Querying the *Lexicon der indogermanischen Verben* in the LiLa Knowledge Base: Two Use Cases

Valeria Irene Boano, Marco Passarotti and Riccardo Ginevra
Università Cattolica del Sacro Cuore
Milan, Italy
valeriairene.boano01@icatt.it, marco.passarotti@unicatt.it, riccardo.ginevra@unicatt.it

Abstract

This paper presents two use cases of the etymological data provided by the *Lexicon der indogermanischen Verben* (LIV) after their publication as Linked Open Data and their linking to the LiLa Knowledge Base (KB) of interoperable linguistic resources for Latin. The first part of the paper briefly describes the LiLa KB and its structure. Then, the LIV and the information it contains are introduced, followed by a short description of the ontologies and the extensions used for modelling the LIV's data and interlinking them to the LiLa ecosystem. The last section details the two use cases. The first case concerns the inflection types of the Latin verbs that reflect Proto-Indo-European stems, while the second one focusses on the Latin derivatives of the inherited stems. The results of the investigations are put in relation to current research topics in Historical Linguistics, demonstrating their relevance to the discipline.

Keywords: Linked Open Data, Latin, Indo-European

1. Introduction

In recent years, the Linked Open Data (LOD) paradigm has been increasingly applied to linguistic (meta)data to achieve their interoperability, leading to a constant growth of the Linguistic Linked Open Data Cloud.[1] Linguistic resources that are part of the cloud include textual corpora, lexicons, dictionaries and more. Among the resources interlinked in the Cloud are those published in the LiLa Knowledge Base (KB),[2] which contains several textual and lexical resources for the Latin language, published as LOD and linked with each other.

With regard to textual resources, LiLa includes so far more than 3,5M words from several Latin corpora, in both Classical Latin and Medieval Latin. Among them, are the corpus of Classical texts LASLA[3] (CIRCSE, 2022; Fantoli et al., 2022), the *Index Thomisticus* treebank (CIRCSE, 2006-2024; Cecchini et al., 2018), which contains works of Thomas Aquinas, and *UDante* (CIRCSE, 2021b; Cecchini et al., 2020), which is a Universal Dependencies[4] treebank for Dante Alighieri's Latin works. As for the lexical resources, LiLa currently includes, among others, the Lewis and Short Latin-English dictionary (CIRCSE, 2021a; Lewis and Short, 1879), the derivational lexicon *Word Formation Latin* (CIRCSE, 2018; Litta et al., 2019), and a resource of morphological principal parts of Latin words, *PrinParLat* (CIRCSE, 2023b; Pellegrini, 2023).

Moreover, LiLa interlinks etymological information from two reference dictionaries: the *Etymological dictionary of Latin and other Italic Languages* (CIRCSE, 2020a; de Vaan, 2008), which focusses on Latin and other Italic languages, and the *Lexicon der indogermanischen Verben* (LIV) (CIRCSE, 2023a; Rix, ed., 2001), which features reconstructed Proto-Indo-European (PIE) verbal roots and details their developments in the attested Indo-European (IE) daughter languages, including Latin. The etymological relations between Latin words and their ancestors in PIE provided by the latter have been recently linked to LiLa (Boano et al., 2023): their integration in the KB helps to put the information contained in the LIV in relation to the one provided by other linguistic resources. To achieve this in the (recent) past different linguistic resources were consulted one at time, and their data were integrated later. Now, thanks to the interoperability among resources made possibile by LiLa, this same process can be achieved automatically and it is made fully replicable.

This paper aims to show the advantages that linking the LIV to LiLa provides in approaching two research questions of Historical Linguistics. After introducing the overall architecture of the LiLa KB (Section 2) and the process performed to interlink the LIV into LiLa (Section 3), in Section 4 the paper details the two use cases, showing how the LIV's information can be queried and exploited in the LiLa KB to address the research questions concerned.

[1] https://linguistic-lod.org/.
[2] https://lila-erc.eu/.
[3] https:/www.lasla.uliege.be/cms/c_8508894/fr/lasla.
[4] https://universaldependencies.org/.

LDL: Linked Data in Linguistics @ LREC-COLING 2024, pages 22–31
20-25 May, 2024.

2. The LiLa Knowledge Base

The LiLa KB provides FAIR linguistic resources (Wilkinson et al., 2016) published as LOD. The syntactic interoperability between the resources of the KB is ensured by the use of the Resource Description Framework (RDF) data model (Lassila and Swick, 1998). The semantic interoperability (Ide and Pustejovsky, 2010) instead is achieved by the use of a few vocabularies widely used for the publication of linguistic resources as LOD, including the Ontolex Lemon model,[5] the OLiA ontology (Chiarcos and Sukhareva, 2015) and the Ontolex lexicography module.[6] The connection between the resources interlinked in the KB is achieved via the so-called Lemma Bank (LB) (CIRCSE, 2019-2024).[7] The LB is a set of more than 200k lemmas, which was originally created from the database of the morphological analyser LEMLAT (Passarotti et al., 2017), and which is constantly extended whenever a new linguistic resource requires a new lemma to be included in the KB.

