Proceedings of the Linked Data in Linguistics Workshop @ LREC2022, pages 26–34
Marseille, 24 June 2022

Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Margherita Fantoli[1], Marco Passarotti[2], Francesco Mambrini[2], Giovanni Moretti[2], Paolo Ruffolo[2]
[1]Katholieke Universiteit Leuven, [2]Università Cattolica del Sacro Cuore
[1]Oude Markt 13, 3000 Leuven, Belgium. [2]Largo Gemelli 1, 20123 Milan, Italy
margherita.fantoli@kuleuven.be,
{marco.passarotti, francesco.mambrini, giovanni.moretti}@unicatt.it, paolo.ruffolo@posteo.eu

Abstract

This paper describes the process of interlinking the 130 Classical Latin texts provided by an annotated corpus developed at the LASLA laboratory with the LiLa Knowledge Base, which makes linguistic resources for Latin interoperable by following the principles of the Linked Data paradigm and making reference to classes and properties of widely adopted ontologies to model the relevant information. After introducing the overall architecture of the LiLa Knowledge Base and the LASLA corpus, the paper details the phases of the process of linking the corpus with the collection of lemmas of LiLa and presents a federated query to exemplify the added value of interoperability of LASLA's texts with other resources for Latin.

Keywords: Linguistic Linked Open Data, Corpora, Latin

1. Introduction

Scholars of Latin are particularly lucky when it comes to the availability of online linguistic resources. A long tradition of computational approaches and cutting-edge digital editing projects results today in an abundance of textual and lexical resources scattered on the web. Although high-quality linguistic resources are nowadays freely accessible online, in most cases they are stored in separate silos and enhanced with layers of linguistic annotation following different criteria and tagsets.

Among the several linguistic resources today available for Latin[1], the CIRCSE Research Center in Milan[2] and the LASLA laboratory in Liège[3] (Laboratoire d'Analyse Statistique des Langues Anciennes) have developed a number of manually validated lexical resources and annotated corpora. The CIRCSE has built, among others, the Word Formation Latin (WFL) derivational lexicon (Litta and Passarotti, 2019), a set of sentiment lexicons (Sprugnoli et al., 2020b) and a few syntactically annotated corpora, including the Index Thomisticus Treebank (IT-TB) (Passarotti, 2019) and the UDante Treebank(Cecchini et al., 2020). The LASLA has produced a manually verified lemmatized and morphosyntactically annotated corpus of more than 1.5 million words mainly belonging to Classical Latin literature (see Section 3).

As mentioned, one of the limitations that currently affect linguistic resources is their sparsity and diversity for what concerns data formats, annotation guidelines and sets of tags adopted. In order to overcome such limitation, the CIRCSE Research Center has developed the LiLa Knowledge Base, with the objective of making distributed linguistic resources for Latin interact through the application of the principles of the Linked Data paradigm (see Section 2).

In their work with digital resources for Latin, LASLA and CIRCSE share a large set of common features, but also show a number of differences. Each research center is dedicated to the development of high-quality, manually created or verified linguistic resources for ancient languages. They both endeavor to comply with the high-quality standards of existing – traditional – resources, such as dictionaries. Finally, both CIRCSE and LASLA combine interest for the lexical and the morphological/syntactic information encoded in texts and words.

However, since the Sixties the LASLA has mainly focused on annotating a corpus of Classical Latin and Ancient Greek literature, and has valued consistency and continuity with respect to internal criteria more than fitting the standards de facto built by the research community working on linguistic resources (like, for instance, those adopted by the Universal Dependencies initiative[4]). Moreover, the integration of Natural Language Processing (NLP) tools into the LASLA corpora (like the tagger Collatinus[5]) was always made with reference only to the LASLA schema of annotation.

Through the LiLa Knowledge Base, instead, CIRCSE supports the web-based interoperability between lexical and textual resources for Latin according to standards widely adopted in the Linguistic Linked Open Data community. Furthermore, the resources currently interlinked in LiLa include annotated corpora (like the IT-TB) that feature texts from the Medieval era, which are outside the chronological boundaries of the LASLA collection.

[1] For an overview of the linguistic resources currently available for Latin see (Passarotti et al., 2020).

[2] https://centridiricerca.unicatt.it/circse_index.html

[3] http://web.philo.ulg.ac.be/lasla/

[4] https://universaldependencies.org/

[5] https://outils.biblissima.fr/fr/collatinus-web/

In spite of the different approaches pursued by the two centers in the past, the idea of combining the high-quality textual data annotation of LASLA with the interoperability provided by LiLa's adoption of the Linked Data paradigm appears potentially very fruitful. With its dense network of other lexical and textual resources, LiLa is indeed capable of opening new avenues of research for scholars working on Latin texts, whose everyday work is strictly bound to the possibility of collecting empirical evidence from texts from different eras, genres and places.

As a consequence, LASLA and CIRCSE have decided to join their forces to interlink LASLA's Classical Latin texts with the LiLa Knowledge Base. This paper describes how such interlinking was performed. After introducing the LiLa Knowledge Base (Section 2) and the LASLA corpus (Section 3), the paper details the process of linking the texts into LiLa (Section 4) and presents a query that can be performed on the interlinked data as a way to exemplify the added value of interoperability of LASLA's texts with other resources for Latin (Section 5).

