Harmonizing Divergent Lemmatization and Part-of-Speech Tagging Practices for Latin Participles through the LiLa Knowledge Base

Marco Passarotti and Federica Iurescia and Paolo Ruffolo
Università Cattolica del Sacro Cuore
CIRCSE Research Centre
Largo Gemelli, 1, 20123 Milan, Italy
{marco.passarotti,federica.iurescia,paolo.ruffolo}@unicatt.it

Abstract

This paper addresses the challenge of divergent lemmatization and part-of-speech (PoS) tagging practices for Latin participles in annotated corpora. We propose a solution through the LiLa Knowledge Base, a Linked Open Data framework designed to unify lexical and textual data for Latin. Using lemmas as the point of connection between distributed textual and lexical resources, LiLa introduces hypolemmas — secondary citation forms belonging to a word's inflectional paradigm — as a means of reconciling divergent annotations for participles. Rather than advocating a single uniform annotation scheme, LiLa preserves each resource's native guidelines while ensuring that users can retrieve and analyze participial data seamlessly. Via empirical assessments of multiple Latin corpora, we show how the LiLa's integration of lemmas and hypolemmas enables consistent retrieval of participle forms regardless of whether they are categorized as verbal or adjectival.

1 Introduction

Lemmatization and part-of-speech (PoS) tagging constitute fundamental steps in many natural language processing (NLP) workflows, including information retrieval, machine translation, and sentiment analysis (Manning and Schutze, 1999; Jurafsky and Martin, 2025). Lemmatization is the process of reducing a word to its canonical form (or lemma), while PoS tagging entails assigning discrete grammatical categories (e.g., Verb, Noun, Adjective) to tokens in a text. Together, these tasks provide a structured linguistic representation that enables downstream algorithms to handle lexical variation systematically.

Despite the apparent straightforwardness of these tasks, significant variability arises when moving across different annotation schemes and corpora. One source of variability is the choice of annotation guidelines for morphological categories such as participles. In some corpora, participles – morphologically derived verb forms that can function as adjectives (e.g., *broken window*), nouns (e.g., *the breaking of the law*), or as parts of periphrastic verb tenses (e.g., *has broken*) – are consistently lemmatized under the corresponding verb root (e.g., *break*) (see, for Latin, Busa (1974–1980)). Other corpora treat such forms as belonging to the adjective category when they occur in attributive or predicative positions, lemmatizing them separately (e.g., *broken*) (see, for English, Marcus et al. (1993). These divergent lemmatization practices stem from different theoretical perspectives on morphological and syntactic categories, as well as from the practical goals of corpus designers.

A similar issue affects PoS tagging decisions. For instance, the Penn Treebank guidelines (Marcus et al., 1993) tend to annotate verb-derived adjectives such as *broken* or *burnt* as adjectives (with tag: JJ) when used attributively (*broken glass*, *burnt toast*), whereas the Universal Dependencies framework (De Marneffe et al., 2021) may tag these forms as VERB with the accompanying feature for participles (`VerbForm=Part`), or as ADJ depending on their syntactic function.

These differences can significantly impact the consistency of corpora used in training NLP systems. Models trained on one annotation scheme may struggle to generalize effectively to data labeled under a different scheme (Atwell et al., 2000). In the context of lemmatization, inconsistent treatment of participles can complicate tasks such as vocabulary alignment and cross-lingual transfer (McDonald et al., 2011). Moreover, variations in lemmatization and PoS tagging guidelines impede the comparability of results across distinct corpora, thereby influencing empirical linguistic research.

Such annotation discrepancies underscore the need for clear and consistent guidelines in lemmatization and PoS tagging. Nevertheless, accomplishing this task is not straightforward. Even within the same language, deciding whether a participial form should be considered purely verbal or adjectival can depend on its syntactic position, degree of lexicalization, and the morphological tradition followed by linguists or corpus designers (Aronoff and Fudeman, 2022). In highly inflected languages, such as Czech, or Latin, these decisions become even more complex because participial forms often carry additional morphological information related to gender, number, and case. The ongoing development of universal annotation frameworks like Universal Dependencies seeks to mitigate some of these inconsistencies by promoting cross-linguistic standards (De Marneffe et al., 2021). However, adapting such frameworks to diverse linguistic phenomena remains a non-trivial undertaking, and the tension between theoretical adequacy and practical utility persists.

Addressing these challenges demands the development and adoption of more harmonized annotation frameworks, to integrate heterogeneous resources while preserving their unique annotation guidelines. In this paper, the divergent criteria employed for lemmatization and PoS tagging of participles in multiple Latin corpora are empirically examined in a few corpora and a solution to harmonize the divergent annotation practices is proposed.

After presenting some issues of divergent lemmatization and PoS tagging in Latin corpora (Section 2), the paper introduces the corpora under consideration as part of the LiLa Knowledge Base of interoperable resources for Latin (Section 3). By exploiting the interoperability among the corpora facilitated by their publication in LiLa, an empirical assessment is conducted to determine the extent of divergence in lemmatization and PoS tagging of participles across the corpora under investigation (Section 4). Section 5 demonstrates how the modeling based upon an extensive collection of Latin lemmas employed by LiLa enables the harmonization of diverse annotation practices for participles without enforcing a single, uniform approach. Finally, Section 6 concludes the paper, sketching the future work.

