Linking the *Lexicala Latin-French Dictionary* to the LiLa Knowledge Base

Adriano De Paoli[1], Marco Passarotti[2], Paolo Ruffolo[2], Giovanni Moretti[2], Ilan Kernerman[3]

[1]Università degli Studi di Siena, Italy
[2]CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy
[3]Lexicala by K Dictionaries, Nitsane Oz, Israel
Correspondence: a.depaoli2@student.unisi.it

Abstract

This paper presents the integration of the *Lexicala Latin–French Dictionary* into the LiLa Knowledge Base of linguistic resources for Latin made interoperable through their publication as Linked Open Data. The entries of the dictionary are linked to the large collection of Latin lemmas of LiLa (Lemma Bank), enabling interaction with the other resources published therein. The paper details the data modelling process, the linking methodology, and a couple of practical use cases, showing how interlinking resources via LOD can support advancement in (multilingual) linguistic research.

1 Introduction

Over the past two decades, numerous linguistic resources have been developed for a wide range of languages. In particular, resources for Latin have expanded substantially, resulting in the creation of many annotated corpora, such as treebanks (including five published under the Universal Dependencies initiative; see de Marneffe et al., 2021), as well as additional textual and lexical resources of both born–digital and non–digital origins.

Among the many resources for Latin, Father Busa's *Index Tomisticus* (initiated in 1949) was pioneering in the field (Busa, 1974-1980), comprising 11 million words from the *opera omnia* of Thomas Aquinas. Another noteworthy contribution is the textual corpus developed by the LASLA Laboratory at the University of Liège,[1] which includes 130 Classical Latin texts (Fantoli et al., 2024) by major authors such as Caesar, Cicero, Horatius, and Ovid, totaling over 1.7 million words (Denooz, 2004).

A fundamental limitation of most linguistic resources for Latin (and, in fact, for many languages) is their isolation from each other, functioning as 'silos' that impede data interaction. Establishing interoperability among distributed linguistic resources is currently one of the primary goals in computational linguistics. This objective is now more attainable thanks to the extensive work carried out by the research community devoted to Linguistic Linked Open Data (LLOD).[2] One particularly significant initiative in this area was *Nexus Linguarum*,[3] a COST Action concluded in 2024, whose main aim was to «promote synergies across Europe between linguists, computer scientists, terminologists, and other stakeholders in industry and society, in order to investigate and extend the area of linguistic data science».[4]

For Latin specifically, since 2018 the LiLa project[5] has pursued the goal of making the numerous available linguistic resources mutually interoperable by following the principles of the Linked Open Data (LOD) paradigm (Berners-Lee et al., 2001). This approach ensures that their (meta)data adhere to the FAIR principles.[6] To achieve this, LiLa has developed a LOD Knowledge Base (KB), whose core is a large collection of lemmas linking tokens in textual resources to entries in lexical resources. Through this KB, federated queries can be executed on the interconnected resources via the SPARQL query language.[7]

LiLa's efforts to integrate diverse linguistic resources position Latin at the forefront of LLOD initiatives. Such efforts are particularly crucial for historical languages, which lack native speakers and newly produced texts, necessitating sustained reliance on available data. Moreover, Latin data are

[1] https://www.lasla.uliege.be/cms/c_8508894/fr/lasla

[2] https://linguistic-lod.org/

[3] https://nexuslinguarum.eu/

[4] https://nexuslinguarum.eu/the-action/the-action-objetives/

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

[6] https://www.go-fair.org/fair-principles/

[7] https://www.w3.org/TR/rdf-sparql-query/

dispersed across numerous resources, reflecting its extensive diachronic (spanning over two millennia) and diatopic (across Europe) ranges. Latin also remains widely taught worldwide, resulting in a large number of bilingual dictionaries.[8] Enhancing the interoperability of these dictionaries will benefit not only research on Latin texts and language but also the broader LLOD community, as Latin can function as a bridge to other languages, thereby expanding multilingual interoperability through LOD.

Several bilingual dictionaries have already been incorporated into the LiLa KB, including the *Lewis and Short Dictionary* for English[9] (Mambrini et al., 2022), Velez's *Index Totius Artis* for Portuguese[10] (Dezotti et al., 2024), and the *Latinitatis medii aevi lexicon Bohemorum* (*Dictionary of Medieval Latin in the Czech Lands*) for Czech[11] (Gamba et al., 2024).

This paper details the modeling and linking of the first bilingual Latin–French dictionary — *Lexicala Latin–French Dictionary* (LLFD) — to the LiLa Knowledge Base. Section 2 provides an overview of LiLa's architecture. Section 3 introduces the lexical resource, outlining its key features and structure. Section 4 explains how the dictionary's (meta)data were modeled and integrated into the KB. Section 5 then presents two query examples demonstrating the interoperability of this dictionary with other linked resources within LiLa. Finally, Section 6 offers concluding remarks and outlines directions for future work.

2 The LiLa Knowledge Base

LiLa (Linking Latin) is a large KB of more than 30 Latin resources interlinked on the Web by fitting the principles of the LOD paradigm.[12] The core of the Lila KB consists of a large collection of more than 130,000 Latin lexical items for a total of approximately 215,000 lemmas (the so–called *Lemma Bank*), to which the entries from lexical resources and the tokens from textual resources are linked by using a vocabulary of (meta)data description based upon some of the most widely adopted ontologies in LLOD, as Ontolex[13] for lexical resources, NIF,[14] ConLL–RDF (Chiarcos and Fäth, 2017) and Powla (Chiarcos, 2012) for corpus annotation, OLiA[15] for linguistic annotation, DCMT[16] and LIME[17] (Fiorelli et al., 2015) for metadata.

