Umanistica Digitale – ISSN: 2532-8816 – n. 23, 2026: Special Issue
M. Passarotti, E. Litta – *Querying the LiLa Knowledge Base of Interoperable Resources for Latin*
DOI: http://doi.org/10.60923/issn.2532-8816/23506

The LiLa Knowledge Base

Querying Interoperable Latin Resources

Marco Passarotti

Università Cattolica del Sacro Cuore
marco.passarotti@unicatt.it

Eleonora Litta

Università Cattolica del Sacro Cuore
eleonoramaria.Litta@unicatt.it

Abstract

The proliferation of digital linguistic resources for Latin has created significant research opportunities, yet their potential is often constrained by a lack of interoperability. Developed in isolation, many corpora, dictionaries, and lexicons use heterogeneous formats, query languages, annotation criteria and tag sets, hindering integrated data analysis. The *LiLa* (Linking Latin) project addresses this challenge by creating a Knowledge Base of interconnected resources built on Linked Open Data principles. At its core is a lemma-based architecture, where a central Lemma Bank harmonizes divergent lemmatization practices across different sources, enabling seamless data integration. This paper introduces the fundamental structure of the *LiLa Knowledge Base*, which employs standard ontologies like OntoLex-Lemon and models data using RDF. We demonstrate the practical value of this interoperable ecosystem through a series of ready-to-use SPARQL queries. These use cases showcase how researchers can perform complex, cross-resource analyses, such as comparing lexical inventories between Classical and Medieval texts, examining word formation, or tracing semantic concepts across multiple corpora and dictionaries. By linking previously fragmented data, *LiLa* not only streamlines scholarly inquiry but also establishes a new paradigm for creating comprehensive, interconnected digital ecosystems for (not only) historical languages, making Latin a model for linguistic resource interoperability.

Keywords: Latin; Linked Open Data; Interoperability; Linguistic Resources; SPARQL

La proliferazione di risorse linguistiche digitali per il latino ha creato significative opportunità di ricerca, ma il loro potenziale è spesso limitato da una mancanza di interoperabilità. Sviluppati in modo isolato, molti corpora, dizionari e lessici utilizzano formati, linguaggi di query e criteri di

annotazione eterogenei, ostacolando un'analisi integrata dei dati. Il progetto LiLa: Linking Latin *affronta questa sfida attraverso la creazione di una Knowledge Base di risorse interconnesse basata sui principi dei Linked Open Data. Il fulcro di LiLa è un'architettura centrata sul lemma, in cui una 'Lemma Bank' armonizza le diverse pratiche di lemmatizzazione tra le varie fonti, consentendo l'integrazione dei dati. Questo articolo descrive la struttura fondamentale della* Knowledge Base *di* LiLa*, che impiega ontologie standard come OntoLex-Lemon e modella i dati utilizzando RDF. Il valore pratico di questo ecosistema interoperabile è evidenziato attraverso una serie di query SPARQL rese disponibili all'uso. Queste query mostrano come i ricercatori possano eseguire analisi complesse e trasversali tra le risorse, come confrontare inventari lessicali tra testi classici e medievali, esaminare la formazione delle parole o tracciare concetti semantici attraverso più corpora e dizionari. Attuando il collegamento di dati precedentemente sparsi,* LiLa *non solo ottimizza il lavoro di ricerca, ma stabilisce anche un nuovo paradigma per la creazione di ecosistemi digitali completi e interconnessi per le lingue storiche (e non solo), rendendo il latino un modello per l'interoperabilità delle risorse linguistiche.*

Parole chiave: Latino; Linked Open Data; Interoperabilità; Risorse Linguistiche; SPARQL

1. Introduction

The study of the Latin language has, since its earliest stages, been grounded in the analysis of empirical evidence, primarily in the form of ancient texts. These texts (spanning Classical and Medieval literature, as well as legal, religious, and scientific documents) constitute an unparalleled record of Roman culture and civilization and their subsequent developments. Owing to its long history and enduring influence, Latin provided both the structural foundation for the Romance languages and has been for centuries a principal vehicle for the transmission of knowledge throughout the Western world. A comprehensive understanding of its complexity and richness has always depended on the availability of different kinds of instruments enabling the systematic exploration and analysis of these textual witnesses.

Over the centuries, numerous lexical resources for Latin have been developed, including dictionaries, thesauri, and lexicons. These have constituted indispensable tools for the study of Latin language and literature. As early as the Renaissance, the renewed interest in Classical antiquity gave rise to influential lexicographical works such as Ambrogio Calepino's *Dictionarium Latinae Linguae* [2] and Robert Estienne's *Thesaurus Linguae Latinae* [8]. Initially conceived in print form, these works granted scholars structured access to essential information on Latin vocabulary, grammar, and usage, establishing authoritative reference points for generations of Latinists.

The advent of digital technologies has brought about a profound transformation in Latin studies. In recent decades, extensive digital collections of Latin texts have been created, providing unprecedented, immediate access to a broad range of sources. Notable examples include the *Perseus Digital Library*,[1] which offers a substantial corpus of Classical texts accompanied by analytical and translation tools; the *Archivio della Latinità Italiana del Medioevo* (*ALIM*), a

1 https://www.perseus.tufts.edu.

repository of Medieval Latin texts,[2] and the *Monumenta Germaniae Historica* (*MGH*), a critical collection of Medieval Latin sources of major importance for historical and philological research.[3] These digital resources have expanded the scope of inquiry, enabling detailed textual analyses and the identification of interconnections that would have been difficult to detect through traditional means.