2. The LiLa Knowledge Base

The "LiLa - Linking Latin" project[6] aims to reach interoperability between the wealth of existing lexical and textual resources that have been developed in the last decades for Latin. One of the main problems that LiLa intends to solve is the fact that such resources and tools are often characterized by different conceptual and structural models, which makes it difficult for them to interact with one another.

To this goal, LiLa has undertaken the creation of an open-ended Knowledge Base, following the principles of the Linked Data paradigm[7]. All content involved or referenced in the linguistic resources connected in LiLa is made unambiguously findable and accessible by assigning an HTTP Uniform Resource Identifier (URI) to each data point. Data reusability and interoperability between resources are achieved by establishing links between different URIs and by using web standards such as: [a] the RDF data model, which is based on triples: (i) a predicate-property connects (ii) a subject (a resource) with (iii) its object (another resource, or a literal) (Lassila and Swick, 1998); and [b] SPARQL, a query language specifically devised for RDF data.

Furthermore, the LiLa Knowledge Base makes reference to classes and properties of already existing ontologies to model the relevant information. The main ones are POWLA for corpus data (Chiarcos, 2012), OLiA for linguistic annotation (Chiarcos and Sukhareva, 2015), and Ontolex-Lemon for lexical data (Buitelaar et al., 2011; McCrae et al., 2017).

Within this framework, LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and NLP tools. Consequently, the architecture of the LiLa Knowledge Base is highly lexically based (Figure 1), grounding on a simple, but effective assumption that strikes a good balance between feasibility and granularity: textual resources are made of (occurrences of) words ("tokens"), lexical resources describe properties of words (in "lexical entries"), and NLP tools process words (producing "NLP outputs")[8].

The core of the Knowledge Base is the so-called Lemma Bank,[9] a collection of about 200,000 Latin lemmas – defined as the canonical form of a lexical item, i.e. its citation form – taken from the database of the morphological analyzer LEMLAT (Passarotti, M. et al., 2020) (Passarotti et al., 2017). Interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma.

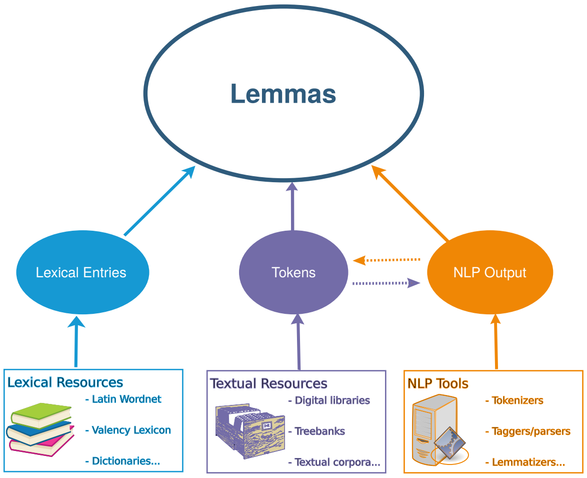

Figure 1: The architecture of LiLa

3. LASLA

The Latin section of the LASLA corpus contains nowadays 2,500,000 semi-automatically annotated tokens: for every token of the corpus, the automatic annotation has been manually verified by a Latin scholar. A significant part of the corpus (more than 1.7M tokens) will be soon released for free download.

The LASLA corpus features mainly Classical Latin literary texts, both poetical and in prose[10]. The earliest author in the corpus is Plautus (III-II century BC) and the latest Apuleius (II century AD; to be released soon). The data available for sharing and linked to the LiLa

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

[7] `https://www.w3.org/DesignIssues/LinkedData.html`

[8] In Figure 1, the arrows going from and to the node for "NLP Output" represent the fact that tokens that are the output of a specific NLP tool (a tokenizer) become the input of further tools (like, for instance, a syntactic parser).

[9] `http://lila-erc.eu/lodview/data/id/lemma/LemmaBank`.

[10] `http://web.philo.ulg.ac.be/lasla/textes-latins-traites/`.

Lemma	LASLA Index	French
cubitus	1	Le coude (the elbow)
cubitus	2	L'action d'être couché (the act of lying down)

Table 1: Example of an homographic lemma from the LASLA dictionary

Knowledge Base include 130 works of 21 different authors.

The linguistic information available in the corpus consists of lemmatization, morphological tagging and an additional syntactic layer for verbs. The choice of the lemma in the LASLA corpus is based on the Forcellini dictionary (Facciolati, J. and Forcellini, E., 1771). A sequence of alphanumerical tags encodes the morphological description of the word form and some syntactic features[11]. The annotation guidelines are those provided by (Philippart de Foy, 2014).

A partial list of the lemmas included in the LASLA texts is available in the so-called LASLA dictionary[12]. The LASLA dictionary is an essential resource to address homographic lemmas, which are distinguished in the dictionary by the use of an index. In particular, the index "N" is assigned to proper nouns and "A" is assigned to adjectives derived from proper nouns (e.g. *Romanus*, "Roman, of Rome"). If one of two homographic lemmas in the LASLA dictionary is a proper noun (or an adjective derived from a proper noun), the index "N" (resp. "A") allows to disambiguate. For instance, the lemma *urbs* meaning the city of Rome is assigned index "N", whereas the lemma *urbs* meaning a generic city is not assigned any index. Similarly, in case one of two homographic lemmas is a proper noun and the other is a derived adjective (e.g. the person's name *Latinus* and the adjective *Latinus* "of the Latium"), they are assigned respectively indices "N" and "A". If none of the homographic lemmas is a proper noun, they are simply distinguished through sequential numbers.

