The Services of the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Marco Passarotti, Francesco Mambrini, Giovanni Moretti
Università Cattolica del Sacro Cuore
Milan, Italy
{marco.passarotti,francesco.mambrini,giovanni.moretti}@unicatt.it

Abstract

This paper describes three online services designed to ease the tasks of querying and populating the linguistic resources for Latin made interoperable through their publication as Linked Open Data in the LiLa Knowledge Base. As for querying the KB, we present an interface to search the collection of lemmas that represents the core of the Knowledge Base, and an interactive, graphical platform to run queries on the resources currently interlinked. As for populating the KB with new textual resources, we describe a tool that performs automatic tokenization, lemmatization and Part-of-Speech tagging of a raw text in Latin and links its tokens to LiLa.

Keywords: Latin, Linked Open Data, SPARQL

1. Introduction

Over the past two decades, the scientific community that focuses on Linguistic Linked Open Data (LLOD) has worked in two main closely connected directions. First, it has developed numerous vocabularies and ontologies for representing various types of linguistic (meta)data as Linked Open Data (LOD) (Khan et al., 2022). Secondly, these vocabularies and ontologies have been applied to (meta)data extracted from various linguistic resources for publishing them as LOD: the LLOD Cloud (Cimiano et al., 2020, 29-41)[1] provides a synoptic view of the resources published so far.

One challenge that the LLOD community must now address is to make the interoperable (meta)data of the resources easily accessible and fully exploitable. Such task is challenging as it must fit the needs and expertise of diverse user communities besides computer scientists and computational linguists. However, this challenge is unavoidable, especially because many semantic web technologies (like RDF, OWL or SPARQL) have a (not entirely undeserved) reputation of being too abstruse or hard to learn for the general public.

The current availability of projects like the LiLa Knowledge Base (KB)[2], which has published several lexical and textual resources for the Latin language as LOD, or, more in general, the increasing success of the LOD paradigm in the Digital Humanities communities (Khan et al., 2022, 991-2) has highlighted the need to enable also specialists from areas like Classics to access and query the resources, as well as to encourage the production of new LOD-compliant resources.

While developing LiLa, we built a number of services to address such needs. After introducing the LiLa KB (Section 2), this paper describes those services, all developed as web applications with the backend managed via servlets and the interface developed using the React javascript framework. The source code for all applications is published in Github under an open-source license. As for querying the KB, we present an interface to search the collection of lemmas that represents the core of the KB (Section 3.1), and an interactive, graphical platform to run queries on the resources interlinked therein (Section 3.2). As for populating the KB with new textual resources, we describe a tool that performs automatic tokenization, lemmatization and Part-of-Speech tagging of a raw text in Latin and links its tokens to LiLa (Section 4). Finally, we draw some conclusion and sketch future works (Section 5).

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 (McCrae et al., 2017) and POWLA for corpus data (Chiarcos, 2012). As for the latter, the Resource Description Framework (RDF) is the data model used to describe information in terms of triples (McBride, 2004).

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

[1] https://linguistic-lod.org/llod-cloud
[2] https://lila-erc.eu

LDL: Linked Data in Linguistics @ LREC-COLING 2024, pages 75–83
20-25 May, 2024.

Bank (Mambrini et al., 2023) (CIRCSE, 2019-2024), a collection of around 200,000 lemmas taken from the database of the morphological analyzer LEMLAT (Passarotti et al., 2017) and constantly extended. A `lila:Lemma`[3] is a subclass of `ontolex:Form`[4], whose individuals are the inflected forms of a lexical item. In particular, the lemma is a form that can be linked to a `ontolex:LexicalEntry`[5] via the property `ontolex:canonicalForm`[6], which identifies the form that is canonically used to represent a lexical entry. To overcome divergent lemmatization criteria that may possibly be adopted in resources, LiLa exploits three key properties. The symmetric property `lila:lemmaVariant`[7] 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`[8] registers different spellings or graphical variants (called "written representations") 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`[9] is defined.