2 Lemmatization and PoS Tagging in Latin Corpora

Latin, as a highly inflected language, presents numerous challenges for the design and implementation of lemmatization and PoS tagging schemes in annotated corpora. Available Latin resources often diverge in how they treat morphological categories, leading to inconsistencies and reduced interoperability across corpora. A primary source of variation lies in the criteria for determining both the lemma and the PoS of morphologically complex forms.

Like for many other languages, one notable point of discrepancy in Latin corpora is the treatment of participles. Depending on the corpus or annotation scheme, participles may be categorized as adjectives or verbal forms.

In certain corpora, like for instance the *Index Thomisticus* corpus (Busa, 1974–1980) and Treebank (Passarotti et al., 2019), participles are mostly lemmatized under their verbal dictionary entry (e.g., *laudo* for any participial forms of ‘to praise’), reflecting the view that participles are primarily verbal derivatives.[1]

Conversely, other resources, including the *Opera Latina* corpus by LASLA (Denooz, 2004) and the large repository *Corpus Corporum*[2] treat participles as distinct lemmas when they exhibit syntactic properties characteristic of adjectives, thereby assigning them an independent lemma (e.g., *laudatus* - perfect participle of *laudo* - as a standalone entry when functioning attributively). Nonetheless, the boundary between verbal and adjectival functions often remains subtle.

These differing conventions can yield inconsistent lexical representations and hamper comparative analyses across datasets.

3 The LiLa Knowledge Base

LiLa (Linking Latin) is a Linked Open Data (LOD) Knowledge Base (KB) developed to promote interoperability across a broad spectrum of textual and lexical resources for Latin (Passarotti et al., 2020).[3] Its conceptual model revolves around two primary components:

1. the Lemma Bank,[4] a collection of approximately 200,000 Latin lemmas (canonical citation forms of lexical items) published as LOD

[1] The *Index Thomisticus* corpus lemmatizes participles always under the verb and never under the adjective. Only a limited set of fully lexicalized nominalized participles are lemmatized under the noun, like *aduentus* ‘arrival’. Instead, the *Index Thomisticus* Treebank includes a few participle forms lemmatized under the adjective, mostly when technical terms of Thomas Aquinas’s philosophy are concerned, like *efficiens* ‘efficient’, lit. ‘executing, accomplishing’.

[2] https://mlat.uzh.ch/home

[3] http://lila-erc.eu

[4] http://lila-erc.eu/data/id/lemma/LemmaBank

and originating from the LEMLAT 3.0 morphological analyzer (Passarotti et al., 2017);

2. a set of linguistic resources for Latin published as LOD and interconnected through the Lemma Bank, including corpora, lexica, and dictionaries.[5]

As new resources are integrated, the Lemma Bank is continually expanded, while resources link back to the Lemma Bank by connecting their lexical entries in lexical resources and individual word occurrences (tokens) in textual resources to the corresponding lemma in the LiLa Lemma Bank.

The LiLa KB leverages several established ontologies to represent the (meta)data of interlinked linguistic resources. Chief among these are POWLA for corpus data (Chiarcos, 2012), OLiA for linguistic annotation (Chiarcos and Sukhareva, 2015), and Ontolex-Lemon for lexical data (McCrae et al., 2017). In addition, LiLa employs its own ontology[6] to model lemmas in the Lemma Bank as instances of the class `lila:Lemma`,[7] defined as a subclass of `ontolex:Form`.[8] The class `lila:Lemma` has a specific subclass `lila:Hypolemma`,[9] whose instances are citation forms that belong to a word's regular inflectional paradigm but receive a different PoS tag or degree of comparison than their 'most canonical' lemma, including participles, gerundives, deadjectival adverbs, and comparatives (see Section 5).

For lexical resources, each lexical entry, modeled using the class `ontolex:LexicalEntry`,[10] is connected to its corresponding lemma in the Lemma Bank through the property `ontolex:canonicalForm`.[11] With respect to textual resources, tokens are represented as instances of the class `Terminal`[12] in the POWLA ontology and linked to their corresponding lemma in the Lemma Bank via the property `lila:hasLemma`.[13]

Among the textual resources currently interlinked in the LiLa KB are those examined in this study, selected for their manually verified lemmatization and PoS tagging. Specifically, they include:

- the corpus *Opera Latina* by LASLA, which collects approximately 1.7M tokens from Classical Latin texts (Fantoli et al., 2024);[14]

- the *Index Thomisticus* Treebank (ITTB) (Passarotti et al., 2019), which features the entire text of Thomas Aquinas' *Summa contra Gentiles* for a total of more than 375K tokens enhanced with syntactic annotation according to two styles (Mambrini et al., 2022):[15] the Universal Dependencies one and another resembling that of the analytical layer of the Prague Dependency Treebank (Bamman et al., 2008);

- the UDante treebank, which includes the Latin texts of Dante Alighieri annotated according to the Universal Dependencies style (55K) (Passarotti et al., 2021);[16]

- the CIRCSE Latin Library,[17] a collection of a few Classical and Medieval Latin texts for a total of more than 900K tokens, namely: *Pharsalia* (approx. 67K tokens)[18] by Lucan, the autobiography *Vita Caroli* of the emperor of the Holy Roman Empire Charles IV (18K) (Gamba et al., 2024),[19] *Epistulae ex Ponto* (25K)[20] and *Tristia* (28K)[21] by Ovid (Alagni et al., 2024), *Confessiones* (92K),[22] *De Trinitate* (131K)[23] and *De Civitate Dei* (330K)[24] by Augustine;

[5] The full list of resources currently interlinked in LiLa is available at https://lila-erc.eu/data-page/.

