous matches (1:N), and no matches (1:0) (see Table 2).

Single matches are automatically validated and account for 90% of the total links. This indicates strong performance for the proposed method, particularly in its ability to distinguish homographs. However, this unsupervised linking raises at least two issues. First, derivative forms such as participles may be lemmatized and tagged inconsistently in the source data, as either verbal or adjectival. In this case, LiLa's ontology class of Hipolemma – a subclass of `lila:Lemma` used to represent a citation form belonging to a word's regular inflectional paradigm that receives a different PoS tag or degree of comparison than its most canonical lemma – solves the problem, as tokens can be unified under a shared representational framework (Passarotti et al., 2025). Second, some 1:1 mappings may hide a potential 1:N mapping, considering that later Latin developments might contain homographs that are absent from the current LiLa lemma database. This remains an open issue for future investigation.

Ambiguous matches and linking candidates account for fewer than 10% of the lemmas and require manual validation. In most cases, disambiguation can be performed by inspecting the token form. For instance, *frondes* (sent_id 11089, token 3) and *frontem* (sent_id 75866, token 2) are both lemmatized as `frons`, yet they are clearly forms of two distinct lexemes, as shown by their

[3] Namely, the token `magnanimnos` (sent_id 86312, token 5), corrected to `magnanimos`. While issues with morphological features (Gamba and Zeman, 2023a) and dependency relations (Gamba and Zeman, 2023b) have been reported, no official release has addressed them yet.

[4] There are six problematic lemma strings: `calendar`, `calendar.expression`, `expression`, `greek.expression`, `monetary`, `monetary.expression` and `FIXME`. The issue is reported in https://github.com/UniversalDependencies/UD_Latin-PROIEL/issues/1.

Table 3: New lemma entries and written representations (wr) added to the LiLa Lemma Bank, categorized by part of speech (pos).

pos	new wr	new entries
PROPN	83	35
ADJ	13	8
ADV	4	3
NOUN	6	4
NUM	1	0

morphological bases *frond-* and *front-* (meaning 'leafy branches' and 'the forehead', respectively). In other cases, however, the token form itself is ambiguous, and disambiguation requires semantic analysis to identify distinct morphological bases. A representative example is *deserunt*, consistently lemmatized as `desero` but corresponding to different verbs: one derives from the base **se-* and means 'to sow', as in Palladius' *De agri cultura* (sent_id 159953, token 6); the other derives from the base *ser-* and means 'to abandon', as in Cicero's *De officiis* (sent_id 86136, token 1).

Both single and multiple linking candidates are inspected manually, with decisions made on a case-by-case basis. Of the 300 unmatched lemmas, 130 corresponded to lemmas already present in the Lemma Bank, the mismatch being due to inconsistencies in the corpus lemmatization. The remaining 170 lemmas required additions to the Lemma Bank, resulting in 107 new written representations for existing lemmas and 50 entirely new lemma entries, most of which are proper nouns (see Table 3). Furthermore, 13 words were left unlinked as they are Aramaic transcriptions, mostly from the Gospel of Mark (e.g., in the phrases "Heloi Heloi lama sabacthani" and "talitha cumi").

Finally, modeling the citation hierarchy requires converting the source data into standardized strings. For instance, from the source attribute-value pair `Ref=MATT_1.1`, a new pair is created consisting of the attribute `CitationHierarchy` and the value "Evangelium_Matthaei,Capitulum_1,Versiculus_1". In addition, the original `Ref` value is also standardized, in this case to `Vulg.Matt.1.1`, using abbreviations drawn from standard Latin dictionaries (Glare, 1968; Lewis and Short, 1879).

Figure 1 illustrates the conversion and linking process using the nominal phrase "philosophiae praeceptis" from Cicero's *De officiis* 1.1. The work node is linked to the corpus as its SubDocument and to the annotation layers as their Document. Each annotation layer connects to the tokens (directly or through subunits) via specific properties. The tokens are in turn linked to the appropriate Lemma nodes and to a dependency relation node (UD DepRel `nmod`)—as either the dependent (*philosophiae*) or the head (*praeceptis*).

5. Querying Interlinked Resources

The PROIEL Latin treebank data are now part of the LiLa Knowledge Base and can be accessed via an interactive search interface[5] or through a SPARQL endpoint[6]. Both access methods enable users to explore connections between the treebank and the many linguistic resources interlinked in LiLa.[7] To illustrate this potential, we present a query-based use case focused on sentiment analysis.