The LASLA dictionary provides also further information, like the French translation of Latin words, to help the annotators of the corpus and its users in choosing the right homographic lemma (see Table 1).

The creation of the corpus started in 1961 with the foundation of the LASLA, and is still going on nowadays. Textual annotation is performed both via an online semi-automatic web-interface where annotators choose, for every word, the correct analysis among those proposed by the software, and through a standalone tagger with a post-correction interface (Verkerk et al., 2020).

The LASLA corpus is searchable along the different linguistic categories on the Opera Latina website[13]. In addition, the HyperbaseWeb portal, developed at the "UMR 7320 : Bases, Corpus, Langage" of the Université de Nice[14], allows to search and perform some statistical analysis of the corpus as a whole, as well as of specific thematic subsections of it (e.g., historiographical, poetic, dramatic texts).

4. LASLA in LiLa

This section details the process undertaken to perform the linking of the LASLA corpus in the LiLa Knowledge Base.

As said, the LiLa project adopts the assumption that the lemma, i.e. the form of a word's inflectional paradigm that is used to index a lexical entry in a dictionary or to lemmatize a corpus, is a gateway to connect the different resources. The Ontolex-Lemon ontology provides a convenient model to formalize this assumption and to express most of the relevant properties of lemmas used in standard Latin lexicography or in the practice of corpus annotation (McCrae et al., 2017).

The lemma in LiLa is defined as a subclass of the class `Form` of Ontolex,[15] which includes all forms that are potentially used (or usable) as citation forms for lexical entries or to lemmatize corpus tokens. Each of them is defined by a series of object and data properties. In particular, the Ontolex-Lemon data property 'written representation' (WR)[16] registers the different spellings or graphical variants of one lemma[17]. All forms in Ontolex-Lemon must have at least one WR; lemmas can also have a special type of representation that we define in the LiLa ontology, namely the 'prosodic representation'[18], where we register the quantity (long or short) of the form's vowels. Vowel quantity (which is generally not marked in the corpora) is often crucial to disambiguate words, such as *pŏpulus* ("people", with short *o*) and *pōpulus* ("poplar", with long *o*)[19].

[11] More precisely, the LASLA corpus indicates whether a verb belongs to a main clause or a subordinate clause. For subordinate verbs, it shows which is the subordinating element that introduces the clause.

[12] `http://cipl93.philo.ulg.ac.be/LaslaEncodingInitiative/Files/lasladic.pdf`. The list has not been updated yet with the lemmas found in the latest additions to the LASLA corpus.

[13] `http://web.philo.ulg.ac.be/lasla/opera-latina/`

[14] `http://hyperbase.unice.fr/hyperbase/`

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

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

[17] Note however that, whenever two or more spellings entail also a change in the inflectional paradigm of a word, these are not registered as WRs of the same lemma; instead, we create as many different lemmas as we need to account for all the inflectional paradigms (Passarotti et al., 2020).

[18] `https://lila-erc.eu/lodview/ontologies/lila/prosodicRepresentation`

[19] `http://lila-erc.eu/data/id/lemma/118463`; `http://lila-erc.eu/data/id/lemma/118501`

The Part of Speech (POS) and the inflectional category are other properties that provide decisive contributions to disambiguation. For this reason, all the forms in the Lemma Bank of LiLa are annotated with tags from the Universal POS tagset (Petrov et al., 2012) and are classified according to their inflectional paradigm. The list of inflectional classes is inspired by the traditional grammars of Latin and is the one used by the morphological analyzer LEMLAT (Passarotti et al., 2017).

Like all annotated corpora, LASLA registers lemmatization with a string identifying the canonical form attached to the token. For instance, the token *uiuamus* ('let us live', 1st-person plural subjunctive present) is lemmatized with the string 'uiuo'; the same goes also for POS tagging, where the tag (e.g. VERB) is also encoded as a string. Like several other corpora, the string used for lemmatization in LASLA occasionally includes disambiguation indexes: in the case of *populus* mentioned above, LASLA uses the indexes 1 and 2 to distinguish between "the people" ("populus1") and "the poplar" ("populus2").

'Linking' a corpus to LiLa means converting the string-based annotation recorded for each corpus token into a link to a lemma in the Lemma Bank. In turn, the process entails the identification of the correct lemma corresponding to the lemmatization string registered in the corpus. The POS tag and the inflectional class attached to the tokens, when this information is available as it is the case with LASLA, are features that can help in disambiguating many of the cases where the lemma string is not sufficient. Such workflow implies three steps:

1. to align the POS tagset and the inflectional classes used in the source corpus (LASLA) and in the LiLa Lemma Bank;
2. to align the indexed strings of the homographic lemmas in the source to the correct lemma in the Lemma Bank;
3. to match the lemma and POS tag strings in the source with the WRs and the POS tags in the Lemma Bank to identify the candidates.