The LiLa Knowledge Base has already a wide coverage in terms of interlinked resources, including corpora, and dictionaries. Among the former are the *Opera Latina* corpus by LASLA, which features 130 Classical Latin texts (Fantoli et al., 2022), and two dependency treebanks, namely the *Index Thomisticus* Treebank, which comprises texts by Thomas Aquinas (1225–1274) (Mambrini et al., 2022) (CIRCSE, 2006-2024), and the *UDante* treebank, which encompasses Medieval Latin works written by Dante Alighieri (Passarotti et al., 2021) (CIRCSE, 2021b). Among the latter are the bilingual Latin-English dictionary by Lewis and Short, whose primary focus is on Classical Latin (Mambrini et al., 2021a) (CIRCSE, 2021a), and the *Dictionary of Medieval Latin in the Czech Lands*, a lexical resource that collects the Latin vocabulary (provided with translations into Czech) as it emerged in Eastern Europe during the Middle Ages (Gamba et al., 2023) (CIRCSE, 2023a). Currently, the LiLa RDF graph includes a total of more than 80 million triples, which can be queried from the SPARQL endpoint of the KB, where a few ready-made queries are provided[10].

3. Querying LiLa

This Section describes two services for querying, respectively: a) the Lemma Bank (3.1), and b) the textual resources and a selection of lexical resources currently linked to the LiLa KB (3.2).

3.1. The Lemma Bank Query Interface

The Lemma Bank query interface[11] allows users to interrogate the collection of Latin lemmas utilized in LiLa to interlink the linguistic resources published therein.

Relevant lemmas from the Lemma Bank can be selected based on various filters, including the lemma string, the presence of a specific affix (either prefix or suffix), the connection with a lexical base, the gender (for nouns), the part of speech (PoS), and the inflectional category. The lemma string search is performed by entering the desired string in a free text-box that supports regular expressions. The values for the other filters are provided through a dropdown menu.

The Lemma Bank query interface was designed to keep the search for lemmas as light as possible, by breaking down the query into blocks. Such query decomposition ensures that the minimum number of null results is obtained, by recalculating dynamically the values of all the fixed-value boxes every time the user adds a value in the query. For instance, if the lemmas of the verbs of the second conjugation are selected, the system cascades a series of SPARQL queries that update the values of the fixed-value boxes (i.e., prefix, suffix, gender, PoS, and inflectional category) only with those values that are compatible with the lemmas of the verbs of the second conjugation. The results of the query are then obtained by concatenating the selected values into a single SPARQL query, which can be downloaded.

Results are presented in the form of an alphabetically ordered list of lemmas, which can be downloaded along with the SPARQL query that produced it. For each lemma in the list, its written representation(s) and its PoS are shown, followed by two kinds of icons:

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

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

[5] http://www.w3.org/ns/lemon/ontolex#LexicalEntry

[6] http://www.w3.org/ns/lemon/ontolex#canonicalForm

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

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

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

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

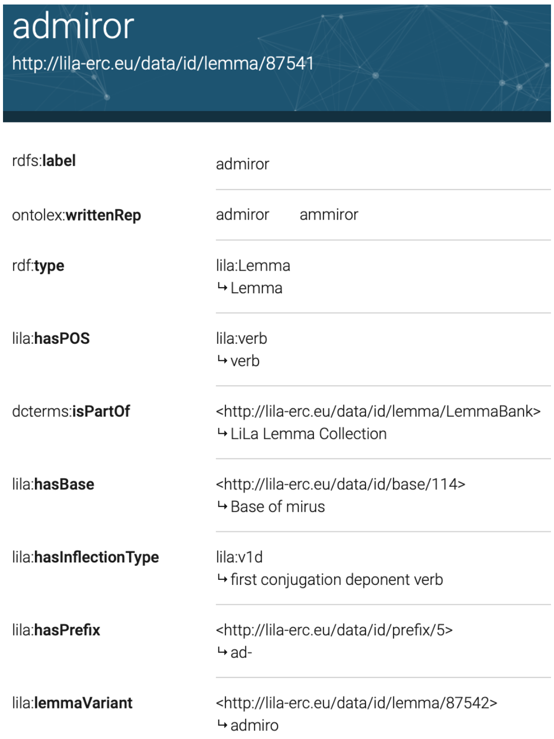
[11] https://lila-erc.eu/query/; https://github.com/CIRCSE/LiLa_LB_QueryInterface.

