ISSN 2499-4553

IJCoL

Italian Journal
of Computational Linguistics

Rivista Italiana
di Linguistica Computazionale

Volume 10, Number 1
june 2024

aA
ccademia
university
press

editorial board

editorial office

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

AILC
Associazione Italiana di
Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 9791255000983

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_10_1

aAccademia
university
press

Accademia University Press è un marchio registrato di proprietà
di LEXIS Compagnia Editoriale in Torino srl

IJCoL

Volume 10, Number 1
june 2024

CONTENTS

Publishing the Dictionary of Medieval Latin in the Czech Lands as Linked Data in the LiLa Knowledge Base

Federica Gamba[*]
Università Carolina

Marco C. Passarotti [**]
Università Cattolica del Sacro Cuore

Paolo Ruffolo[**]
Università Cattolica del Sacro Cuore

The article presents the process of linking the Dictionary of Medieval Latin in the Czech Lands to the LiLa Knowledge Base, which adopts the Linked Data paradigm to make linguistic resources for Latin interoperable. First, we provide an overview of the architecture of the LiLa Knowledge Base and of the Dictionary; then, we detail the stages of the process of linking the Dictionary to the collection of Latin lemmas that represents the core of LiLa. Once completed the linking process, we demonstrate to what extent the publication of the Dictionary as Linked Data proves beneficial by presenting the linking to LiLa of a new text from the same area and period covered by the Dictionary. In conclusion, a few queries illustrate how interoperability allows for full exploitation of a set of interlinked Latin resources.

1. Introduction

Many digital linguistic resources are nowadays available for Latin, making it a particularly privileged language among the historical ones. However, most often those resources are scattered, with their sparsity representing a substantial hindrance to the full exploitation of the information they contain. To overcome the sparsity of resources, stored in separate silos, the CIRCSE Research Center in Milan, Italy, started the LiLa - Linking Latin project[2] (2018-2023), which built a Knowledge Base to make all existing textual and lexical resources for Latin interoperable by adopting the four principles of the Linked Open Data (LOD) paradigm (Berners-Lee, Hendler, and Lassila 2001): 1) use URIs as names for things; 2) use HTTP URIs so that people can look up those names; 3) when someone looks up a URI, provide useful information; 4) include links to other URIs, so that they can discover more things.[3]

[*] Faculty of Mathematics and Physics - Malostranské náměstí 25, 118 00 Prague, Czechia.
E-mail: `gamba@ufal.mff.cuni.cz`

[**] CIRCSE Research Center - Largo Agostino Gemelli 1, 20123 Milan, Italy.
E-mail: `{marco.passarotti,paolo.ruffolo}@unicatt.it`
This article is an extended version of a paper by the same authors entitled *Linking the Dictionary of Medieval Latin in the Czech Lands to the LiLa Knowledge Base*, published in the Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics (Nov 30 — Dec 02, 2023, Venice, Italy), CEUR Workshop Proceedings.

[2] `https://lila-erc.eu/`.

[3] `https://www.w3.org/wiki/LinkedData`.

The LiLa Knowledge Base has already a wide coverage in terms of resources interlinked. Classical Latin is naturally well-represented, as proved by the *Opera Latina* corpus by LASLA, which includes 130 Classical Latin texts (Fantoli et al. 2022), and by the Lewis and Short dictionary (Lewis and Short 1879), whose primary focus is on Classical Latin. Later stages of Latin are found as well in the Knowledge Base; for instance, the *Index Thomisticus* Treebank (Passarotti 2019) comprises texts by Thomas Aquinas (1225–1274), the UDante treebank (Cecchini et al. 2020) encompasses Medieval Latin works written by Dante Alighieri, and the Computational Historical Semantics corpus (Geelhaar et al. 2020) includes e.g. the *Decretum Gratiani*, a collection of canon law compiled in the XII century CE.

While the LiLa Knowledge Base already extends over a large temporal range, its spatial coverage is not as wide. So far, no resource from the Eastern Europe areas where Latin was spoken has been linked. For this reason, we decided to link to LiLa the *Dictionary of Medieval Latin in the Czech Lands*, a lexical resource that aims at collecting the Latin vocabulary as it emerged in that area during the Middle Ages. The resource encompasses a late variety of Latin (1000-1500 CE), strongly tied to a specific geographical area. These two levels of variability, along the temporal and spatial axes, make it extremely interesting to link such a resource to the Knowledge Base, as we expect it to contribute to enlarge the amount of lemmas stored in the collection of Latin lemmas that represents the core part of the whole architecture of LiLa.

The article is organised as follows. Section 2 introduces the LiLa Knowledge Base. Section 3 describes the *Dictionary*. Section 4 outlines the process of linking the *Dictionary* to LiLa. Section 5 investigates the contribution of the *Dictionary* made interoperable while linking a new text from comparable period and area. Section 6 shows the added value of interoperability of Latin resources in LiLa by presenting three queries on the *Dictionary* interlinked.

2. The LiLa Knowledge Base

The LiLa Knowledge Base (Passarotti et al. 2020) achieves interoperability between linguistic resources for Latin, by adopting a set of ontologies widely used to model linguistic information, as well as Semantic Web and Linked Data standards. Among the former, OLiA is used to model linguistic annotation (Chiarcos and Sukhareva 2015), Ontolex-Lemon for lexical data (Buitelaar et al. 2011; McCrae et al. 2017) and POWLA for corpus data (Chiarcos 2012). As for the latter, the Resource Description Framework (RDF) (Lassila and Swick 1998) is a data model used to describe information in terms of triples, consisting of: (1) a predicate-property that connects (2) a subject (i.e. a resource) with (3) its object (another resource or a literal). Data recorded in the form of RDF triples are queried via the SPARQL query language (Prud'Hommeaux and Seaborne 2008).

The architecture of the LiLa Knowledge Base is highly lexically-based, as it exploits the lemma as the most productive interface between resources and tools. Indeed, its core is the so-called Lemma Bank, a collection of around 200,000 lemmas taken from the database of the morphological analyser LEMLAT (Passarotti et al. 2017) and constantly extended. A `lila:Lemma`[4] is a subclass of `ontolex:Form`[5], whose individuals are the inflected forms of a lexical item. In particular, the lemma is a form that can be linked

[4] `https://lila-erc.eu/lodview/ontologies/lila/Lemma`.

[5] `http://www.w3.org/ns/lemon/ontolex\#Form`.

to a `ontolex:lexicalEntry`[6] via the property `ontolex:canonicalForm`[7], which identifies the form that is canonically used to represent a lexical entry.

To overcome divergent lemmatisation criteria that may possibly be adopted in different resources, LiLa exploits three key properties. The symmetric property `lila:lemmaVariant`[8] connects different forms of the same lexical item that can be used as lemmas for that item, like for verbs with an active and a deponent inflection (e.g., *sequo* and *sequor* 'to follow'). The property `ontolex:writtenRep`[9] registers different spellings or graphical variants of one lemma, like for instance *conditio* and *condicio* 'condition'. For forms that can be reduced to multiple lemmas like participles – that can be considered either part of the verbal inflectional paradigm or as independent lemmas – a special sub-class of `lila:Lemma` called `lila:hypolemma`[10] is defined.

3. The Dictionary of Medieval Latin in the Czech Lands

The *Dictionary of Medieval Latin in the Czech Lands*[11] is a lexical resource developed at the Department of Medieval Lexicography of the Institute of Philosophy of the Czech Academy of Sciences. It aims to collect the vocabulary of Medieval Latin as it was used in the Czech lands from about 1000 CE, when Latin writing began in the area, to 1500 CE. In light of this aim, the *Dictionary* features three types of entries:

- Vocabulary taken from Classical Latin without any semantic change during the Middle Ages. The meaning of words is illustrated only by source citations with a translation. E.g., *labellum* 'small lip'.
- Vocabulary taken from Classical Latin with changes. This type of entry is composed of two parts: first, ancient meanings are listed; then, the + sign introduces Medieval developments (such as syntactical alternations, new phrases, meanings of the word coined in Medieval times). E.g., *falcatus* 'curved' + 'shod'.
- Vocabulary that emerged during the Middle Ages. Such entries are marked with an asterisk (∗). The heading of the entry is followed by information between square brackets [ ] about etymology and references to other dictionaries. E.g., *emicamen* [*emicare*] 'splendour, clarity'.