[6] http://lila-erc.eu/ontologies/lila/

[7] http://lila-erc.eu/ontologies/lila/Lemma

[8] http://www.w3.org/ns/lemon/ontolex#Form

[9] http://lila-erc.eu/ontologies/lila/Hypolemma

[10] http://www.w3.org/ns/lemon/ontolex#LexicalEntry

[11] http://www.w3.org/ns/lemon/ontolex#canonicalForm

[12] http://purl.org/powla/powla.owl#Terminal

[13] http://lila-erc.eu/ontologies/lila/hasLemma

[14] http://lila-erc.eu/data/corpora/Lasla/id/corpus

[15] http://lila-erc.eu/data/corpora/ITTB/id/corpus

[16] http://lila-erc.eu/data/corpora/UDante/id/corpus

[17] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus

[18] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Pharsalia

[19] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Vita%20Caroli

[20] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/P.%20Ovidii%20Epistulae%20ex%20Ponto

[21] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/P.%20Ovidii%20Tristia

[22] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Confessiones

[23] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/De%20Trinitate

[24] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/De%20Ciuitate%20Dei

- the corpus CLaSSES, a digital resource which gathers non-literary Latin texts (inscriptions, writing tablets, letters) of different periods and provinces of the Roman Empire (47K) (De Felice et al., 2023);[25]

- chapter VII of *Liber Abbaci*, a historic treaty on arithmetic written in 1202 by Leonardo Fibonacci (30K) (Grotto et al., 2021).[26]

4 Assessing Divergences through LiLa

To investigate lemmatization divergences among the six corpora under examination, we begin by selecting relevant tokens using LiLa[27] — namely, those linked via the property `lila:hasLemma` to a lemma in the Lemma Bank with PoS = VERB or to a hypolemma with PoS = ADJ.[28] We then perform minimal preprocessing, removing tokens that are linked to an ADJ hypolemma but are not participles, specifically gerundives (hypolemmas ending in .**ndus*, e.g., *laudandus* 'to be praised'), and comparatives (hypolemmas ending in .*-*or*), e.g., *citerior* 'further' (see Section 5). Conversely, we retain tokens lemmatized as participles, regardless of their grade or PoS features. For instance, we include comparative and superlative forms of both present and perfect participles (e.g., *promptiores* 'the more attentive (ones)',[29] *abstrusior* 'more recondite',[30] *diligentissimo* '(to) the most attentive (one)',[31] *desideratissima* 'the most desired'),[32] and adverbs derived from participles (e.g., *affluenter* 'abundantly',[33] or *fortunate* 'fortunately').[34]

Next, tokens are normalized by lowercasing, removing diacritics, and replacing *j* with *i*, and *v* with *u*. We also remove enclitics by leveraging the lemmatization available in LiLa; for instance, any token listing *que* 'and'[35] among its lemmas has the enclitic -*que* removed.

From these preprocessed items, two lists of normalized types are compiled: (i) types linked to a lemma with PoS = VERB, and (ii) types linked to a hypolemma with PoS = ADJ. Types linked to a VERB lemma require further preprocessing, as they may include verb forms that are not participles. To filter out these non-participial forms, these types are processed with the LEMLAT morphological analyzer for Latin (Passarotti et al., 2017). Only forms recognized as participles are retained, and any remaining homographs (e.g., *amatis*, which can be either a perfect participle form or the first-person plural present active indicative of *amo* 'to love') are resolved through manual verification.

For each type, we record the total number of tokens across the six corpora and the distribution within each corpus.

These lists are compared to identify shared types, representing participles that exhibit divergent lemmatization strategies in the corpora. An illustrative example is *abundans*, the present participle of the first-conjugation verb *abundo* 'to overflow', which appears under the hypolemma *abundans* (ADJ) in nine occurrences from the *Opera Latina* corpus, and under the lemma *abundo* (VERB) in one occurrence from *Opera Latina*, one from the UDante Treebank, and one from the CIRCSE Latin Library.

Types linked to an ADJ hypolemma that do not appear in the VERB-linked type list are participles consistently associated with a participle hypolemma across all corpora. Conversely, types linked to a VERB lemma that do not appear in the ADJ-linked type list are participles always lemmatized with a verbal lemma.

As an initial overview of the data, Table 1 reports the number of participle tokens (both overall and per corpus) associated with a VERB lemma or an ADJ hypolemma. In all corpora, the majority of participle tokens are lemmatized under the VERB lemma, although the relative proportion of ADJ lemmas varies — from approximately 15:1 in the CIRCSE Latin Library to about 3:1 in the ITTB. Looking at the total of participle tokens lemmatized as VERB versus those as ADJ, the proportion is 5:1 (128,325 vs 26,162). However, this figure may be misleading because the presence of a few participle tokens with exceptionally high frequencies

[25] `http://lila-erc.eu/data/corpora/CLaSSES/id/corpus`

[26] `http://lila-erc.eu/data/corpora/CorpusFibonacci/id/corpus`

[27] See the SPARQL queries (1) and (2) in the Appendix.

[28] The LiLa Lemma Bank uses the Universal PoS tagset (Petrov et al., 2011).