Despite the breadth of available digital materials, one of the principal challenges in Latin studies remains the fragmentation of resources. Most of these have been developed independently, employing heterogeneous formats and standards, which complicates data integration and hinders their full exploitation. This lack of interoperability limits the possibility of forming a coherent and comprehensive empirically based representation of Latin language, literature, and culture. In response to this issue, the *LiLa: Linking Latin* project was launched in 2018 with the aim of creating an infrastructure to interlink and enable interaction among Latin linguistic resources by adopting principles and technologies developed for the Semantic Web.[4]

The *LiLa* project is founded on the principle that a comprehensive account of Latin can only be achieved by making the greatest possible quantity of empirical data interoperable. The scholarly value of digital resources lies not solely in their availability, but in their capacity to communicate with one another, forming an integrated data ecosystem that can support research more effectively. To this end, *LiLa* adopts the Linked Open Data paradigm, which facilitates the standardized publication and interlinking of data on the Web, thereby ensuring their findability, accessibility, interoperability and reusability ([26]; [1]).

While a substantial body of Latin texts, lexicons, and dictionaries already exists in digital form, the central challenge is not the creation of additional resources but the removal of technical and conceptual barriers to their integration. *LiLa* addresses this by constructing a lemma-centred Knowledge Base that functions as a central hub connecting heterogeneous resources through unique identifiers for linguistic entities such as lemmas, inflected forms, and senses.

The *LiLa Knowledge Base* is characterized by a modular, flexible architecture capable of accommodating diverse linguistic resources and adapting to evolving research requirements. Since 2018, it has progressively integrated an array of materials, including digital dictionaries, annotated corpora, and specialized lexicons, thereby establishing an expanding ecosystem of interconnected data. This interoperability not only streamlines access to information but also enables new forms of linguistic analysis, affording scholars the means to investigate Latin from innovative perspectives.

The present study introduces the core architecture of the *LiLa Knowledge Base* and demonstrates practical applications of the interoperability among the resources published therein, by presenting concrete use cases through a few queries on the resources interconnected.

2 https://alim.unisi.it.

3 https://www.mgh.de.

4 https://lila-erc.eu.

2. The Fundamental Architecture of the *LiLa Knowledge Base*

The *LiLa Knowledge Base* (*KB*) [18] facilitates interoperability among linguistic resources for Latin by employing established ontologies for linguistic data modelling in conjunction with Semantic Web and Linked Open Data (LOD) standards. For the former, the Ontologies of Linguistic Annotation (OLiA) [4] are adopted for the representation of linguistic annotations, OntoLex-Lemon [16] for lexical data, and POWLA [3] for corpus data. For the latter, the Resource Description Framework (RDF) [15] serves as the underlying data model, representing information as subject–predicate–object triples.

The *LiLa KB* adopts a strongly lexically based architecture, in which the lemma functions as the principal interface between heterogeneous resources. Its central component is the 'Lemma Bank',[5] a repository of approximately 200,000 lemmas sourced from the morphological analyser LEMLAT [17] and continuously expanded. In the *LiLa* ontology, a *LiLa:Lemma*[6] is defined as a subclass of *ontolex:Form*,[7] the latter representing the inflected forms of a lexical entry. The lemma, in turn, is linked to a *ontolex:LexicalEntry*[8] via the property *ontolex:canonicalForm*,[9] which designates the form conventionally used to represent that lexical entry in a lexical resource.

To reconcile divergent lemmatization practices across resources, *LiLa* employs three key means. The symmetric property *LiLa:lemmaVariant*[10] associates alternative lemma forms of the same lexical item (e.g., active and deponent verb forms such as *sequo* and *sequor* 'follow'). The property *ontolex:writtenRep*[11] encodes written representations, i.e., orthographic variants (e.g., *conditio* vs. *condicio* 'condition'). Additionally, the subclass *LiLa:Hypolemma*[12] is introduced to represent forms that may be assigned to more than one lemma in lemmatized resources, such as participles, which can be analysed either as part of a verb's inflectional paradigm or as independent lexical entries.

The *LiLa KB* integrates a broad range of interlinked resources, including corpora and lexical databases. The corpora include the *Opera Latina* corpus from LASLA, comprising 130 Classical Latin texts [9], and two dependency treebanks: the *Index Thomisticus* Treebank, containing works by Thomas Aquinas (1225–1274) [12], and the *UDante* Treebank, comprising Medieval Latin works by Dante Alighieri [19]. The lexical resources include the bilingual Latin–English dictionary by Lewis and Short, primarily covering Classical Latin [13], and the *Dictionary of Medieval Latin in the Czech Lands*, which documents Latin vocabulary used in Eastern Europe during the Middle Ages and provides Czech translations [10].

5 http://lila-erc.eu/data/id/lemma/LemmaBank.

6 http://lila-erc.eu/ontologies/lila/Lemma.

7 http://www.w3.org/ns/lemon/ontolex#Form.

8 http://www.w3.org/ns/lemon/ontolex#LexicalEntry.

9 http://www.w3.org/ns/lemon/ontolex#canonicalForm.

10 http://lila-erc.eu/ontologies/lila/lemmaVariant.

11 http://www.w3.org/ns/lemon/ontolex#writtenRep.

12 http://lila-erc.eu/ontologies/lila/Hypolemma.

At present, the *LiLa* 'RDF graph' contains more than 240,000 lexical entries and approximately 12 million word occurrences (tokens), for a total of over 140 million triples. These are accessible via the KB's 'SPARQL access point', which also offers a set of predefined example queries.[13]

3. Querying the *LiLa Knowledge Base*

This section presents a set of queries that can be run on the *LiLa* SPARQL access point. These queries are available in the list of precompiled examples accessible by clicking on the 'Examples' button on the access point's web page.[14] The queries are here organized into categories according to the number and type of linguistic resources involved, starting from queries targeting a single resource (which really do not exploit the interoperability made possible by *LiLa*). The names reported for the subsections are exactly those assigned to the queries in the list available on the *LiLa* 'SPARQL access point', where they can be found and run.