- if the lemma is linked to a lexical entry of (a) the derivational lexicon *Word Formation Latin* (Pellegrini et al., 2021) (CIRCSE, 2018), (b) a manually checked subset of the *Latin WordNet* enhanced with valency information taken from the *Latin Vallex* lexicon (Mambrini et al., 2021b) (CIRCSE, 2020b) (CIRCSE, 2023b), (c) the *LatinAffectus* polarity lexicon (Sprugnoli et al., 2020) (CIRCSE, 2020a), or (d) the Lewis and Short dictionary, an icon for each of these resources opens a window that provides an overview of the information reported by the lexical entry for the lemma in the resource selected (e.g., the derivational cluster of the lemma from *Word Formation Latin*);

- two icons show the triples connected to the selected lemma in the LiLa KB, respectively presenting the triples in a datasheet and in a network-like graphical representation, where nodes are individuals (e.g., the lemma) and edges are properties connecting individuals[12].

Figure 1 shows the datasheet for the verb *admiror* 'to admire', presenting the triples where *admiror* is in the domain (i.e., it is the subject of the property). Among the information shown in the datasheet is that the lemma: (a) has 2 written representations (*admiror* and *ammiror*), (b) pertains to the lexical base of *mirus*, which connects the lemmas in the Lemma Bank that share this base (property `lila:hasBase`[13]), (c) is a first conjugation verb (property `lila:hasInflectionType`[14]), (d) is formed with the prefix *ad-* (property `lila:hasPrefix`[15]), and (e) has as lemma variant the first conjugation not deponent form *admiro* (property `lila:lemmaVariant`[16]).

In the bottom of the datasheet, the inverse relations for the lemma are shown, namely those where the lemma is in the range (i.e., it is the object of the property). These are the cases where the lemma is linked to: (a) a lexical entry in a lexical resource (property `ontolex:canonicalForm`), (b) an hypolemma (property `lila:isHypolemma`[17]), (c) a lemma variant (property `lila:lemmaVariant`), or (d) a token in a textual resource (property `lila:hasLemma`[18]). By clicking on the URI of a token linked to the lemma, its datasheet is shown, where also the sentence-based context of the token and its citation reference is provided.



admiror

http://lila-erc.eu/data/id/lemma/87541

rdfs:**label**	admiror
ontolex:**writtenRep**	admiror ammiror
rdf:**type**	lila:Lemma ↳ Lemma
lila:**hasPOS**	lila:verb ↳ verb
dcterms:**isPartOf**	<http://lila-erc.eu/data/id/lemma/LemmaBank> ↳ LiLa Lemma Collection
lila:**hasBase**	<http://lila-erc.eu/data/id/base/114> ↳ Base of mirus
lila:**hasInflectionType**	lila:v1d ↳ first conjugation deponent verb
lila:**hasPrefix**	<http://lila-erc.eu/data/id/prefix/5> ↳ ad-
lila:**lemmaVariant**	<http://lila-erc.eu/data/id/lemma/87542> ↳ admiro

Figure 1: The datasheet for *admiror*.

3.2. The LiLa Interactive Search Platform

The LiLa Interactive Search Platform (LISP)[19] is an interactive graphical interface to perform SPARQL queries on the textual resources and a subset of the lexical resources interlinked in the LiLa RDF triple store.

Like the Lemma Bank query interface, LISP relies on a SPARQL endpoint, although it works on a larger scale, performing searches on all the graphs present in the LiLa triple store. The interface of LISP was developed in react-js and it replicates the macro structure of the graphs of the resources interlinked in LiLa, representing graphically the connections between them via nodes (for the Lemma Bank and the resources) and directed edges (for their relations). Such network-like representation helps the user to select the nodes that make up the search and to visualize the various levels on which to act to refine the results of the query.

For example, to retrieve in a selection of the corpora interlinked in LiLa all the tokens of those lem-

[12] Graphical representations are shown using the LodLive navigator (Camara et al., 2012).

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

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

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

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

[17] `http://lila-erc.eu/ontologies/lila/isHypolemma`

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

[19] `https://lila-erc.eu/LiLaLisp/`; `https://github.com/CIRCSE/LiLa_LISP`.

mas that feature certain properties reported by a corresponding entry in a specific lexical resource, LISP completes the path from the node for the tokens to that for the lexical resource in question. In particular, by applying a Depth-first search algorithm on the descriptive tree of the LiLa graphs, LISP adds the nodes for the Documents[20] and for the Lemma Bank along the path. Like for the Lemma Bank query interface, the values of each node restrict the configurable values of the others in the query. To reduce the amount of data obtainable by querying the entire LiLa triple store, each node contains only the instances of the class it represents. Then, each node executes a SPARQL query that recovers the data by concatenating backwards the descriptive SPARQL queries of all the nodes present in the generated tree.