4.1. Matching POS and Inflectional Classes

Most of the LASLA POS tags, described in the documentation of the LASLA dictionary, show a 1:1 correspondence with those of LiLa, as detailed in Table 2

LASLA POS	LiLa POS
Verb	VERB
Adjective	ADJ
Adverb: generic, relative, interrogative, negative, int/neg	ADV
Preposition	ADP
Substantive	NOUN
Proper noun (i.e. Noun + Index N)	PROPN
Coordinating Conjunction	CCONJ
Subordinating Conjunction	SCONJ
Interjection	INTJ
Numeral	NUM

Table 2: 1:1 mapping between LASLA and LiLa POS

Although the great majority of the lemmas labeled with these POS tags in the LASLA corpus are assigned the corresponding Universal POS tag in the Lemma Bank, some exceptions do hold, due to the different criteria of application of POS tags in the two resources. For instance, the names of populations are tagged as proper nouns in LASLA, while they are assigned the POS tag for adjectives in the Lemma Bank (see Section 4.3.1 for the treatment of these exceptions).

A particularly compelling case of mismatch between the POS tags of LASLA and those of LiLA is represented by those words that are labeled as pronouns in LASLA and either as Determiners (DET) or as Pronouns (PRON) in LiLa. For instance, words in the category "Indefinite Pronoun" in LASLA can be tagged either as PRON in LiLa (e.g. *aliquis*, "somebody"), or as DET (e.g. *aliquantulus*, "small, little"). The issue is closely related to the fact that the difference between the tags PRON and DET in the Universal POS tagset is still fuzzy. The Universal tag DET is assigned to those words "that modify nouns or noun phrases and express the reference of the noun phrase in context"[20]. Pronouns, instead, are defined as terms that "substitute for nouns or noun phrases, whose meaning is recoverable from the linguistic or extralinguistic context"[21]. However, the UD guidelines report that it is not simple to draw a line between DETs and PRONs[22].

In Latin, as well as in several other languages, some words can be used both as DET and PRON according to the definitions given above. For instance, the lemma *is* can be used both as PRON (meaning "that person") and as DET (e.g., *eo loco*, "that place"). The LASLA tagset conflates both categories under the label "Pronoun", which covers both usages. Such uncertainty is reflected in the documentation provided by the LASLA dictionary, where the label "Pronoun" alternates with "Pronoun/Adjective". In Lila, instead, the tag DET is assigned when both usages are possible (as with *is*), while PRON is assigned to those words that can be used only as pronouns (like *aliquis*, "somebody", which has a distinct adjectival form: *aliqui*).

To sum up, the tags "Pronoun" and "Pronoun/Adjective" of LASLA were matched with either DET or PRON of the Lemma Bank.

As for the inflectional classes, the tagsets of LASLA

[20] `https://universaldependencies.org/u/pos/DET.html`

[21] `https://universaldependencies.org/u/pos/PRON.html`

[22] "It is not always crystal clear where pronouns end and determiners start. [...] Language-specific documentation should list all determiners (it is a closed class) and point out ambiguities, if any" (`https://universaldependencies.org/u/pos/DET.html`).

and LiLa can be easily aligned, except for the names of Greek origin following an irregular inflection. While LiLa makes use of a separate tag for the "irregular" nouns of each declension (like, for instance, for the second declension irregular nouns), the LASLA tagset includes two broad categories "Anomalous" and "Greek declension" covering the nouns of any declension. As a consequence, in these cases, there is a many-to-many correspondence between the two tagsets. In addition, in the LASLA corpus many words are alternatively tagged as Greek declension and as "regular" declension based on the inflection of single word forms. For instance, the proper noun *Orestes* is assigned in the LASLA corpus alternatively the tag for the third declension in the case of forms that are inflected according to the paradigm used also for any other Latin word (e.g. accusative *Orestem*), and that for the Greek declension in the case where the Greek ending is used (as in the accusative form of Greek origin *Oresten*). While linking the two resources, the lemmas affected by this issue were treated manually (see Section 4.3.2).

4.2. Handling Homography

As said, homography is addressed in LASLA by using indices. Information that allows readers to identify the indexed lemmas is provided in the LASLA dictionary. For instance, there are two third declension neuter nouns *tempus* in Latin, respectively meaning "time" and "temple" (the side of the head near the eye), as it is recorded in the LASLA dictionary. These words are identified respectively as "tempus1" and "tempus2" in the LASLA corpus.

The work to link these strings to the correct entry in the Lemma Bank can only be made manually, by matching the lexicographic information in the LASLA dictionary with that provided by the array of lexical resources currently linked to LiLa Knowledge Base.

For instance, information that allows to disambiguate the two nouns with WR "tempus" is found in the WFL lexicon linked to LiLa (Litta et al., 2019), which assigns to each of the two lemmas *tempus* in question its respective derivatives. The information provided by WFL proves particularly helpful when two homographic lemmas formed with the same prefix are derived from two different base verbs. For instance, this is the case of the two verbs *contingo* in the Lemma Bank, both formed with prefix *cum* and respectively meaning "to happen" and "to dye". WFL informs that one verb derives from *tango* ("to touch") and the other from *tingo* ("to wet, moisten, bathe").

