

The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin

Marco C. Passarotti¹ 

CIRCSE, Università Cattolica del Sacro Cuore, Milan, Italy
marco.passarotti@unicatt.it

Flavio M. Cecchini 

CIRCSE, Università Cattolica del Sacro Cuore, Milan, Italy
flavio.cecchini@unicatt.it

Greta Franzini 

CIRCSE, Università Cattolica del Sacro Cuore, Milan, Italy
greta.franzini@unicatt.it

Eleonora Litta 

CIRCSE, Università Cattolica del Sacro Cuore, Milan, Italy
eleonoramaria.litta@unicatt.it

Francesco Mambrini 

CIRCSE, Università Cattolica del Sacro Cuore, Milan, Italy
francesco.mambrini@unicatt.it

Paolo Ruffolo 

CIRCSE, Università Cattolica del Sacro Cuore, Milan, Italy
paolo.ruffolo@posteo.net

Abstract

The *LiLa: Linking Latin* project was recently awarded funding from the European Research Council to build a Knowledge Base of linguistic resources for Latin. LiLa responds to the growing need in the fields of Computational Linguistics, Humanities Computing and Classics to create an interoperable ecosystem of resources and Natural Language Processing tools for Latin. To this end, LiLa makes use of Linked Open Data practices and standards to connect words to distributed textual and lexical resources via unique identifiers. In so doing, it builds rich knowledge graphs, which can be used for research and teaching purposes alike. This paper details the architecture of the LiLa Knowledge Base and presents the solutions found to address the challenges raised by populating it with a first set of linguistic resources.

2012 ACM Subject Classification Information systems → Ontologies; Information systems → Graph-based database models; Information systems → Semantic web description languages; Applied computing → Digital libraries and archives; Applied computing → Annotation

Keywords and phrases Latin, Linguistics, Linked Open Data, NLP, Metadata, Graph

Funding *Marco C. Passarotti*: [ERC-2017-COG]

Flavio M. Cecchini: [ERC-2017-COG]

Greta Franzini: [ERC-2017-COG]

Eleonora Litta: [ERC-2017-COG]

Francesco Mambrini: [ERC-2017-COG]

Paolo Ruffolo: [ERC-2017-COG]

Acknowledgements The LiLa 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.

¹ Corresponding author.



1 Introduction

Despite the proliferation and the increasing coverage of linguistic resources for many languages, the interoperability issues imposed by their different formats severely limits their potential for exploitation and use. Indeed, linking linguistic resources to one another would maximize their contribution to, and use in, linguistic analysis at multiple levels, be those lexical, morphological, syntactic, semantic or pragmatic.

The objective of the *LiLa: Linking Latin* project (2018-2023)² is to connect and, ultimately, exploit the wealth of linguistic resources and Natural Language Processing (NLP) tools for Latin developed thus far, in order to bridge the gap between raw language data, NLP and knowledge description [4, p. 111]. Latin is an optimal use case for this kind of research for two reasons: (a) the diachrony and diversity of the language present complex challenges for NLP; (b) an interconnected network of the numerous linguistic resources currently available for Latin would greatly support both research and learning communities, including historians, philologists, archaeologists and literary scholars.

LiLa addresses this challenge by building a Linked Data Knowledge Base of linguistic resources (e.g., corpora, lexica, ontologies, dictionaries, thesauri) and NLP tools (e.g., tokenizers, lemmatizers, PoS-taggers, morphological analyzers and dependency parsers) for Latin currently available from different providers under various licences. This paper details the architecture of the LiLa Knowledge Base and presents the solutions found to address the challenges raised by populating it with a first set of linguistic resources.

2 The LiLa Knowledge Base

In order to achieve interoperability between resources and tools, LiLa makes use of a set of Semantic Web and Linguistic Linked Open Data standards. These include ontologies to describe linguistic annotation (OLiA [3]), corpus annotation (NIF [6], CoNLL2RDF [2]) and lexical resources (Lemon [1], Ontolex³). The Resource Description Framework (RDF) [7] is used to encode graph-based data structures to represent linguistic annotations in terms of triples. The SPARQL language is used to query the data recorded in the form of RDF triples [12].