[29] Lemmatized under *promptus* (`http://lila-erc.eu/data/id/hypolemma/35758`) in the *Liber Abbaci*.

[30] Lemmatized under *abstrudo* (`http://lila-erc.eu/data/id/lemma/87036`) in the CIRCSE Latin Library.

[31] Lemmatized under *diligens* (`http://lila-erc.eu/data/id/hypolemma/12447`) in the *Opera Latina* corpus.

[32] Lemmatized under *desidero* (`http://lila-erc.eu/data/id/lemma/98900`) in CLaSSES.

[33] Lemmatized under *affluo* (`http://lila-erc.eu/data/id/lemma/88030`) in the ITTB.

[34] Lemmatized under *fortunatus* (`http://lila-erc.eu/data/id/hypolemma/17176`) in UDante.

[35] `http://lila-erc.eu/data/id/lemma/131416`

	TOTAL	LASLA	ITTB	UDante	CIRCSE	CLaSSES	Fibonacci
VERB	128,325	79,086	15,888	1,564	30,975	667	145
ADJ	26,162	17,603	5,715	425	2,236	168	15

Table 1: Number of participle tokens by PoS assignment.

can skew the interpretation of the results.

To provide a more nuanced perspective, Table 2 presents a type-based distribution of lemmatization of participles by PoS. In particular, it lists the total number of participle types and tokens consistently assigned to the same PoS (either ADJ or VERB) across all corpora, as well as those that are sometimes lemmatized as VERB and sometimes as an ADJ hypolemma. The number of hapax forms is also reported.

Focusing on types, Table 2 confirms that most participles are consistently lemmatized as VERB in the corpora, but it additionally reveals a sizable number of types (and tokens) with inconsistent PoS assignment. Among the 22,851 total types, 2,202 exhibit inconsistent PoS, corresponding to 41,173 tokens. It should be noted that many types that are consistently assigned to a given PoS (either VERB or ADJ) are hapax forms, which necessarily excludes them from the inconsistent VERB/ADJ category because at least two tokens are required for a type to show inconsistent assignment.

For the participle types t that fall under the category VERB/ADJ in Table 2, we calculate the entropy of PoS assignment:

$$H(t) = -log_2(p_V(t)) - log_2(p_A(t))$$

where:

$$p_V(t) = \frac{f_V(t)}{f_V(t) + f_A(t)}$$

$$p_A(t) = \frac{f_A(t)}{f_V(t) + f_A(t)}$$

$f_V(t)$ and $f_A(t)$ are the number of tokens lemmatized as VERB or ADJ respectively for the type t. We estimate an overall *index of homogeneity* as the average of $H(t)$. $H(t)$ is normalized with values in the range of the interval [0,1], where $H(t) = 1$ is maximum entropy, i.e., 50% VERB and 50% ADJ, and $H(t) = 0$ is minimum entropy, i.e., 100% VERB and 0% ADJ, or 0% VERB and 100% ADJ.[36]

Using the values reported in Table 2, the average entropy of PoS assignment to participle tokens in the examined corpora is $H(t) = 0.76$. This moderately high value indicates that, for tokens whose types belong to the VERB/ADJ category, no single PoS assignment clearly predominates. Specifically, these VERB/ADJ types account for 23,136 tokens labeled as VERB and 18,037 tokens labeled as ADJ.

Having established the overall extent of inconsistent PoS assignment for participle types across the investigated corpora, Tables 3 and 4 present the distribution of participle types, tokens and hapax per corpus according to (in)consistent PoS assignment. These tables illustrate the degree of (in)consistency in participle PoS assignment within each individual corpus.

An examination of the data in Tables 3 and 4 indicates that no Latin corpus under consideration exhibits completely consistent PoS assignment for participle forms. Apart from the Fibonacci corpus — which, due to its limited size, exerts minimal influence on the overall findings — ITTB and CIRCSE yield the smallest proportions of participle types that are invariably assigned the ADJ category. The proportion of participle types that fall within the VERB/ADJ category varies among corpora: it is approximately 2% in ITTB, 4% in CIRCSE and 8% in LASLA. Table 5 provides the average entropy, $H(t)$, of PoS assignment for participle tokens in each corpus. Consistent with the proportions described above, the ITTB and CIRCSE corpora exhibit the lowest average entropy values, indicating the lowest degree of uncertainty in PoS assignment for participles.

This variability in PoS assignment (and by extension, lemmatization) for participles is unsurprising, given the inherently hybrid nature of participles, which can function as both nominal and verbal forms. The Universal Dependencies documentation about the `VerbForm` feature (i.e., form of verb or deverbative)[37] states that “some verb forms in some languages actually form a gray zone between

[36] Since the word types considered are those whose tokens show different PoS assignment, maximum and minimum entropy is never found.

[37] `https://universaldependencies.org/u/feat/VerbForm.html`

Category	No. Types [No. Hapax]	No. Tokens
VERB only	18,623 [13,497]	105,189
ADJ only	2,026 [1,320]	8,125
VERB/ADJ	2,202 [0]	41,173
TOTAL	22,851 [14,799]	154,487
VERB/ADJ (VERB)		23,136
VERB/ADJ (ADJ)		18,037

Table 2: Number of participle types [hapax] and tokens by (in)consistency of PoS assignment.