On the left part of the screen, the platform features a few buttons organized in three areas. From top to bottom, they are the following:

- area for textual resources, which can be queried by Authors, Corpora, Documents, and Tokens;
- area for the Lemma Bank;
- area for lexical resources. Currently, it includes *Word Formation Latin*, *LatinAffectus*, *Latin WordNet*, the Lewis and Short dictionary and *Latin Vallex*.

LISP helps to combine information taken from different resources, by filtering their (meta)data, using the buttons from the three areas described above. For instance, by using the Documents, one can make a selection of the works (or sections of works) to query. Once works are selected, one can add information taken from a lexical resource, thus narrowing the query further. Typically, the last button to use is that of tokens, as it shows the list of tokens in the works selected that present the lexical properties taken from the lexical resources interlinked. As mentioned, the query is represented graphically in network-like fashion, showing the complete query path leading to tokens, according to the Lemma Bank based architecture of the LiLa KB.

Figure 2 shows the graphical representation of a query that searches in the documents whose authors are Catullus (taken from the LASLA corpus), Thomas Aquinas (from the *Index Thomisticus* Treebank), or Dante (from *UDante*). The node for the authors is linked to that for the tokens by the node for the Documents, which is connected to the lexical resources by passing through the Lemma Bank. The lexical resources provide lexical information to restrict furthermore the tokens to search. In the example, two resources are used: from *Word Formation Latin* the deverbal verbs formed with the prefix *de-* are selected; from *Latin Vallex* those words that have an Addresse in at least one of their valency frames (passing through the node for the *Latin WordNet*, as the two resources share the lexical entries). This query results in 1,225 tokens, which LISP presents as an alphabetically ordered list, where each token is followed by the title of the work in which it occurs (see Figure 3). By clicking on a token, its datasheet is shown, where its full reference and a KWIC-like visualization is provided.

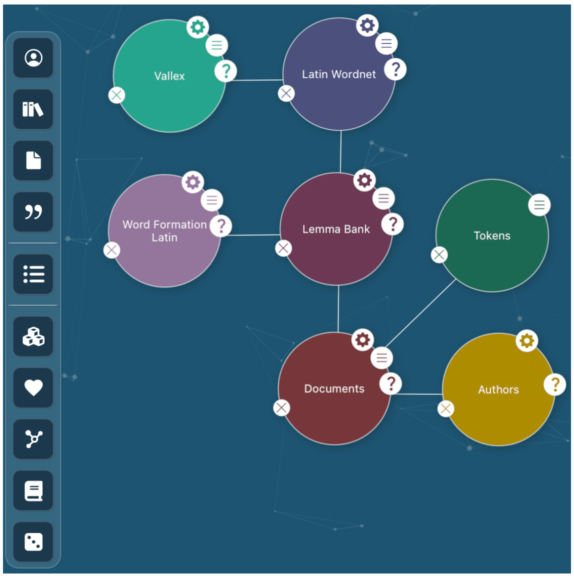
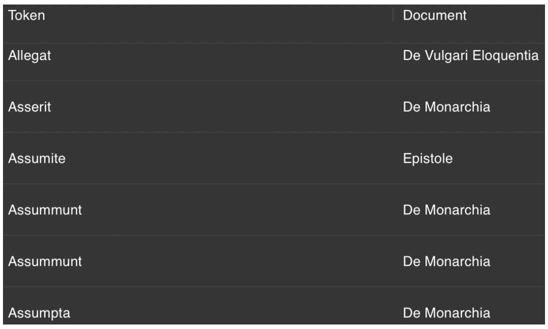


Figure 2: A graphical query in LISP.



Token	Document
Allegat	De Vulgari Eloquentia
Asserit	De Monarchia
Assumite	Epistole
Assummunt	De Monarchia
Assummunt	De Monarchia
Assumpta	De Monarchia

Figure 3: Results of a query in LISP.