The LiLa Knowledge Base is lexically-based and strikes a balance between feasibility and granularity: textual resources are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words. **Lemma** is the key node type in LiLa. A Lemma is an (inflected) **Form** conventionally chosen as the citation form of a lexical item. Lemmas occur in **Lexical Resources** as canonical forms of lexical entries. Forms, too, can occur in lexical resources, for instance in a lexicon containing all of the forms of a language (for instance, [13]). The occurrences of Forms in real texts are **Tokens**, which are provided by **Textual Resources**. Texts in Textual Resources can be different editions or versions of the same work (e.g., the numerous editions of the *Orator* by Cicero, which may be available from different Textual Resources). Finally, **NLP tools** process either Forms, regardless of their contextual use (e.g., a morphological analyzer), or Tokens (e.g., a PoS-tagger).

² <https://lila-erc.eu/>

³ <https://www.w3.org/community/ontolex/>

2.1 Harmonizing Different Lemmatization Strategies

Because the lemma serves as the optimal interface between lexical resources, annotated corpora and NLP tools, the core of the LiLa Knowledge Base is a collection of citation forms. Interoperability can be achieved by linking the entries in lexical resources and the corpus tokens pointing to the same lemma.

The task of building and organizing a repository of lemmas that may serve as a hub in such an architecture is, however, complicated by the fact that different corpora, lexica or tools for Latin may adopt different strategies to solve the conceptual and linguistic challenges posed by lemmatization. These include:

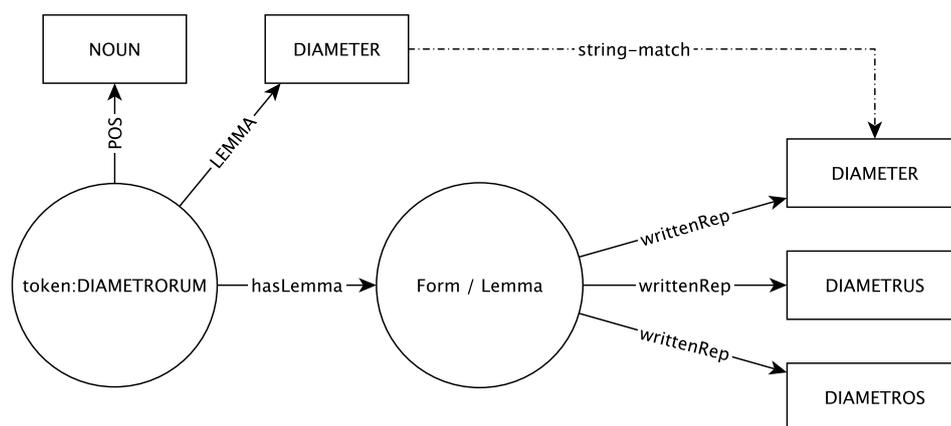
- different citation forms of the same word, resulting from interchange in (a) graphical representation (*voluptas* vs. *uoluptas*, “satisfaction”), (b) spelling (*sulphur* vs. *sulfur*, “brimstone”), (c) ending (*diameter* vs. *diametros* vs. *diametrus*, “diameter”) or (d) the paradigmatic slot representing the lemma (*sequor*, “to follow”, first person singular of the passive/deponent present indicative vs. *sequo*, first person singular of the active present indicative, attested in some lexicographical sources);
- the existence of homographic lemmas, like *occido* (*occīdo* < *ob* + *caedo*, “to strike down”) vs. *occido* (*occīdo* < *ob* + *cado*, “to fall down”);
- ambiguity in choosing the lemma: certain forms, such as participles or deadjectival adverbs, can be considered either part of the inflectional paradigm of verbs or adjectives, or independent lemmas provided with an autonomous entry in lexical resources;
- polythematic words, for which missing forms are taken from other stems, as is the case for *melior* used as a comparative of *bonus* (i.q. the English “good” and “better”).

When dealing with homographs, corpora may choose to index the different entries, but, generally, the string of the lemma is not disambiguated. Participles can either be lemmatized under the main verb, or have a dedicated participial lemma, which in turn may be used systematically or only when the participle has grown into an autonomous lexical item (e.g. *doctus*, “learned”, morphologically the past participle of *doceo*, “to teach”). Deadjectival adverbs (e.g. *aequaliter*, “evenly” from *aequalis*, “equal”) or peculiar forms such as comparatives (both regular and irregular) are sometimes subsumed under the (positive degree of the) adjective, or given a self-standing lemma.

