LiLa: Linking Latin

A Knowledge Base of Linguistic Resources & NLP Tools for Latin

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

MOTIVATION & METHOD

A collection of interoperable linguistics resources and NLP tools for Latin described with the same vocabulary of knowledge description

Interlinking as a form of interaction

2018-2023
https://lila-erc.eu

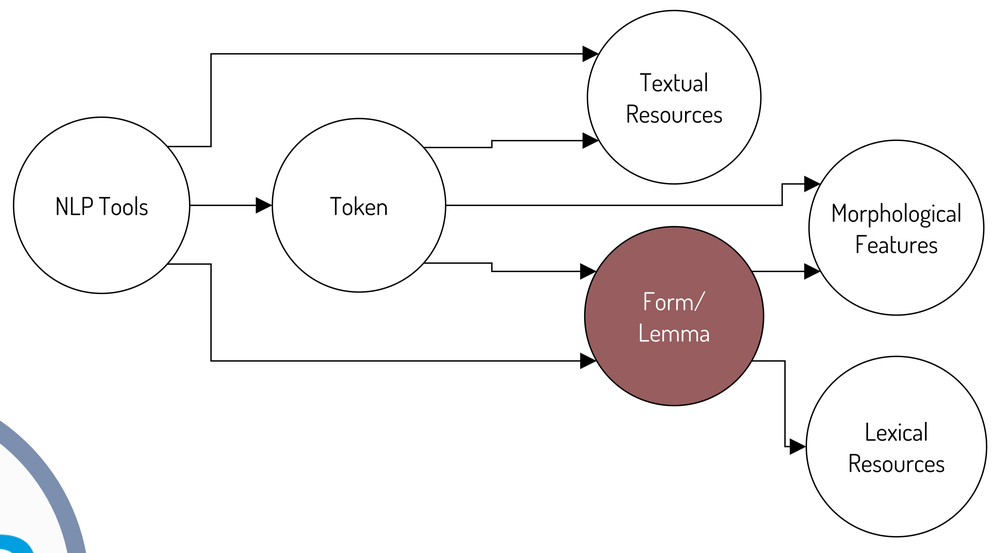
Despite the proliferation and the increasing coverage of linguistic resources for many languages, the interoperability issues imposed by their different formats severely limits their potential for exploitation and use. **Interlinking linguistic resources would maximise their contribution to, and use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

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.

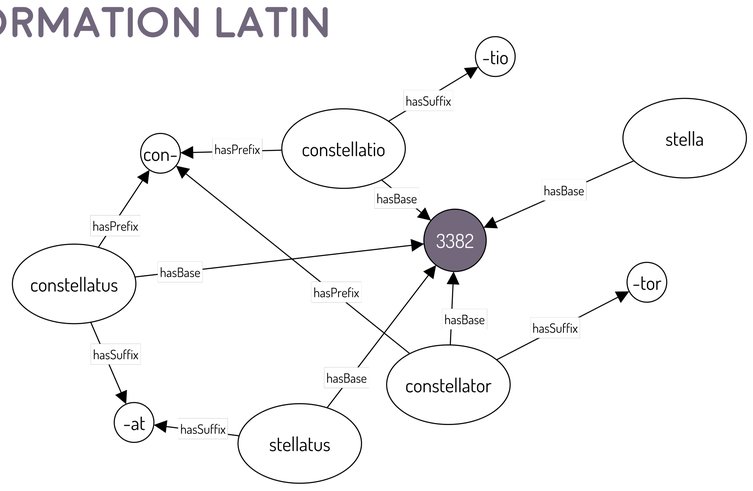
STRUCTURE

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).

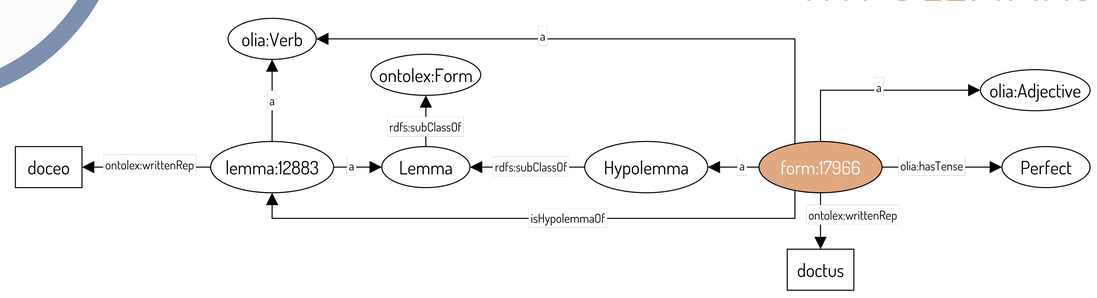
LEMMA REFERENCE: Lexical basis of the Latin morphological analyser LEMLAT (ca. 155,000 lemmas).