4. Populating LiLa. The Text Linker

The LiLa Text Linker[21] is a web application designed to assist users in the every step of the workflow to produce RDF editions of Latin texts fully

[20] Documents are single works, or sections of works (e.g., books). Corpora are collections of Documents.

[21] https://lila-erc.eu/LiLaTextLinker/; https://github.com/CIRCSE/LiLa_TextLinker.

integrated with the LiLa KB. The Text Linker integrates components to perform the text-processing stage, the manual editing and the creation of the RDF output.

The workflow starts from the raw text of a Latin work[22]. In the text-processing stage, after a minimal normalization step that takes care of spelling conventions such as the use of characters *u* or *j* for *v* and *i*, the input is lemmatized and PoS-tagged with the help of a custom model for the UDPipe (v.1.3) annotation pipeline (Straka and Straková, 2017). The ad-hoc model was trained on approximately 3,400,000 tokens, including data from 4 of the Latin treebanks distributed in Universal Dependencies (*Index Thomisticus* Treebank, *PROIEL*, *Perseus*, and *UDante*)[23], the *Opera Latina* published by LASLA, the Latin text database *Computational Historical Semantics*, which is part of the Latin Text Archive[24], and a series of lemmatized works curated by the CIRCSE, either published[25], or in publication[26]. Data were harmonized as for both lemmatization criteria and PoS tagging, using the Universal PoS tagset (Petrov et al., 2011).

	Prec.	Recall	F1	AligndAcc
UPOS	94.02	94.02	94.02	94.02
Lemmas	93.70	93.70	93.70	93.70

Table 1: Performance of the ad-hoc model used by the LiLa Text Linker.

This corpus was randomly partitioned into a training (70%), development (20%) and test (10%) set. We evaluated the performances of the model on the test set. The results are reported in Table 1.

In a second step, the lemmatized tokens are matched against the lemmas in the LiLa KB. The Text Linker's matching algorithm is set to be strict, returning only candidates whose lemma string and PoS-tag fully match the output of the annotation via the UDPipe model.

The result of the lemmatization and linking phase is returned to the users, who have the opportunity to perform any manual edits or correction that they desire. A screenshot of the interface is shown in Figure 4. The tokens in the text are coloured according to the results from the previous stage: tokens that were matched with one single entry of the LiLa KB are visualized in green. Grey is used for tokens that were matched to more than one candidate; tokens in orange could not be matched.

By clicking on any linked token in the text, it is always possible to modify the automatic match by removing the suggested link and searching for candidates in the KB manually. In case of ambiguous matches (tokens in gray), it is also possible to select the appropriate candidate (or search for the right lemma by unlinking any of the proposed options), thus manually turning a 1:many match into a 1:1. Figure 4 shows an example of this process: the right pane of the interface shown in the screenshot is triggered by clicking on the ambiguous word *litora*, which is automatically assigned lemma *litus* and PoS NOUN[27]. For all the matching lemmas in the LiLa KB, the interface displays a series of information (including the senses for the *Latin WordNet* and the Lewis and Short dictionary, if available). These data are retrieved via a chain of SPARQL queries to the LiLa triple store executed in the background. By selecting one of the lemma candidate, users have the opportunity to save the link. A pie chart on the top-right corner visualizes the statistics of the matching phase, showing the number of unique, ambiguous or missing matches; the counts are updated after any manual intervention of the editor.

The lemmatization and linking process can also be performed using a REST API for the service. The API returns a JSON output with the tokenized and sentence-split text. For each token, the output includes the PoS-tag, the lemma string produced by UDPipe and the list of URIs of all candidates for matching in the LiLa KB. It is also possible to use the API via the Language Resource Switchboard of the CLARIN consortium (Zinn, 2018), where the tool can be selected from the menu of the lemmatizers for Latin[28].

Once that the users are satisfied of the results, they can use the Text Linker to export the text as RDF. In order to generate a RDF serialization, the interface requires a series of metadata, which the users can enter by filling the short form shown in Figure 5.

[22] At the moment, the application only accepts simple text (txt) as input. A future development could be to support also other formats and standards that are commonly used for digital editions, including in particular TEI-compliant XML. On TEI see https://tei-c.org/.