Given the challenges and the degree of variation raised by different lemmatization strategies for Latin, our approach is to be as descriptive and inclusive as possible: our aim is to collect as many word forms as may be used for lemmatization and attempt to model their relations. To do so, we rely on a series of ontologies for lexical resources to describe the word forms used in lemmatization, and turn to the Web Ontology Language (OWL) for ontologies to model the relations between them ([9]).

Building on the Ontolex ontology, we define a Lemma as a Form of a word. In this way, lexical resources compiled using the Ontolex or Lemon formalism can already be connected to our collection. Forms have one or more written representations and are linked to one or more Parts of Speech (PoS). PoS are linked to the appropriate OLiA concepts, and we plan to represent the most widespread Latin PoS-tagging tagsets via dedicated OLiA ontologies.

The relations between the lemma and the other forms of the same word are defined horizontally, i.e. via direct relations between forms. While the architecture is ready to accommodate all of the attested or morphologically possible inflected forms of a lexical item, it is currently being populated only with those forms that are potentially used as lemmas, thus shaping LiLa’s previously mentioned core.



■ **Figure 1** Connecting tokens and lemmas with different written representations in LiLa.

The reference list of Latin lemmas is taken from that provided by the Latin morphological analyzer Lemlat [11]. Specifically, following the practice of Lemlat, we define a special subclass of lemmas, called “hypolemmas”, to harmonize different strategies for the lemmatization of participles. Hypolemmas are defined as forms of the inflectional paradigm of a word that may be used in annotated corpora or by NLP tools to lemmatize certain forms instead of the main lemma, i.e. the nominal inflected forms of verbal paradigms (participles, gerunds, gerundives, supines). As a result, we have generated hypolemmas for all the canonical forms of present, future and perfect participles and have connected them with their main (verbal) lemma via a subclass of the property “Form variant” of the Lemon ontology.⁴ Thus, for instance, the present participle *subsistens*, “taking a stand” is hypolemma of the main lemma *subsisto*, “to take a stand”. The same subclass is also used for alternative paradigmatic slots representing that lemma.

Systematic graphical variations (e.g. *u/v*) are preprocessed automatically, whereas changes in spelling and ending are managed as different written representations of the same lemma. For instance, Figure 1 shows how the token *diametrorum* (provided by a textual resource) is connected to LiLa via the lemma. In its source text, *diametrorum* is assigned a PoS (NOUN) and a lemma (*diameter*). A string match is found between the string used to lemmatize the token and one of the three written representations of a LiLa Lemma. On the basis of this string match, the token *diametrorum* is connected to the lemma *diameter* via the relation `hasLemma`.

2.2 Linguistic Resources in LiLa

The linguistic resources currently linked in the LiLa Knowledge Base are stored in a triplestore using the Jena framework; the Fuseki component exposes the data as a SPARQL end-point accessible over HTTP. The current prototype of the LiLa RDF triplestore database connects the following resources: (a) the collection of lemmas provided by Lemlat, (b) the morphological derivation lexicon Word Formation Latin (WFL) [8], (c) the PROIEL Latin Treebank [5] in its Universal Dependencies (UD) version (release 2.3)⁵ and (d) the *Index Thomisticus* Treebank in both its UD 2.3 and original format [10].

⁴ <https://www.lemon-model.net/lemon-cookbook/node17.html>

⁵ <http://universaldependencies.org/>

An example SPARQL query traversing all of these resources might search for all tokens (a) whose lemma is a noun including the suffix *-(t)or* for *nomina agentis / instrumenti* (sources: Lemlat and WFL), (b) that are assigned dependency relation *nsubj* (nominal subject), and (c) that depend directly on a node of a verb in the UD tree of the sentence in which they occur (source: PROIEL UD 2.3). The output provides the list of all noun/verb couples resulting from the query, sorted in descending order of frequency (see code below).⁶