The decision to create a Lemma Bank as the pivot component of LiLa was aimed at finding a «good balance between feasibility and granularity» while interlinking the resources (Passarotti et al., 2020). Within the LiLa–specific ontology,[18] the class `lila:Lemma`[19] — a subclass of `ontolex:Form`[20] — is defined as «a Form that is linked to a LexicalEntry via the property 'canonical form'» of Ontolex[21] (Passarotti et al., 2020). Following this structural choice it is possible to link all the lexical resources compiled using the Ontolex formalism to LiLa: each lemma can be used as a connection point among the different resources stored in LiLa, ensuring interaction and interoperability.

As far as textual resources are concerned, occurrences of words in texts (tokens) are modelled as instances of the class `Terminal`[22] in the ontology Powla. Tokens are linked to their corresponding lemma in the Lemma Bank of LiLa by the property `lila:hasLemma`.[23]

3 Lexicala Latin–French Dictionary

LLFD[24] is a bilingual dictionary aimed at French–speaking learners of Latin at a beginner or intermediate level, developed by K Dictionaries. The company creates lexical resources for and across different languages, which «enable infinite ways of extracting components and implementing

[8] https://en.wikipedia.org/wiki/Instruction_in_Latin

[9] http://lila-erc.eu/data/lexicalResources/LewisShort/Lexicon

[10] http://lila-erc.eu/data/lexicalResources/LatinPortuguese/Velez/Lexicon

[11] http://lila-erc.eu/data/lexicalResources/LexiconBohemorum/Lexicon

[12] For the full list of the Latin resources currently interlinked in LiLa, see https://lila-erc.eu/data-page/.

[13] https://www.w3.org/2016/05/ontolex/

[14] https://persistence.uni-leipzig.org/nlp2rdf/

[15] https://acoli-repo.github.io/olia/

[16] https://www.dublincore.org

[17] https://art.uniroma2.it/lime/

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

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

[20] http://www.w3.org/ns/lemon/ontolex#lexicalForm

[21] Entries in lexical resources are modeled as instances of the class `ontolex:LexicalEntry` (http://www.w3.org/ns/lemon/ontolex#LexicalEntry). The URI of the property `ontolex:canonicalForm` is the following: http://www.w3.org/ns/lemon/ontolex#canonicalForm.

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

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

[24] The source data for the resource were provided (in JSON–LD format) by K Dictionaries free of charge as part of an agreement with Università Cattolica del Sacro Cuore to publish LLFD as LOD in the LiLa KB.

them for machine translation, spellchecking, text annotation, speech recognition, semantic technologies, knowledge management, language learning, online dictionaries, and more».[25]

The foundations of LLFD were laid by Marjorie Jean during her internship program with K Dictionaries in 2006, as part of her Master degree in Lexicography and Natural Language Processing at the University of Lille 3, with the lexicographic consultation of Pierre Corbin and Ilan Kernerman. Following graduation, Jean was hired by K Dictionaries to compile the full dictionary, which she finally co-edited with Chantal Guglielmi, and which was published in print in 2014 by Assimil as *Dictionnaire Assimil Kernerman Latin–Français* (Jean and Guglielmi, 2014). LLFD is part of K Dictionaries' *Global series* «of multi–layer cross–lingual lexical datasets»,[26] a project started in 2005 when K Dictionaries teamed up with Assimil, a prominent French publisher of foreign language learning materials, focusing on the creation of a series of bilingual learner's dictionaries for French speakers.

Today, this series includes more than two dozen languages: the resources for each language are developed independently, covering its main semantic, syntactic and grammatical aspects, while the underlying framework and technical infrastructure is the same for all languages. The monolingual layer of each language can be used on its own or as a core component for adding translations in different languages to create bilingual pairs in a multilingual network.

As far as LLFD is concerned, the resource contains more than 12,000 entries and 2,000 multiword expressions, which are enriched with 21,000 examples of usage and 38,000 translations. The headwords are selected from Latin of the Classical era, especially from the period ranging from the 1st century BC to the 1st century AD. The authors most represented in the examples are among the best known of the Latin literature, particularly those who lived from the end of the Republican period to the first century of the Empire, including Cicero, Caesar, Sallust, Livy and Seneca for prose, and Virgil, Horace and Ovid for poetry.

In 2014, K Dictionaries began to experiment with linguistic linked data (Klimek and Brümmer, 2015; Bosque-Gil et al., 2016), and the *Global series* was utilized in the development of the Ontolex-Lemon lexicography module *lexicog* (Bosque-Gil et al., 2019).[27]

4 Modelling and Linking the Dictionary

4.1 Modelling the Data

The process of modelling LLFD focused on representing the lexicological and lexicographic content contained in its entries. To describe both types of information, we used classes and properties taken from the Ontolex–Lemon and *lexicog*.

Ontolex–Lemon was used to describe the lexicological part of the entries, while *lexicog* was adopted to represent the lexicographic content, following what has been done in other cases, like, for instance, Mambrini et al., 2022 and Dezotti et al., 2024.