Moreover, the *Dictionary* relies on a differential method to capture all divergences – at several linguistic layers – of Medieval Latin vocabulary inherited from the ancient era as compared with the Classical norms. Indeed, it records language phenomena not attested in the 8th edition (and later unchanged editions) of Georges' Latin-German Lexicon (Georges and Georges 1913).

The material the *Dictionary* is built upon amounts today to ca. 800,000 excerpt sheets, assembled from various sources of Czech provenance (diplomatical, official, belles-lettres, scientific literature, etc.). What is particularly valuable is that not only edited texts served as a source to build the *Dictionary*, but also several manuscripts

[6] `http://www.w3.org/ns/lemon/ontolex\#LexicalEntry`.

[7] `http://www.w3.org/ns/lemon/ontolex\#canonicalForm`.

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

[9] `http://www.w3.org/ns/lemon/ontolex\#writtenRep`.

[10] `https://lila-erc.eu/lodview/ontologies/lila/Hypolemma`.

[11] The Czech title is *Slovník středověké latiny v českých zemích*; the Latin one *Latinitatis medii aevi lexicon Bohemorum*.

and old prints from Czech and foreign libraries were used. The excerption of sources has been carried out from 1934, when the project of the *Dictionary* started, until the 1970s. In 1977 the first fascicle was published, illustrating editorial principles and lists of sources and abbreviations. Overall, the electronic database (Ctibor and Nývlt 2021) is built upon, and comprises, the three volumes prepared by Silagiová and colleagues (Silagiová et al. 2018, 2019, 1995 to 2016).

So far, letters A-M are covered, for a total of 48,452 entries. 24,943 out of these are full entries (provided with meanings, definitions, grammatical information, examples), whereas 23,509 are references that point to full entries (see 3.1). Fascicle 24, encompassing entries beginning with N, is currently under preparation.

The *Dictionary* is accessible through a dedicated website[12] and can be downloaded from the LINDAT/CLARIAH-CZ research infrastructure[13] as a compressed set of XML files.

3.1 XML Files

We provide a brief overview of the structure of the XML files of the *Dictionary*, as those data are relevant for the process of modeling information and linking the entries to the LiLa Knowledge Base. The lexical entry for the adjective *exquisitus* 'exquisite' (Figure 1) will serve as an example of the XML files of the resource.

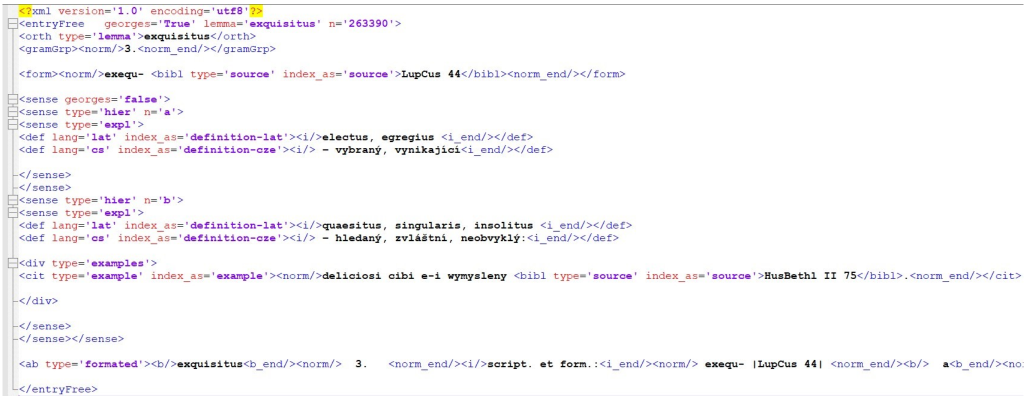
The whole entry is encoded as the value of an `entryFree` element, which contains a single unstructured entry in any kind of lexical resource, such as a dictionary or lexicon. Core information about the entry is provided through attributes: the lemma is given, together with a numerical unique identifier assigned to it; `georges='True'` or `'False'` specifies whether an entry for the same lemma is found or not in Georges' dictionary. Optionally, `hom_nr` distinguishes homographs, and `type='reference'` denotes that the entry is just a reference to a different one; for instance, the dummy entry for *geniculor* 'to bend the knee' is just a reference to its active counterpart *geniculo*, which, in light of that, is the only full entry of the two (with meanings, grammatical information, etc.). Then, in the `orth` element the lemma is stated once again as a value; `orth` includes the attribute `type` either with value `'lemma'`, if it is a full entry, or with value `'ref_all'`, if it is a reference.

Following the lemma, the `gramGrp` element encodes grammatical information about the lexical item, roughly corresponding to its Part of Speech (POS) and (possibly) its inflectional category. In the case of *exquisitus*, the value `<gramGrp> 3. </gramGrp>` indicates that it is an adjective of the first class, i.e. with three distinct endings for the three genders (*exquisitus, -a, -um*, respectively for the forms of masculine, feminine and neuter singular nominative).

The `sense` elements (possibly more than one for a same entry) capture the different meanings of a lexical item. For each `sense`, a definition `def` is provided both in Latin and in Czech, with the Czech one corresponding to a translation of the Latin counterpart. Some examples are listed as well, together with their source. The label *script. et form.* is used to record orthographic and morphological variants (e.g., *exequisitus* for *exquisitus*), while the label *metr.* for metrical ones.

[12] `http://lb.ics.cas.cz`.

[13] `http://hdl.handle.net/11234/1-4792`.



```
<?xml version='1.0' encoding='utf8'?>
<entryFree   georges='True' lemma='exquisitus' n='263390'>
<orth type='lemma'>exquisitus</orth>
<gramGrp><norm/>3.<norm_end/></gramGrp>

<form><norm/>exequ- <bibl type='source' index_as='source'>LupCus 44</bibl><norm_end/></form>

<sense georges='false'>
<sense type='hier' n='a'>
<sense type='expl'>
<def lang='lat' index_as='definition-lat'><i/>electus, egregius <i_end/></def>
<def lang='cs' index_as='definition-cze'><i/> - vybraný, vynikající<i_end/></def>

</sense>
</sense>
<sense type='hier' n='b'>
<sense type='expl'>
<def lang='lat' index_as='definition-lat'><i/>quaesitus, singularis, insolitus <i_end/></def>
<def lang='cs' index_as='definition-cze'><i/> - hledaný, zvláštní, neobvyklý:<i_end/></def>

<div type='examples'>
<cit type='example' index_as='example'><norm/>deliciosi cibi e-i wymysleny <bibl type='source' index_as='source'>HusBethl II 75</bibl>.<norm_end/></cit>

</div>

</sense>
</sense></sense>

<ab type='formated'><b/>exquisitus<b_end/><norm/>  3.   <norm_end/><i/>script. et form.:<i_end/><norm/> exequ- |LupCus 44| <norm_end/><b/>  a<b_end/><no

</entryFree>
```

Figure 1
XML file of the *Dictionary* entry for *exquisitus* 'exquisite'.

4. Linking the Dictionary to LiLa

This Section describes the process of linking the *Dictionary of Medieval Latin in the Czech Lands* to the LiLa Knowledge Base. So far, we have been working only with full entries (i.e., excluding those with `type='reference'`).

As mentioned in Section 2, in LiLa the lemma works as interface between interlinked resources. In light of the pivotal use of lemmas, the core operation at the base of the linking process is to perform a string match between the tuples *(lemma, POS)* in the resource to be linked and the lemmas and their POS in the LiLa Lemma Bank. The goal is to retrieve the correct lemma in the Lemma Bank corresponding to the lemma/POS used in the entry of the *Dictionary*.