■ **Listing 1** A SPARQL query in LiLa.

```
PREFIX: <http://lila-erc.eu/data/ontologies/lemlat-base#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX conll:
  <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#>

SELECT ?headlab ?deplab (count(*) as ?tot) WHERE {
  SERVICE <http://lila-erc.eu:3030/lemlat/sparql> {
    ?suff a :Suffix .
    ?suff rdfs:label "-(t)or" .
    ?lemma :hasSuffix ?suff .
    ?lemma ontolex:writtenRep ?deplab . }
  ?tok :hasLemma ?lemma .
  GRAPH <http://lila-erc.eu:3030/corpora/data/la-proiel-ud> {
    ?tok conll:EDGE "nsubj" .
    ?tok conll:HEAD ?head .
    ?head conll:UPOS "VERB" . }
  ?head :hasLemma ?l
  SERVICE <http://lila-erc.eu:3030/lemlat/sparql> {
    ?l ontolex:writtenRep ?headlab . } }
GROUP BY ?headlab ?deplab
ORDER by desc(?tot)
```

3 Conclusion

In this paper, we have introduced the architecture of the LiLa Knowledge Base, which is being built in accordance with the Linked Data paradigm to foster interoperability between linguistic resources for Latin. In particular, we focused on the challenges introduced by the harmonization of the different lemmatization strategies adopted by annotated corpora.

Given the central role of the Lemma in LiLa, the project is developing a strategy to automatically PoS tag and lemmatize the (many) corpora of Latin texts that are still free of this level of linguistic annotation. Indeed, despite the availability of NLP tools (and trained models) for automatic PoS tagging, lemmatization, morphological analysis and dependency parsing, their large-scale application to Latin textual resources is severely limited by their low degree of portability across two millennia of language change. This large diachronic and diatopic span serves as a perfect use-case for the development, application, and testing of solutions capable of providing equally good accuracy rates for all Latin “types”.

⁶ The query can be run at <https://lila-erc.eu/data/> using the `/corpora` endpoint.

References

- 1 Paul Buitelaar, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. Ontology lexicalisation: The lemon perspective. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence*, pages 33–36, 2011.
- 2 Christian Chiarcos and Christian Fäth. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*, pages 74–88, Cham, 2017. Springer International Publishing. URL: https://link.springer.com/content/pdf/10.1007%2F978-3-319-59888-8_6.pdf.
- 3 Christian Chiarcos and Maria Sukhareva. OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 6(4):379–386, 2015. URL: <http://www.semantic-web-journal.net/content/olia-%E2%80%93-ontologies-linguistic-annotation>.
- 4 Thierry Declerck, Piroska Lendvai, Karlheinz Mörth, Gerhard Budin, and Tamás Váradi. Towards linked language data for digital humanities. In *Linked Data in Linguistics*, pages 109–116. Springer, 2012.
- 5 Dag TT Haug and Marius Jøhndal. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, 2008.
- 6 Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating NLP using Linked Data. In *12th International Semantic Web Conference, Sydney, Australia, October 21-25, 2013*, 2013. URL: https://svn.aksw.org/papers/2013/ISWC_NIF/public.pdf.
- 7 Ora Lassila, Ralph R. Swick, World Wide, and Web Consortium. Resource description framework (rdf) model and syntax specification, 1998.
- 8 Eleonora Litta, Marco Passarotti, and Chris Culy. *Formatio formosa est*. Building a Word Formation Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016), Napoli, Italy, December 5-7, 2016*, volume Vol-1749, pages 185–189, 2016. URL: <http://ceur-ws.org/Vol-1749/paper32.pdf>.
- 9 Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.
- 10 Marco Passarotti. Theory and practice of corpus annotation in the index thomisticus treebank. *Lexis*, 27(A):5–23, 2009.
- 11 Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, volume 133, pages 24–31. Linköping University Electronic Press, 2017. URL: <http://www.ep.liu.se/ecp/article.asp?issue=133&article=006&volume=>.
- 12 Eric Prud’Hommeaux, Andy Seaborne, et al. Sparql query language for rdf. w3c. *Internet: https://www.w3.org/TR/rdf-sparql-query/[Accessed on February 27th, 2019]*, 2008.
- 13 Paul Tombeur. *Thesaurus formarum totius Latinitatis: a Plauto usque ad saeculum XXum*. Turnhout: Brepols, 1998.