The `ontolex:LexicalEntry` class was used to model single lexical entries in the dictionary. Each instance of this class must be linked to at least one instance of the class `ontolex:Form`, possibly its lemma, via the property `ontolex:canonicalForm`. The total number of lexical entries in LLFD is 12,003.

Regarding the lexicographic contents of the dictionary, the class for lexicographic entries `lexicog:Entry` was used to describe «the structural element that represents a lexicographic article or record as it is arranged in a source lexicographic resource».[28] A `lexicog:Entry` includes one or more `lexicog:LexicographicComponent`, defined as «a structural element that represents the (sub–)structures of lexicographic articles providing information about entries, senses or sub–entries».[29] A `lexicog:LexicographicComponent` links to one or more instances of the class `ontolex:LexicalEntry` or `ontolex:LexicalSense`[30] via the property `lexicog:describes`, which «relates a lexicographic component to an element that represents the actual information provided by that component in the lexicographic resource».[31]

As for the senses of individual lexical entries conveyed by the definitions provided by the dictionary, these are represented as instances of the

[25] https://lexicala.com/k-dictionaries/

[26] https://lexicala.com/dictionaries/

[27] The LLFD data was later converted to the *lexicog* module along with the other *Global series* resources.

[28] http://www.w3.org/ns/lemon/lexicog#Entry

[29] http://www.w3.org/ns/lemon/lexicog#LexicographicComponent

[30] http://www.w3.org/ns/lemon/ontolex#LexicalSense

[31] http://www.w3.org/ns/lemon/lexicog#describes

class ontolex:LexicalSense. Following Ontolex–Lemon, they are linked to the corresponding ontolex:LexicalEntry via the property ontolex:sense.[32] Each sense is the lexicalization of a more general ontolex:LexicalConcept[33] to which a sense is related by the property ontolex:isLexicalizedSenseOf.[34]

Whenever provided by the dictionary, lexical senses are linked to their usage example(s) by the property lexicog:usageExample.[35] Examples are modeled as instances of the class lexicog:UsageExample.[36]

4.2 Linking to the LiLa Knowledge Base

To link the entries of LLFD to the LiLa KB, the first step involved mapping the Part–of–Speech (PoS) tagset used in the dictionary to the one adopted by the LiLa Lemma Bank.[37] This was a straightforward step, as the PoS tagset of the dictionary is more fine–grained than the one of the Lemma Bank. The citation forms of the dictionary entries were then standardized by replacing *j* with *i* and *v* with *u*, and by removing diacritics, in accordance with the Lemma Bank's convention.

Subsequently, a string–matching procedure was applied to identify correspondences between the lemmas in the Lemma Bank and those in the dictionary. This procedure followed a three–stage approach: first, both the lemma and its associated PoS were matched; second, for all unmatched entries, only the lemma string was considered, irrespective of the PoS; and third, the Levenshtein edit distance was applied to the remaining unmatched entries, yielding candidate links that underwent manual verification.

The matching results were classified into four categories, each corresponding to a distinct type of outcome:

1. *single matches* (1:1): cases in which the initial matching step identifies a unique <lemma, PoS> pair in the Lemma Bank that corresponds to the dictionary entry;

2. *ambiguous matches* (1:N): cases arising in the first matching step where multiple <lemma, PoS> pairs in the Lemma Bank correspond to the dictionary entry;

3. *partial matches*: cases resulting from the second matching step, further divided into: *single partial matches* (1:1p): a single candidate lemma in the Lemma Bank matches the dictionary entry, ignoring PoS; *ambiguous partial matches* (1:Np): multiple candidate lemmas in the Lemma Bank match the dictionary entry, ignoring PoS;

4. *no matches* (1:0): cases in which no candidates from the Lemma Bank match the dictionary entry.

Table 1 provides an overview of the outcome of the matching process. Notably, over 80% of the dictionary entries fall into the 1:1 category. This result is consistent with the figures found while linking the *Lewis and Short Dictionary* for English as well as the *Index Totius Artis* for Portuguese to the LiLa KB.[38] On the contrary, the numbers for the *Latinitatis medii aevi lexicon Bohemorum* for Czech are very different, most likely due to the peculiar variety of Latin represented therein, covering the vocabulary of Medieval Latin as used in the Czech lands since the beginnings of Latin writing in this area (from about 1,000 AD) to 1,500 AD.[39]

The partial matches were examined to assess data quality, recognizing that linking dictionary entries to the Lemma Bank solely on the basis of lemmas may yield incorrect correspondences. Two types of partial matches were inspected manually: (i) single partial matches such as *mille* 'one thousand', which is categorised as an Adjective in the dictionary but as a Numeral in the Lemma Bank,[40] and (ii) ambiguous partial matches such

[32] http://www.w3.org/ns/lemon/ontolex#sense

[33] http://www.w3.org/ns/lemon/ontolex#LexicalConcept

[34] http://www.w3.org/ns/lemon/ontolex#isLexicalizedSenseOf

[35] http://www.w3.org/ns/lemon/lexicog#usageExample

[36] http://www.w3.org/ns/lemon/lexicog#UsageExample

[37] The LiLa Lemma Bank uses the Universal PoS tagset (Petrov et al., 2012) and employs a slightly modified subset of Lemlat's morphological labels (Passarotti et al., 2017) for inflectional categories.