The LB constitutes LiLa's core structure and the crossroads between all the resources part of the KB. Interoperability is achieved by linking tokens provided by textual corpora and entries in lexical resources to their corresponding lemma in the LB.

Whenever possible, lemmas, tokens and lexical entries are represented and published as LOD by means of classes and properties from the Ontolex Lemon core module. Each `ontolex:LexicalEntry`[8] of each lexical resource is linked via the property `ontolex:canonicalForm`[9] to the corresponding `lila:Lemma`[10] in the LB. A `lila:Lemma` is a subclass of the class `ontolex:Form`,[11] namely a word's citation form. The simple link established between a `lila:Lemma` and the corresponding `ontolex:LexicalEntry` ensures the interoperability between the lexical resources part of LiLa. As for the tokens of the corpora interlinked in LiLa, they are connected to the LB via the property `lila:hasLemma`.[12]

The `lila:Lemma` also carries morphological information, such as the gender and the inflection type. Some lemmas are also assigned derivational information about prefixes, suffixes and lexical bases: at the time of writing, the derivational information recorded in the LiLa LB regards Classical Latin words only, while the coverage for the Medieval Latin is significantly lower (Pellegrini et al., 2022).

The LiLa KB can be queried via a SPARQL endpoint,[13] via a user-friendly interface[14] and via an interactive search platform.[15]

3. The LIV and its Modelling

Etymology can be broadly defined as "the branch of linguistics which deals with determining the origin of words and the historical development of their form and meanings" (OED, s.v. *etymology*, *n.*). The LIV is the reference etymological dictionary for verbs attested in the ancient IE languages. It was curated by Helmut Rix and first published in 1998 by Reichert Verlag (Rix, ed., 1998). A second edition appeared in 2001, with the additions and corrections by Martin Kümmel and Helmut Rix (Rix, ed., 2001). This dictionary contains information regarding the PIE verb and its development in the IE languages: it details the etymology of verbs attested in IE languages by tracing them back to reconstructed PIE verbs. In particular, the LIV contains three main types of lexical items:

- Reconstructed PIE verbal roots. They constitute the entries of the dictionary, and are provided with their phonological structure and broad lexical meaning. A verbal root is the part of a word that "carries the core of the meaning, the idea of a situation, which is recognisable in all forms derived from the root" (Rix, ed., 2001, p. 5, my translation).
- Reconstructed PIE verbal stems. They consist of the verbal root processed with affixes and they encode aspectual information.
- Word forms attested in IE languages. The LIV lists word forms for several IE languages: they can be traced back to the corresponding PIE stems and are provided with their attested meaning.

As by agreement with the LIV's publisher, only the relations established between these elements were modelled and linked to LiLa. For the modelling, we decided to use the `lemonEty` extension of the Ontolex Lemon model (Khan, 2018), which was developed precisely for representing etymological information. `lemonEty` provides three key classes:

[5] https://www.w3.org/2016/05/ontolex/.
[6] https://www.w3.org/2019/09/lexicog/.
[7] http://lila-erc.eu/data/id/lemma/LemmaBank.
[8] http://www.w3.org/ns/lemon/ontolex#LexicalEntry.
[9] http://www.w3.org/ns/lemon/ontolex#canonicalForm.
[10] http://lila-erc.eu/ontologies/lila/Lemma.
[11] http://www.w3.org/ns/lemon/ontolex#Form.
[12] https://lila-erc.eu/ontologies/lila/hasLemma.
[13] https://lila-erc.eu/sparql/.
[14] https://lila-erc.eu/query/.
[15] https://lila-erc.eu/LiLaLisp/.

