The LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. Architecture and Current State

Marco Passarotti, Eleonora Litta, Flavio Massimiliano Cecchini, Matteo Pellegrini, Giovanni Moretti, Paolo Ruffolo, Giulia Pedonese
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Gemelli, 1 - 20123 Milan, Italy
marco.passarotti@unicatt.it

Abstract

This abstract presents the architecture and the current state of the LiLa Knowledge Base, i.e. a collection of multifarious linguistic resources for Latin described with the same vocabulary of knowledge description and interlinked according to the principles of the so-called Linked Data paradigm. This abstract presents the architecture of the LiLa Knowledge Base, whose core consists of a large collection of Latin lemmas, serving as the backbone to achieve interoperability between the resources by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. Moreover, the abstract details the linguistic resources for Latin currently interlinked through LiLa and informs about how the Knowledge Base can be queried.

1 Introduction and Motivation

Over the past two decades the research area dedicated to building, improving and evaluating linguistic resources has seen substantial growth and, today, covers a wide span of languages and language varieties. Despite the increase in the quantity and coverage of linguistic resources, most of these are locked in data silos, which prevents users from honing both their individual and joint potential in interoperable ways. Indeed, linguistic data and metadata today are scattered in distributed resources, thus failing to provide a comprehensive overview of the annotations available in these separate collections. One of the main challenges at the present time is interlinking the motley amount of linguistic data and metadata stored in the resources developed over the past five decades of Computational Linguistics and empirical language studies (Chiarcos et al., 2012, p. 1).

A current approach to interlinking linguistic resources takes up Linked Data principles, so that «it is possible to follow links between existing resources to find other, related data and exploit network effects» (Chiarcos et al., 2013, p. iii). According to the Linked Data paradigm, data in the Semantic Web (Berners-Lee et al., 2001) are interlinked through connections that can be semantically queried, so as to make the structure of web data better serve the needs of users.

One subfield that has enjoyed particular prosperity over the past decade is that devoted to ancient languages. Owing to their key role in accessing and understanding the so-called Classical tradition, Latin and Ancient Greek are among the main beneficiaries. Thanks to international efforts, several textual and lexical resources as well as Natural Language Processing (NLP) tools are currently available for Latin. Despite the launch of a number of projects for automatic extraction of structured knowledge from ancient sources in the last decade, much like other languages, linguistic resources and tools for Latin often live in isolation, a condition which prevents them from benefiting a large research community of historians, philologists, archaeologists and literary scholars.

To this end, the LiLa: Linking Latin project (2018-2023)[1] is building a Knowledge Base of linguistic resources for Latin based on the Linked Data paradigm, i.e. a collection of multifarious, interlinked data sets described with the same vocabulary of knowledge description, by using common data categories and ontologies. This abstract presents the architecture of the LiLa Knowledge Base, details the linguistic resources for Latin currently interlinked through LiLa and informs about how the Knowledge Base can be queried.

2 The Architecture of LiLa

Given the presence and role played by lemmatization in various linguistic resources and the good accuracy rates achieved by the best performing lemmatizers for Latin (up to 95.30%, as per (Eger et al.,

[1] https://lila-erc.eu/

2015))[2], LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and NLP tools. Consequently, the LiLa Knowledge Base is highly lexically based, grounding on a simple but effective assumption that strikes a good balance between feasibility and granularity: lexical resources describe properties of words (in lexical entries), textual resources are made of occurrences of words in texts (tokens), and NLP tools process words, producing outputs.

The core of the LiLa Knowledge Base consists of a large collection of Latin lemmas (called Lemma Bank)[3]: interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma (Passarotti et al., 2020) (Figure 1).

3 Linguistic Resources in LiLa

The linguistic resources currently interlinked through the LiLa Knowledge Base are the following.

Textual resources:

- Index Thomisticus Treebank (Passarotti, 2019): the largest syntactically annotated corpus for Latin available, featuring more than 400,000 tokens from works of Thomas Aquinas (including the entire text of *Summa contra Gentiles*). The treebank is available in LiLa in both its formats: the original one (Bamman et al., 2008) and its conversion into *Universal Dependencies* (Cecchini et al., 2018).

- UDante: the opera omnia in Latin of Dante Alighieri (approximately 50,000 tokens) enhanced with syntactic annotation according to the *Universal Dependencies* style (Cecchini et al., 2020).

- *Querolus sive Aulularia*: the text of an anonymous short comedy from late antiquity.

- LASLA Corpus: a large corpus collecting approximately 1,7 million words from Classical and Late Latin texts lemmatized and morphologically tagged (Verkerk et al., 2020).

Lexical resources:

- Word Formation Latin: a Latin lexicon where lexical entries (around 30,000) are related by wordformation rules (Litta et al., 2019).

- Etymological Dictionary of Latin & the Other Italic Languages: a dictionary featuring proto-Indoeuropean and proto-Italic reconstructed forms to explain the history of 1,400 Latin forms (Mambrini and Passarotti, 2020).

- Latin Vallex 2.0 + Latin WordNet: a manually checked subset of the Latin WordNet, where each sense of a word is assigned a valency frame (Mambrini et al., 2021b).

- Index Graecorum Vocabulorum in Linguam Latinam Translatorum: a list of 1,763 Ancient Greek loanwords in the Latin language published in 1874 by classical scholar Günther Alexander E. A. Saalfeld (Franzini et al., 2020).