3.1 Queries on Single Resources

3.1.1 "LemmaBank – Affixes-Noun"

This query is run on the 'Lemma Bank'. Although the 'Lemma Bank' is not strictly a lexical resource (as it does not contain lexical entries), but rather a collection of word forms that can be used as canonical forms of citation (lemmas) in various Latin resources, it nonetheless contains highly valuable lexical information — such as the distribution of parts of speech (PoS) and inflectional classes in Latin vocabulary. In particular, the 'Lemma Bank' encodes derivational morphological information, represented through two types of LOD resources [22]: (a) Lexical bases, which group lemmas belonging to the same derivational morphological family,[15] and (b) Affixes,[16] with two subclasses: Prefixes,[17] and Suffixes.[18]

The query calculates the number of times an affix is used in lemmas with PoS NOUN in the 'Lemma Bank'. The query starts from a variable for the lemma (?lemma), which is the subject of the property LiLa:hasPOS, whose object is a NOUN (represented by the named individual LiLa:noun).[19] The lemma must also be the subject of either the property LiLa:hasPrefix or the property LiLa:hasSuffix, whose object is a variable called aff, for affix (?aff). Basically, this part of the query selects in the Lemma Bank those lemmas that are assigned PoS NOUN and are linked to an affix. Then, the query focusses on affixes, stating that the ?aff variable is assigned a label, by making it the subject of the property rdfs:label, with object the variable for the label of the affix (?afflab). Moreover, the query states that ?aff is of a certain type (indefinite in the query

13 https://lila-erc.eu/sparql/.

14 The LiLa API is accessible at https://lila-erc.eu/sparql/lila_knowledge_base/sparql.

15 http://lila-erc.eu/ontologies/lila/Base.

16 http://lila-erc.eu/ontologies/lila/Affix.

17 http://lila-erc.eu/ontologies/lila/Prefix.

18 http://lila-erc.eu/ontologies/lila/Suffix.

19 In the Lemma Bank, each lemma is assigned a PoS tag, following the universal tagset by [24].

and represented by the variable ?type), i.e., it belongs to a certain class, which, for ?aff can be either LiLa:Prefix or LiLa:Suffix.

The output is a table with four columns:

1. The URI of the affix (aff)
2. The label of the affix (afflab)
3. The affix type (Prefix or Suffix)
4. The frequency, i.e., the number of times the affix is used in the Lemma Bank

The table is sorted in descending order of frequency, showing that the suffix *–(t)io(n)* is the most used affix in morphologically derived Latin nouns.

3.1.2 "Latin Wordnet – Lemmas and synsets"

This query produces the list of lemmas (and, for each lemma, its lexical base) associated with a specific synset (or one of its hypernyms) in the version of Latin WordNet published in *LiLa* [14]. The specific synset targeted here is "02593467-v", a verbal synset with gloss "reign; have sovereign power".[20]

The output is a table with four columns:

1. The URI of the lemma
2. The label of the lemma
3. The URI of the lemma's lexical base
4. The URIs of the WordNet synsets linked to the lemma

The query consists of three UNION blocks. The first block finds all the lemmas that directly evoke the given synset, the second block finds lemmas that evoke synsets which are hypernyms (superordinate concepts) of the given one, and the third block expands the result to all verbal lemmas with the same morpho-derivational family of the hypernyms selected by the previous UNION block.

This query effectively reconstructs the morpho-semantic lexical network surrounding a Latin word provided with an entry in the Latin WordNet and shows how *LiLa*'s interoperability potentials can make the bridge between lexical semantics and morphological structure in Latin explicit.

3.2 Queries on Multiple Textual Resources

3.2.1 "LemmaBank and Corpora – Harmonized lemmatization by lemma (VERB) and connected hypolemmas (ADJ)"

This query retrieves word types using a harmonized lemmatization, i.e., regardless of whether, in a textual resource, a token is lemmatized under a lemma with PoS VERB or under one of its

20 WordNet 3.1 version: http://lila-erc.eu/data/lexicalResources/LatinWordNet/id/LexicalConcept/02593467-v.

hypolemmas with PoS ADJ. Hypolemmas are a subclass of *LiLa:Lemma* used to represent citation forms that belong to a word's regular inflectional paradigm but receive a different PoS tag (e.g., participles or gerundives, tagged ADJ, and deadjectival adverbs, tagged ADV) or degree of comparison than the most canonical lemma (e.g., *inferus* – *inferior* 'lower'). The introduction of hypolemmas in the *LiLa* 'Lemma Bank' allows to harmonize differing lemmatization practices across linguistic resources. The case of participles is particularly notable, as they may be lemmatized either under the verbal or the adjectival lemma [21].

By focusing on six corpora published in *LiLa*[21], the query joins (via the UNION pattern) those tokens in the corpora that are linked to a lemma with PoS VERB in the 'Lemma Bank' with those that are linked to one of its hypolemmas with PoS ADJ. The query produces a table with five columns:

1. The label of the word type
2. The label of the lemma or hypolemma
3. The URI of the lemma
4. The PoS of the lemma (VERB) or hypolemma (ADJ)
5. The token frequency associated with the word type

For instance, the results of the query show that the present participle type *volens* 'want' is lemmatized as a verb form in 34 tokens and as an adjective in 16 tokens in the corpora investigated.