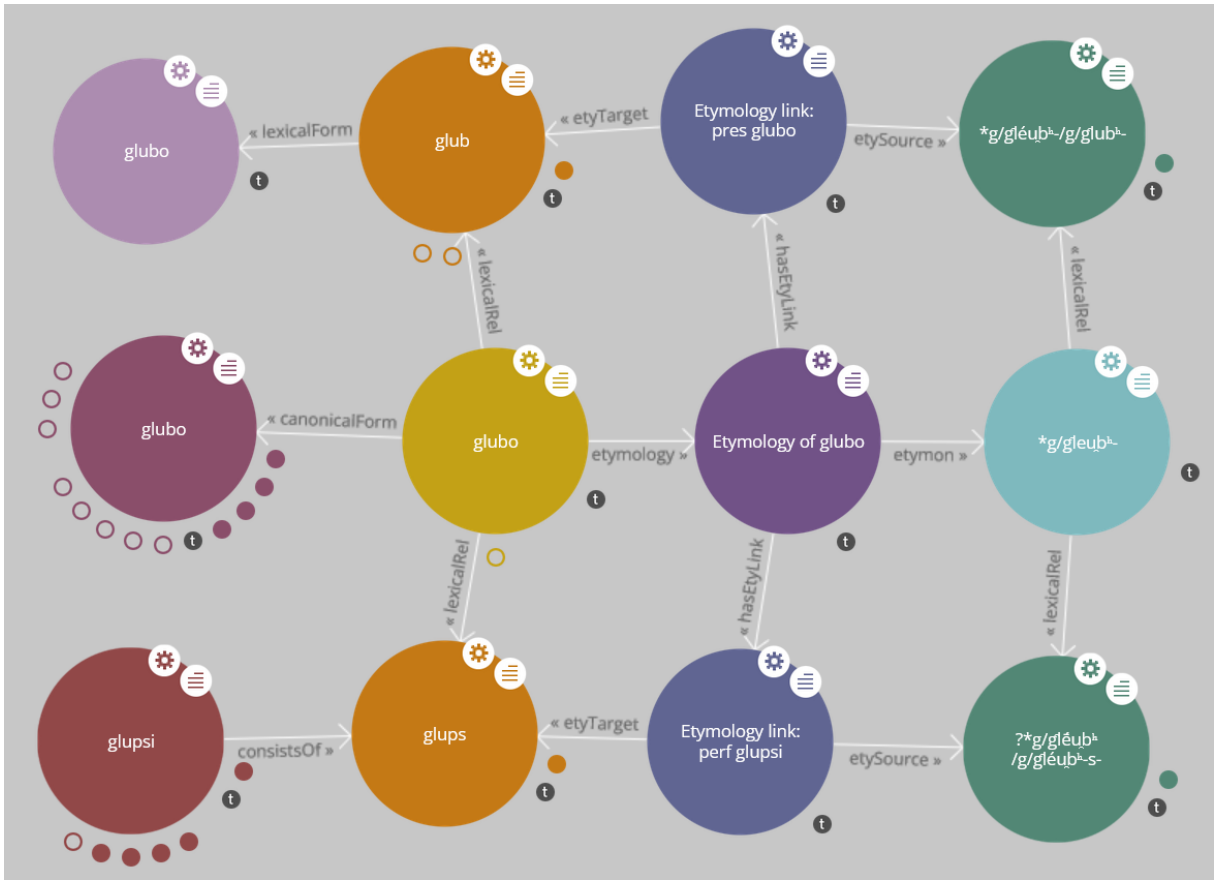

Figure 1: The model of the LIV etymological relations, with respect to the verb *glubo*.

- `lemonEty:Etymon`:[16] this class is a subclass of `ontolex:LexicalEntry` and contains all the lexical items of the source language that are introduced to explain the etymology of the target language;
- `lemonEty:Etymology`:[17] this class "reifies the whole process of etymological reconstruction as scientific hypothesis" (Passarotti et al., 2020, p. 22);
- `lemonEty:EtyLink`:[18] this class is used to connect linguistic items from the source language to the target language.

We modelled the PIE roots provided by the LIV (e.g. PIE **h₃emh₃-*, underlying the Latin verb *amo* 'to love') as instances of the class `lemonEty:Etymon`, since they are items of the source language (in this case, PIE) and are introduced "to describe the origin and history of another Lexical Entry".[19] Moreover, since `lemonEty:Etymon` is a subclass of `ontolex:LexicalEntry`, this allowed us to preserve the structure of the LIV, which treats the roots as lexical entries.

The `lemonEty:EtyLink` class was used to model the relation between a PIE stem and its corresponding Latin stem. The LIV provides in fact the Latin first-person present and first-person perfect word forms, which are traditionally used to represent all the forms derived from the present and the perfect stems. For this reason, we were able to include the Latin stems as part of the model: in particular, we reused the individuals of the class `Stem`[20] provided by *PrinParLat* (CIRCSE, 2023b), which is a collection of principal parts of Latin morphological paradigms already interlinked in the LiLa KB. When the Ontolex Morph module (Chiarcos et al., 2022) will be released, the PIE stems, instead, will be represented as instances of the class `morph:Morph`:[21] this class is used to repre-

[16] `http://lari-datasets.ilc.cnr.it/lemonEty#Etymon`..
[17] `http://lari-datasets.ilc.cnr.it/lemonEty#Etymology`.
[18] `http://lari-datasets.ilc.cnr.it/lemonEty#EtyLink`.
[19] `http://lari-datasets.ilc.cnr.it/lemonEty`.
[20] `https://lila-erc.eu/lodview/ontologies/prinparlat/Stem`.
[21] At the time of writing, the URIs provisionally point to the Morph's GitHub page (`https://github.com/`

sent all those elements of morphological analysis which are below the word level.