The string match results in three possible outcomes: a) only one matching lemma/POS is found in the Lemma Bank; b) more than one matching lemma/POS is found, resulting in an ambiguity due to homography; c) no matching lemma/POS is found, as the couple is not present in the Lemma Bank.

The first outcome is overall straightforward and does not raise particular issues. The second one, i.e. multiple matches found, requires disambiguation to be performed. To this aim, grammatical information about inflectional classes can be exploited, although it does not always guarantee a full resolution of the ambiguity; Subsection 4.1 elaborates on this. The third kind of outcome of the string match, i.e. missing matches, is the most interesting; firstly, because it entails enlarging the Lemma Bank with new canonical forms of citation, and secondly because it allows to reflect about the peculiar aspects of the variety of the Latin vocabulary represented in the *Dictionary*, by focusing on those lexical items provided by the *Dictionary* that result as out-of-vocabulary with respect to the current Lemma Bank of LiLa.

4.1 Aligning Grammatical Information

In order to automatically disambiguate multiple matches, we exploit the grammatical information provided by the *Dictionary* in the `gramGrp` element. However, this information is not encoded in a fully standardised way, thus requiring an alignment to be performed. Indeed, we need to define a set of heuristics to align grammatical categories as they are encoded in the *Dictionary* and the set of tags employed in LiLa, which is based on the Universal POS tagset (Petrov, Das, and McDonald 2012) and expanded with inflectional categories. As an illustration, the word *acus* 'needle' has *-us, f.* as `gramGrp`, i.e. the genitive ending and the gender. From that we can generalise and establish a correspondence between the genitive ending in *-us* together with the gender, as found in the *Dictionary*, and a NOUN with inflectional class `n4`[14] in LiLa.

In most cases, grammatical information provided by the *Dictionary* is sufficiently fine-grained to provide all elements needed to disambiguate the multiple linking to the Lemma Bank, as it roughly consists of POS and inflection class, like in the case of *acus*. Yet, sometimes only information corresponding to POS is available. Several substantives are marked just as *subst.* (e.g., *deptar*, type of medicinal plant), which makes it non-trivial, if possible at all, to infer an inflectional category.

4.2 Linking to the Lemma Bank

After aligning the two tagsets, we proceed to link the *Dictionary* entries to the Lemma Bank. The one-to-one matches, i.e. lemmas in the *Dictionary* that match with just one lemma in the LiLa Lemma Bank with respect to both lemma and POS, have been considered validated. The following Subsections discuss the two other scenarios, namely one-to-many and one-to-zero matches.

4.2.1 One-to-Many

The string match on lemma and POS results in 831 ambiguous matches. Therefore, we add inflectional class as a further constraint; however, 445 lemmas still remain ambiguous and were inspected manually. For instance, for *lacertus* a correspondence in the Lemma Bank is found with *lacertus* 'upper arm' and *lacertus* 'lizard; a seafish', both NOUNs of the second declension (inflectional class `n2`). Only the manual checking of the meaning can thus allow to retrieve the correct match.

4.2.2 One-to-Zero

After performing the string match on lemma and POS, no match is found in the LiLa Lemma Bank for 10,276 lemmas. Among those, we automatically handle adverbs, verbs and *pluralia tantum* to find out whether they could be linked to the Lemma Bank respectively as hypolemmas of an adjective, lemma variants of a corresponding verb with opposite voice (active if deponent and vice versa), or lemma variant of a noun in singular form. By defining a set of heuristics applied automatically, we find that: (a) 92 adverbs can be linked to the adjective they are derived from (e.g., *homagialiter* - *homagialis* 'of homage'); (b) 18 verbs can be linked to their counterpart with opposite voice (e.g., *attaedio* - *attaedior* 'to bore'); (c) 80 plural forms can be linked to their singular equivalent (*moscilli* - *moscillus* 'little habit').

A closer look at lemmas that remain unmatched raises interesting insights, allowing for some linguistic considerations. First, clear evidence of areal contact is provided by forms like *bosako, -onis* and *kamennikko, -onis*. As the spelling reveals, these forms are the result of a contact with the language that was spoken in the area at that time, namely Old Czech. Indeed, *bosako* comes from the Czech form *bosák*, denoting a monk that by virtue of the rule has to walk barefoot, while *kamenniko* 'stonemason' derives from *kameník*. Additionally, several lemmas pertain to very specific domains. Consider e.g. *ascoa*, a sea

[14] `n4` corresponds to fourth declension nouns.

animal, *igenecha*, a type of quadruped[15], or *cinapus*, a species of fish, as an example of vocabulary of fauna. Flora is found as well: e.g., *elipurgis*, corresponding to *Cynoglossum officinale*, *bulboquilon*, 'mandrake', and *atomana*, a herb. Similar forms evidently display the specificity of some domains covered by the *Dictionary of Medieval Latin in the Czech Lands*.

Table 1 shows the distribution of Universal POS tags for 1:0 lemmas unmatched with respect to lemma only, which amount to 8,845. The total of 8,779 displayed in the Table excludes i) 40 initially unmatched lemmas that after manual inspection were found to be linked to lemmas already included in the Lemma Bank; ii) 26 lemmas that it was chosen not to link at all to the Knowledge Base, most often because the amount of available information in the *Dictionary* is minimal (for instance, not even the meaning is clear).

Table 1
POS distribution of unmatched lemmas.

UPOS	count
ADJ	1,822
ADP	2
ADV	490
DET	3
INTJ	4
NOUN	5,616
NUM	5
PRON	2
PROPN	119
SCONJ	1
VERB	715
Total	8,779

4.3 Results

The string match on lemmas and POS tags results in 55.5% one-to-one mappings; for 3.3% of entries more than one possible match was found, while for 41.2% no match was retrieved. The amount of lemmas that are not found in the Lemma Bank reflects the nature of the *Dictionary*, and especially its temporal, geographical and domain specificity. For comparison purposes, consider, for instance, that the process of linking the bilingual Latin-English dictionary by Lewis and Short, which is focused on Classical Latin, resulted in only 9% of unmatched lemmas (Mambrini et al. 2021). The percentage of no-match entries increases to 70% in the case of the *Neulateinische Wortliste* by Ramminger (Iurescia et al. 2023), which covers a time range spanning between 1300 and 1700 CE and features entries mirroring contemporary changes in the society, e.g. *typographus* 'typographer'.

Figure 2 shows an example of an entry of the *Dictionary* (*exquisitus*) linked to the LiLa Knowledge Base. The (yellow) node in the center of Figure 2 is the `ontolex:lexicalEntry` for *exquisitus*, which is linked via the property

[15] Possibly the common genet.

`lime:entry`[16] to the node that represents the entire *Dictionary* (an individual of the class `lime:lexicon`[17]) and to the corresponding lemma in the Lemma Bank via the property `ontolex:canonicalForm`. The lexical entry works as gateway to all the information associated to it in the resource. For instance, Figure 2 shows how the two meanings associated to *exquisitus* in its entry in the *Dictionary* are modeled. The two definitions provided by the resource (in Latin and in Czech) are linked to the lexical entry as individuals of the class `ontolex:lexicalSense`[18] via the property `ontolex:sense`[19]. Each sense is the specific lexicalisation of a more general `ontolex:lexicalConcept`[20], to which the sense is linked via the property `ontolex:isLexicalizedSenseOf`[21].

Although not visible in Figure 2, the lemma *exquisitus* in the Lemma Bank is linked, via `ontolex:canonicalForm`, to the entries for *exquisitus* in several other lexical resources and, via the property `lila:hasLemma`[22], to its occurrences (tokens) in the textual resources interlinked in LiLa[23] .

5. *Vita Caroli*: Linking a New Text

Following the linking of the *Dictionary of Medieval Latin in the Czech Lands* to the LiLa Knowledge Base, we now aim to evaluate to what extent this newly incorporated resource enhances the linking process of another text which shares temporal and geographical coordinates with the *Dictionary*, thereby exhibiting a comparable variety of Latin.