- LatinAffectus: a lexicon that assigns a prior sentiment score to a selection of more than 2,500 Latin adjectives and nouns (Sprugnoli et al., 2020a).

- Lewis & Short Latin-English dictionary: a Latin-English bilingual dictionary curated by Ch. T. Lewis and Ch. Short, and published by Harper and Oxford University Press in 1879 (Mambrini et al., 2021a).

4 Querying LiLa

The LiLa Knowledge Base can be queried through a SPARQL endpoint at `https://lila-erc.eu/sparql/`, where a number of pre-compiled queries are available. Federated queries can be run on all the textual and lexical resources currently interlinked in the Knowledge Base.

The collection of Latin lemmas of LiLa can be accessed at `https://lila-erc.eu/query/`. Lemmas can be searched by string of characters (also using regular expressions), 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.[4]

[2] For the state of the art in automatic lemmatization and PoS tagging for Latin, see the results of the first edition of EvaLatin, a campaign devoted to the evaluation of NLP tools for Latin (Sprugnoli et al., 2020b).

[3] `https://lila-erc.eu/lodview/data/id/lemma/LemmaBank`

[4] The Turtle files of the resources interlinked in LiLa are available at `https://github.com/CIRCSE`.

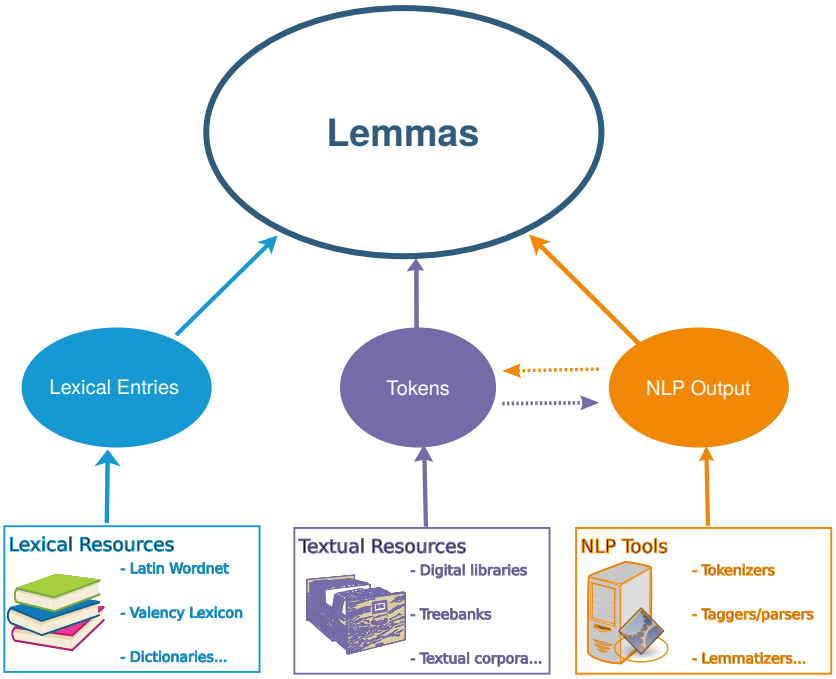

Figure 1: The architecture of LiLa.

References

David Bamman, Marco Carlo Passarotti, Roberto Busa, and Gregory Crane. 2008. The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin. In *LREC 2008*, pages 71–76. ELDA.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american*, 284(5):34–43.

Flavio M Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020. Udante: First steps towards the universal dependencies treebank of dante's latin works. In *Seventh Italian Conference on Computational Linguistics*, pages 1–7. CEUR-WS. org.

Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 27–36, Brussels, Belgium. Association for Computational Linguistics.

Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P. McCrae. 2013. Linguistic linked open data (LLOD). introduction and overview. In *Proc. of the 2nd Workshop on Linked Data in Linguistics*, pages i – xi, Pisa, Italy. ACL.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012. Introduction and overview. In *Linked Data in Linguistics*, pages 1–12. Springer.

Steffen Eger, Tim vor der Brück, and Alexander Mehler. 2015. Lexicon-assisted tagging and lemmatization in latin: A comparison of six taggers and two lemmatization methods. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 105–113.

Greta Franzini, Federica Zampedri, Marco Passarotti, Francesco Mambrini, and Giovanni . 2020. Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Seventh Italian Conference on Computational Linguistics. Bologna, Italy, March 1-3, 2021*, pages 1–6, Bologna. CEUR-WS.org.

Eleonora Litta, Marco Passarotti, and Francesco Mambrini. 2019. The treatment of word formation in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 35–43, Prague, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

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). Milan, Italy, January 26-28, 2022*, pages 1–7, Milan, Italy. CEUR Workshop Proceedings.

Francesco Mambrini and Marco Passarotti. 2020. Representing etymology in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 20–28, Marseille, France. European Language Resources Association.

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.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. In *Digital Classical Philology*, number 10 in Age of Access? Grundfragen der Informationsgesellschaft, pages 299–320, Berlin, Germany; Boston, MA, USA. De Gruyter Saur.

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*, LVIII(1):177–212.

Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020a. Towards the modeling of polarity in a Latin knowledge base. In *WHiSe 2020 Workshop on Humanities in the Semantic Web 2020*, pages 59–70, Heraklion, Greece. CEUR.

Rachele Sprugnoli, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020b. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Philippe Verkerk, Yves Ouvrard, Margherita Fantoli, and Dominique Longrée. 2020. Lasla and collatinus: a convergence in lexica. *Studi e Saggi Linguistici*, 58(1):95–120.