[23] https://universaldependencies.org/

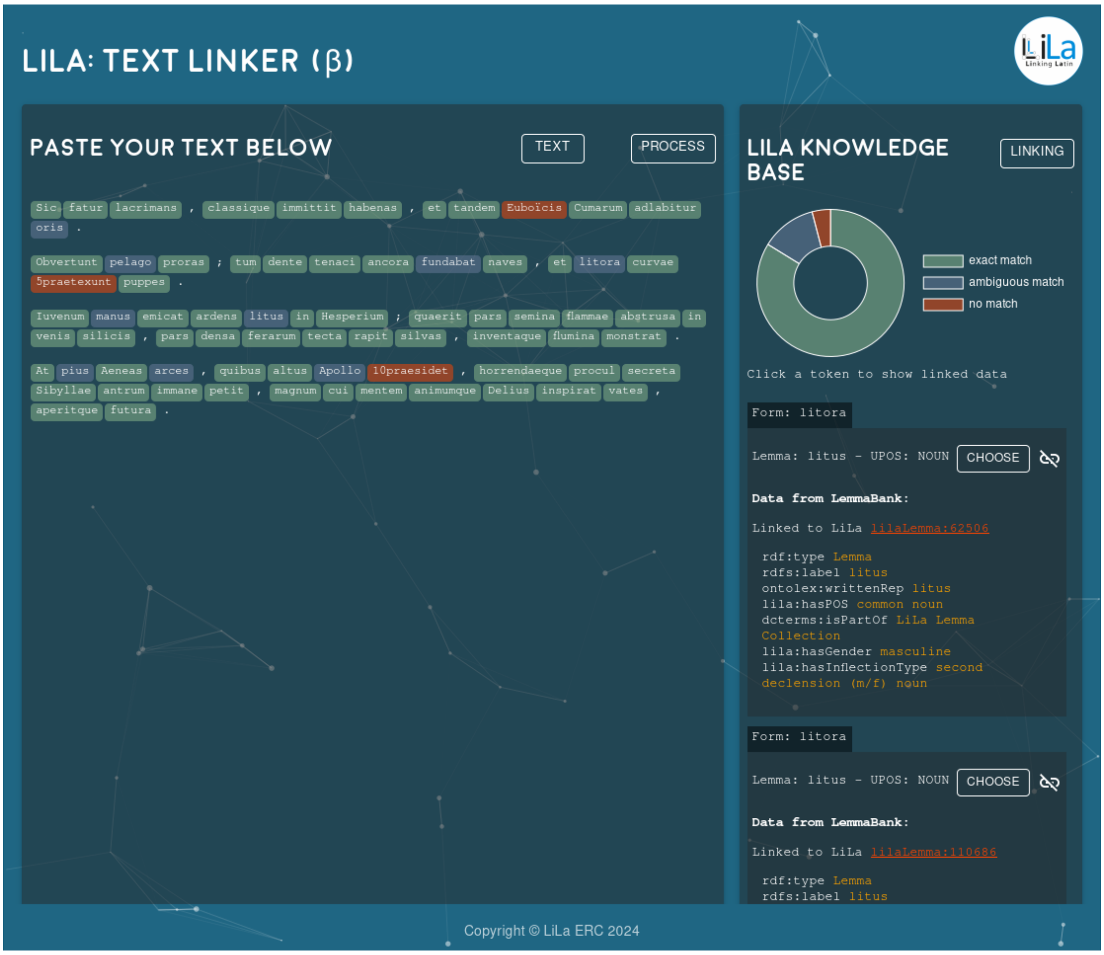
[24] https://lta.bbaw.de

[25] Augustine's *Confessions* (https://github.com/CIRCSE/AugustiniConfessiones), Sabellicus' *De Latinae Linguae Reparatione* (https://github.com/CIRCSE/Sabellicus).

[26] Avianus' *Fabulae*, Cicero's *De Divinatione*.

[27] There are three lemmas *litus* (NOUN) in the Lemma Bank: http://lila-erc.eu/data/id/lemma/110686 (meaning: 'a landing place'), http://lila-erc.eu/data/id/lemma/62506 (meaning: 'a servant'), and http://lila-erc.eu/data/id/lemma/111141 (meaning: 'a smearing').

[28] At the moment, the integration is still in progress, and the tool is only available in the testing interface of the Switchboard: https://beta-switchboard.clarin.eu/.



