▶ Issues in Building the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Marco Passarotti,
Francesco Mambrini,
Università Cattolica del Sacro Cuore, Milan, Italy, marco.passarotti@unicatt.it

Purpose: This abstract presents the architecture and the current state of the LiLa Knowledge Base (https://lila-erc.eu), i.e., a collection of multifarious linguistic resources for Latin described with the same vocabulary of knowledge description, by using common data categories and ontologies developed by the Linguistic Linked Open Data (LLOD) community according to the principles of the Linked Data paradigm.

Design/methodology/approach: LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora, and NLP tools. The core of the LiLa Knowledge Base consists of a large collection of Latin lemmas (called Lemma Bank): interoperability is achieved by linking, via (L)LOD data categories and ontologies, all those entries in lexical resources and tokens in corpora that point to the same lemma.

Findings: The textual resources currently interlinked through LiLa include: two dependency treebanks (*Index Thomisticus* Treebank, UDante), Fibonacci's *Liber Abbaci*, and a large corpus of Classical Latin texts (LASLA corpus). Lexical resources include: a manually checked subset of the Latin WordNet, a sentiment lexicon (Latin Affectus), a derivational lexicon (Word Formation Latin), a Latin-English Dictionary (Lewis & Short), a list of Greek loans in Latin, and an etymological dictionary.

Research limitations/implications: Limitations of representation of linguistic (meta)data in the currently available LLOD models and ontologies; issues of harmonization of different lemmatization strategies and PoS tagging; problems of automatic linking of lexical and textual resources; and problems related to extracting information from linguistic resources published as LLOD.

Practical implications: The LiLa Knowledge Base can be queried through a SPARQL endpoint at https://lila-erc.eu/sparql/, where a few pre-compiled queries are available. The Lemma Bank of LiLa can be accessed at https://lila-erc.eu/query/. Lemmas can be searched by string of characters, Part of Speech, affix, lexical base, inflectional category, and gender (for nouns). Results are provided both in data sheet fashion 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. The Turtle files of the resources interlinked in LiLa are available at https://github.com/CIRCSE.

Originality/Value: The LiLa Knowledge Base makes it possible to perform que-

ries on interoperable, distributed linguistic resources for Latin published on the web: this represents a terrific advancement in the way such resources can be used, particularly by Classicists, who need a steady confrontation with the empirical evidence provided by textual and lexical resources. Building the LiLa Knowledge Base represents a large-scale use-case where to apply and test the currently available vocabularies developed by the LLOD community, as well as developing new ones.

Keywords: *linguistic resources, Linguistic Linked Open Data, Latin*

Research type: Research paper.