A second resource exploited to get the information that leads to correct disambiguation of homographic lemmas is the Latin-English dictionary (Lewis, Ch. and Short, Ch., 1879) (L&S), which is now partially linked to the LiLa Knowledge Base (Mambrini et al., 2021). The definition and translation provided by L&S can be used to distinguish homographic lemmas. One example is given by the two homographic verbs of the third conjugation *sero*. In LiLa, the link with the dictionary provides a translation of the two lemmas ("to sow, to plant", "to join and bind together"). The LASLA dictionary distinguishes them using the verbal paradigm, i.e. by indicating that the perfect indicative is *serui* for one lemma ("sero2"), and *seui* for the other ("sero3"). In total, 2,118 LASLA homographic lemmas were linked manually to the LiLa Lemma Bank by exploiting the linguistic information found in the two resources.

Type of Match	No of Lemmas
1:1	19,543
1:0	3,369
1:N	932
TOTAL	23,844

Table 3: Number of lemmas per type of match (LASLA to LiLa)

4.3. Linking LASLA to the Lemma Bank

Once that the POS tags used by LASLA and LiLa were aligned and the homographic lemmas were manually matched, we proceeded to link all the other, non-homographic lemmas of LASLA to those of the LiLa Lemma Bank. The linking was based on: [a] the form of the lemma from LASLA and the value(s) of the Ontolex-Lemon data property 'written representation' from LiLa, and [b] their POS. The results of the match are shown in Table 3.

The one-to-one matches were considered validated, as one LASLA lemma matches both the form and the POS of exactly one LiLa lemma. The steps taken to perform the linking of the one-to-zero and the one-to-many matches are described in the following Sections.

4.3.1. One-to-zero Matches

First we considered the 3,369 LASLA lemmas where no match for the tuple $(form, POS)$ was found with the (WR, POS) tuples of the LiLa Lemma Bank.

A relevant source of mismatch was the fluctuating distinction between nouns and proper nouns in the two resources. For this reason, we decided to conflate the two categories. After conflation, we were able to match 298 LASLA lemmas to exactly one LiLa lemma, while 25 lemmas showed a one-to-many correspondence and 3,046 lemmas still remained unmatched. For instance the lemma *babylonicum*, "textiles from Babylonia", originally tagged as proper noun in LASLA and noun in LiLa, was matched correctly after conflation.

Out of the 25 one-to-many matches, 14 were once again disambiguated automatically on the basis of their inflectional class. For instance, the third declension neuter noun *bacchanalia* of LASLA matched with two neuter proper nouns *bacchanalia* in the Lemma Bank, respectively of the third and of the second declension[23].

[23] http://lila-erc.eu/data/id/lemma/405; http://lila-erc.eu/data/id/lemma/404

Based on the correspondence between the tagsets for inflectional classes used by the two resources, the match with the latter was discarded.

For the remaining 11 lemmas, it was necessary either to proceed with manual disambiguation or to add the missing lemmas in the Lemma Bank. The former was the case of e.g. the proper noun *annus* of LASLA, which matched with the two nouns *annus* in LiLa, one meaning "year" and the other, more commonly spelled *anus*, meaning "posteriors"[24].

To handle the remaining one-to-zero 3,046 lemmas, we removed the constraint on the POS, thus extracting the lemmas that matched exclusively on the level of LASLA form and LiLa WR. As a result, 1,031 lemmas were matched automatically with exactly one LiLa lemma and were manually validated. 59 lemmas showing a one-to-many match were manually disambiguated. For instance, the LASLA adverb *attamen* ("but yet") corresponds to the subordinating conjunction *attamen* in the Lemma Bank and not to the noun *attamen*[25], which is a Late Latin term meaning "impurity".

Finally, the 1,956 lemmas still remaining were the ones showing no match between the form of the lemma in LASLA and a WR in the Lemma Bank. These cases were tackled by enriching the LiLa Lemma Bank with the missing lemmas (mostly, proper nouns).

4.3.2. One-to-many Matches

This category includes 932 lemmas of LASLA that yield a positive match with more than one lemma in the Lemma Bank, based on the WR and the POS. For instance, the verb *alleuo* ("to lift up") in LASLA can be paired with two verbs with WR *alleuo* in LiLa (respectively meaning "to lift up" and "to make smooth"[26]).

By adding the constraint of inflectional class, we improved the rate of 1:1 matches by 364. 460 still matched multiple lemmas, while 108 lemmas resulted in an empty match based on the new constraints. This set of no-matches is mostly caused by the problematic mapping of the Greek declension. These cases have been solved by manually validating the link to the lemma with the correct inflection class in LiLa.

The 460 remaining multiple matches are mainly due to two reasons. First, in several cases a lemma in LASLA was linked to two or more lemmas that are connected via the the symmetric property 'lemma variant', defined in the LiLa ontology.[27]. The property is used to connect forms of the same lexical item that fill different cells of the inflectional paradigm and can both be used alternatively as lemmas for that item (Passarotti et al., 2020). For instance, the LASLA lemma *specus* ("cave") matches both with the LiLa masculine/feminine lemma and with its neuter lemma variant[28]. As the use of the 'lemma variant' property makes the two forms practically equivalent, this case is not problematic.

The second source of ambiguous matches is the diachronic range covered by the LiLa's Lemma Bank. The LASLA corpus features Classical Latin texts only, whereas the Lemma Bank is built also over Late and Medieval Latin lexical resources, which might contain lemmas with the same POS and inflectional class of a Classical Latin lemma, but with different meaning. One example is given by the noun *conditor*: LASLA has only the Classical Latin lemma ("founder", from the verb *condo*), whereas LiLa includes also the Late Latin lemma ("the seasoner", from the verb *condio*), thus resulting in a case of homography. These matches were manually disambiguated.