This query highlights one of the main added values of the *LiLa KB*, specifically in relation to the way the data of the Lemma Bank have been modelled. Since lemmatization is a necessary condition for linking a textual resource to the Lemma Bank, and thus making it interoperable with the other resources in *LiLa*, the fact that different corpora adopt different lemmatization criteria becomes a significant issue when lemmas are used as connecting elements between the tokens provided by multiple corpora. To address this problem, *LiLa*'s approach is not prescriptive but harmonizing: rather than introducing additional lemmatization guidelines beyond those already adopted in the various corpora, *LiLa* provides a framework that allows the different existing lemmatization criteria to interact without enforcing a new one. By relying on hypolemmas, as well as lemma variants and written representations, it is possible to retrieve all occurrences of the same lexeme, regardless of how they are lemmatized in the corpora. It is worth noting that the Lemma Bank is a dynamic collection of citation forms, which is continuously expanded and updated in a resource-driven fashion, that is, by reflecting (and harmonizing) the citation forms found in the resources progressively integrated into *LiLa*.

21 (1) Opera Latina; (2) Index Thomisticus Treebank, (3) UDante Treebank; (4) CLaSSES, a digital resource which gathers non-literary Latin texts (inscriptions, writing tablets, letters) of different periods and provinces of the Roman Empire [6]; (5) the CIRCSE Latin Library, a collection of a few Classical and Medieval Latin texts for a total of more than 900K tokens; (6) chapter VII of Liber Abbaci, a historic treaty on arithmetic written in 1202 by Leonardo Fibonacci [11].

3.2.2 "UDante and IT-TB – Lemmas of a specific base in UDante and IT-TB"

This query lists those lemmas occurring both in the *UDante* Treebank and in the *Index Thomisticus* Treebank that share a specific lexical base in the Lemma Bank —in the default query, the base for the lemma *loquor* 'speak' is used.[22] The query traverses the graph of the *LiLa KB* by collecting all the tokens in the two textual resources that are linked to a lemma in the 'Lemma Bank' that is assigned the same lexical base of *loquor* (using the property *LiLa*:hasBase).

The output table has two columns:

1. The lemma label
2. The label of the work in which the lemma occurs at least once

For instance, the results of the query inform that one of the lemmas sharing the same lexical base of *loquor* is *eloquium* 'speech', which occurs in two works of Dante Alighieri (*De Monarchia* and *Epistole*) and in *Summa contra Gentiles* by Thomas Aquinas.

3.2.3 "CLaSSES – Lemmas in CLaSSES and Caesar's De bello gallico"

This query compares the lexicon of the corpus *CLaSSES* with that of Caesars' *De bello gallico* taken from the *Opera Latina* corpus. It lists lemmas that occur at least once in both datasets.

The query searches for those tokens from the two sources that are linked to the same lemma in the Lemma Bank (represented by the variable ?lemmaLiLa) via the property LiLa:hasLemma. The table produced in output contains four columns:

1. The lemma URI in the *LiLa* 'Lemma Bank'
2. The lemma label
3. The number of occurrences of the lemma in *CLaSSES*
4. The number of occurrences of the lemma in *De bello gallico*

The table is sorted in descending order of frequency for the occurrences in *De bello gallico*.

This is one of the simplest, yet most useful types of queries enabled by *LiLa*. Before their publication as LOD within *LiLa*, the data and the metalinguistic annotations (including lemmatization) of the two textual resources queried in the present example had to be accessed through separate interfaces, since the two corpora were not interoperable. Now, thanks to *LiLa*, it is sufficient to submit a SPARQL query to the *LiLa* 'graph', which, moreover, is open-ended and not limited to the two resources in question.

3.3 Queries on Multiple Lexical Resources

3.3.1 "Lexicon Bohemorum – Lemmas with natura not found in Lewis & Short"

This query compares two lexical resources. It retrieves lemmas from the *Dictionary of Medieval Latin in the Czech Lands* whose definitions include the word *natura* 'nature', then compares these lemmas with those in the dictionary by Lewis and Short, extracting those that occur only in the

22 http://lila-erc.eu/data/id/lemma/110805.

Bohemian dictionary (by using the pattern MINUS). Finally, for each lemma of this list, it reports its occurrence counts in each *LiLa* corpus.

The output table has four columns:

1. The URI of the lemma
2. The label of the lemma
3. The number of occurrences of the lemma in a *LiLa* corpus
4. The label of the corpus name

For instance, the adjective *accidentalis* 'accidental' is present in the *Dictionary of Medieval Latin in the Czech Lands*, but not in the Lewis and Short one. Indeed, among the textual resources published as LOD in *LiLa*, the lemma occurs only in corpora featuring Medieval Latin works, namely the *Index Thomisticus* Treebank (63 occ.), *UDante* 2, and the *Digital Library of Late-Antique Latin Texts* (DigilibLT) 1 occ.,[23] a repository of 375 secular prose texts written in Latin in Late Antiquity (from the second to the seventh century AD).

This query illustrates one of the guiding assumptions behind the design of the *LiLa* architecture: that linguistic resources, whether lexical or textual, share the common feature of dealing with words. Textual resources collect occurrences of words, while lexical resources provide different types of information about them. The query essentially transforms into a computationally replicable empirical experiment a practice that humanistic scholars have carried out for centuries: consulting (and comparing) dictionaries and searching in texts for the occurrences of a set of words derived from the use (and comparison) of dictionaries. In the specific case of the query presented here, isolating the lemmas of a dictionary of Medieval Latin (which in this case also refers to a real linguistic variety) with respect to those of a dictionary of Classical Latin highlights the lexical features specific to the Medieval variety in contrast to the Classical one: the finding of the occurrences of the "Medieval" lemmas in corpora permitted by *LiLa* makes the whole process of lexical selection and analysis of word usage straightforward.