One of the lexical resources interlinked in LiLa is the Latin Affectus lexicon (Sprugnoli et al., 2020), a prior-polarity lexicon of Latin lemmas. It supports sentiment classification of LiLa-linked documents by identifying so-called sentiment words (Liu, 2012). In practice, one can use the citation layer to restrict the query scope (e.g., to a book or chapter), follow links from citations to tokens and from tokens to lemmas, and then retrieve polarity information from the lexicon. This information can be used to compute (a) the number of polarity-marked lemmas, (b) the sum of their polarity values, and (c) a normalized score obtained by dividing (b) by (a).[8]

As an example, we apply this method to Books 3 and 4 of Cicero's *Letters to Atticus*, included in UD_Latin-PROIEL. These books correspond to two distinct moments in Cicero's political career: his banishment from Rome (58 BCE) and his return (57 BCE). Moreover, Cicero explicitly foregrounds his emotional state, inviting the recipient (and thus the reader) to attend to it.[9]

Each letter in Books 3 and 4 is classified as expressing positive or negative sentiment, and the resulting sequence is used to visualize the sentiment flow of each book (Figs. 2 and 3).

Although a full interpretation of the results is beyond the scope of this paper, some patterns are noteworthy. The sentiment flow of Book 3 seems to reflect Cicero's emotional instability during his flight from Rome: letters with positive scores convey hope (e.g., letters 16 and 18), whereas those

[5] https://lila-erc.eu/LiLaLisp/.

[6] https://lila-erc.eu/sparql/.

[7] A complete list of resources is available at https://lila-erc.eu/data-page/.

[8] The query script run on the LiLa SPARQL endpoint can be found here: https://github.com/lucascdz/psm/blob/main/LiLa_sentiment_analysis_Cicero_ad_Atticum_3_and_4.rq

[9] In Book 4, letter 3: "[I would like] that you may see from my letters how I am taking events and what are my feelings and my general state of existence" (Winstedt, 1912, p.273).

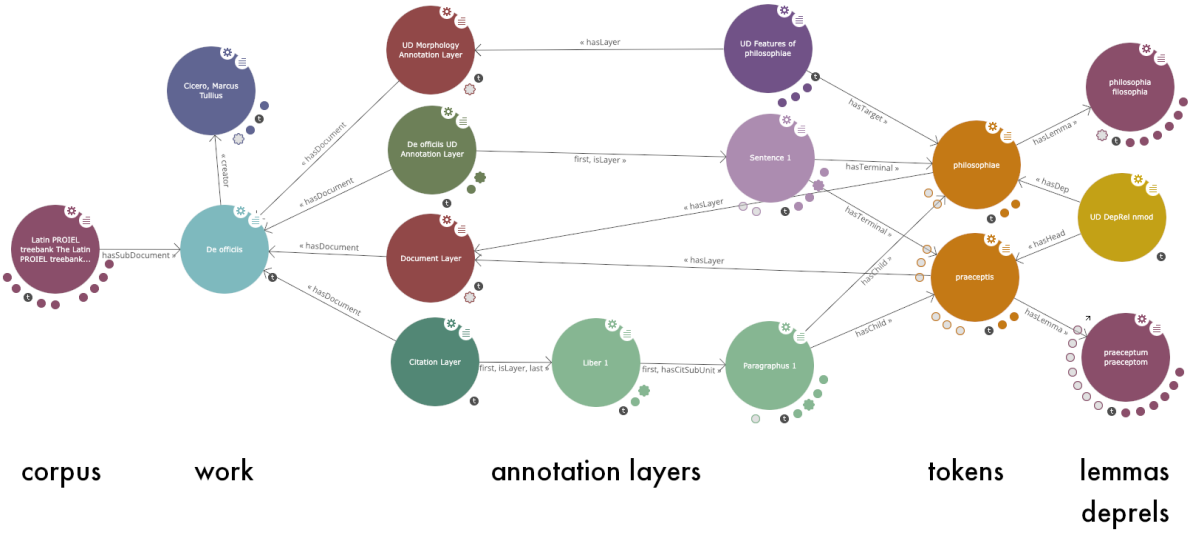

Figure 1: The nominal phrase “philosophiae praeceptis” as RDF triples in the LiLa Knowledge Base.