Category	CLaSSES	LASLA	CIRCSE
VERB only	343 (660) [267]	14,853 (69,160) [7,166]	8,412 (27,433) [4,716]
ADJ only	87 (161) [63]	2,207 (9,272) [1,123]	392 (1,317) [256]
VERB/ADJ	4 (14) [0]	1,472 (18,257) [0]	346 (4,461) [0]
VERB/ADJ (VERB)	(7)	(9,926)	(3,542)
VERB/ADJ (ADJ)	(7)	(8,331)	(919)

Table 3: Number of participle types (tokens) [hapax] by (in)consistency of PoS assignment per corpus. First set.

verbs and other parts of speech (nouns, adjectives and adverbs). For instance, participles may be either classified as verbs or as adjectives, depending on language and context”.[38]

As shown by the data presented in the preceding tables, the presence of such a gray zone in PoS assignment considerably complicates information retrieval from annotated corpora, as different lemmas and PoS tags must be queried to capture all forms within a verb's inflectional paradigm. A potential solution would be to enforce highly stringent annotation guidelines. For instance, one might mandate that all participles be assigned exclusively the verbal lemma and VERB PoS, irrespective of their syntactic function. In practice, however, no corpus under investigation adopts such an approach, as demonstrated, because it conflicts with the fact that PoS labels tend to reflect the function of a word in discourse — that is, its contextual rather than purely lexical or morphological properties. As an illustrative example, consider the type *confusa* ‘mingled’, a perfect participle form of the third conjugation verb *confundo* ‘to mingle’, which exhibits an entropy value of $H(\textit{confusa}) = 0.99$. This value is derived from the following distribution: out of 43 total tokens, 20 are assigned PoS ADJ (1 in CIRCSE, 19 in LASLA), whereas 23 are assigned PoS VERB (1 in ITTB, 10 in CIRCSE, and 12 in LASLA).

To address the challenges of PoS assignment for participles in Latin corpora, the LiLa KB has developed a strategy that harmonizes the various criteria followed by these corpora without introducing a new annotation framework. Although designed for Latin corpora, this solution is language-independent and can be applied to any language for which a LOD collection of lemmas and hypolemmas is made available.

5 Harmonizing Divergences through LiLa

This Section describes the methodology used in the LiLa Knowledge Base to reconcile discrepancies in the annotation of participles, which may be labeled as either adjectives or verbs in different textual resources.

To address this issue, the Lemma Bank makes use of the class `lila:Hypolemma`, a subclass of `lila:Lemma` (see Section 3), to represent citation forms that belong to a word's regular inflectional paradigm but receive a different PoS tag or degree of comparison than their ‘most canonical’ lemma.

Typical examples of hypolemmas include participles and gerundives (assigned PoS ADJ but linked to lemmas with PoS VERB) as well as deadjectival adverbs (assigned PoS ADV but linked to lemmas with PoS ADJ). A limited set of comparative adjectives (e.g., *exterior* from *exter* ‘external’, or *posterior* from *posterus* ‘next’) is also recorded as hypolemmas with PoS ADJ linked to lemmas with the same PoS. These forms are typically treated as canonical citation forms in Latin corpora, rather

[38] For one of the most recent pieces of evidence on the challenges presented by this gray zone, see `https://github.com/UniversalDependencies/docs/issues/1088#issuecomment-2722358950`.

Category	**ITTB**	**UDante**	**Fibonacci**
VERB only	2,506 (15,576) [1,276]	1,086 (1,554) [862]	77 (145) [48]
ADJ only	211 (4,280) [51]	216 (392) [148]	9 (15) [7]
VERB/ADJ	59 (1,747) [0]	7 (43) [0]	0 (0) [0]
VERB/ADJ (VERB)	(312)	(10)	(0)
VERB/ADJ (ADJ)	(1,435)	(33)	(0)

Table 4: Number of participle types (tokens) [hapax] by (in)consistency of PoS assignment per corpus. Second set.

Corpus	**avg H(*t*)**
CLaSSES	0.94
UDante	0.88
LASLA	0.78
CIRCSE	0.76
ITTB	0.7

Table 5: Average entropy of PoS assignment to participles tokens by corpus.

than being lemmatized under their positive-degree forms.

In the Lemma Bank, hypolemmas are connected to their corresponding lemmas via the symmetric properties `lila:hasHypolemma`[39] and `lila:isHypolemma`.[40]

For example, the lemma *armo* 'to furnish with weapons' (VERB)[41] is linked via the properties `lila:hasHypolemma/lila:isHypolemma` to three hypolemmas (ADJ): the participles *armans* (present tense), *armatus* (perfect tense), and *armaturus* (future tense).

In the textual resources examined in this study, there are currently 76 occurrences of the different inflected forms of the perfect participle *armatus* (e.g., *armatas*, *armati*, *armato*) linked to the lemma *armo*, and 265 occurrences linked to the hypolemma *armatus*. The modeling approach employed in LiLa facilitates the reconciliation of these divergent lemmatization practices across multiple corpora by linking the participle forms to the Lemma Bank. Regardless of whether a perfect participle form of *armo* is treated as an adjective (lemma *armatus*) or a verb (lemma *armo*) in individual corpora, its occurrences can be uniformly retrieved and integrated via a SPARQL query that traverses the LiLa knowledge graph. This query identifies tokens from different corpora linked, via the property `lila:hasLemma`, either to a lemma with PoS VERB or to a hypolemma with PoS ADJ, which are in turn connected through the properties `lila:hasHypolemma/lila:isHypolemma`.[42]