[38] Out of 38,693 entries of the *Lewis and Short Dictionary* linked to LiLa, 31,142 are 1:1 matches (80.5%), 2,998 are 1:N matches (7.7%), and 4,553 are 1:0 matches (11.8%). That the percentage of 1:0 matches is higher for the *Lewis and Short* than for LLFD may be due to the fact that the former was linked to an older version of the Lemma Bank, thus provided with a lower number of lemmas. *Index Totius Artis*: 1:1 = 4,093 (86.7%), 1:N = 368 (7.8%), 1:0 = 262 (5.5%).

[39] *Latinitatis medii aevi lexicon Bohemorum*: 1:1 = 13,838 (55.5%), 1:N = 827 (3.3%), 1:0 = 10,278 (41.2%).

[40] http://lila-erc.eu/data/id/lemma/112335

Match type	NoE	%
Total	**12,003**	**100.0%**
1:1	9,779	81.5%
1:N	764	6.4%
1:0	917	7.6%
Partial matches	543	4.5%
1:1p	438	3.6%
1:Np	105	0.9%

Table 1: Results of the matching process

Match type	NoE	%
Total	**12,003**	**100.0%**
Single matches	10,923	91.0%
Ambiguous matches	823	6.9%
No matches	257	2.1%

Table 2: Matching results after refining

as *capito* 'a man with a big head', for which the PoS assigned in the dictionary is Adjective, while three distinct lemmas were available in the Lemma Bank (`capito_NOUN`, `capito_VERB`, and `capito_PROPN`). In this latter case, excluding the verb narrowed the possibilities to two plausible lemmas; to identify the most appropriate correspondence, meanings were verified in other LiLa–linked resources (specifically, by consulting the *Lewis and Short Dictionary*). As a result, `capito_NOUN`[41] 'one that has a large head, big–headed' was selected over the lemma referring to the Roman cognomen *Capito, –onis*.

Only five instances of incorrect linking were detected overall — three involving single partial matches and two involving ambiguous partial matches. These errors stemmed from the absence of the relevant lemmas in the Lemma Bank, which have since been incorporated.[42]

Additional heuristics were implemented to refine automatic linking in cases of ambiguity. For verbs, the inflected forms provided in the dictionary's source JSON–LD file (modeled as `ontolex:Form` instances) were used to repeat the matching process. For example, this strategy enabled the lexical entry *adgero* (VERB) — which includes the inflected forms *adgero, adgeris, adgessi, adgestum, adgerere* in LLDF — to be correctly linked in the LiLa Lemma Bank to the third–declension verb *adgero, –ere*,[43] 'to bear' rather than to the first-declension verb *adgero, –are* 'to heap up'.[44] For nouns exhibiting ambiguity in gender, inflectional class (e.g., distinguishing second– from fourth–declension forms ending in *–us*), or number (*pluralia tantum*), disambiguation relied on the gender, the genitive form, or the presence of plural indicators in the dictionary's source file. By applying these procedures, 46.7% (357 out of 764) of ambiguous matches were successfully resolved and linked to the appropriate lemmas.

Heuristics were also employed for no matches, to further refine the results and to automatically propose potential linking candidates. Since more than 70% of the unmatched lexical entries were inflected verbal forms (with the dictionary providing no additional information about the verb), the morphological analyzer for Latin Lemlat (Passarotti et al., 2017) was used to derive canonical citation forms from the inflected ones (e.g., *curro* 'to run' from the perfect tense form *cucurri*). This approach enabled approximately 65% of the previously unmatched inflected forms to be linked to the corresponding verbal lemma in the Lemma Bank; the remaining cases underwent manual verification.

This process of data linking refinement led to the numbers shown in Table 2. As can be seen, the increase in percentage in the case of 1:1 matches is remarkable (from 81.5% to 91.0%).

Following the application of heuristics for automatically assigning lemmas to ambiguous and unmatched entries, the next step involved manually disambiguating the remaining ambiguous matches (823 cases).

These ambiguous instances arise when morphological features (e.g., PoS, inflectional category, or gender) alone are insufficient to distinguish between multiple candidate lemmas in the Lemma Bank as the canonical form of the corresponding dictionary entry. In such circumstances, the semantic information assigned to lemmas in the Lemma Bank is utilized. This information is drawn from a set of lexical resources already interconnected within the LiLa KB, including five bilingual dictio-

[41] http://lila-erc.eu/data/id/lemma/92703

[42] The erroneously linked entries were *laevum* (ADV) 'to the left', *fines* (NOUN) 'borders', and *he* (INTJ) 'ah!' (a variant of the existing lemma *ha*) for single partial matches; and *eventus* (ADJ for the participle form of the verb *evenio*) 'happened', *olor* (NOUN) 'odor' (an alternative form of *odor, odoris*) for ambiguous partial matches.