Each PIE stem is also linked to the class `prinparlat:StemType`:[22] new individuals were added to this class in order to include all the PIE stem types provided by the LIV. The latter are: present, aorist, perfect, causative, iterative, causative-iterative, desiderative, intensive, fientive and essive. Each of these categories expresses a specific grammatical or lexical aspect.

Both PIE and Latin stems were connected to the `lemonEty:EtyLink` via the properties `lemonEty:etySource` and `lemonEty:etyTarget`, respectively. Each Latin stem was also linked to the corresponding Latin form. For the perfect, we reused the form provided by *PrinParLat*. For the present, instead, we generated it from scratch, since *PrinParLat* supplies only the third-person present form.

Finally, the `lemonEty:Etymology` class stands as a central crossroads between all the LIV lexical items: it reifies the generic etymological relation between the Latin `ontolex:LexicalEntry` and the PIE root, while also being linked to the two etymological links.

Figure 1 shows the model applied to the case of the verb *glubo* 'to peel'. On the left side is the lexical entry *glubo*, linked to the LiLa lemma via the property `ontolex:canonicalForm`. The two *PrinParLat* Latin stems are linked to the lexical entry via the property `vartrans:lexicalRel`.[23] The Latin stems are the starting point of two connections: one with the Latin forms, the present *glubo* and the perfect *glupsi*,[24] and the other with the two etymological links.[25] These reify the etymological relation between the Latin stems and the corresponding PIE stems (**g/ĝléu̯bʰ-/g/ĝlubʰ-*, underlying the Latin present stem, and *?*g/ĝléu̯bʰ/g/ĝléu̯bʰ-s-*, underlying the Latin perfect stem), which are displayed on the right side of the picture. The PIE root (**g/ĝléu̯bʰ-*) is linked to both of them.[26] In the central part of the graph is the `lemonEty:Etymology` class, connected with the lexical entry, the PIE root and the two etymological links.[27]

4. Case Studies

Thanks to the creation of a total of 385 lexical entries and to their linking to the LB, the etymological information provided by the LIV was included in LiLa. The integration of the LIV's information into LiLa allows to put it in relation to that provided by the other resources that are part of the KB. The RDF data can be queried by means of the SPARQL query language.[28]

Querying the LIV in LiLa allows to enhance the quality of research in the field, by providing new insights about the relations of attested Latin word forms with reconstructed PIE roots and stems. This section illustrates two case studies made possible by the interoperability of the LIV with other resources in LiLa: the first use case regards the inflection types of the Latin verbs inherited from PIE, while the second one investigates the derivatives of PIE stems in Classical Latin.

4.1. An Investigation about the Lemmas' Inflection Types

When investigating the etymological relationship holding between Latin verbs and their ancestors in PIE, a question that emerges regards their inflection type. In particular, some inflection types seem to be more common among Latin verbs that are inherited from PIE, and less common among verbs that cannot be traced back to PIE stems (Weiss, 2020). The linking of the LIV to the LB can be effectively exploited to answer this question, as it

`ontolex/morph`).

[22] https://lila-erc.eu/lodview/ontologies/prinparlat/StemType.

[23] http://www.w3.org/ns/lemon/vartrans#lexicalRel.

[24] The perfect form, provided by *PrinParLat* is linked via the property `morph:consistsOf` (https://ontolex.github.io/morph/consistsOf), while the present form, created from scratch, is linked via the property `ontolex:lexicalForm`. This does not constitute an inconsistency, rather it is a choice imposed by economic reasons: in fact, whenever a relation is already expressed by a property (in this case `consistsOf`), it is not necessary to represent it again with another one (`lexicalForm`), since this would result in redundancy.

[25] The etymological links are connected with the source element and the target element via the properties `lemonEty:etySource` (http://lari-datasets.ilc.cnr.it/lemonEty#etySource) and `lemonEty:etyTarget` (http://lari-datasets.ilc.cnr.it/lemonEty#etyTarget), respectively.

[26] The property that links the PIE stems and the PIE root is `vartrans:lexicalRel`, mirroring the relationship between the lexical entry and the Latin stems.

[27] The lemonEty ontology defines specific properties to link the `Etymology` to these elements, namely `lemonEty:etymology` (http://lari-datasets.ilc.cnr.it/lemonEty#etymology), `lemonEty:etymon` (http://lari-datasets.ilc.cnr.it/lemonEty#etymon) and `lemonEty:hasEtyLink` (http://lari-datasets.ilc.cnr.it/lemonEty#hasEtyLink).