Figure 1 provides a graphical representation of how a textual occurrence of the plural accusative feminine form *armatas* is linked to the hypolemma *armatus*, which, in turn, is connected to the lemma *armo*. This arrangement parallels the linking of future and present participles to the same lemma. The token *armatas*[43] is drawn from Vergil's *Georgica*, as indicated in the figure by the link between the token and the Document Layer of this text via the property `powla:hasLayer`.[44]

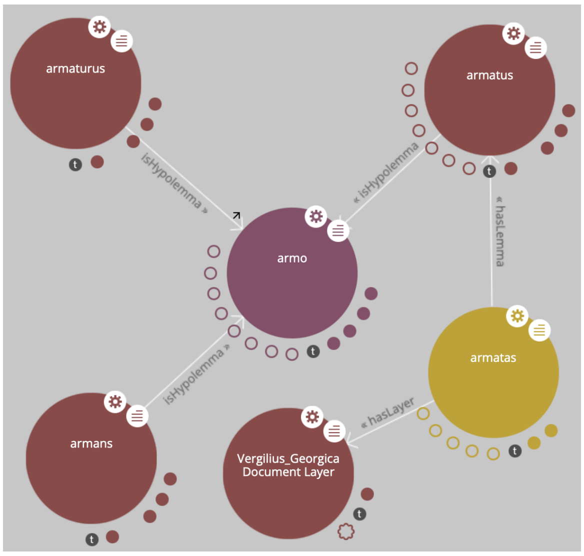

Figure 1: A token (*armatas*) linked to a participle hypolemma (*armatus*) in the LiLa Lemma Bank.

The LiLa Lemma Bank modeling does not include the harmonization of nominalized participle

[39] http://lila-erc.eu/ontologies/lila/hasHypolemma

[40] http://lila-erc.eu/ontologies/lila/isHypolemma

[41] http://lila-erc.eu/data/id/lemma/90036

[42] The SPARQL query (3) reported in the Appendix generalizes this search, retrieving word types by harmonized lemmatization, i.e., regardless of whether a token is lemmatized to a lemma with PoS VERB, or to one of its hypolemmas with PoS ADJ.

[43] http://lila-erc.eu/data/corpora/Lasla/id/corpus/VergiliusGeorgica/Vergilius_Georgica_VerGeor1.BPN_t_0001719

[44] http://purl.org/powla/powla.owl#hasLayer

forms with their corresponding base verbs. Instead, these forms are recorded as separate lemmas, independent of the verbal lemma from which they originate. For example, in the Lemma Bank *intellectus* 'intellect' is listed as a lemma with PoS NOUN, distinct from its base verb *intelligo* 'to understand'. This decision reflects the fact that fully lexicalized nominalizations typically appear as independent entries in dictionaries and, in most cases, receive PoS tag NOUN in corpus annotation.

However, challenges may arise when PoS and lemma assignment in a corpus are determined on a contextual basis rather than a strictly lexical one. Such challenges occur, for instance, when a participle form is used as a noun in a given context, but this nominalization is not sufficiently lexicalized to warrant its own dictionary entry. In these scenarios, the LiLa approach typically links such occurrences with their corresponding participle, recorded as a hypolemma with PoS ADJ, rather than creating a distinct lemma for the nominalization in the Lemma Bank. This is the case of a token like *mendicantem* 'beggar' (present participle of *mendico* 'to go begging') in the following sentence drawn from Plautus' *The Captives*:[45] [...] *ne patri,* [...] *decere uideatur magis, me saturum seruire apud te* [...] *potius quam illi* [...] *mendicantem uiuere* '[...] otherwise it might seem more appropriate to my father that I should be a well-fed slave at your place, [...] rather than [...] live as a beggar back there'.[46]

6 Conclusion

This study has highlighted the challenges posed by divergent lemmatization and PoS tagging schemes for Latin participles in annotated corpora. By demonstrating how these discrepancies can be addressed via the LiLa Knowledge Base, we show that heterogeneous annotation practices — whether stemming from theoretical approaches or from the practical aims of corpus designers — hinder interoperability among resources. Through LiLa's Lemma Bank and the notion of hypolemma, it is possible to unify tokens annotated as either verbal or adjectival participles under a shared representational framework, preserving corpus-specific practices while enabling cross-resource integration.

Rather than enforcing a single "correct" solution, LiLa's graph-based design allows researchers to explore and compare multiple annotation strategies across corpora with minimal manual intervention. In so doing, it promotes data interoperability, and provides a robust platform for linguistic research and NLP applications. Ultimately, this approach underscores the value of LOD methodologies in bridging divergent annotation practices and advancing the broader goal of accessible and reusable linguistic resources.

In future research, we aim to extend our analysis to include nominalized participle forms, which may be documented as independent entries and lemmas in both lexical and textual resources, as well as in the Lemma Bank. After collecting the set of nominalized participle tokens from corpora and corresponding entries from the lexical resources published in LiLa, we will apply the same analytical methodology outlined in this study. This will allow us to assess the degree of consistency in the treatment of nominalized participles across different linguistic resources.

Finally, given the language-independent nature of LiLa's strategy for harmonizing PoS assignment divergences in participles, we hope that other languages will adopt the same architecture. In particular, building and publishing collections of lemmas and hypolemmas as LOD for different languages is crucial for enabling distributed linguistic resources to interoperate in the Semantic Web. A pertinent example is offered by the LiITA Knowledge Base, which has recently implemented a Lemma Bank to enhance LOD-based interoperability across Italian linguistic resources (Litta et al., 2024).[47]

Acknowledgments

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.