Sic fatur lacrimans , classique immittit habenas , et tandem Euboïcis Cumarum adlabitur oris .

Obvertunt pelago proras ; tum dente tenaci ancora fundabat naves , et litora curvae 5praetexunt puppes .

Iuvenum manus emicat ardens litus in Hesperium ; quaerit pars semina flammae abstrusa in venis silicis , pars densa ferarum tecta rapit silvas , inventaque flumina monstrat .

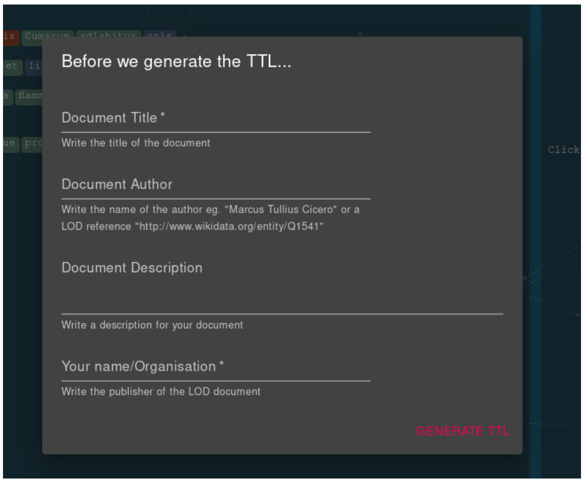
At pius Aeneas arces , quibus altus Apollo 10praesidet , horrendaeque procul secreta Sibyllae antrum immane petit , magnum cui mentem animumque Delius inspirat vates , aperitque futura .

Click a token to show linked data

```
Form: litora

Lemma: litus - UPOS: NOUN

Data from LemmaBank:

Linked to LiLa lilaLemma:62506

  rdf:type Lemma
  rdfs:label litus
  ontolex:writtenRep litus
  lila:hasPOS common noun
  dcterms:isPartOf LiLa Lemma
  Collection
  lila:hasGender masculine
  lila:hasInflectionType second
  declension (m/f) noun

Form: litora

Lemma: litus - UPOS: NOUN

Data from LemmaBank:

Linked to LiLa lilaLemma:110686

  rdf:type Lemma
  rdfs:label litus
```

Copyright © LiLa ERC 2024

Figure 4: Correcting the lemmatization/linking output with the LiLa Text Linker.



Before we generate the TTL...

Document Title *
Write the title of the document

Document Author
Write the name of the author eg. "Marcus Tullius Cicero" or a LOD reference "http://www.wikidata.org/entity/Q1541"

Document Description

Write a description for your document

Your name/Organisation *
Write the publisher of the LOD document

Figure 5: Metadata for the RDF output of the LiLa Text Linker.

5. Conclusion and Future Work

One of the main challenges for the LLOD world is to make fully exploitable the wealth of data and metadata from linguistic resources that, over the last decade, has been made interoperable through the application of the principles of the Linked Data paradigm.

In this paper, we have presented some services of the LiLa KB, developed with the aim of enabling scholars to make the most out of the interactions between the Latin resources made available by the KB. Indeed, more specifically, the challenge concerns the impact that the computational treatment of linguistic data can and should have on Classical language studies. For this impact to occur, it is necessary for digital resources and computational analysis tools to be made more easily accessible, and for computational skills to be provided to humanists, especially classicists. The LiLa services described in this paper allow classicists to collect empirical results that could not be obtained previously. They represent a good showcase demonstrating the utility of interoperability between different linguistic resources. The hope is that classicists not only use the services but also strive to go beyond, becoming autonomous in both querying and publishing linguistic data.

To this goal, testing and improving usability is a key factor. Of the three tools, the LiLa Text Linker has been demoed and showcased to a series of events for professionals in the Digital Humanities, including the *LinkedPast 6* workshop (2020),[29] and a dedicated tutorial at the 2nd Conference of the European Association for Digital Humanities (EADH21).[30] The other two, on the other hand, are still to be presented to the wider public. In the future, we intend to monitor the users more closely and to run usability tests for the interfaces involving representatives from the different communities of our target users.