We select *Vita Caroli* 'Life of Charles'[24], the autobiography that the emperor of the Holy Roman Empire Charles IV wrote in Latin during the early years of his reign. The text, which represents one of the most famous historiographic works written in Medieval Bohemia, dates back to about 1360 CE. The work narrates events up to 1340, with additions up to 1346, covering his youth and the beginning of his reign. It holds significance particularly for historical understanding of this period, functioning as a literary monument of the reign of Charles IV and representing a rare example of autobiographical literature from a Medieval ruler. The narrative blends personal anecdotes with Charles IV's official duties, military campaigns, and diplomatic initiatives, alongside substantial contemplations on religious and moral themes. Structurally, it comprises twenty chapters.

The main motivation behind the selection of *Vita Caroli* therefore stems from its status as one of the most representative texts of Latin spoken in the Czech lands. Other candidate texts showing a degree of significance comparable to that of *Vita Caroli* include: a) a book of law of the city of Brno, pivotal for Czech municipal law (1355-1357); b) *Chronica Boemorum*, a famous chronicle authored by Cosmas of Prague (1045-

[16] `http://www.w3.org/ns/lemon/lime\#entry`.

[17] `http://www.w3.org/ns/lemon/lime\#Lexicon`.

[18] `http://www.w3.org/ns/lemon/ontolex\#LexicalSense`.

[19] `http://www.w3.org/ns/lemon/ontolex\#sense`.

[20] `http://www.w3.org/ns/lemon/ontolex\#LexicalConcept`.

[21] `http://www.w3.org/ns/lemon/ontolex\#isLexicalizedSenseOf`.

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

[23] For the full list of the resources currently made interoperable through LiLa, see `https://lila-erc.eu/data-page/`.

[24] Available at `http://dlb.ics.cas.cz/browser?path=/1/20/45/46//47&text=47`.

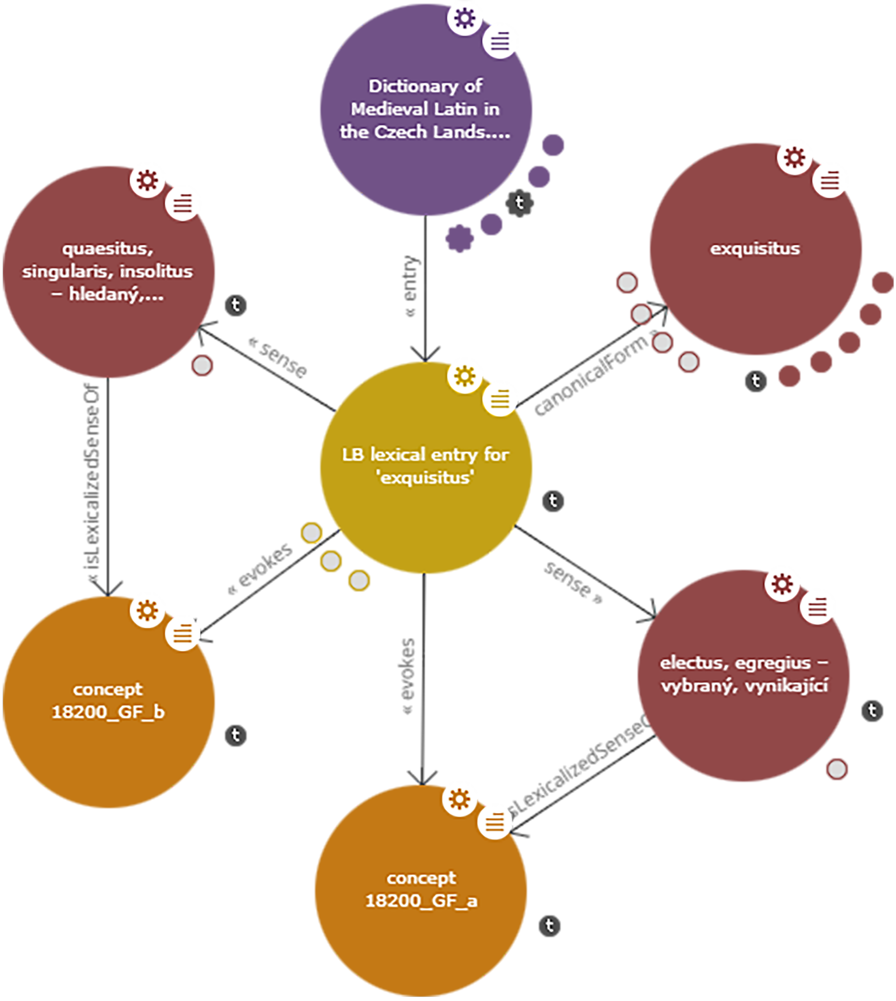

Dictionary of Medieval Latin in the Czech Lands....
quaesitus, singularis, insolitus – hledaný,...
exquisitus
« entry
« sense
canonicalForm »
LB lexical entry for 'exquisitus'
« isLexicalizedSenseOf
« evokes
sense »
electus, egregius – vybraný, vynikající
concept 18200_GF_b
« evokes
sLexicalizedSenseO
concept 18200_GF_a

Figure 2
The entry for *exquisitus* after being linked to LiLa.

1125); c) a visitation protocol[25] of an archdeacon with innumerable bohemisms (1379-82); d) a philosophico-theological commentary on Peter Lombard's Sentences written by Jan Hus (1407-9); e) treatises on construction and use of the astrolabe, written by a Prague scholar, with lots of technical vocabulary (1407). Nevertheless, the challenges

[25] *Protocollum visitationis archidiaconatus Pragensis annis 1379-1382 per Paulum de Janowicz, archidiaconum Pragensem, factae.*

encountered in acquiring data for these texts constitutes an additional reason for the choice of *Vita Caroli*, which is more easily retrievable and readily available[26].

5.1 Linking Coverage

Linking *Vita Caroli*[27], a text that shares spatial and temporal coordinates with the *Dictionary*, can offer interesting insights on the extent to which the *Dictionary* can actually contribute when linking new, non-Classical texts. In order to assess that, we analyse the resulting linking coverage in three distinct settings. We examine the percentages of 1:1, 1:N, and 1:0 matches when the string match is performed against:

1. B: only the base version of the Lemma Bank, which encompasses three Latin dictionaries: (Georges and Georges 1913–1918), (Glare 1982), (Gradenwitz 1904), for a total of 40,341 lexical entries and 43,408 lemmas;
2. B+O+D: the base Lemma Bank, plus lemmas from the *Onomasticon* (Forcellini 1940) (25,700 lexical entries and 25,700 lemmas) and *DuCange* (Du Cange 1883-1887), a Latin glossary from the western Middle Ages (64,902 lexical entries and 64,556 lemmas)[28];
3. All: the complete Lemma Bank, which includes also the *Dictionary*, for a total of 164,888 lemmas.

Table 2
String match results on tokens (total of 15,181).

	All	**B + O + D**	**B**
1:1	84.75%	82.49%	79.62%
1:N	6.37%	6.08%	4.20%
1:0	8.88%	11.43%	16.18%

Table 3
String match results on lemmas (total of 2,904).

	All	**B + O + D**	**B**
1:1	66.86%	64.26%	60.02%
1:N	5.75%	5.68%	3.59%
1:0	27.39%	30.06%	36.39%

[26] Acquiring the above-mentioned texts presents several issues; for instance, their transcriptions resulting from OCR include text together with notes, critical apparatus, and metadata, making it not straightforward to extract the raw text only.

[27] The linking process was performed by using the LiLa TextLinker service available at `https://lila-erc.eu/LiLaTextLinker`. The TextLinker is an online tool that provides automatic lemmatisation and POS tagging in order to prepare a raw Latin text for inclusion into the LiLa Knowledge Base (Passarotti, Mambrini, and Moretti 2024). Neither a resolution of 1:0 nor of 1:N was performed.