3.3.2 "Velez's Dictionary – Lemmas with the base dico"

This query intersects a lexical resource with data from the Lemma Bank. The resource is the *Index Totius Artis*, a small 17th-century Latin–Portuguese dictionary [5]. It finds entries whose lemmas are linked in the Lemma Bank to a specific lexical base—here, *dico* 'say'.

The output table has four columns:

1. The entry label from the *Index Totius Artis*
2. The definition of one of the lemma's senses
3. The base label from the 'Lemma Bank'
4. The URI of the lexical sense

23 https://digiliblt.uniupo.it/.

3.4 Queries on Multiple Lexical and Textual Resources

3.4.1 "LatinAffectus – Sentiment in UDante"

This query links a textual resource, the *UDante* Treebank, with *LatinAffectus*, a lexicon containing a polarity value for prior senses of Latin words [25]. It lists lemmas occurring in Dante's Latin texts that have a negative polarity value in LatinAffectus.

The query starts from selecting the lexical entries of *LatinAffectus* (variable ?lexentry), which are modelled as objects of the property lime:entry,[24] with subject the URI of the *LatinAffectus* resource.[25] The entries are linked to their lemma in the Lemma Bank by the property ontolex:canonicalForm and to their prior sense via the property ontolex:sense.[26] Polarity is modelled with the Marl ontology, using the property marl:hasPolarity.[27] Finally, the tokens from UDante linked to the lemmas selected are found.

Figure 1 shows the path of the query with an example: the token *litigia* 'dispute' is taken from one of the *Epistole* contained in the UDante Treebank.[28] The token is linked to the lemma *litigium/litigiom* (two written representations) via the property LiLa:hasLemma. The lemma is connected to the lexical entry *litigiom/litigium* of LatinAffectus via ontolex:canonicalForm.

The output table has three columns:

1. The URI of the lemma
2. The label of the lemma
3. The number of occurrences of the lemma in the *UDante* Treebank

24 http://www.w3.org/ns/lemon/lime#entry.

25 http://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon.

26 http://www.w3.org/ns/lemon/ontolex#sense.

27 https://www.gsi.upm.es/ontologies/marl/.

28 Epistula 1, Paragraphus 7. URI of the token: http://lila-erc.eu/data/corpora/UDante/id/corpus/Epistole/CiteStructure/I/Paragraphus_7/s10t28.

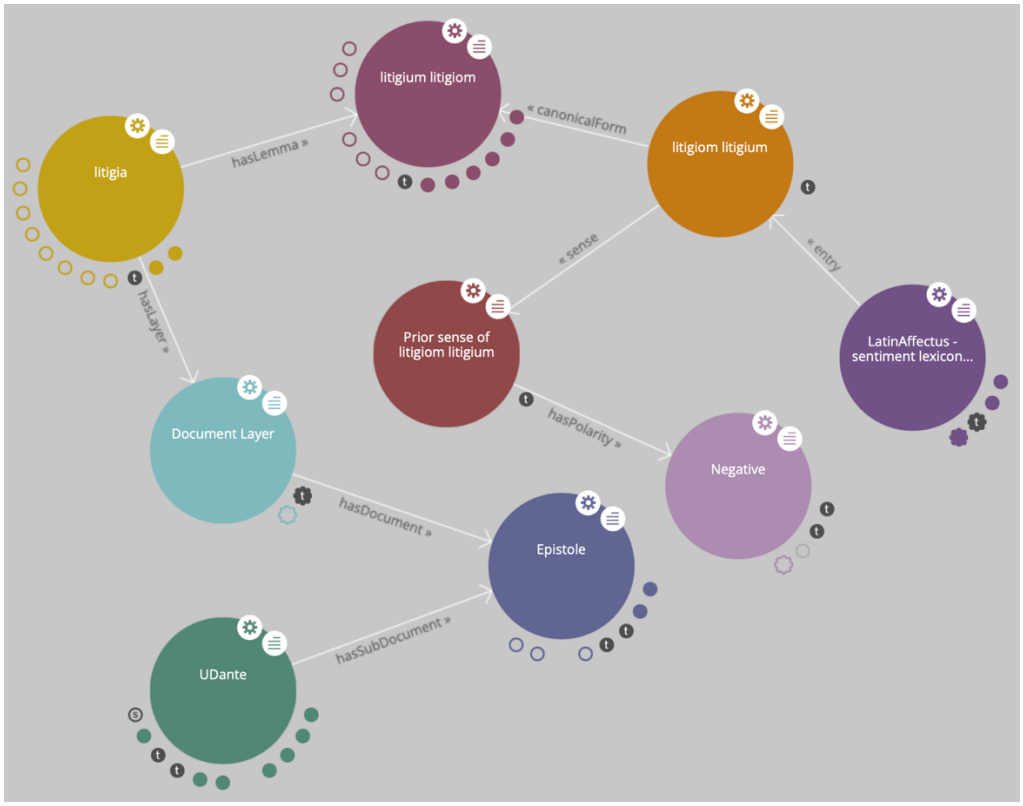

Figure 1 - A token from the UDanteTreebank (*litigia*) linked to a lemma (*litigium*/*litigiom*) in the Lemma Bank with a lexical entry in LatinAffectus.