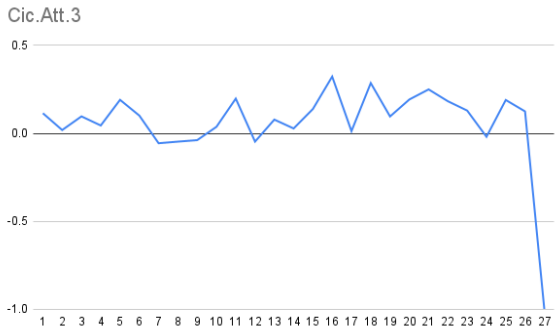

Figure 2: Sentiment classification of the letters from the third book of Cicero’s *Letters to Atticus*.

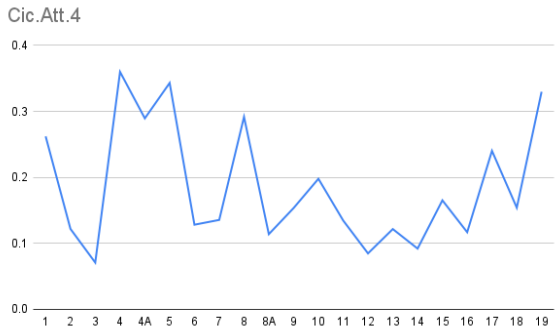

Figure 3: Sentiment classification of the letters from the fourth book of Cicero’s *Letters to Atticus*.

with negative scores express intense despair (notably letters 7–9).[10] By contrast, all letters in Book 4 receive positive scores. The lowest score (letter 3) corresponds to a report of harassment by political enemies, but it does not appear to undermine Cicero’s confidence.[11] The subsequent letter (letter 4) receives the highest score, reflecting his joy at learning that Atticus was coming to Rome. Further queries could move toward close reading, such as listing positive and negative lemmas in each letter.

Many other connections can be explored. Although space does not permit a full discussion, it is worth noting the possibility of comparing, for instance, two historians such as Caesar and Tacitus with respect to stylistic matters – for example, how they express negative meaning, whether by using the negative prefix *in-* or adverbs such as *non* – as well as lexical usage, including how many lemmas they share and which words they use to refer to non-Roman populations.

6. Conclusion

Treebanks are essential resources for the study of historical languages such as Latin. In this paper, we described the process of modelling the UD_Latin-PROIEL treebank as Linked Open Data and linking it to the LiLa Knowledge Base of interoperable linguistic resources for Latin. To this end, we addressed the challenges of identifying appropriate lemma URIs for ambiguous and previously unmatched lemmas. Although time-consuming, the data-curation process benefits both the corpus – by enabling improvements to the lemmati-

[10] An account of Cicero’s use of language of emotions in the letters written during his exile is given by Evangelou (2024).

[11] He concludes the episode by declaring: “So far as my mind is concerned, I am as strong as ever I was even in my most palmy days, if not stronger” (Winstedt, 1912, p.281).

zation layer – and the LiLa Lemma Bank, which is enriched with new entries and written representations. Overall, this work expands the possibilities for searching, assessing, and reusing treebank data within a broader ecosystem of interlinked resources.

Among the main outcomes of this work are the following:

- **Enhanced querying:** the corpus can be queried by leveraging other LiLa resources as reference points.
- **Improved reuse:** the linked data support the automatic creation of new resources, for instance by combining corpus and lexical data to derive text-specific vocabularies for educational purposes.

As such, the linked data produced here can be valuable to Latin researchers and teachers, advanced students of Latin, and computational linguists working on historical languages.

In summary, this work represents a substantial contribution to the set of Linked Open Data resources for Latin, increasing by 48% the amount of syntactically annotated corpus data available in LiLa and improving the diachronic coverage of the linked corpora by incorporating Classical and Late Latin materials. Reported inconsistencies in the source data may be addressed in future versions as new releases of UD_Latin-PROIEL become available. Given the strong performance of the proposed linking method, a natural next step is to apply it to additional Latin treebanks in order to expand the coverage of LiLa's textual resources.

7. Acknowledgements

The "LiLa - Linking Latin" project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

We are grateful to Paolo Ruffolo for his assistance with data insertion into the Lemma Bank, without which this work would not have been possible.

8. Bibliographical References

David Bamman and Gregory Crane. 2006. The design and use of a latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78.

David Bamman and Gregory Crane. 2008. The logic and discovery of textual allusion. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakesh.

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.

David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. Guidelines for the syntactic annotation of latin treebanks (v. 1.3). *Tufts University Digital Library*.

Patrick J. Burns. 2023. LatinCy: Synthetic trained pipelines for latin NLP.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020. 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, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in converting the index Thomisticus treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium. Association for Computational Linguistics.

Flavio Massimiliano Cecchini and Giulia Pedonese. 2022. A treebank-based approach to the supprema constructio in dante's Latin works. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 51–58, Marseille, France. European Language Resources Association.