Finally, for 4 lemmas showing a multiple match, we performed a manual disambiguation on the level of their single tokens[29]. In these cases, the LASLA corpus contains a single lemma for two LiLa lemmas that are homographic and cannot be distinguished on the basis neither of linguistic features (like the POS or the inflectional class), nor of formal features (like the plural vs singular form). Given that the distinction between the two LiLa lemmas is exclusively semantic, only the meaning of their single occurrences in the LASLA corpus can be used to link to the correct LiLa lemma.

4.4. Results

The publicly shared part of the Latin section of the LASLA corpus is now entirely linked to the LiLa Knowledge Base. In total, 1,738,435 tokens from LASLA are now connected to the LiLa Knowledge Base via the lemmas of the Lemma Bank. Manual linking by one expert annotator was necessary for 3,791 lemmas, for a total of ca. 50 hours of work. Figures 2 and 3 visualize some of the information attached to tokens from the corpus.

The LASLA corpus, its texts and the tokens are modeled using the POWLA ontology (Chiarcos, 2012). Figure 2 shows an example of a document (the philosophical dialogue "Of Friendship" (*De Amicitia*) by Cicero, pink node in the middle of the figure), i.e. one of the 130 works in the corpus. The document is subdivided in a series of structural units, that are grouped in three layers. The sentence and citation layers (light blue node on the top and bottom left) aggregate respectively all the sentences and the structural units (in this case, the numbered paragraphs) that make up the text.

[24] `http://lila-erc.eu/data/id/lemma/89129`; `http://lila-erc.eu/data/id/lemma/89365`

[25] `http://lila-erc.eu/data/id/lemma/91078`; `http://lila-erc.eu/data/id/lemma/32914`

[26] `https://lila-erc.eu/data/id/lemma/88348`; `https://lila-erc.eu/data/id/lemma/88385`

[27] `https://lila-erc.eu/lodview/ontologies/lila/lemmaVariant`.

[28] `http://lila-erc.eu/data/id/lemma/125318`; `http://lila-erc.eu/data/id/lemma/125319`

[29] *clauiger* ("club-bearing"), *insomnium* ("dream"), *myrrheus* ("of myrrh"), *propola* ("forestaller").

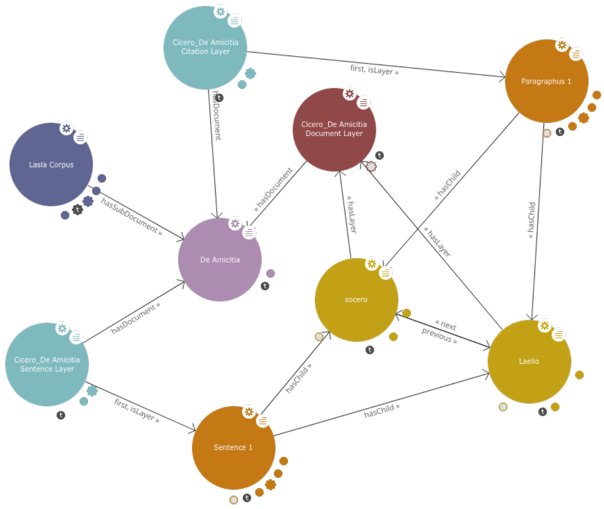

Figure 2: A LASLA token in LiLa: structural relations

The document layer (red node) links directly to the tokens. Two of them from the first paragraph, *socero* ("father in law") and the word that immediately precedes it (*Laelio*, "Laelius") are reproduced in the figure (yellow nodes).

Figure 3 represents some of the lexical information that the network of resources linked to LiLa allows to uncover for the same token *socero* from Cicero's *De Amicitia*. The lemma from the Lemma Bank is represented in the center (purple node), with the three WRs attested for the form (*socer*, *socrus*, *socerus*). The lemma is used as canonical form for entries in a series of lexical resources, two of which are reported in the Figure. In the bottom part, an entry in an etymological dictionary (Mambrini and Passarotti, 2020) accounts for the hypothetical origin of the word from the reconstructed Proto-Italic root **swekuro-* (de Vaan, 2008)[30]. The entry in the Latin-English dictionary L&S documents the two senses of the word, i.e. the main one ("father in law") and the transferred sense ("own child's father in law", properly *consocer*). Figure 3 visualizes the latter. On the top part of the Figure, we also see two lemmas linked to the same derivational family of *socer*: *consocrus* (a variant of *consocer*), composed with the preposition *cum* "with" (lit. "one who is father in law with"), and *prosocer* ("wife's grandfather").

5. Querying LASLA in LiLa

A query interface for the Lemma Bank can be accessed at `https://lila-erc.eu/query/`. Lemmas can be searched by string of characters (also using regular expressions), POS, affix, lexical base, inflectional category, and gender (for nouns). Results are provided both as data sheet and in a network-like graphical visualization. The entries in lexical resources and the tokens in corpora linked to each lemma in LiLa are reported as well[31].