[43] http://lila-erc.eu/data/id/lemma/88073

[44] http://lila-erc.eu/data/id/lemma/88074

naries,[45] two etymological dictionaries,[46] and the Latin WordNet.[47]

Additional lexicographic sources used in constructing the Lemma Bank (but not published as LOD) provide another means of disambiguation, specifically through entries collated from two dictionaries[48] for Classical and Late Latin, as well as one glossary for Medieval Latin.[49]

Finally, the 'lexical bases' recorded in the Lemma Bank represent a further strategy for disambiguating homographic lemmas. A lexical base is a class in the LiLa ontology[50] whose instances denote a 'morpheme of a word that is neither a prefix nor a suffix' (Passarotti et al., 2020). In the Lemma Bank, lexical bases link lemmas sharing the same lexical ancestor (i.e., belonging to the same derivational family) via the `lila:hasBase` property.[51]

An illustrative example is the third–declension verb *occido*, which may correspond to a lemma derived from the lexical base *caedo* (*occīdo*,[52] 'to strike down') or another derived from *cado* (*occĭdo*,[53] 'to fall down'). In this case, the dictionary provides information on the lexical base, thus facilitating disambiguation in conjunction with the Lemma Bank data. Conversely, for the noun `colum`, which can denote either 'a straining vessel, a colander'[54] or the 'colon'[55] (part of the human body/member of a verse), the correct match was identified by consulting the definition in the *Lewis and Short Dictionary*.

With regard to the 257 cases in which no matching lemma was identified, manual intervention concentrated on the dictionary entries that remained unmatched after heuristic procedures were applied. Entries corresponding to lemmas not yet present in the Lemma Bank were introduced as new lemmas; for example, the noun *deducta* 'amount deducted from an inheritance and abandoned by the heir', a term attested in Cicero's works.

By contrast, entries that were merely graphical variants of already existing lemmas were incorporated as new written representations of those lemmas using the `ontolex:writtenRep`[56] property (e.g., *Olympus* 'Mount Olympus' vs. *Olimpus*).

In cases where the difference from an existing lemma pertained to the inflectional category and/or the specific cell of the inflectional paradigm used by the dictionary, a new lemma was occasionally created in the Lemma Bank and linked to the existing lemma via the symmetric property `lila:lemmaVariant`.[57] This approach was adopted, for instance, when the dictionary employed nominative plural forms of nouns as lemmas (e.g., *facultas* 'ability, skill' [singular; from the Lemma Bank] vs. *facultates* 'goods, property' [plural; from the dictionary]) or when the dictionary listed nouns with alternate inflectional classes (e.g., *Atrida* 'Atreides, son of Atreus', featuring a Latin first–declension ending, vs. *Atrides*, displaying a Greek ending explicitly labeled in the Lemma Bank tagset).

Multiword expressions were linked to the semantically more specific lemma; for example, *Esquilinus mons* 'Mount Esquiline' was connected to the lemma *esquilinus* 'pertaining to the Esquiline'. This is due to the strict constraint by Ontolex-Lemon that a `ontolex:lexicalEntry` must be linked to no more than one canonical form (i.e., lemma).

In some instances, entries from LLFD were deliberately not linked to any lemma for two principal reasons. First, certain entries describe derivational morphemes (e.g., `ex-prefix` or `dis-par`), which in the Lemma Bank are classified as `Affixes`[58] (Passarotti et al., 2020), and thus are not associated with any lexical entry in the KB's interconnected

[45] The *Lewis and Short Dictionary* (Latin–English) and three Latin–Portuguese dictionaries — Velez, Fonseca (http://lila-erc.eu/data/lexicalResources/LatinPortuguese/Fonseca/Lexicon), and Cardoso (http://lila-erc.eu/data/lexicalResources/LatinPortuguese/Cardoso/Lexicon) — as well as the Latin–Czech *Latinitatis medii aevi lexicon Bohemorum* (*Dictionary of Medieval Latin in the Czech Lands*).

[46] The *Lexicon Der Indogermanischen Verben* (http://lila-erc.eu/data/lexicalResources/LIV/Lexicon) (Boano et al., 2023) and the *Etymological Dictionary of Latin and the other Italic Languages* (http://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon) (Mambrini and Passarotti, 2020).

[47] http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon (Franzini et al., 2019).

[48] *Oxford Latin Dictionary* (Glare, 2012) and *Ausführliches lateinisch–deutsches Handwörterbuch* (Georges, 1998).

[49] *Glossarium mediae et infimae latinitatis* (du Cange et al., 1883–1887).

[50] http://lila-erc.eu/ontologies/lila/Base

[51] http://lila-erc.eu/ontologies/lila/hasBase

[52] http://lila-erc.eu/data/id/lemma/114585

[53] http://lila-erc.eu/data/id/lemma/114586

[54] http://lila-erc.eu/data/id/lemma/94963

[55] http://lila-erc.eu/data/id/lemma/97826

[56] http://www.w3.org/ns/lemon/ontolex#writtenRep

[57] http://lila-erc.eu/ontologies/lila/lemmaVariant

[58] http://lila-erc.eu/lodview/ontologies/lila/Affix

No matches	NoE	%
Total	**257**	**100.0%**
New lemmas	17	6.6%
Written representations	16	6.2%
Lemma variants	40	15.6%
Inflected forms	54	21.0%
Multiword expressions	10	3.9%
Not linked	114	44.4%
Typographical errors	6	2.3%

Table 3: Distribution of no matches entries

resources. Second, some entries do not represent full word forms or lemmas but only partial forms (e.g., *advors–* as an alternative spelling of *advers–*). In Ontolex, Forms are grammatical realizations of words (or of any other class of lexical entries) that possess at least one written representation: in the LiLa ontology, Lemmas are treated as a subclass of `ontolex:Form`, selected as the canonical citation form of a lexical item (Passarotti et al., 2020). Accordingly, partial forms are excluded from linking.