References

Aurora Alagni, Francesco Mambrini, and Marco Passarotti. 2024. Lifeless winter without break: Ovid's exile works and the LiLa knowledge base. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 4–12, Pisa, Italy. CEUR Workshop Proceedings.

[45] https://lila-erc.eu/data/corpora/Lasla/id/corpus/PlautusCaptiui/Plautus_Captiui_PlCapt.BPN_t_0002418

[46] Text and translation of this excerpt are drawn from De Melo (2011).

[47] https://www.liita.it

Mark Aronoff and Kirsten Fudeman. 2022. *What is Morphology?* John Wiley & Sons.

ES Atwell, George Demetriou, John Hughes, Amanda Schiffrin, Clive Souter, and Sean Wilcock. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, 24:7–23.

David Bamman, Marco Passarotti, Roberto Busa, Gregory R Crane, et al. 2008. The Annotation Guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank: the Treatment of some specific Syntactic Constructions in Latin. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). May 28-30, 2008, Marrakech, Morocco*, pages 71–76.

Roberto Busa. 1974–1980. *Index Thomisticus*. Frommann-Holzboog, Stuttgart - Bad Cannstatt, Germany.

Christian Chiarcos. 2012. POWLA: Modeling Linguistic Corpora in OWL/DL. In *The Semantic Web: Research and Applications. 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27–31, 2012, Proceedings*, number 7295 in Lecture Notes in Computer Science, pages 225–239, Berlin/Heidelberg, Germany. Springer.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386.

Irene De Felice, Lucia Tamponi, Federica Iurescia, and Marco Passarotti. 2023. Linking the Corpus CLaSSES to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy*, pages 1–7.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational linguistics*, 47(2):255–308.

Wolfgang David Cirilo De Melo. 2011. *Amphitryon ; The Comedy of Asses ; The Pot of Gold ; The Two Bacchises ; The Captives*. Plautus 1, Ed. 2011. Harvard University Press, Cambridge, Mass.

Joseph Denooz. 2004. Opera Latina : une base de données sur internet. *Euphrosyne*, 32:79–88.

Margherita Fantoli, Marco Passarotti, Dominique Longrée, et al. 2024. Lemmas in Dialogue: Linking the LASLA Corpus to the LiLa Knowledge Base. *Recent Trends and Findings in Latin Linguistics: Volume I: Syntax, Semantics and Pragmatics. Volume II: Semantics and Lexicography. Discourse and Dialogue*, pages 297–314.

Federica Gamba, Marco Passarotti, and Paolo Ruffolo. 2024. Publishing the Dictionary of Medieval Latin in the Czech Lands as Linked Data in the LiLa Knowledge Base. *Italian Journal of Computational Linguistics*, 10(1):95–116.

Francesco Grotto, Rachele Sprugnoli, Margherita Fantoli, Maria Simi, Flavio Massimiliano Cecchini, and Marco Passarotti. 2021. The Annotation of Liber Abbaci, a Domain-Specific Latin Resource. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021). Milan, Italy, January 26-28, 2022*, pages 176–183. Accademia University Press.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Eleonora Litta, Marco Passarotti, Paolo Brasolin, Giovanni Moretti, Valerio Basile, Andrea Di Fabio, and Cristina Bosco. 2024. The Lemma Bank of the LiITA Knowledge Base of Interoperable Resources for Italian. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 517–522, Pisa, Italy. CEUR Workshop Proceedings.

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.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

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 587–597, Brno, Czech Republic. Lexical Computing CZ s.r.o.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source Transfer of Delexicalized Dependency Parsers. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 62–72.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, volume 133, pages 24–31, Gothenburg. Linköping University Electronic Press.

Marco Passarotti, Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, et al. 2021. UDante. L'annotazione sintattica dei testi latini di Dante. *Studi Danteschi*, 86:309–338.

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*, LVIII(1):177–212.

Marco Passarotti et al. 2019. The Project of the Index Thomisticus Treebank. *Digital classical philology. Ancient Greek and Latin in the digital revolution*, 10:299–319.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A Universal Part-of-Speech Tagset. *arXiv preprint arXiv:1104.2086*.

A Appendix

(1)

SPARQL query to retrieve types lemmatized to lemmas with PoS VERB (endpoint: https://lila-erc.eu/sparql/):

```
PREFIX rdfs: <http://www.w3.org
   /2000/01/rdf-schema#>
PREFIX lila: <http://lila-erc.eu/
   ontologies/lila/>
PREFIX dc: <http://purl.org/dc/
   elements/1.1/>
PREFIX rdf: <http://www.w3.org
   /1999/02/22-rdf-syntax-ns#>
PREFIX powla: <http://purl.org/
   powla/powla.owl#>

SELECT distinct ?corpora_title ?
   token1_label ?lemma1_label (
   count(?token1) as ?nToken1)
WHERE
{
  VALUES ?corpora {
    <http://lila-erc.eu/data/
       corpora/CIRCSELatinLibrary
       /id/corpus>
    <http://lila-erc.eu/data/
       corpora/UDante/id/corpus>
    <http://lila-erc.eu/data/
       corpora/Lasla/id/corpus>
    <http://lila-erc.eu/data/
       corpora/CorpusFibonacci/id
       /corpus>
    <http://lila-erc.eu/data/
       corpora/CLaSSES/id/corpus>
    <http://lila-erc.eu/data/
       corpora/ITTB/id/corpus>
  }
  ?lemma1 rdf:type lila:Lemma ;
      lila:hasPOS lila:verb ;
      rdfs:label ?lemma1_label .
      ?token1 lila:hasLemma ?
         lemma1 ;
      rdf:type powla:Terminal ;
      powla:hasLayer ?
         DocumentLayer1 ;
      rdfs:label ?token1_label .
  ?DocumentLayer1 powla:
     hasDocument ?Document1 .
  ?Document1 ^powla:
     hasSubDocument ?corpora .
  ?corpora dc:title ?
     corpora_title .
    }
order by ?token1_label
```