Work	Author	Tokens	Neg x100
Medea	Seneca	5,700	7.4
Phaedra	Seneca	7,281	7.27
Phoenissae	Seneca	4,182	7.17
De Ira	Seneca	22,541	7.02
Thyestes	Seneca	6,321	6.64
De Constantia	Seneca	5,323	6.63

Table 4: Works with highest x100-frequency of negative words in LASLA

A SPARQL endpoint is also available at `https://lila-erc.eu/sparql/`, to query the contents of all the textual and lexical resources currently interlinked in the Knowledge Base. A number of pre-compiled queries is provided, including a query that counts the number of occurrences of those tokens from the LASLA corpus that are linked to a lemma of the Lemma Bank connected to a lexical entry provided with a negative polarity in the *Latin Affectus* lexicon (Sprugnoli et al., 2020a)[32].

As it is to be expected, this query returns words like: *hostis* "enemy" (2,109 occurrences), *mors* "death" (1,555), *periculum* "danger" (1,299), *grauis* "heavy, grievous" (1,232), or *malum* "evil" (1,220).

If we disaggregate the results by the different documents, we can rank the texts by the relative frequency of negative terms. Table 4 reports the 6 highest results, excluding fragmentary works that are too short to be meaningful. Not surprisingly, 4 out 6 slots are taken by tragedies of Seneca, the only tragic poet represented in the corpus. It is very interesting to note, however, that the other two works in the table, the moral treatises "On Wrath" (*De Ira*) and "On the Firmness of the Wise" (*De Constantia Sapientium*), are also authored by Seneca. The presence of the former text is certainly accounted for by the high occurrence of the word referring to the subject (*ira* "wrath", 242 occurrences, 1.07 x100 words). The latter treatise, on the other hand, is concerned with the ability of the Stoic philosopher to withstand abuse and suffering.

The first text not written by Seneca to figure in the list is only found at rank number 12; the work is "The Conspiracy of Catilina" (*De coniuratione Catilinae*, 5.45 negative words x100) by the historian Sallust, an essay dedicated to an infamous political plot that is certainly lavish of many sinister details about the protagonists and the moral decadence of the Roman society.

[30] As usual in historical linguistics, the star is used to mark unattested forms reconstructed with the help of the comparative method.

[31] The Turtle files of the resources interlinked in LiLa are available at `https://github.com/CIRCSE`.

[32] The pre-made queries can also be downloaded at `https://github.com/CIRCSE/SPARQL-queries`.