Christian Chiarcos. 2012. Powla: Modeling linguistic corpora in owl/dl. In *The Semantic Web: Research and Applications*, pages 225–239, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christian Chiarcos and Maria Sukhareva. 2015. Olia – ontologies of linguistic annotation. *Semantic Web*, 6(4):379–386.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon model for ontologies: Final community group report, 10 may 2016. Technical report, Ontology-Lexicon Community Group under the W3C Community Final Specification Agreement (FSA).

Gabriel Evangelou. 2024. Loss of self, desperation, and glimmers of hope in cicero's letters from exile. In Ioannis Deligiannis, editor, *Cicero in Greece, Greece in Cicero*, pages 31–54. De Gruyter, Berlin, Boston.

Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

Peter GW Glare. 1968. *Oxford Latin Dictionary*. Clarendon Press, Oxford.

Dag Haug. 2015. Treebanks in historical linguistic research. In Carlotta Viti, editor, *Perspectives on historical syntax*, pages 187–202. John Benjamins Publishing Company.

Dag Trygve Truslew Haug. 2010. Proiel guidelines for annotation.

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.

Timo Korkiakangas and Marco Passarotti. 2011. Challenges in annotating medieval latin charters. *Journal for Language Technology and Computational Linguistics*, 26(2):105–116.

C. Lewis and C. Short. 1879. *A Latin Dictionary*. Clarendon Press, Oxford.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Francesco Mambrini. 2016. The ancient greek dependency treebank: Linguistic annotation in a teaching environment. In G. Bodard and M. Romanello, editors, *Digital Classics Outside the Echo-Chamber*, pages 83–99. Ubiquity Press, London.

Francesco Mambrini, Marco Passarotti, Giovanni Moretti, and Matteo Pellegrini. 2022. The index Thomisticus treebank as linked data in the LiLa knowledge base. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4022–4029, Marseille, France. European Language Resources Association.

Brian McBride. 2004. The resource description framework (rdf) and its vocabulary description language rdfs. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, pages 51–65. Springer Berlin Heidelberg, Berlin, Heidelberg.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*, pages 587–97, Brno. Lexical Computing CZ.

Barbara McGillivray and Alessandro Vatri. 2015. Computational valency lexica for latin and greek in use: a case study of syntactic ambiguity. *Journal of Latin Linguistics*, 14(1):101–126.

Joakim Nivre. 2008. Treebanks. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics*, volume 1, pages 225–41. Walter de Gruyter, Berlin/New York.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Marco Passarotti. 2019. The project of the index thomisticus treebank. In Monica Berti, editor, *Digital Classical Philology*, pages 299–320. De Gruyter Saur, Berlin, Boston.

Marco Passarotti, Federica Iurescia, and Paolo Ruffolo. 2025. Harmonizing divergent lemmatization and part-of-speech tagging practices for Latin participles through the LiLa knowledge base. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 103–114, Vienna, Austria. Association for Computational Linguistics.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the LiLa Knowledge Base of linguistic resources for Latin. *Studi e Saggi Linguistici (SSL)*, 58(1):177–212.

Marco Carlo Passarotti, Flavio Massimiliano Cecchini, Rachele Sprugnoli, G Moretti, et al. 2021. Udante. syntactic annotation of dante alighieri's latin texts. *Studi Danteschi*, 86(1):309–338.

Rachele Sprugnoli, Francesco Mambrini, Giovanni Moretti, and Marco Passarotti. 2020. Towards the modeling of polarity in a latin knowledge base. In *Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020) co-located with 15th Extended Semantic Web Conference (ESWC 2020). Heraklion, Greece, June 2, 2020*, pages 59–70. CEUR-WS.

Eric Otto Winstedt. 1912. *Cicero: Letters to Atticus; with an English translation*, volume 1. W. Heinemann, London.

9. Language Resource References

Dag Haug. 2025. *UD_Latin-PROIEL*. Universal Dependencies (UD), r2.17.

Federica Iurescia and Federica Gamba and Flavio Massimiliano Cecchini and Francesco Mambrini and Giovanni Moretti and Marco Passarotti and Paolo Ruffolo. 2025. *UD_Latin-CIRCSE*. Universal Dependencies (UD), r2.17.

Rachele Sprugnoli and Giovanni Moretti and Marco Passarotti and Daniela Corbetta and Andrea Peverelli. 2020. *CIRCSE/Latin_Sentiment_Lexicons: First release*. Zenodo.