In the event of typographical errors in the source (e.g., *conservarix* instead of *conservatrix* 'preserver, keeper'), the entry was corrected and subsequently linked to the corresponding lemma in the Lemma Bank.

Table 3 provides quantitative insights into the linking process for dictionary entries that initially yielded no matches during automated procedures. Excluding those forms intentionally left unmatched for the aforementioned reasons — representing approximately half of the no matches (44.4%) — the largest share of newly added lemmas in the Lemma Bank consists of inflected forms (21.0%) and lemma variants (15.6%). The prevalence of inflected forms is attributable to the dictionary's inclusion of numerous such variants, reflecting a common practice in similar resources. See, for instance, the lexical entry *faxim* in the Velez Latin–Portuguese Dictionary, which is linked to the canonical form *facio* (Dezotti et al., 2024).

Regarding lemma variants, the majority pertain to nouns whose meanings diverge between singular and plural forms, e.g., *carceres* 'the barrier at the starting point of a racecourse' vs. *carcer* 'prison', which LLFD lists as separate entries.

5 Use Cases

As a result of the linking process described in the previous Section, LLFD has been integrated into the LiLa KB[59] and interconnected with the other linguistic resources available therein. The LOD publication of the dictionary in LiLa enables users to query its data through the LiLa SPARQL endpoint.[60] The following Subsections present two examples of basic SPARQL queries that demonstrate the added value of the integration of the dictionary with various resources in the LiLa KB. These examples illustrate how the interoperability of resources published as LOD facilitates data exploration and enhances empirically–based linguistic research.

5.1 Corpus Occurrences of Lemmas with a Specific Definition in the Dictionary

The SPARQL query presented in this Subsection serves as a useful tool for French–speaking high school students (among others) aiming to enhance their comprehension of the Latin language.

As part of a pre–compiled set of queries available through the LiLa SPARQL endpoint, this query adopts a comparative approach, integrating data from a lexical resource (LLFD) and a selection of Latin texts spanning different historical periods.

The query is structured in three steps:

1. Retrieval of lexical entries with a specific definition.
   The query first identifies those lexical entries in LLFD (using the property `lime:entry`[61]) whose definitions contain the French verb *enlever* 'to remove'. This is achieved by selecting those dictionary entries (individuals of the class `ontolex:LexicalEntry`) that possess at least one sense (class `ontolex:LexicalSense`) which is linked via the property `skos:definition`[62] to a literal value equal to 'enlever';

2. Selection of corresponding lemmas in the Lemma Bank.
   Next, the query retrieves lemmas from the Lemma Bank that are associated with the lexical entries identified in the previous step. This is accomplished by selecting lemmas linked to these dictionary entries through the property `ontolex:canonicalForm`. In particular, the query focuses on lemmas that contain ei-

[59] https://lila-erc.eu/data/lexicalResources/Lexicala/Lexicon/Lexicon

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

[61] http://www.w3.org/ns/lemon/lime#entry

[62] http://www.w3.org/2004/02/skos#definition

ther the prefix *de–** or *a(b)–* 'away from', as specified by the property `lila:hasPrefix`;[63]

3. Identification of lemma tokens in corpora. Finally, the query searches for the tokens of the selected lemmas within five Latin corpora linked to LiLa, using the property `lila:hasLemma`, thus allowing for further comparative analysis of their usage across different texts.

The five corpora concerned are the following:

- the corpus *Opera Latina* by LASLA, which collects approximately 1.7M tokens from Classical Latin texts (Fantoli et al., 2024);[64]
- the UDante treebank, which includes the Latin texts of Dante Alighieri annotated according to the Universal Dependencies style (55K) (Passarotti et al., 2021);[65]
- the CIRCSE Latin Library,[66] a collection of a few Classical and Medieval Latin texts for a total of more than 900K tokens, namely: *Pharsalia* (approx. 67K tokens)[67] by Lucan, the autobiography *Vita Caroli* of the emperor of the Holy Roman Empire Charles IV (18K),[68] *Epistulae ex Ponto* (25K)[69] and *Tristia* (28K)[70] by Ovid, *Confessiones* (92K),[71] *De Trinitate* (131K)[72] and *De Civitate Dei* (330K)[73] by Augustine;
- 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);[74]
- chapter VII of *Liber Abbaci*, a historic treaty on arithmetic written in 1202 by Leonardo Fibonacci (30K) (Grotto et al., 2021).[75]

A total of 470 distinct word types were returned by the query, amounting to 5,634 tokens. A substantial proportion of these word types occurs in the LASLA corpus (393 out of 470), followed by the CIRCSE Latin Library (160). This distribution is likely a consequence of the larger size of these text collections. The most frequently represented lemma is *aufero*[76] 'to take off' (103 types; 1,618 tokens), followed by *detraho*[77] 'to draw off' (72; 672), *abduco*[78] 'to lead one away' (50; 181) and *demo*[79] 'to withdraw' (46; 159).

Figure 1 illustrates a token[80] of the verb *demo* linked to its lemma in the Lemma Bank, which, in turn, is linked to its corresponding lexical entry in LLFD and to one of its senses provided therein, namely a sense encompassing the word *enlever*.

5.2 Dictionary Coverage of the Classical Latin Lexicon

In this Subsection, we present a use case that compares the entries of LLFD with the lexical items in the *Opera Latina* corpus, a set of Classical Latin texts already interlinked in LiLa. The goal is to assess the dictionary's coverage of the Classical Latin lexicon by counting the number of tokens and lemmas in the corpus that lack corresponding lexical entries in the dictionary — i.e., are not assigned a lexical entry there.