[28] The SPARQL queries performed to obtain the results presented in this paper can be found at https://github.com/CIRCSE/SPARQL-queries-LIV.

allows to quickly and easily identify the inflection type of each Latin verb listed in the LIV. More precisely, the LB records information about the inflection type of each lemma, represented using the property `lila:hasInflectionType`.[29] By counting the number of lemmas for each verbal inflection type, it is possible to compare the predominant inflection types of the entire LB (table 1) with those of the Latin verbs listed in the LIV (table 2).

Inflection Type Label	Number of lemmas
First conjugation	9530
Third conjugation	3398
First conjugation deponent	1019
Fourth conjugation	922
Second conjugation	823

Table 1: The inflection types of the LiLa LB.

Inflection Type Label	Number of lemmas
Third conjugation	172
Second conjugation	80
First conjugation	28
Fourth conjugation	19
Third conjugation deponent	16

Table 2: The inflection types of the Latin lexical entries in LIV.

As Tables 1 and 2 show, the distributions of the inflection types in the LB and in the LIV are very different. In the LB, the first conjugation is predominant, and only around one third of all verbs belong to the third conjugation. Among the Latin lexical entries in the LIV, however, the proportion is reversed: the number of third conjugation verbs is six times higher than that of first conjugation verbs. These data quantitatively confirm what is stated in Michael Weiss standard work, the *Outline of the historical and comparative grammar of Latin* (Weiss, 2020): "the 3rd and 4th Conjugations […] are the main repository of present stem formations inherited from Proto-Indo-European" (p. 404).

Since the information regarding the various PIE stem types was included in the modelling of the LIV (as described in Section 3), it is possible to refine the SPARQL query, and consequently to extend the investigation, by taking this information into account. In particular, for each inflection type, it is possible to count the number of lemmas reflecting a certain PIE stem type. The *Outline of the historical and comparative grammar of Latin* (Weiss, 2020) gives detailed information about the sources of each Latin conjugation. For instance, PIE so-called causative-iterative and iterative stems are usually reflected in Latin by second conjugation verbs (p. 403). By restricting the results of the query to the lemmas derived from a determined stem type, it is possible to quantitatively confirm this statement.

Inflection Type Label	Number of lemmas
Second conjugation verb	5
First conjugation deponent verb	1
First conjugation verb	1

Table 3: The inflection types of the Latin reflexes of LIV causative-iterative stems.

Inflection Type Label	Number of lemmas
Second conjugation verb	13
First conjugation verb	6
Third conjugation verb	2
First conjugation deponent verb	1

Table 4: The inflection types of the Latin reflexes of LIV iterative stems.

Tables 3 and 4 show the results of the queries for the causative-iterative stems and for the iterative stems respectively. As expected, the second conjugation is predominant in both cases. These results can be considered statistically significant, since the p-value, indicating the inter-dependence between the inflection type and the PIE stem type, was calculated to be lower than 0.05. The queries performed to obtain these results are simple, but give an empirical confirmation of what is stated in the *Outline of the historical and comparative grammar of Latin*.

4.2. An Investigation about the PIE Stem Types and their Derivatives

A further research question relevant to Historical Linguistics that may be investigated with the aid of the LIV's linking in LiLa, is whether the derivatives of Latin verbs that may be traced back to PIE stems feature specific affixes depending on their underlying PIE stem type. The derivational information that is recorded in the LiLa KB can be exploited to answer this question, too. Indeed, the lemmas of the LB are linked via the properties `lila:hasBase`,[30] `lila:hasPrefix`[31]

[29] http://lila-erc.eu/ontologies/lila/hasInflectionType.

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

[31] http://lila-erc.eu/ontologies/lila/hasPrefix.

and `lila:hasSuffix`[32] to their derivational bases, their prefixes and their affixes, respectively. By putting this information in relation to the LIV data, it is possible to answer the question previously outlined.[33]

The query counts the number of LiLa lemmas that are derived by means of a specific Latin affix and whose Latin base may be traced back to a specific PIE stem type. What emerges from the results is that some of the most frequent prefixes and suffixes in the entire LB are also the most frequent in the LIV derivatives. However, some affixes that are not in the top five ranking of the LB appear in the first five positions for the derivatives of certain PIE stems.

Prefix Label	Number of lemmas
con-	1992
e(x)-	1438
in (negation)-	1346
de-	1146
in (entering)-	1131

Table 5: The five most frequent prefixes in the Classical Latin lemmas of the LB.

Prefix Label	Number of lemmas
in (negation)-	12
prae-	10
in (entering)-	10
pro-	10
con-	7

Table 6: The five most frequent prefixes in the Classical Latin lemmas reflecting PIE desiderative stems.