Beyond the specific technical skills required to write a SPARQL query, the formulation of any query to be executed on an RDF graph demands a solid understanding of the modelling underlying the data to be retrieved. In the case of the query under discussion, Figure 1 illustrates the path that the query follows within the graph to identify the target positive cases. Both the formal nature of the SPARQL language and the need to understand the modelling of the individual resources published in *LiLa* may constitute an obstacle for non-specialist researchers. To address this, the *LiLa project*, in addition to providing a set of ready-to-use SPARQL queries (some of which are presented here), has developed two online services designed to support query composition [20]: (a) the Lemma Bank query interface, which allows users to interrogate the collection of Latin lemmas utilized in *LiLa* to interlink the linguistic resources published therein;[29] and (b) the *LiLa* Interactive Search Platform (LISP), an interactive graphical interface to perform SPARQL queries on the textual resources and a subset of the lexical resources interlinked in the *LiLa* graph.[30]

29 https://lila-erc.eu/query/.

30 https://lila-erc.eu/LiLaLisp/.

3.4.2 "Latin Wordnet – Specific synset in UDante and IT-TB"

This query lists lemmas linked to a lexical entry of the Latin WordNet published in *LiLa* that belong to a specific synset—again the one with gloss 'reign; have sovereign power"—and that occur at least once in both the *UDante* Treebank and the *Index Thomisticus* Treebank. It includes also the lemmas belonging to one of the hypernyms of the synset selected (via the property wn:hypermyn) and those sharing the same lexical base.[31]

The table has three columns:

1. The URI of the lemma
2. The label of the lemma
3. The work label in which the lemma occurs

Figure 2 illustrates an example of two lemmas (*regno* 'reign' and *rego* 'rule') that 1) in Latin WordNet are associated with the relevant synset or with one of its hypernyms (in the Figure, with the synset glossed as 'exercise authority over; as of nations'),[32] and 2) occur in the UDante Treebank and/or in the *Index Thomisticus* Treebank. In particular, Figure 2 shows the token *regnaverunt*,[33] which occurs in the text of *Summa contra Gentiles*, a work of Thomas Aquinas featured by the *Index Thomisticus* Treebank.

In summary, this query retrieves all verbal lemmas morpho-derivationally or semantically linked to the Latin WordNet synset describing the concept "to reign" and identifies the documents in the *Index Thomisticus* Treebank and *UDante* corpora where those lemmas appear.

31 https://globalwordnet.github.io/schemas/wn#hypernym.

32 http://lila-erc.eu/data/lexicalResources/LatinWordNet/id/LexicalConcept/02592711-v.

33 http://lila-erc.eu/data/corpora/ITTB/id/token/005.SCG*LB4.CP-8++3.N.16.29-5.32-1W11.

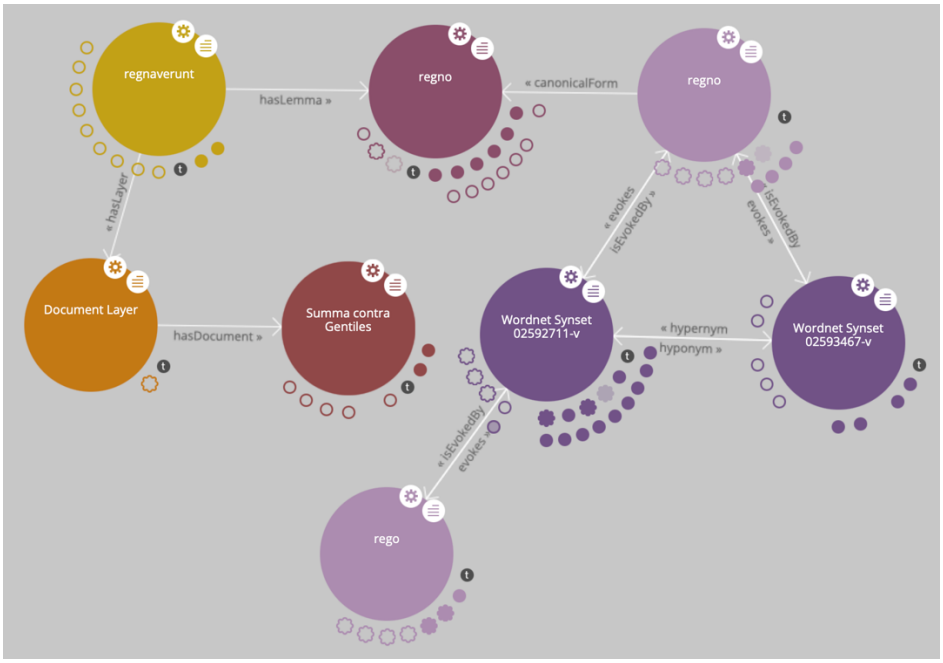

Figure 2 - A token from the Index Thomisticus Treebank (*regnaverunt*) linked to a lemma (*regno*) in the Lemma Bank with a lexical entry in the Latin WordNet.

3.4.3 "Lexicala – Occurrences of words defined with enlever in 5 corpora"

This query links five textual resources (*CIRCSE Latin Library*, *CLaSSES*, *Opera Latina*, *Fibonacci*, *UDante*) with a Latin–French Dictionary developed by Lexicala by K Dictionaries [17]. The query finds lemmas linked to the dictionary's entries whose definition contains the French word *enlever* 'remove' and that are formed with either the prefix *de-* or *a(b)-** (information taken from the Lemma Bank). The output table has three columns:

1. The label of the token
2. The number of occurrences of the token in one of the five corpora
3. The label of the corpus name

The query is structurally simple: it extracts from five corpora all the tokens linked to a lemma in the Lemma Bank that is connected, via the property LiLa:hasPrefix, to one of the two prefixes in question. Despite its simplicity, the query demonstrates the potential of publishing multiple resources as LOD within *LiLa*. In fact, not only were the five corpora originally non-interoperable, but none of them includes annotation about word formation. Their publication in *LiLa* enables interaction among them and enriches the query with lexical information that is otherwise natively unavailable, exploiting at best the contribution of each resource.