[28] The number of lexical entries and lemmas for B, O, and D comes from LEMLAT (Passarotti et al. 2017). A lexical entry can include more than one lemma. The total number of lemmas for D is lower than for lexical entries due to the removal of D entries duplicated in B, or O.

A closer look to the unmatched lemmas in the B+O+D and All settings reveals how leveraging the complete Lemma Bank (which includes also the *Dictionary*) results in the linking of 76 more lemmas. Among those, approximately one third (25 entries) consists of adverbs. Yet, particularly interesting is the presence within this set of several terms related to religion, such as the nouns *antipapa* 'antipope', *archiepiscopatus* 'archbishopric, *cappella* 'chapel', *capellanus* 'chaplain', and *carnispriuium*, i.e. the period from the Saturday preceding the Fiftieth Sunday to the following Tuesday or the entire period of time from the Epiphany to Ash Wednesday. Some of these terms are exclusively found in the *Dictionary* (e.g., *capellanus*), while others are present in the *Neulateinische Wortliste* as well (e.g., *antipapa*). Additionally, we can also observe several geographical terms, often related to the Italian territory. While some of them can indeed be found in the *Dictionary* (*coloniensis* 'Colognian', *florentinus* 'Florentine'), others are not actually contributed to the Lemma Bank by the *Dictionary*. This is the case of the noun *Lombardia*, the Italian region Lombardy, and the adjective *papiensis*, 'from/of Pavia', city in northern Italy. Such cases can be easily identified as they start with a letter in the range N-Z; as mentioned in Section 3, the *Dictionary* currently covers only the range A-M. For this reason, a complete assessment of the contribution of the *Dictionary* will only be feasible when its coverage is extended to include the range N-Z as well.

Overall, the percentages in Tables 2 and 3 illustrate the efficacy of expanding the Lemma Bank to encompass terms from non-Classical Latin texts in facilitating the linking of new texts to the Knowledge Base. In particular, the *Dictionary* proves beneficial for the linking of *Vita Caroli*, as evidenced by the lemmas just presented.

5.2 Analysis of Missing Matches

In addition to the linking coverage discussed in the previous Subsection, we also investigate the cases for which no match is found by the TextLinker run on a Medieval Latin text from the Czech lands such as *Vita Caroli*. We select a sample of 400 tokens categorised as unmatched, approximately corresponding to half of the total. Table 4 and Figure 3 summarise the reasons that have been observed (manually) to account for the unmatched forms.

Table 4
Number of observations per issue in the evaluation sample.

Error type	Count
lemmatisation	181
missing lemma	141
different wr	88
POS attribution	87
ADV mismatch	3
other	1
total	501

On average, an instance is found to exhibit more than one issue. Indeed, the evaluation sample contains 501 issues while amounting to 400 forms, thus revealing co-occurring ones. The categories that have been derived as accounting for 1:0 matches are the following:

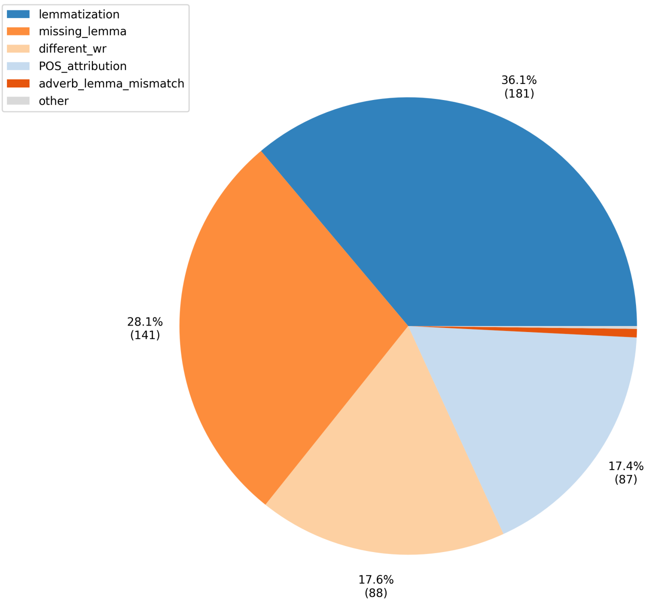


Figure 3
Distribution of issues from Table 4.

- Lemmatisation: the token is lemmatised incorrectly by the TextLinker, thus no correct match can be retrieved.
- Missing lemma: the lemma is missing altogether in the Lemma Bank.
- Different wr: different written representation (see Section 2). The spelling of the lemma differs from the Classical one, preventing the TextLinker from finding a match.
- POS attribution: the token is assigned an incorrect POS tag by the TextLinker, thus no correct match can be retrieved. While sometimes the POS attribution is nevertheless accurate, although not matching the one chosen in the Lemma Bank (e.g., NOUN-ADJ alternation), sometimes the POS is actually incorrect. For instance, this often happens when tokens that are sentence-initial and thus capitalised are interpreted by the TextLinker as PROPNs.
- ADV mismatch: the form is an adverb, that the TextLinker correctly tags as ADV. Yet, the adjective from which the adverb derives is assigned as lemma, instead of the adverbial form itself. In this case the lemmatisation cannot be considered incorrect, but it results in the TextLinker not finding a match.
- Other: one issue that does not fall into any other category. Specifically, the form found in the text is *voluntante*, which appears to be a possible spelling

mistake for *voluntate* 'will' (abl.sg.). This reading is supported by other editions.

The cases of different spellings appear to be particularly interesting in regard to the variability of the Latin language, highlighting nuances in temporal and spatial usage. The category of different written representations ('different wr') includes several phenomena (Table 5 and Figure 4), listed hereafter alongside an example:

- *c* for *t*: for instance, *puericia* for *pueritia* 'childhood'.
- Monophtong: the diphtong is simplified into a single vowel, as in *scintille* for *scintillae*, nominative plural of *scintilla* 'spark'.
- *K* for *c*: e.g. *Karinthia* for *Carinthia*.
- *H* simplification: *h* is omitted, as in *Boemia* for *Bohemia*.
- Simplified double consonant: the Classical double consonant is simplified into a single one, as in *anuncio* for *annuntio* 'to announce' (co-occurring with the replacement of *t* with *c*).
- *V* for *b*, as in *Bravancia* for *Brabantia* (co-occurring with the replacement of *t* with *c*) 'Brabant'.

Table 5
Different phenomena classified as different written representations.

Phenomenon	Count
c for t	29
monophthong	15
k for c	4
h simplification	3
simplified double consonant	1
v for b	1
other	4
total	57

In 8% of the cases of the evaluation sample, for a total of 32 occurrences, the lack of retrieved matches for a lemma is actually due to the co-occurrence of a different graphical variant and a lemmatisation error. In other words, in a non-negligible number of cases the reason why no match is found for the lemma – although the lemma itself is included in the Lemma Bank – is that the form slightly differs from the standard, Classical one in terms of spelling, thus resulting in being misleading for the TextLinker, which as a consequence fails the lemmatisation. In the case of the token *puericie* 'of the childhood', for instance, the proposed lemma is *puericies*, since the ending in *-e* is interpreted by the TextLinker as a sign that the term belongs to the fifth declension. However, in this case *e* is actually a simplification for *ae*, the ending for genitive singular in the first declension, and the token should have been properly lemmatised as *puericia*[29].

[29] We can also observe the *c/t* variation.