Prefix Label	Number of lemmas
con-	24
ad-	21
e(x)-	19
re-	14
pro-	13

Table 7: The five most frequent prefixes in the Classical Latin lemmas reflecting PIE fientive stems.

[32] http://lila-erc.eu/ontologies/lila/hasSuffix.

[33] As described in Section 2, the derivational information recorded in LiLa currently regards only a subset of the Medieval Latin lemmas: for this reason, the results of all the queries including derivational information were restricted to Classical Latin lemmas only.

For instance, the prefix *pro-* is not among the most frequent of the LB (Table 5), but it is the fourth most frequent prefix for the derivatives of Latin lemmas reflecting PIE desiderative stems (Table 6) and in fifth position for the derivatives of Latin lemmas reflecting PIE fientive stems (Table 7). With regard to the suffixes, an interesting example of the same phenomenon is *-ment*, which is not among the top five suffixes of the LB (Table 8), but is in the fourth position for the derivatives of Latin reflexes of PIE fientive stems (Table 9). These data point to a close association between Latin affixes and specific PIE stem types. Thus, the queries performed open new perspectives about the relation between the Latin affixes and the derivatives of the PIE stems, suggesting that the semantic meaning carried by the stem influenced the choice of the affix involved in the derivational process. Indeed, each PIE stem type originally encoded a specific grammatical or lexical aspect, that is, they expressed the duration or the manner of the action (Meier-Brügger, 2003, pp. 164 ff.), as do the various prefixes and suffixes used in Latin to derive new words. This hypothesis may be further investigated since these results can be considered statistically significant: indeed, the p-value indicating the inter-dependence between the affixes involved in the derivational process and the PIE stem type underlying the lemma was calculated to be lower than 0.05.

Suffix Label	Number of lemmas
-(t)io(n)	2961
-(t)or	1837
-ari	1449
-(i)t	1381
-i	1258

Table 8: The five most frequent suffixes in the Classical Latin lemmas of the LB.

Suffix Label	Number of lemmas
-sc	63
-id	30
-ul	18
-ment	17
-(i)t	17

Table 9: The five most frequent suffixes in the Classical Latin lemmas reflecting PIE fientive stems.

To delve more into the matter, it is possible to calculate the percentage of the presence of each affix in the derivatives of Latin lemmas reflecting PIE stems compared to its total occurrences in the LB. The results show that a good part of the Classical Latin derivatives may be traced back to PIE

present stems: more precisely, with regard to both prefixes and suffixes, the percentage of derivatives that can ultimately be traced back to a PIE present stem often exceeds the threshold of 50%. As an example, table 10 shows the first five results for the suffixes involved in the derivational processes concerning Latin reflexes of PIE present stems.

Suffix label	Number of lemmas	Percentage
-(i)t	922	66,76%
-(i)es	80	57,55%
-(t)ur	127	55,22%
-or	92	55,09%
-men/min	171	50,89%

Table 10: The suffixes of Latin derivatives reflecting PIE present stems and their percentage on the total.

Suffix label	Number of lemmas	Percentage
-id	125	34,92%
-sc	103	15,37%
-(i)t	68	4,92%
-i	57	4,53%
-(t)io(n)	68	2,30%

Table 11: The suffixes of the derivatives descending from a PIE essive stem and their percentage on the total.

On the other hand, for the other PIE stem types, the situation is different: they usually cover less than 10% of the derivatives formed with a specific affix, and sometimes their percentages do not even reach the frequency threshold.[34] However, there is one outstanding result: the 34,92% of the LB's lemmas containing the suffix *-id* is derived from a Latin reflex of a PIE essive stem. This suffix is over two times more frequent than the second-ranked one, *-sc*, pointing to a special relation between the Latin adjectives in *-idus* and the Latin reflexes of PIE essive stems.

This special relation may be understood within the context of the so-called Caland system (Rau, 2009; Nussbaum, 1999). This PIE system consisted in a set of formal and semantic relationships between words, based on the alternation of specific affixes. The words involved were not derivatives of each other: rather "the word formation process is called recategorisation, i.e. the part of speech changes, but not the semantic content" (Balles, 2003, p. 10, my translation). The system has been inherited by many IE languages, including Latin. In the latter, however, it was remodelled following language-specific linguistic patterns. In particular, a Latin set of Caland formations usually features an adjective (e.g. *calidus* 'hot' or *liquidus* 'fluid'), a noun (e.g. *calor, -ōris* 'heat' or *liquor, -ōris* 'fluidity'), an essive verb (e.g. *caleō, -ēre* 'to be hot' or *liqueō, -ēre* 'to be fluid'), an inchoative verb (e.g. *calēscō, -ere* 'to become hot' or *liquēscō, -ere* 'to become fluid') and a factitive verb (e.g. *calefaciō, -ere* 'to make hot' or *liquefaciō, -ere* 'to make liquid'). These *-idus* adjectives and essive verbs, which have long been recognized as part of the Caland system in Latin, exactly correspond to the derivatives in *-id* and the Latin reflexes of PIE essive stems identified thanks to the LIV's linking to the LiLa KB. The results of the queries thus quantitatively confirm a relation which has long been noted and discussed within Historical Linguistics: this suggests that other results may also provide relevant information, which may be used to demonstrate new substantial relationships between the Latin affixes and the PIE stems.