Another important aspect that we want to explore is that of the adaptability of the software. The suite of tools that we presented here was designed specifically for the LiLa knowledge base; therefore, it is not ready to be used "out of the box" with data modeled according to other ontologies or structured differently from the LiLa paradigm. However, due to the way our tools were developed, we expect that only limited effort would be required to adapt the software to other projects, especially those that adopt the community standard Ontolex-Lemon. The fact that the tools work with linked data and are (mostly) based on interactions with a SPARQL endpoint is crucial in ensuring adaptability. More specifically, the LiLa query interface and the LiLa Lisp interface retrieve their data via SPARQL and can be re-modulated to query different triple stores. The LiLa Text Linker is the only tool that, at the moment, relies on an SQL database for reasons of efficiency; that application too, however, can be modified to interface with a triple store in order to increase its portability. Such aspects of portability must still be tested concretely, and any requirement for adapting the tools to different data must still be documented properly.

Beside linking new lexical and textual resources and keeping on expanding the coverage of the Lemma Bank, we also plan to update the trained model of the Text Linker, using a larger training set and version 2 of UDPipe. Furthermore, in LISP we will add access to further lexical resources, such as the *Lexikon der indogermanischen Verben* (Zimmer, 2002) (Boano et al., 2023) (CIRCSE, 2023c), by generalizing the query process that we already developed for querying similar resources. Indeed, so far the facets that describe the nodes used in LISP have been developed ad-hoc for each single resource included in the platform. However, in the near future, we expect to reuse the nodes as modeling templates for adding more resources.

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

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 CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30—Dec 02, 2023, Venice, Italy*, pages 1–7. CEUR-WS.

Diego Valerio Camarda, Silvia Mazzini, and Alessandro Antonuccio. 2012. Lodlive, exploring the web of data. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 197–200.

Christian Chiarcos. 2012. Powla: Modeling linguistic corpora in owl/dl. In *Extended Semantic Web Conference*, pages 225–239. Springer.

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

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham.

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 Linked Data in Linguistics Workshop@ LREC2022*, pages 26–34.

Federica Gamba, Marco C Passarotti, and Paolo Ruffolo. 2023. Linking the dictionary of medieval latin in the czech lands to the lila knowledge base. In *Proceedings of CLiC-it 2023: 9th Italian Conference on Computational Linguistics*, pages 1–8. CEUR Workshop Proceedings.

Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P McCrae, et al. 2022. When linguistics meets web

[29] https://lila-erc.eu/linked-pasts-6-activity/.

[30] https://lila-erc.eu/eadh-2021/.

technologies. recent advances in modelling linguistic linked data. *Semantic Web*, 13(6):987–1050.

Francesco Mambrini, Eleonora Litta, Marco Passarotti, and Paolo Ruffolo. 2021a. Linking the lewis & short dictionary to the lila knowledge base of interoperable linguistic resources for latin. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, pages 214–220. Accademia University Press.

Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021b. Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Further with Knowledge Graphs*, volume 53 of *Studies on the Semantic Web*, pages 16–28.

Francesco Mambrini, Marco Passarotti, Giovanni Moretti, and Matteo Pellegrini. 2022. The index thomisticus treebank as linked data in the lila knowledge base. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4022–4029.

Francesco Mambrini, Marco Carlo Passarotti, et al. 2023. The lila lemma bank: A knowledge base of latin canonical forms. *JOURNAL OF OPEN HUMANITIES DATA*, 9(28):1–5.

Brian McBride. 2004. The resource description framework (rdf) and its vocabulary description language rdfs. In *Handbook on ontologies*, pages 51–65. Springer.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

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, Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, et al. 2021. Udante. l'annotazione sintattica dei testi latini di dante. *Studi Danteschi*, 86:309–338.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58(1):177–212.

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

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

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)*, volume 2695, pages 59–70, Heraklion, Greece. ceur-ws.org.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Stefan Zimmer. 2002. *Lexikon der indogermanischen Verben (LIV)*. Walter de Gruyter GmbH & Co. KG.

Claus Zinn. 2018. Squib: The language resource switchboard. *Computational Linguistics*, 44(4):631–639.

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

CIRCSE. 2020b. *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. 2023a. *Dictionary of Medieval Latin in the Czech Lands*. CIRCSE Research Centre. PID https://github.com/CIRCSE/LexiconBohemorum.

CIRCSE. 2023b. *Latin WordNet (revised version)*. CIRCSE Research Centre. PID https://doi.org/10.5281/zenodo.7561689.

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