3.4.4 "PrinParLat – Lemmas inflected like laedo in LASLA"

'PrinParLat' is a lexical resource providing the principal parts, i.e., a minimal set of a word's forms from which all other forms in its inflectional paradigm can be deduced, of Latin verbs

[23], organized into fine-grained inflectional classes. This query searches the *Opera Latina* corpus by LASLA for lemmas belonging to the class of *laedo* 'hurt', whose principal parts are *laedere*, *laesi*, and *laesum*.

The output table has two columns:

1. The URI of the lemma
2. The label of the lemma

'PrinParLat' comprises 535 inflectional microclasses, a level of granularity in the categorization of Latin inflectional morphology that is not provided by the annotation of any textual corpus, including *Opera Latina*. The interoperability between 'PrinParLat' and the corpora published in *LiLa* makes it possible to observe the distribution of these microclasses in data provided by different textual resources.

3.4.5 "IT-TB – Conflictual objects"

This query combines several of *LiLa*'s major components: syntactic annotation from a treebank (in two annotation styles) with information from two lexical resources, namely Latin WordNet (semantics) and *LatinAffectus* (sentiment/polarity). It searches in the *Index Thomisticus* Treebank for occurrences of common or proper nouns that, in a dependency tree, are linked via the direct object relation—labelled *obj* in the Universal Dependencies schema[34] and *Obj* in the *Prague Dependency Treebank* one[35]—to a governing verb associated in the Latin WordNet with either the synset glossed "cause to seem less serious; play down"[36] or the synset with gloss "fight against or resist strongly".[37] The noun in turn governs an adjective with positive polarity taken from *LatinAffectus*.

The output table has 5 columns:

1. The URI of the noun token
2. The label of the adjective token with positive polarity
3. The label of the noun
4. The label of the verb on which the noun depends
5. The wordnet synset associated with the verb

This query, combining syntactic, lexical-semantic, morphological and affective information, identifies cases like verb *repugno* 'fight against', linked to Latin WordNet synset "fight against or resit strongly", governing a noun like *scripturae* 'scriptures', which in turn is modified by a positive adjective such as *divinae* 'holy'.

34 https://universaldependencies.org/u/dep/obj.html.

35 https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/ch03s02x04.html.

36 https://en-word.net/id/oewn-00866139-v](https://en-word.net/id/oewn-00866139-v.

37 https://en-word.net/id/oewn-01093838-v.

4. Conclusions

The *LiLa KB* has emerged as the paradigmatic use case for demonstrating the effectiveness of applying the LOD principles to linguistic resources, together with a set of ontologies specifically designed for the LOD representation of linguistic (meta)data. As a result, Latin is now the language with the highest degree of interoperability among its linguistic resources, and the *LiLa* architecture stands as a model for the development of comparable KBs for other languages. Recently, a Lemma Bank for Italian has been developed, forming the core component of the *LiITA KB* of interoperable linguistic resources for Italian.

LiLa is a graph that aggregates explicit information (i.e., structured data) concerning different types of linguistic resources for Latin, enriched with multiple layers of metalinguistic annotation (such as lemmatization). At present, we are witnessing the rapid proliferation of Large Language Models (LLMs), which accumulate vast amounts of implicit knowledge—that is, meaningful correlations between data that are not explicitly recorded in forms such as annotations or knowledge graphs.

The field of computational linguistics, and particularly its subdomain of Natural Language Processing (NLP), is currently reflecting on the role that explicit linguistic resources, such as those published in *LiLa*, can play in an environment dominated by LLMs. For languages with a closed textual corpus, such as Latin, the limited availability of training data and the scholarly emphasis on the meticulous study of individual texts (regarded as cultural artefacts rather than mere empirical evidence) will continue to necessitate curated and annotated textual collections, as well as dictionaries, lexica, and thesauri. By contrast, for modern languages, the prevailing trend in NLP is to amass as much textual data as possible—often without direct human inspection, relying instead on superficial data cleaning—and to employ it for the unsupervised training of LLMs.

This trend is progressively distancing NLP from linguistics. On the one hand, it highlights the importance of linguistic resources (and their development) for metalinguistic research grounded in empirical evidence; on the other hand, it carries the risk of detaching the processed object—language itself—from the scholars who understand it most deeply.

The intersection between explicit and implicit knowledge, and thus between linguistics and NLP, should be sought in projects that develop large-scale, interoperable datasets of linguistic resources. In addition to "traditional" linguistic resources, we now also have at disposal explicit knowledge graphs derived from decades of linguistic resource development. The *LiLa KB* is among the best examples of this approach. Explicit knowledge networks can be exploited to enhance and specialize the performance of LLMs through fine-tuning and/or Retrieval-Augmented Generation (RAG). The recently launched COST Action GOBLIN (*Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs*)[38] aims to "provide a large-scale, high-quality, cross-domain and multilingual knowledge graph technology that is free to use, reuse, and redistribute." A multilingual knowledge graph will represent an exceptional resource for realising and evaluating the convergence between metalinguistic competence and unsupervised NLP.

38 https://www.cost.eu/actions/CA23147/.

Specifically for *LiLa*, beyond the KB itself—which provides textual and lexical information in an interoperable format—the collection of ready-to-use SPARQL queries available through the *LiLa* endpoint (several examples of which are presented here) represents an important information source for training models in automatic text-to-SPARQL conversion. Just as it is now possible to perform coding by describing in natural language the intended outcome of a script to an LLM, the composition of SPARQL queries can be facilitated by using an LLM, bypassing what for many scholars from the humanities is the seemingly insurmountable obstacle of fully manual query formulation. However, current LLMs still show considerable limitations in text-to-SPARQL tasks targeting specific graphs such as *LiLa*, as they often lack prior knowledge of the modelling choices (in terms of classes and properties) adopted in those graphs. Training an LLM on a dataset of SPARQL queries specifically designed for a given graph can equip it to perform advanced text-to-SPARQL conversion on that graph.