To perform this analysis, two queries were formulated. These queries share an identical first step but differ in the second. The general structure of the queries is as follows:[81]

1. Selecting the tokens from the *Opera Latina* corpus.

[63] http://lila-erc.eu/ontologies/lila/hasPrefix
[64] http://lila-erc.eu/data/corpora/Lasla/id/corpus
[65] http://lila-erc.eu/data/corpora/UDante/id/corpus
[66] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus
[67] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Pharsalia
[68] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Vita%20Caroli
[69] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/P.%20Ovidii%20Epistulae%20ex%20Ponto
[70] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/P.%20Ovidii%20Tristia
[71] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Confessiones
[72] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/De%20Trinitate
[73] http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/De%20Ciuitate%20Dei
[74] http://lila-erc.eu/data/corpora/CLaSSES/id/corpus
[75] http://lila-erc.eu/data/corpora/CorpusFibonacci/id/corpus
[76] http://lila-erc.eu/data/id/lemma/90671
[77] http://lila-erc.eu/data/id/lemma/99047
[78] http://lila-erc.eu/data/id/lemma/86867
[79] http://lila-erc.eu/data/id/lemma/98553
[80] Specifically, the token is a present infinitive form *demere* from Seneca's *Ad Lucilium Epistulae Morales*. The property `skos:definition` is not shown in the Figure, due to limitations of the LodLive visualization (`http://lodlive.it`).
[81] The queries can be found among the pre–compiled available at the LiLa SPARQL endpoint.

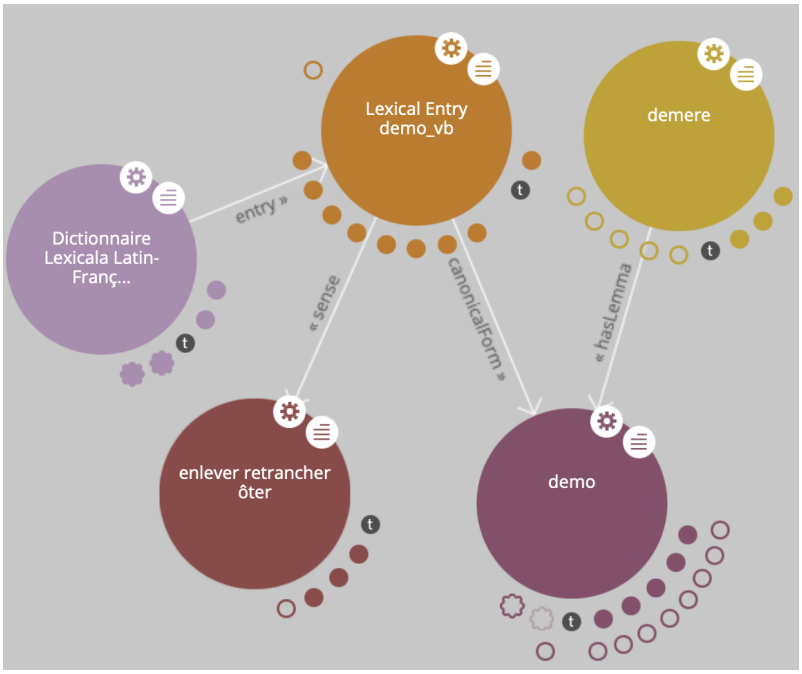

Figure 1: Linking a token from a corpus to a lexical entry of LLFD in the LiLa Knowledge Base.

The query retrieves those tokens (instances of the type `powla:Terminal`) that are linked, via `powla:hasLayer`,[82] to the Document Layer[83] of a `powla:Document`,[84] itself part of the *Opera Latina* corpus (through `powla:hasSubDocument`[85]). Tokens are connected to lemmas in the Lemma Bank via the property `lila:hasLemma`;

2. Excluding tokens and lemmas linked to LLFD entries.
   Two queries use the `MINUS` function to exclude from the current results any tokens in the corpus that are linked to lemmas in the Lemma Bank that, in turn, do not have a corresponding lexical entry in the dictionary (i.e., they are not connected via the property `ontolex:canonicalForm` to an LLFD entry). The first query returns the list of such tokens, while the second query returns their lemmas in the Lemma Bank.

The queries reveal that 127,684 tokens (7.3% of the total 1,745,314 tokens in the *Opera Latina* corpus) are linked to a lemma in the Lemma Bank that lacks a corresponding entry in the LLFD. Furthermore, 15,060 of the 24,200 lemmas in *Opera Latina* (62.2%) are not represented in the dictionary.

These findings indicate that, although the dictionary covers nearly 93% of the textual occurrences (tokens), it captures less than 40% of the distinct lemmas present in the corpus. This result empirically confirms that LLFD incorporates the core vocabulary of Classical Latin — accounting for the majority of tokens — while lemmas unattested in the dictionary predominantly belong to a less frequent or non–Classical stratum of the language. For instance, the verb *admetior*[86] 'to measure out to' occurs in Cato and Curtius Rufus (both non–Classical authors) and only once in Cicero, while the adjective *terreus*[87] 'of earth, earthen' is found exclusively in Vergilius's *Georgica* and Varro's *De re rustica*, both of which employ highly poetic or specialized vocabulary.