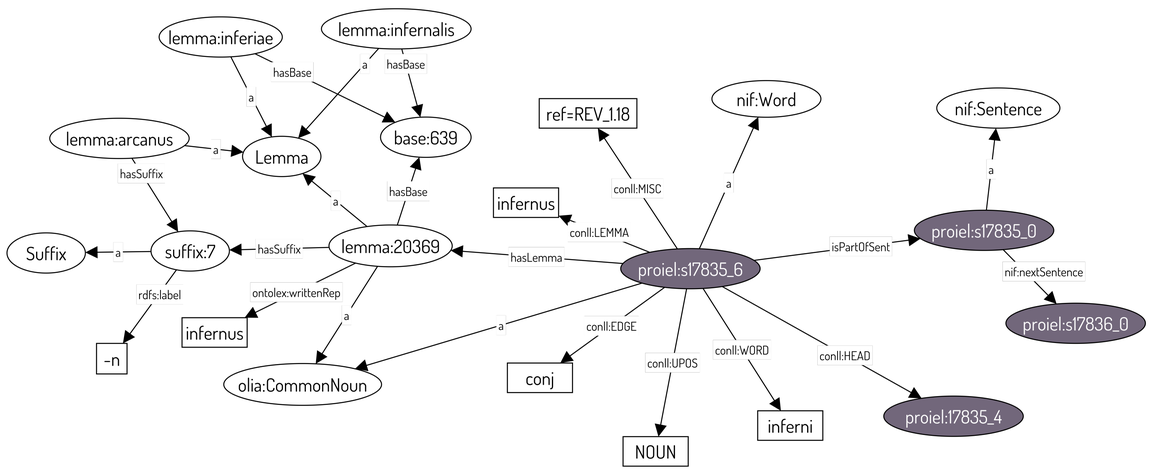
WORD FORMATION LATIN



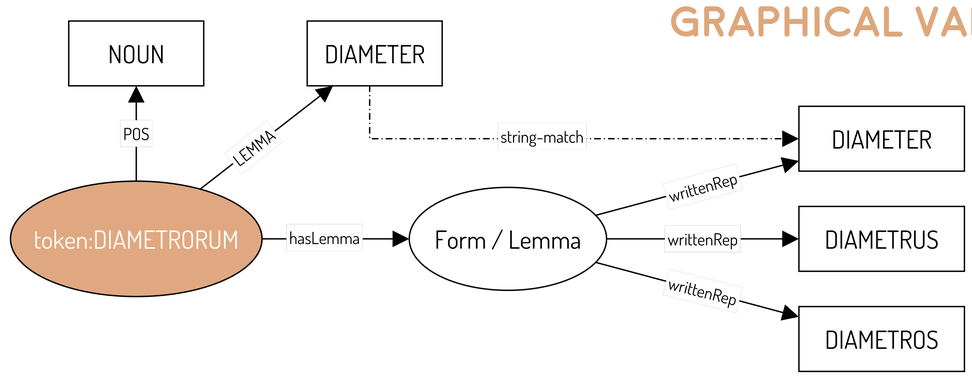
HYPOLEMMAS



CORPORA

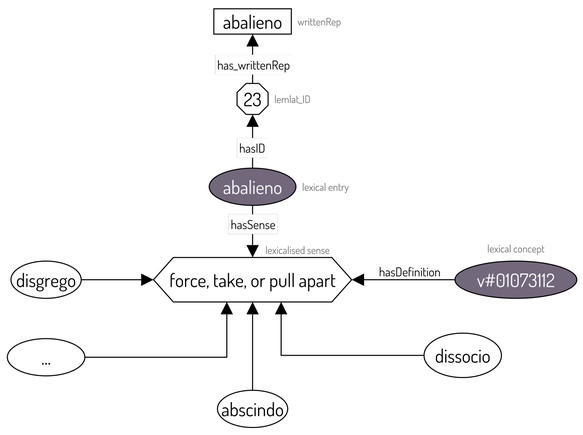


GRAPHICAL VARIANTS



LEXICAL COMPLEXITY

LATIN WORDNET



LINGUISTIC RESOURCES

DH 2019 · 9-12 July · Utrecht, The Netherlands

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.

Poster ID: 239