Finally, it should be emphasized that, by aggregating and making textual and lexical data interoperable, *LiLa* constitutes an extraordinary source of evidence for the construction of lexical entries in new Latin dictionaries and lexica. *LiLa* supports lexicographers by providing a platform to access lemma occurrences across multiple corpora and by making available the lexical properties assigned to that lemma by a wide range of lexical resources. In this sense, *LiLa* does not disrupt the lexicographer's work—or, more broadly, that of Classical scholars—but rather condenses into an open-ended, online, language-independent KB the very material that has historically been central to their research environment, thereby making the process of data identification replicable and reproducible.

References

[1] Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities". *Scientific American* 284 (5): 34–43.

[2] Calepino, Ambrogio. 1502. *Dictionarivm Latinae Lingvae*. Walder.

[3] Chiarcos, Christian. 2012. "POWLA: Modeling Linguistic Corpora in OWL/DL". In *Extended Semantic Web Conference*, 225–239. Springer.

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

[5] Consolin Dezotti, Lucas, Marco Passarotti, and Francesco Mambrini. 2024. "Modelling and Linking an Old Latin–Portuguese Dictionary to the *LiLa Knowledge Base*". In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC–COLING 2024)*, edited by Nicoletta Calzolari et al., 11537–11547. ELRA and ICCL.

[6] De Felice, Irene, 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 (CLiC-it 2023), 1–7. CEUR Workshop Proceedings.

[7] De Paoli, Adriano, Marco Carlo Passarotti, Paolo Ruffolo, Giovanni Moretti, and Ilan Kernerman. 2025. "Linking the Lexical Latin–French Dictionary to the *LiLa Knowledge Base*". In *Proceedings of the 5th Conference on Language, Data and Knowledge*, edited by Mehwish Alam et al, 197–207. UniorPress.

[8] Estienne, Robert. 1532. *Dictionarium, seu Latinae Linguae Thesaurus*. Vol. 2.

[9] Fantoli, Margherita, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. "Linking the LASLA Corpus in the *LiLa Knowledge Base* of Interoperable Linguistic Resources for Latin". In *Proceedings of the Linked Data in Linguistics Workshop* ELRA.

[10] Gamba, Federica, Marco C. Passarotti, and Paolo Ruffolo. 2023. "Linking the Dictionary of Medieval Latin in the Czech Lands to the *LiLa Knowledge Base*". In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics*, 26–34. CEUR Workshop Proceedings.

[11] Grotto, Francesco, 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)*, 176–183. Accademia University Press – Torino.

[12] Mambrini, Francesco, 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,* 4022–4029. ELRA.

[13] Mambrini, Francesco, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2021a. "Linking the Lewis & Short Dictionary to the *LiLa Knowledge Base* of Interoperable Linguistic Resources for Latin". In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, 214–220. Accademia University Press – Torino.

[14] Mambrini, Francesco, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021b. "Interlinking Valency Frames and WordNet Synsets in the *LiLa Knowledge Base* of Linguistic Resources for Latin". In *Further with Knowledge Graphs*, 16–28. Ios Press. https://doi.org/10.3233/SSW210032.

[15] McBride, Brian. 2004. "The Resource Description Framework (RDF) and Its Vocabulary Description Language RDFS". In *Handbook on Ontologies*, 51–65. Springer.

[16] McCrae, John P., Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The OntoLex–Lemon Model: Development and Applications". In *Proceedings of eLex 2017 Conference*, 19–21.

[17] Passarotti, Marco, 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*. Linköping University Electronic Press.

[18] Passarotti, Marco, 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". *Current Approaches in Latin Lemmatization* 58 (1): 177–212.

[19] Passarotti, Marco, Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, et al. 2021. "UDante: L'Annotazione Sintattica dei Testi Latini di Dante". *Studi Danteschi* 86: 309–338.

[20] Passarotti, Marco, Francesco Mambrini, and Giovanni Moretti. 2024. "The Services of the *LiLa Knowledge Base* of Interoperable Linguistic Resources for Latin". In *Proceedings of the 9th Workshop on Linked Data in Linguistics* @LREC–COLING 2024, edited by Christian Chiarcos et al., 309–338. ELRA.

[21] Passarotti, Marco, 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)*, edited by Siyao Peng and Ines Rehbein, 103–114. Association for Computational Linguistics.

[22] Pellegrini, Matteo, Eleonora Litta, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2021. "The Two Approaches to Word Formation in the *LiLa Knowledge Base* of Latin Resources". In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*, 101–109. ATILF.

[23] Pellegrini, Matteo, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2025. "PrinParLat: A Lexicon of Principal Parts of Latin Verbs Linked to the *LiLa Knowledge Base*". *Language Resources and Evaluation*, 1–41. https://doi.org/10.1007/s10579-025-09847-y.

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

[25] Sprugnoli, Rachele, Marco Passarotti, Marinella Testori, and Giovanni Moretti. 2021. "Extending and Using a Sentiment Lexicon for Latin in a Linked Data Framework". In *Proceedings of the Workshops and Tutorials – Language Data and Knowledge 2021 (LDK 2021)*, edited by Sara Carvalho and Renato Rocha Souza, 151–164. CEUR Workshop Proceedings.

[26] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship". *Scientific Data* 3 (1): 1–9.