[34] The frequency threshold was set on the 1% of the total occurrences of the most common affix in the Classical Latin lemmas of the LB.

5. Conclusion and Future Work

The linking of the LIV to the LiLa KB provides new opportunities to explore its etymological data in relation to Latin. The queries and the results shown in this paper confirm that the etymological information included in the LiLa KB can be effectively exploited to acquire new information about the relationship between Latin and PIE lexical items. The queries discussed here could not have been performed without the LIV's linking to the LiLa KB. Thus, the publication of the LIV's etymological relationships as LOD increases the research possibilities in the field, while representing an enhancement of the etymological subset of LiLa and of the LLOD Cloud.

Indeed, the queries and the results discussed in the present paper exemplify only a few of the advantages that the LIV's linking may actually provide. LiLa contains resources that supply information with regard to syntax (e.g. *Latin Vallex 2.0* (CIRCSE, 2020c; Mambrini et al., 2021), morphology (e.g. *PrinParLat* (CIRCSE, 2023b)), semantics (e.g. the Lewis and Short dictionary (CIRCSE, 2021a)) and sentiment analysis (e.g. *LatinAffectus* (CIRCSE, 2020b; Sprugnoli et al., 2020), while also providing different textual corpora, both for Classical and for Medieval Latin. All these layers of information are interoperable with each other and with the LIV. Querying their interconnected data can have a concrete impact on the academic communities of Classicists and Historical Linguists, by allowing them to carry out investigations that were not possible before.

Moreover, two future challenges can be outlined. First, the LIV does not exclusively contain etymological information with regard to Latin, but actually details the etymology of lexical items in many other IE languages: by modelling their data with the same ontologies that we used, it will be possible to enlarge the etymological network and investigate the etymological relationships between several IE languages.

Secondly, the biggest challenge not only for the LIV and LiLa, but for all the linguistic resources published as LOD, will be their integration within the world of the so-called Big Data and Large Language Models (LLMs). LLMs (such as BERT (Devlin et al., 2018) or ChatGpt (Ouyang et al., 2022)) are the future of Computational Linguistics, since they can process huge amounts of raw text, without the need to learn patterns provided by previous annotations. They can achieve very good results in several tasks, such as question answering (Jiang et al., 2021), machine translation (Lewis et al., 2020) and text generation (Li et al., 2022). In this context, the future of annotated linguistic resources is uncertain, given that they may stop being necessary altogether. However, the shift from supervised models to unsupervised machine learning methods constitutes a radical change that cannot be faced without critical thinking: if no annotation is required, the linguist's expertise and the deep analysis of the linguistic data are not required either. The challenge will thus be to preserve the original analytical component of Computational Linguistics, while taking all the benefits that LLMs can offer. In particular, this can be achieved by incorporating the linguistic resources published as LOD into LLMs. The LOD resources are stored in the form of knowledge graphs (KGs): these are able to generate interpretable results and to perform symbolic reasoning (Zhang et al., 2021), thus providing a solution for some of the limitations of LLMs (Biever, 2023). In this view, the linguistic resources published as LOD will hopefully preserve their crucial and innovative role in the discipline by establishing a fruitful relationship with the LLMs: in fact, the quality of the structured data contained in these resources can be reused to fine-tune and provide external knowledge to the LLMs (Zhang et al., 2019; Liu et al., 2021), while also being useful to analyse their results and provide interpretability (Petroni et al., 2019). This will hopefully constitute an opportunity to enhance the LLMs' performance and continue to improve the machine's capabilities with human knowledge.

6. Acknowledgements

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

7. Bibliographical References

Irene Balles. 2003. Die lateinischen *idus*- Adjektive und das Calandsystem. *Indogermanisches Nomen. Derivation, Flexion und Ablaut. Akten der Arbeitstagung der indogermanischen Gesellschaft, Freiburg (2001)*, pages 8–29.

Celeste Biever. 2023. Chatgpt broke the turing test-the race is on for new ways to assess ai. *Nature*, 619(7971):686–689.