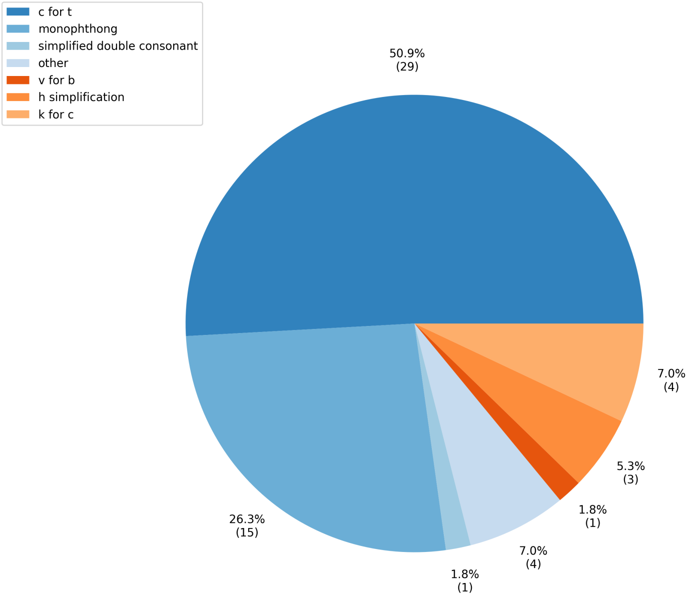


Figure 4
Pie chart showing the distribution of phenomena classified as different written representations.

6. Querying the Dictionary in LiLa

This Section presents three queries to exemplify the added value of interoperability between the resources linked to LiLa[30].

First query: non-classical lemmas featuring *natura* in their definition
The first query, available within a set of pre-compiled queries in the SPARQL endpoint of LiLa, retrieves all those lemmas whose entries in the *Dictionary* include the word *natura* ‘nature’ in their definition(s) and do not occur also in the Lewis and Short dictionary, and returns the number of their occurrences in the textual corpora linked to LiLa. The 11 retrieved lemmas[31] occur in 5 corpora, for a total of 132 occurrences, 5 out of which are found in the Computational Historical Semantics corpus[32], 104 in the *Index Thomisticus* Treebank[33], 4 in UDante[34], 1 in the CIRCSE Latin Library[35] (specifically,

[30] LiLa offers three query services: a query graphical interface (`https://lila-erc.eu/query/`) allows to query the Lemma Bank; the linguistic resources for Latin interlinked in LiLa can be queried either through a SPARQL endpoint (`https://lila-erc.eu/sparql/`) or via the graphical LiLa Interactive Search Platform (LISP: `https://lila-erc.eu/LiLaLisp/`).

[31] *Accidentalis* ‘accidental’, *bestialitas* ‘bestiality’, *connaturalis* ‘connatural’, *connaturalitas* ‘connaturalitas’, *contingentia* ‘contingency’, *eligibilis* ‘eligible’, *finitas* ‘finiteness’, *fumositas* ‘smoking, fume’, *leuiathan* ‘Leviathan (aquatic monster)’, *materialitas* ‘materiality’, *mollitia* ‘softness, weakness’.

[32] `http://lila-erc.eu/data/corpora/CompHistSem/id/corpus`.

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

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

[35] `http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus`. Collection of Latin texts enhanced with different layers of linguistic annotation.

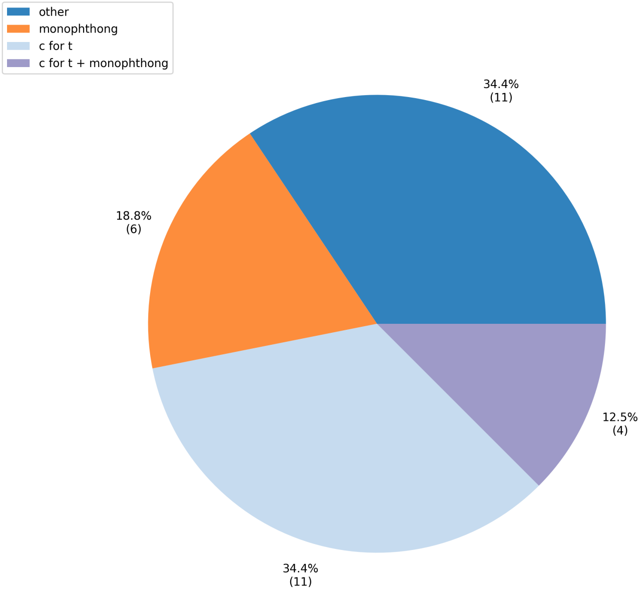
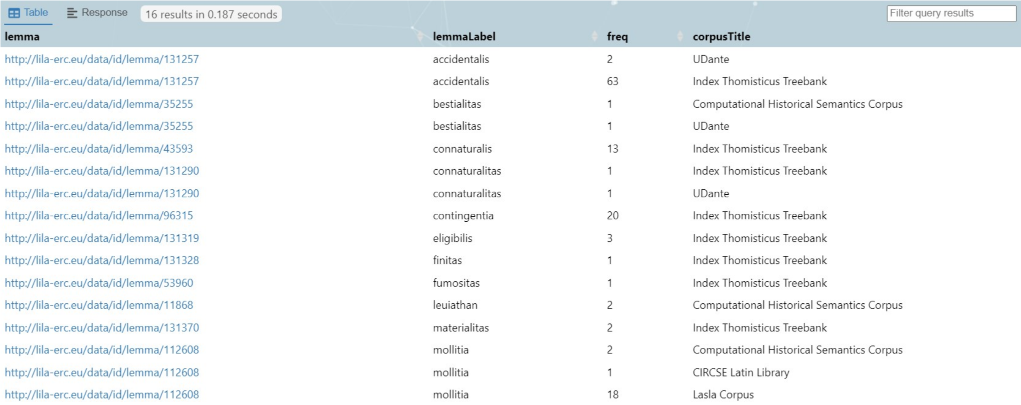


Figure 5
Pie chart showing cases of lemmatisation errors associated to a different spelling.



lemma	lemmaLabel	freq	corpusTitle
http://lila-erc.eu/data/id/lemma/131257	accidentalis	2	UDante
http://lila-erc.eu/data/id/lemma/131257	accidentalis	63	Index Thomisticus Treebank
http://lila-erc.eu/data/id/lemma/35255	bestialitas	1	Computational Historical Semantics Corpus
http://lila-erc.eu/data/id/lemma/35255	bestialitas	1	UDante
http://lila-erc.eu/data/id/lemma/43593	connaturalis	13	Index Thomisticus Treebank
http://lila-erc.eu/data/id/lemma/131290	connaturalitas	1	Index Thomisticus Treebank
http://lila-erc.eu/data/id/lemma/131290	connaturalitas	1	UDante
http://lila-erc.eu/data/id/lemma/96315	contingentia	20	Index Thomisticus Treebank
http://lila-erc.eu/data/id/lemma/131319	eligibilis	3	Index Thomisticus Treebank
http://lila-erc.eu/data/id/lemma/131328	finitas	1	Index Thomisticus Treebank
http://lila-erc.eu/data/id/lemma/53960	fumositas	1	Index Thomisticus Treebank
http://lila-erc.eu/data/id/lemma/11868	leuiathan	2	Computational Historical Semantics Corpus
http://lila-erc.eu/data/id/lemma/131370	materialitas	2	Index Thomisticus Treebank
http://lila-erc.eu/data/id/lemma/112608	mollitia	2	Computational Historical Semantics Corpus
http://lila-erc.eu/data/id/lemma/112608	mollitia	1	CIRCSE Latin Library
http://lila-erc.eu/data/id/lemma/112608	mollitia	18	Lasla Corpus

Figure 6
Overview of results of the first query.

in Augustine's *Confessiones*) and 18 in the *Opera latina* LASLA corpus[36] (Fantoli et al. 2022). The results of the query confirm once again the specificity of the *Dictionary of Medieval Latin in the Czech Lands*. Having excluded Classical lemmas that can also be found in the Lewis and Short dictionary, what remains are mostly lemmas that occur in corpora featuring texts of later stages of Latin: for instance, the texts from the *Index Thomisticus* Treebank and UDante date back respectively to XIII and XIV centuries CE.

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

The only exception is represented by the LASLA corpus, which includes Classical Latin. Yet, occurrences in LASLA are limited to the lemma *mollitia* 'softness, weakness', which is therefore attested in Classical times as well, while all the other lemmas appear to have originated later.