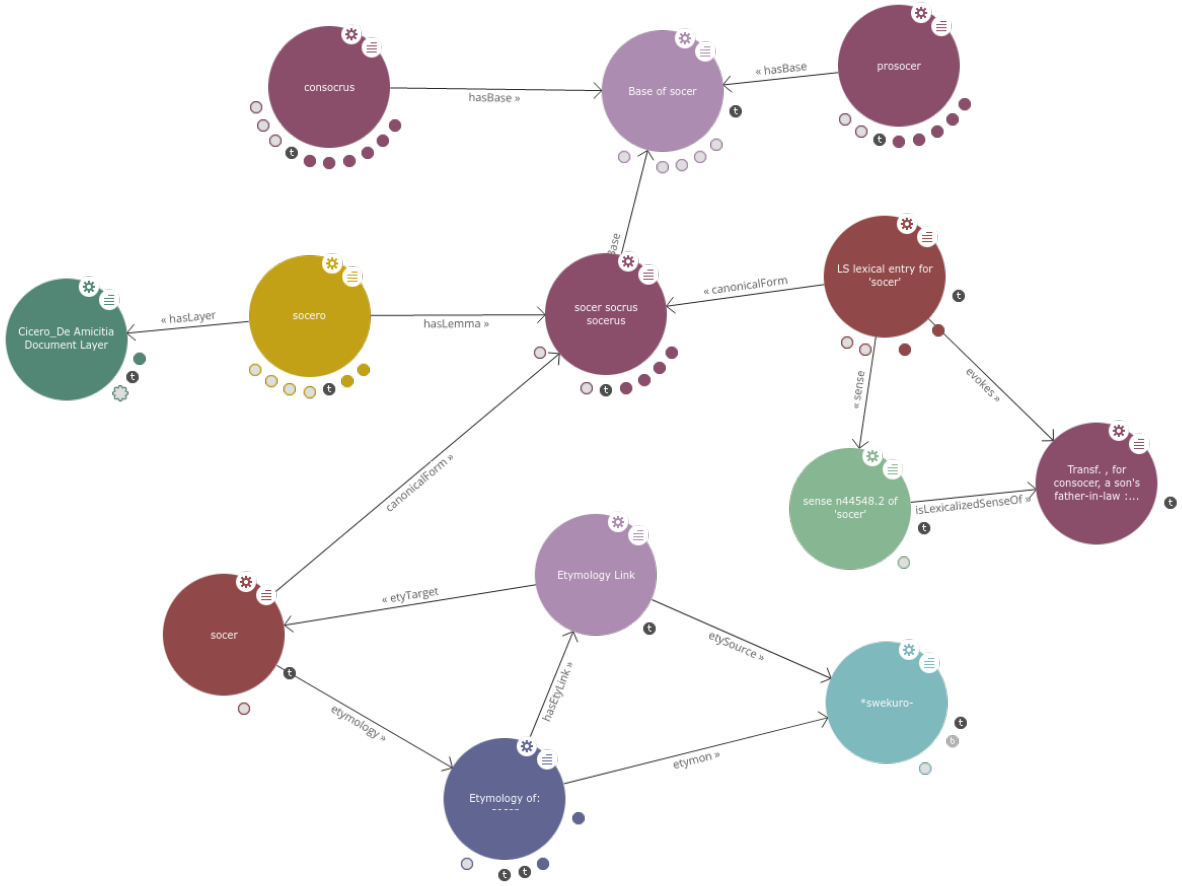

Figure 3: A LASLA token in LiLa: lexical information

6. Conclusion

The (soon) freely available portion of the LASLA corpus (ca. 1.7M tokens) is now linked to the LiLa Knowledge Base[33]. This result is a major achievement for both projects. As for LASLA, making its texts interoperable with other (kinds of) linguistic resources extends the degree of granularity of information extraction from the corpus, by focusing on words with specific lexical properties, and supports comparative research, by collecting relevant occurrences of words from corpora of different era and genre. As for LiLa, beyond enlarging the number and diversity of texts interlinked, the inclusion of the LASLA corpus will favor the use and dissemination of the Knowledge Base among Classicists, who are used to consider LASLA as one of the reference corpora of their community. Indeed, one of the objectives of the "LiLa - Linking Latin" project is to make digital linguistic resources and NLP tools finally become part of the everyday work of Classicists. Such objective can be achieved also by leading to a new level of accessibility those resources that are already well known in that community, to show how much more helpful they can become once made interoperable.

Not only is LiLa based on the principles of the Linked Data paradigm, but it reflects as much as possible the common grounds of the Linguistic Linked Open Data community. Such openness of the (meta)data of the resources interlinked through LiLa impacts the community of Classicists in that the entire process followed to collect the empirical evidence supporting their claims is made repeteable, replicable and reproducible (Cohen-Boulakia et al., 2017). Given the highly empirically-based nature of any linguistic, literary, or philological research on ancient languages, such an aspect is a very valuable added value, which is supposed to impact heavily how research in Classics is performed and published.

An open challenge to the community is represented by the management of the flow-back of information from the LiLa Knowledge Base to resources developed outside the Linked Data paradigm: for example, LASLA users would benefit from the integration of LiLa URIs in the current LASLA database and search interfaces.

7. 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.

[33] https://lila-erc.eu/lodview/data/corpora/Lasla/id/corpus

8. Bibliographical References

Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E., and Declerck, T. (2011). Ontology Lexicalization: The *lemon* Perspective. In *Proceedings of the Workshops-9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*, pages 33–36.

Cecchini, F. M., Sprugnoli, R., Moretti, G., and Passarotti, M. (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. CEUR-WS.org.

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

Chiarcos, C. (2012). POWLA: Modeling Linguistic Corpora in OWL/DL. In Elena Simperl, et al., editors, *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, pages 225–239, Berlin, Heidelberg. Springer.

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Le Bras, Y., Lemoine, F., et al. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, 75:284–298.

de Vaan, M. (2008). *Etymological Dictionary of Latin: and the other Italic Languages*. Brill, Amsterdam.

Lassila, O. and Swick, R. R. (1998). Resource Description Framework (RDF) Model and Syntax Specification.

Litta, E. and Passarotti, M. (2019). (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, et al., editors, *Words and Sounds*, pages 224–239. De Gruyter, Berlin, Boston, 12. Interrogable online at `http://wfl.marginalia.it/`.

Litta, E., Passarotti, M., and Mambrini, F. (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*, pages 35–43, Prague, Czechia, September. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Mambrini, F. and Passarotti, M. (2020). Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Globalex Workshop on Linked Lexicography. LREC 2020 Workshop*, pages 20–28, Paris. European Language Resources Association (ELRA).

Mambrini, F., Litta, E., Passarotti, M., and Ruffolo, P. (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*, Milan, Italy, December. CEUR.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex*, pages 587–597.

Passarotti, M., Budassi, M., Litta, E., and Ruffolo, P. (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.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. (2020). Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58:177–212.

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

Petrov, S., Das, D., and McDonald, R. (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).

Philippart de Foy, C. (2014). Lasla - nouveau manuel de lemmatisation du latin.

Sprugnoli, R., Moretti, G., and Passarotti, M. (2020a). Towards the modeling of polarity in a Latin knowledge base. In Alessandro Adamou, et al., editors, *WHiSe 2020 Workshop on Humanities in the Semantic Web 2020*, pages 59–70, Heraklion, Greece. CEUR.

Sprugnoli, R., Passarotti, M., Corbetta, D., and Peverelli, A. (2020b). Odi et amo. creating, evaluating and extending sentiment lexicons for latin. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3078–3086.

Verkerk, P., Ouvrard, Y., Fantoli, M., and Longrée, D. (2020). L.A.S.L.A. and Collatinus: a convergence in lexica. *SSL*, 1(LVIII):95–120.

9. Language Resource References

Facciolati, J. and Forcellini, E. (1771). *Totius Latinitatis lexicon*. Patavii: typis Seminarii.

Lewis, Ch. and Short, Ch. (1879). *A Latin Dictionary*. Clarendon Press.

Passarotti, M. et al. (2020). *LEMLAT 3.0*. CIRCSE, Università Cattolica del Sacro Cuore, and Zenodo, DOI:10.5281/zenodo.1492134, v. 3.0.