LiLa: LINKING LATIN

A Knowledge Base of Linguistic Resources & NLP Tools

Marco C. Passarotti, Flavio M. Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini, Paolo Ruffolo
CIRCSE Research Centre, Università Cattolica del Sacro Cuore - Milan, Italy

https://lila-erc.eu

RESEARCH QUESTION

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, **interlinking linguistic resources would maximise their contribution to, and use in, linguistic analysis at multiple levels**, be those lexical, morphological, syntactic, semantic or pragmatic.

OBJECTIVE

The objective of the **LiLa: Linking Latin** project (2018-2023) **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**.

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 linguists, philologists, epigraphists and literary scholars.

METHODOLOGY

LEXICAL STRUCTURE

The LiLa Knowledge Base is lexically-based and strikes a balance between granularity and feasibility: 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. 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' of Cicero, which may be available from different Textual Resources). Finally, **NLP tools** process either Forms, regardless of their contextual use (e.g., a morphological analyser), or Tokens (e.g., a PoS-tagger).

LEMMA REFERENCE Lexical basis of the Latin morphological analyser LEMLAT (ca. 155,000 lemmas).

IMPLEMENTATION

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**), corpus annotation (**NIF**, **CoNLL2RDF**) and lexical resources (**Lemon**, **Ontolex**). The Resource Description Framework (**RDF**) is used to encode graph-based data structures to represent linguistic annotations as triples. LiLa triples are stored in a triplestore using the **Jena** framework; the **Fuseki** component exposes the data as a **SPARQL** end-point accessible over HTTP.

RESOURCES CONNECTED THUS FAR

The current prototype of the LiLa RDF triplestore connects: (a) the collection of lemmas provided by the morphological analyser **LEMLAT**, (b) the morphological derivation lexicon **Word Formation Latin** (WFL), (c) the **PROIEL Latin Treebank** (Universal Dependencies (UD) version 2.3) and (d) the **Index Thomisticus Treebank** in both its UD v. 2.3 and original formats.

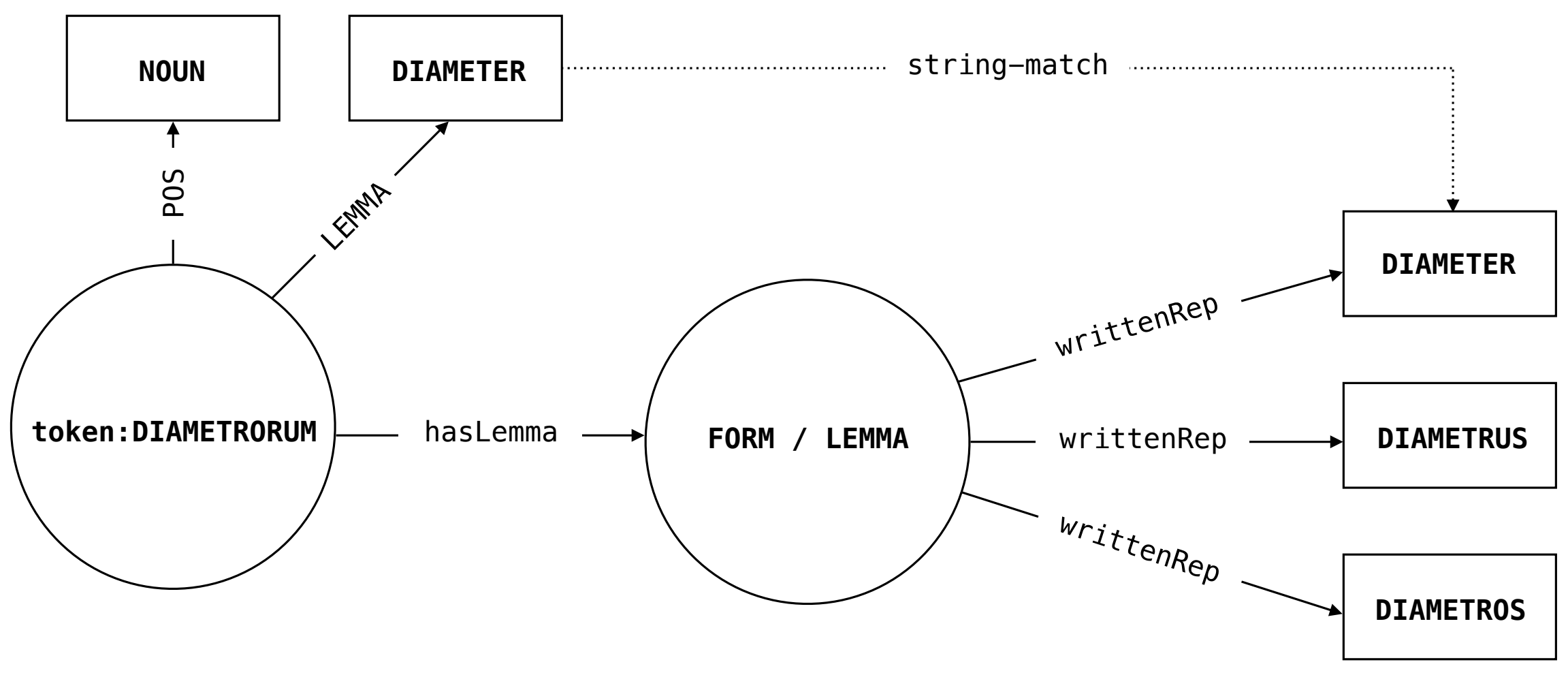
SOLUTIONS TO LINGUISTIC ISSUES

PARTICIPLES LiLa generates **hypolemmas** for all the canonical forms of present, future and perfect participles and **connects them with their main (verbal) lemma via a subclass of the property "Form variant" of the Lemon ontology**. The same subclass is also used for alternative paradigmatic slots representing that lemma.

DEADJECTIVAL ADVERBS These adverbs (e.g., 'aequaliter', "evenly" from 'aequalis', "equal") and peculiar forms such as comparatives (both regular and irregular) are either **subsumed under the (positive degree of the) adjective** or given a **self-standing lemma**.

HOMOGRAPHS Homographs, such as 'occidit' (occīdo = ob+caedo, "to strike down") vs. 'occidit' (occĭdo = ob+cado, "to fall down"), are **lemmatised under both/multiple lemmas unless disambiguated** in the source corpus.

GRAPHICAL VARIATION **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** (see diagram below).



VALUE FOR THE COMMUNITY

LINGUISTICS LiLa can help Latin linguists with very specific linguistic queries, e.g., search for all tokens (a) whose lemma is a noun including the suffix -(t)or; (b) whose dependency relation is '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.

PHILOLOGY LiLa can help philologists look up textual variation across scholarly editions and manuscript transcriptions to support the production of e.g., new editions.

NLP By connecting lexica and corpora to ontologies, LiLa can help NLP researchers better tune tools to the semantics, diachrony and diversity of Latin, as well as improve Information Extraction (IE) and Retrieval (IR) systems.

TRANSLATION STUDIES LiLa's future connection to the Linguistic Linked Open Data cloud and to multilingual semantic networks (e.g., BabelNet) will support queries in relation to other languages.

LDK 2019 · 20-23 May · Leipzig, Germany

Acknowledgements
This 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.

erc

UNIVERSITÀ CATTOLICA del Sacro Cuore