Second query: derivational processes

The following query investigates the derivational process leading to the formation of a new verb via the addition of a preverb to an existing verb (V-to-V), and compares the distribution of such preverbs in the *Dictionary* with those observed in the derivational lexicon Word Formation Latin (WFL)[37] (Litta and Passarotti 2019).

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lime: <http://www.w3.org/ns/lemon/lime#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX lemonVartrans: <http://www.w3.org/ns/lemon/vartrans#>
PREFIX wfl: <http://lila-erc.eu/ontologies/lila/wfl/>

SELECT  ?ruleType ?affixPrefixLabel (count (?lemma) as ?countLemma)
WHERE {
  VALUES ?ruleType {
    <http://lila-erc.eu/ontologies/lila/wfl/Prefixation/VerbToVerb>
  }
  <http://lila-erc.eu/data/lexicalResources/WFL/Lexicon>
  lime:entry ?le .
  ?wflsourceRel lemonVartrans:target ?le;
                wfl:hasWordFormationRule ?rule .
  ?le  ontolex:canonicalForm ?lemma .

  ## uncomment for LexiconBohemorum
  <http://lila-erc.eu/data/lexicalResources/LexiconBohemorum/Lexicon>
  lime:entry ?lexentryBM .
          ?lexentryBM ontolex:canonicalForm ?lemma .

  ?rule rdfs:label ?ruleLabel;
        rdf:type ?ruleType .
  ?rule wfl:involves ?involve .
  ?involve rdfs:label ?affixPrefixLabel
} group by ?ruleType ?affixPrefixLabel
order by  (?affixPrefixLabel)
```

Table 6 shows all the preverbs retrieved in the *Dictionary*, and the most productive ones from WFL (derivatives > 10). It can be observed how the set of the most productive preverbs is consistent across both resources, with a similar quantitative distribution as well. The only observable deviation corresponds to the preverb *re-*, quite represented in WFL with 561 occurrences, but absent altogether in the *Dictionary*.

Similar considerations can be made e.g. about deverbal nouns, by replacing the derivational rule `Prefixation/VerbToVerb` with `Suffixation/VerbToNoun` in the query. Overall, the *Dictionary* appears to adhere closely to the canonical distribution of derivational processes outlined in WFL, indicating a lack of significant linguistic variability in this regard.

[37] The two lines introduced by 'uncomment for LexiconBohemorum' are to be commented out to run the query on WFL only, whereas when uncommented the *Dictionary* is included as well.

Table 6
Results of the query investigating V-to-V derivational processes in the *Dictionary* and in Word Formation Latin.

Dictionary	**countLemma**	**WFL**	**countLemma**
in (entering)-	257	con-	615
con-	255	in (entering)-	611
e(x)-	224	e(x)-	609
de-	195	de-	581
ad-	170	re-	561
dis-	71	ad-	494
a(b)-	62	per-	331
inter-	37	prae-	293
circum-	30	sub-	270
am(b)(i)-	6	dis-	266
in (negation)-	6	ob-	227
ante-	5	a(b)-	213
contra-	3	super-	198
intro-	3	pro-	171
bi-	2	circum-	161
indu/endo/indo-	1	inter-	138
		tra(ns)-	94
		in (negation)-	46
		ante-	24
		se/sed/so-	23
		praeter-	21
		subter-	21
		semi-	18
		intro-	14
		am(b)(i)-	11
		contra-	11

Third query: LASLA and *Vita Caroli* collections of lemmas

The third query aims at investigating the differences which can be observed between the collection of lemmas pertaining to the LASLA corpus and that of the text of *Vita Caroli*. LASLA collection of lemmas is obtained thanks to the query displayed hereafter, which retrieves absolute frequencies of lemmas[38].

```
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#>
prefix lila: <http://lila-erc.eu/ontologies/lila/>
prefix lilacorpora: <http://lila-erc.eu/ontologies/lila_corpora/>
prefix lime: <http://www.w3.org/ns/lemon/lime#>
prefix marl: <http://www.gsi.dit.upm.es/ontologies/marl/ns#>
```

[38] Relative frequencies can be derived by the total count of tokens in LASLA, i.e. 1,744,608.
To obtain the *Vita Caroli* collection of lemmas, in the query the URI of the LASLA corpus reported in `VALUES ?corpora` has to be replaced by the URI of *Vita Caroli* (`http://lila-erc.eu/data/corpora/CIRCSELatinLibrary/id/corpus/Vita%20Caroli`), while the part `/^powla:hasSubDocument` has to be removed.

```
prefix powla: <http://purl.org/powla/powla.owl#>
prefix corpora: <http://lila-erc.eu/ontologies/lila_corpora/>