(2)

SPARQL query to retrieve types lemmatized to hypolemmas with PoS ADJ (endpoint: https://lila-erc.eu/sparql/):

```
PREFIX rdfs: <http://www.w3.org
   /2000/01/rdf-schema#>
PREFIX lila: <http://lila-erc.eu/
   ontologies/lila/>
PREFIX dc: <http://purl.org/dc/
   elements/1.1/>
PREFIX rdf: <http://www.w3.org
   /1999/02/22-rdf-syntax-ns#>
PREFIX powla: <http://purl.org/
   powla/powla.owl#>

SELECT distinct ?corpora2_title ?
   token2_label ?lemma2_label (
   count(?token2) as ?nToken2)
    WHERE
    {
      VALUES ?corpora2 {
    <http://lila-erc.eu/data/
       corpora/CIRCSELatinLibrary
       /id/corpus>
    <http://lila-erc.eu/data/
       corpora/UDante/id/corpus>
    <http://lila-erc.eu/data/
       corpora/Lasla/id/corpus>
    <http://lila-erc.eu/data/
       corpora/CorpusFibonacci/id
       /corpus>
    <http://lila-erc.eu/data/
       corpora/CLaSSES/id/corpus>
    <http://lila-erc.eu/data/
       corpora/ITTB/id/corpus>
  }
  ?lemma2 rdf:type lila:Hypolemma
      ;
      lila:hasPOS lila:adjective
         ;
      rdfs:label ?lemma2_label .
      ?token2 lila:hasLemma ?
         lemma2 ;
      rdf:type powla:Terminal ;
      powla:hasLayer ?
         DocumentLayer2 ;
      rdfs:label ?token2_label .
  ?DocumentLayer2 powla:
     hasDocument ?Document2 .
  ?Document2 ^powla:
     hasSubDocument ?corpora2 .
  ?corpora2 dc:title ?
     corpora2_title .
}
order by ?token2_label
```

(3)

SPARQL query to retrieve types by harmonized lemmatization, i.e, either lemmatized to a lemma with PoS VERB, or to one of its hypolemmas with PoS ADJ (endpoint: https://lila-erc.eu/sparql/):

```
PREFIX lila: <http://lila-erc.eu/
   ontologies/lila/>
PREFIX rdfs: <http://www.w3.org
   /2000/01/rdf-schema#>
PREFIX dc: <http://purl.org/dc/
   elements/1.1/>
PREFIX rdf: <http://www.w3.org
   /1999/02/22-rdf-syntax-ns#>
PREFIX powla: <http://purl.org/
   powla/powla.owl#>
SELECT ?token_label ?lemma_label
   ?lemma ?pos_label (count(?
   token) as ?nToken) WHERE {
  VALUES ?corpora {
    <http://lila-erc.eu/data/
       corpora/CIRCSELatinLibrary
       /id/corpus>
    <http://lila-erc.eu/data/
       corpora/UDante/id/corpus>
    <http://lila-erc.eu/data/
       corpora/Lasla/id/corpus>
  <http://lila-erc.eu/data/
     corpora/CorpusFibonacci/id
     /corpus>
  <http://lila-erc.eu/data/
     corpora/CLaSSES/id/corpus>
  <http://lila-erc.eu/data/
     corpora/ITTB/id/corpus>
}
{
  ?pos rdf:type lila:Verb;
       rdfs:label ?pos_label.
  ?lemma rdf:type lila:Lemma ;
         lila:hasPOS ?pos ;
         rdfs:label ?
            lemma_label .
  ?token lila:hasLemma ?lemma ;
         rdf:type powla:
            Terminal ;
         powla:hasLayer ?
            DocumentLayer ;
         rdfs:label ?
            token_label .
  ?DocumentLayer powla:
     hasDocument ?Document .
  ?Document ^powla:
     hasSubDocument ?corpora .
}
UNION{
  ?pos rdf:type lila:Adjective;
       rdfs:label ?pos_label.
  ?hypolemma rdf:type lila:
     Hypolemma ;
             lila:hasPOS ?pos ;
             rdfs:label ?
                lemma_label .

  ?hypolemma  lila:isHypolemma
     ?lemma.

  ?token lila:hasLemma ?
     hypolemma ;
         rdf:type powla:
            Terminal ;
         powla:hasLayer ?
            DocumentLayer ;
         rdfs:label ?
            token_label .
  ?DocumentLayer powla:
     hasDocument ?Document .
  ?Document ^powla:
     hasSubDocument ?corpora .
}
} group by ?token_label ?lemma  ?
   lemma_label ?pos_label
```