Valeria Irene Boano, Francesco Mambrini, Marco Carlo Passarotti, and Riccardo Ginevra. 2023. Modelling and publishing the *Lexicon der indogermanischen Verben* as linked open data. In *Proceedings of the Ninth Italian Conference on Computational Linguistics*.

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

Flavio Massimiliano Cecchini, 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 (CLiC-it 2020)*, Bologna.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022. Computational morphology with ontolex-morph. *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, 20-25 June 2022, Marseille, France*.

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

Michiel de Vaan. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Brill, Leiden and Boston.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*.

Margherita Fantoli, 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. European Language Resources Association.

Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway? toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Anas Fahad Khan. 2018. Towards the representation of etymological data on the semantic web. *Information*, 9(304).

Ora Lassila and Ralph R. Swick. 1998. Resource Description Framework (RDF) Model and Syntax Specification.

Charlton T. Lewis and Charles Short. 1879. *A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary*. Clarendon Press, Oxford.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, page 7871–7880.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*.

Eleonora Litta, Marco Passarotti, and Francesco Mambrini. 2019. The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019). 19-20 September 2019, Prague, Czechia*, pages 35–43, Prague, Czech Republic. Institute of Formal and Applied Linguistics, Charles University in Prague.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6418–6425.

Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Further with Knowledge Graphs. Studies on the Semantic Web 53*, Amsterdam. IOS Press.

Michael Meier-Brügger. 2003. *Indo-European linguistics*. De Gruyter, Berlin; New York.

Alan Nussbaum. 1999. **Jocidus*: an account of the latin adjectives in *-idus*. *Compositiones indogermanicae: in memoriam Jochem Schindler*, pages 377–419.

OED. *Oxford English dictionary online*. Oxford University Press, Oxford.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

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*, 58(1):177–212.

Matteo Pellegrini. 2023. Flexemes in theory and in practice. *Morphology*, pages 1–35.

Matteo Pellegrini, Marco Passarotti, Eleonora Litta, Francesco Mambrini, Giovanni Moretti, Claudia Corbetta, and Martina Verdelli. 2022. Enhancing derivational information on latin lemmas in the lila knowledge base. a structural and diachronic extension. *The Prague Bulletin of Mathematical Linguistics*, (119):67–92.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller,

and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Jeremy Rau. 2009. *Indo-European nominal morphology: The decads and the Caland system*. Innsbrucker Beiträge zur Sprachwiss, Innsbruck.

Helmut Rix, ed. 1998. *LIV. Lexikon der indogermanischen Verben. Die Wurzeln und ihre Primärstammbildungen*. Reichert Verlag, Wiesbaden.

Helmut Rix, ed. 2001. *LIV. Lexikon der indogermanischen Verben. Die Wurzeln und ihre Primärstammbildungen*, 2nd edition. Reichert Verlag, Wiesbaden.

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

Michael Weiss. 2020. *Outline of the historical and comparative grammar of Latin*. Beech Stave Press.

Mark Wilkinson et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

Jing Zhang, Bo Chen, Lingxi Zhang, Xirui Ke, and Haipeng Ding. 2021. Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open*, 2:14–35.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

8. Language Resource References

CIRCSE. 2006-2024. *The Index Thomisticus Treebank*. CIRCSE Research Centre, ISLRN 105-545-284-528-2.

CIRCSE. 2018. *Word Formation Latin*. CIRCSE Research Centre. PID https://doi.org/10.5281/zenodo.1492327.

CIRCSE. 2019-2024. *The LiLa Lemma Bank*. CIRCSE Research Centre. PID https://doi.org/10.5281/zenodo.8300851.

CIRCSE. 2020a. *Etymological Dictionary of Latin and the Other Italic Languages*. CIRCSE Research Centre. PID https://zenodo.org/records/4147500.

CIRCSE. 2020b. *Latin Affectus*. CIRCSE Research Centre. PID https://doi.org/10.5281/zenodo.4022689.

CIRCSE. 2020c. *Latin Vallex 2.0*. CIRCSE Research Centre. PID https://doi.org/10.5281/zenodo.4032430.

CIRCSE. 2021a. *Charlton T. Lewis and Charles Short. 1879. A Latin Dictionary. Clarendon Press, Oxford*. CIRCSE Research Centre. PID https://github.com/CIRCSE/LewisShort.

CIRCSE. 2021b. *UDante Treebank*. CIRCSE Research Centre. PID https://github.com/CIRCSE/UDante.

CIRCSE. 2022. *LASLA corpus*. CIRCSE Research Centre. PID https://github.com/CIRCSE/LASLA.

CIRCSE. 2023a. *Lexicon der indogermanischen Verben*. CIRCSE Research Centre. PID https://github.com/CIRCSE/LIV.

CIRCSE. 2023b. *PrinParLat*. CIRCSE research centre. PID https://github.com/CIRCSE/PrinParLat.