SELECT  ?lablemma (count(?t) as ?tot)  WHERE {
  VALUES ?corpora {
    <http://lila-erc.eu/data/corpora/Lasla/id/corpus>
  }
  ?t a powla:Terminal ;
       lila:hasLemma ?lemma .
        ?lemma lila:hasPOS ?pos .
  FILTER(?pos IN (lila:verb, lila:noun, lila:adjective,
        lila:proper_noun, lila:adverb))
  ?t powla:hasLayer/powla:hasDocument/^powla:hasSubDocument ?corpora .
  ?lemma rdfs:label ?lablemma
}group by ?lemma ?lablemma
order by desc(?tot)
```

Table 7 presents a comparative analysis of the 30 most common lemmas found in the LASLA corpus and in the text of *Vita Caroli*. The analysis focuses exclusively on nouns, verbs, adjectives, and adverbs, due to their potential to more accurately represent the textual content compared to function words. The table provides valuable insights, revealing distinct linguistic patterns between the texts, particularly shedding light on the specialised domain of *Vita Caroli* and its lexical choices.

Indeed, the frequency distribution in *Vita Caroli* emphasises its thematic concentration on politics and governance, evident from prominent lemmas such as *civitas* 'town/state', *dominus* 'lord', *dux* 'leader/commander', *regnum* 'kingdom', and *rex* 'king'. Notably, many terms denote interpersonal relationships (*frater* for 'brother', *pater* for 'father', *inimicus* for 'enemy'), underscoring the relational nature inherent in political discourse, especially within the context of imperial power and familial dynasties. Furthermore, the list includes personal and geographical proper names (*Boemia*, *Ludovicus*) whose frequency is indeed very high in the whole text.

In contrast, the most frequent lemmas in LASLA lack thematic concentration (due to the large coverage of the Classical Latin literature provided by the corpus), appearing more generic in nature and not clustering around any specific sub-field. This distinction is particularly evident among substantives, with examples such as *animus* 'mind', *causa* 'reason', *locus* 'place', and *dies* 'day', as well as adjectives like *magnus* 'big' and *bonus* 'good'.

7. Conclusions

Linking the *Dictionary* to the LiLa Knowledge Base not only was a further step towards the full exploitation of linguistic resources for Latin, thanks to their interoperability, but also contributed to improve the degree of linguistic diversity represented in LiLa as for three aspects, that are particularly relevant for Latin as a language that was used for centuries all over Europe: (a) diachronic diversity: the *Dictionary* collects a portion of the Latin vocabulary that emerged in Medieval times; (b) diatopic diversity: the lexical resource includes items from a specific area, namely the Czech lands; (c) domain-based diversity: quite frequently the entries of the *Dictionary* belong to very specific domains (e.g., flora and fauna; see Section 4.2.2). The contribution of the lemmas from the *Dictionary* in enlarging the LiLa Lemma Bank is thus considerable both in terms of quantity and in terms of quality, and highlights the importance of linking to the Knowledge Base also resources that feature non-standard varieties of Latin. Such non-standard information could also prove beneficial for the future linking of new lexical

Table 7
Frequency list of the 30 most common lemmas in LASLA and *Vita Caroli* (NOUNs, PROPNs, ADJs, VERBs, ADVs only).

LASLA		**Vita Caroli**	
lemma	**%**	**lemma**	**%**
sum	3,17	*sum*	2.82
non	1,26	*rex*	0.84
res	0,55	*pater*	0.83
possum	0,52	*dico*	0.79
facio	0,44	*ciuitas*	0.72
qua	0,44	*non*	0.69
dico	0,43	*dominus*	0.53
uideo	0,40	*sic*	0.49
magnus	0,35	*habeo*	0.45
do	0,32	*tempus*	0.43
habeo	0,32	*facio*	0.41
enim	0,28	*dux*	0.41
etiam	0,28	*dies*	0.39
homo	0,25	*magnus*	0.37
iamiam	0,25	*possum*	0.34
ut	0,24	*regnum*	0.34
animus	0,23	*uenio*	0.33
uolo	0,224	*tunc*	0.31
tamen	0,22	*uerbum*	0.3
causa	0,19	*deus*	0.28
nam	0,19	*castra*	0.28
locus	0,19	*homo*	0.27
fero	0,18	*Boemia*	0.27
ne	0,18	*nomen*	0.26
nunc	0,18	*ludouicus*	0.26
quidem	0,18	*filius*	0.24
uenio	0,17	*uero*	0.22
publicus	0,17	*uolo*	0.22
dies	0,16	*frater*	0.22
bonus	0,16	*inimicus*	0.22

resources. For instance, the insights into non-standard spelling variants, summarised in Table 5, could possibly be exploited to normalise non-standard entries found in lexical resources to be linked to LiLa in the future.

In the near future, we intend to finalise the linking of the *Dictionary*, by including referencing lemmas besides full entries (see Section 3.1). We also intend to model citations of attestations, i.e. references to other dictionaries where an entry is found, and to sources of examples. Moreover, in line with what was done with the text of *Vita Caroli*, we plan to link to the LiLa Knowledge Base some documents from the same area and period as the *Dictionary*, such as the Czech Medieval sources from the

AHISTO project[39]. However, working with such texts first implies dealing with various difficulties in acquiring the texts themselves (e.g., quality of OCR, transcriptions of texts hardly separable from metadata and notes). Additionally, these documents are currently available only as raw texts, and would need to be lemmatised before the linking. Given the peculiar nature of their Latin variety, conditioned by the Czech language and rich of local proper names, lemmatisation with the currently available trained models for Latin will probably provide low accuracy rates. Once again, this proves the importance of collecting non-standard Latin data (and resources) and investigating to what extent Latin varieties differ.

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. This work was partially supported by the Grant No. 20-16819X (LUSyD) of the Czech Science Foundation (GACR), by project GA UK No. 104924, and by SVV project No. 260 698.
We want to thank Pavel Nývlt for his collaboration in providing the database "The Dictionary of Medieval Latin in Czech Lands", available via the LINDAT/CLARIAH-CZ Research Infrastructure, supported by the Ministry of Education, Youth, and Sports of the Czech Republic (Project No. LM2018101). We also thank Giovanni Moretti for his support during the linking phase.

References

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific american*, 284(5):34–43.

Buitelaar, Paul, Philipp Cimiano, John McCrae, Elena Montiel Ponsoda, and Thierry Declerck. 2011. Ontology lexicalisation: The lemon perspective. In *Proceedings of the Workshops, 9th International Conference on Terminology and Artificial Intelligence*, pages 33–36, Paris, France, November. Ontology Engineering Group - OEG.

Cecchini, Flavio Massimiliano, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin works. In *Seventh Italian Conference on Computational Linguistics*, pages 1–7, Bologna, March 1-3, 2021. CEUR-WS.org.

Chiarcos, Christian. 2012. POWLA: Modeling linguistic corpora in OWL/DL. In *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings 9*, pages 225–239. Springer.

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

Ctibor, Jan and Pavel Nývlt. 2021. On-line Dictionary of medieval Latin in the Czech lands. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Du Cange, Charles Du Fresne. 1883-1887. *Glossarium Mediae et Infimae Latinitatis*. L. Favre, Niort.

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 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34, Marseille, France, June. European Language Resources Association.

Forcellini, Egidio. 1940. *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos. Perin*. Typis Seminarii, Padova.

[39] https://nlp.fi.muni.cz/projekty/ahisto/portal.

Geelhaar, Timo, Alexander Mehler, Benjamin Jussen, Alexander Henlein, Giuseppe Abrami, Daniel Baumartz, Tolga Uslu, et al. 2020. The Frankfurt Latin Lexicon from Morphological Expansion and Word Embeddings to Semiographs. *Studi e saggi linguistici*, 58(1):45–81.

Georges, Karl Ernst and Heinrich Georges. 1913. Ausführliches lateinisch-deutsches Handwörterbuch, 2 vols. *Hannovre/Leipzig: Hahnsche Buchhandlung*.

Georges, Karl Ernst and Heinrich Georges. 1913–1918. *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hahn, Hannover.

Glare, Peter G.W. 1982. *Oxford Latin Dictionary*. Oford University Press, Oxford.

Gradenwitz, Otto. 1904. *Laterculi vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. Hirzel, Leipizig.

Iurescia, Federica, Eleonora Litta, Marco Passarotti, Matteo Pellegrini, Giovanni Moretti, and Paolo Ruffolo. 2023. Linking the Neulateinische Wortliste to the LiLa Knowledge Base of Interoperable Resources for Latin. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 82–87, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Lassila, Ora and Ralph R. Swick. 1998. Resource Description Framework (RDF) model and syntax specification. Available online at `https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/`.

Lewis, Charlton T. and Charles Short. 1879. *A Latin Dictionary*. Clarendon Press, Oxford.

Litta, Eleonora and Marco Passarotti. 2019. (when) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Lemmata Linguistica Latina*, volume 1: Word and Sounds. De Gruyter, Berlin, Boston, pages 224–239.

Mambrini, Francesco, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2021. 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 214–220, December.

McCrae, John Philip, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: development and applications. In *Proceedings of The Fifth Biennial Conference on Electronic Lexicography, eLex 2017*, pages 19–21, Leiden, Netherlands, 19-21 September 2017.

Passarotti, Marco. 2019. The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*. De Gruyter, Berlin, pages 299–319.

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*, pages 24–31, Gothenburg, May.

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. *Studi e Saggi Linguistici*, 58(1):177–212.

Passarotti, Marco, Francesco Mambrini, and Giovanni Moretti. 2024. The services of the LiLa knowledge base of interoperable linguistic resources for Latin. In Christian Chiarcos, Katerina Gkirtzou, Maxim Ionov, Fahad Khan, John Philip McCrae, Elena Montiel Ponsoda, and Patricia Martín Chozas, editors, *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, pages 75–83, Torino, Italia, May. ELRA and ICCL.

Petrov, Slav, 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, May. European Language Resources Association (ELRA).

Prud'Hommeaux, Eric and Andy Seaborne. 2008. SPARQL query language for RDF. *W3C working draft*, 4(January).

Silagiová, Zuzana, Pavel Nývlt, Julie Černá, Hana Florianová, Barbora Kocánová, Hana Šedinová, and Kateřina Vršecká. 2019. Latinitatis medii aevi lexicon Bohemorum – The Dictionary of Medieval Latin in Czech Lands, Volume II (D-H), second, revised edition.

Silagiová, Zuzana, Julie Černá, Hana Florianová, Pavel Nývlt, Hana Šedinová, and Kateřina Vršecká. 2018. Latinitatis medii aevi lexicon Bohemorum – The Dictionary of Medieval Latin in Czech Lands, Volume I (A-C), second, revised edition.

Silagiová, Zuzana, Julie Černá, Dana Martínková, Barbora Kocánová, Markéta Koronthályová, Kateřina Vršecká, Richard Mašek, Jiří Matl, Hana Miškovská, Pavel Nývlt, Hana Šedinová, and Irena Zachová. 1995 to 2016. Latinitatis medii aevi lexicon Bohemorum – The Dictionary of Medieval Latin in Czech Lands, Volume III (I-M). The electronic version has been created by Jan Ctibor.