6 Conclusions and Future Work

In this paper, we have detailed the integration of *Lexicala Latin–French Dictionary* as Linked Open Data (LOD) within the LiLa Knowledge Base. Thanks to LiLa's architecture and its firm grounding in ontologies and models widely adopted by the LOD community, the dictionary has become fully interoperable with a rich ecosystem of other linguistic resources for Latin. These include tex-

[82] http://purl.org/powla/powla.owl#hasLayer
[83] http://purl.org/powla/powla.owl#DocumentLayer
[84] http://purl.org/powla/powla.owl#Document
[85] http://purl.org/powla/powla.owl#hasSubDocument
[86] http://lila-erc.eu/data/id/lemma/87518
[87] http://lila-erc.eu/data/id/lemma/127922

tual corpora totaling over 12 million words, and a number of lexical resources, like a few bilingual dictionaries, a WordNet, and a derivational morphological lexicon.[88] As a result, French-speaking learners of Latin at beginner and intermediate levels, as well as researchers, can now seamlessly traverse a web of interconnected lexical information, dramatically enhancing the utility and reach of the original dictionary.

Moreover, by bringing together multiple bilingual dictionaries and two etymological resources in LiLa, new avenues for multilingual research and cross–linguistic resource linking emerge. The recent development of a Lemma Bank for Italian in the LiITA Knowledge Base[89] (Litta et al., 2024), following the LiLa model, further demonstrates the potential of this interlinking approach. Envisioning a network of similar Lemma Banks for different languages, all interconnected via bilingual dictionaries, points to a substantial leap forward in harnessing linguistic empirical evidence across diverse resources and languages.

The publication of LLFD in LiLa highlights how interconnected data can enrich linguistic research. By adhering to widely recognized LOD best practices, we have ensured that this dictionary can be integrated and reused alongside other resources for Latin and beyond.

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

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American*, 284(5):29–37.

Valeria Irene Boano, Francesco Mambrini, Marco Passarotti, and Riccardo Ginevra. 2023. Modelling and Publishing the "Lexicon der indogermanischen Verben" as Linked Open Data. In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy*, pages 1–7. CEUR Workshop Proceedings.

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Guadalupe Aguado-de Cea. 2016. Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In *Proceedings of GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop*, pages 65–72.

Julia Bosque-Gil, Dorielle Lonke, Jorge Gracia, and Ilan Kernerman. 2019. Validating the OntoLex-lemon Lexicography Module with K dictionaries' Multilingual Data. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference*, pages 726–746.

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

Christian Chiarcos. 2012. POWLA: Modeling linguistic corpora in OWL/DL. In *The Semantic Web: Research and Applications. ESWC 2012*, volume 7295 of *Lecture Notes in Computer Science*, pages 225–239, Berlin, Heidelberg. Springer.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In *Language, Data, and Knowledge*, pages 74–88, Berlin. Springer.

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.

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

Lucas Consolin Dezotti, 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), ELRA and ICCL, Torino, Italia, 2024*, pages 11537–11547.

Charles du Fresne sieur du Cange, bénédictins de la congrégation de Saint-Maur, d. Pierre Carpentier, Johann Christoph Adelung, G. A. Louis Henschel, Lorenz Diefenbach, and Léopold Favre. 1883–1887. *Glossarium mediae et infimae latinitatis*. Favre, Niort, France.

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.

[88] http://lila-erc.eu/data/lexicalResources/WFL/Lexicon

[89] http://liita.it/data/id/lemma/LemmaBank

Manuel Fiorelli, Armando Stellato, John P. McCrae, Philipp Cimiano, and Maria Teresa Pazienza. 2015. LIME: The Metadata Module for Ontolex. In *The Semantic Web. Latest Advances and New Domains. ESWC 2015*, volume 9088 of *Lecture Notes in Computer Science*, pages 225–239, Cham. Springer.

Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signorini, Viviana Ventura, and Federica Zampieri. 2019. Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), 13-15 November 2019*, pages 1–8. Accademia University Press.

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.

Karl Ernst Georges. 1998. *Ausführliches lateinisch-deutsches Handwörterbuch*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany. Reprint of first edition of 1913–1918, Hannover, Germany: Hahnsche Buchhandlung.

Peter Geoffrey William Glare. 2012. *Oxford Latin Dictionary*, 2nd edition. Oxford University Press, Oxford.

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.

Marjorie Jean and Chantal Guglielmi. 2014. *Dictionnaire Assimil Kernerman Latin-Français*. Assimil, Paris. ISBN: 978270056406464.

Bettina Klimek and Martin Brümmer. 2015. Enhancing lexicography with semantic language databases. *Kernerman Dictionary News*, 23:5–10.

Eleonora Litta, Marco Passarotti, Paolo Brasolin, Giovanni Moretti, Francesco Mambrini, 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 Tenth Italian Conference on Computational Linguistics (CLiC-it 2024). Pisa, Italy, December 4-6, 2024*, pages 1–6. CEUR Workshop Proceedings.

Francesco Mambrini, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2022. 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). Milan, Italy, January 26-28, 2022*, pages 1–7. CEUR Workshop Proceedings.

Francesco Mambrini and Marco Passarotti. 2020. Representing etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 20–28, Marseille, France. European Language Resources Association (ELRA).

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*, 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 (SSL)*, 58(1):177–212.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).