Wordforms and Meanings: an Updated Report on the LiLa Project

Marco Passarotti, Flavio Massimiliano Cecchini, Eleonora Litta, Francesco Mambrini, Giovanni Moretti, Giulia Pedonese, Matteo Pellegrini, Paolo Ruffolo, Rachele Sprugnoli, Marinella Testori
Università Cattolica del Sacro Cuore di Milano, Italy - {nome.cognome}@unicatt.it

ABSTRACT
This contribution presents the current status of the ERC project "LiLa: Linking Latin", the main objective of which is to connect and exploit the wealth of existing linguistic resources for Latin by making them interoperable, through the creation of a Knowledge Base following Linked Data standards. We describe the textual and lexical resources linked to the Knowledge Base and the ways in which it is possible to query and explore them.

KEYWORDS
Linguistic resources, Latin, Semantic Web.

POSTER

1. INTRODUCTION
Linguistic resources are machine-readable collections of language data and descriptions. Thanks to international efforts, several resources as well as Natural Language Processing (NLP) tools are currently available for ancient languages, including Latin. Linguistic resources are usually classified in two main categories depending on the kind of content they contain: (a) textual resources, such as written corpora, featuring either partial or full texts which may differ in genre, author or time period and (b) lexical resources like lexica, dictionaries and terminological databases providing information on lexical items for one or more languages including definitions, translations and morphological properties.
However, despite the increase in their quantity and coverage, linguistic data and metadata today are scattered in isolated resources, preventing users (in particular those from the humanities, such as historians, philologists, archaeologists and literary scholars) from honing both their individual and joint potential across platforms.
A current approach to making linguistic resources interact takes up Linked Data principles ([2];[3]), according to which data in the Semantic Web ([1]) are interlinked through connections that can be semantically queried so that the structure of web data can better answer to the needs of users.
With this in mind, the "LiLa: Linking Latin project" (2018-2023: https://lila-erc.eu) was awarded funding from the European Research Council (ERC) to build a Knowledge Base (KB) of linguistic resources for Latin following the Linked Data paradigm: the KB is a collection of diverse, interlinked data sets described with the same vocabulary of knowledge description that uses common data categories and ontologies ([10]). Given the presence and role played by lemmatization in various linguistic resources and the good accuracy rates achieved by state-of-the-art lemmatizers for Latin (up to 95.30% ([7]))[1], LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and NLP tools. Accordingly, the LiLa KB is highly lexically based, grounding on the simple postulation that strikes a good balance between feasibility and granularity: textual resources are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words. This granted, the heart of the LiLa KB consists of a large collection of Latin lemmas called Lemma Bank, currently comprising of more than 130,000 canonical forms: interoperability is attained by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. The linguistic properties of the Latin lemmas in LiLa are expressed as RDF triples using the LiLa ontology semantics.

[1] Such high rates of automatic lemmatization of Latin should be taken with a grain of salt. Indeed, performances of stochastic NLP tools heavily depend on the training set on which their models are built, and so decrease when they are applied to out-of-domain texts. This problem is particularly challenging for Latin owing to its wide diachrony (spanning two millennia), genre diversity (ranging from literary to philosophical, historical and documentary texts) and diatopy (Europe and beyond). 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 ([12]).

This abstract introduces the current status of the LiLa KB, focussing on the textual and lexical resources that were interlinked so far thanks to their association to the collection of lemmas of LiLa[2].

2. RESOURCES

In this section we provide a brief description of the resources linked so far via the LiLa KB covering different linguistic aspects (from morphology to syntax and semantics) and different time periods (from Late Antiquity to the Middle Ages) of Latin linguistic material. More specifically, the textual resources currently available are the Index Thomisticus Treebank (ITTB) containing the works by Thomas Aquinas, the corpus of Latin texts by, or disputedly attributed to, Dante Alighieri (UDante), the text of the comedy "Querolus sive Aulularia" and the eighth chapter of the "Liber Abaci", a mathematical treatise by Fibonacci. All these corpora are annotated following the Universal Dependencies framework ([4]): the last two resources are annotated with Part-of-Speech tags and lemmas whereas ITTB and UDante also contain syntactic information.

For what lexical resources are concerned, the LiLa KB currently contains: a collection of Proto-Italic and Proto-Indo-European reconstructed forms taken from the "Etymological Dictionary of Latin and the other Italic Languages" ([5]), the LatinAffectus sentiment lexicon, a collection of Ancient Greek loanwords in the Latin language extracted from the "Index Graecorum vocabulorum in linguam Latinam translatorum quaestiunculis auctus" ([11]), around 1800 manually checked entries of the Latin WordNet mapped onto Princton WordNet 3.0, a valency lexicon for Latin and a derivational morphology lexicon. In order to achieve interoperability, all these resources are modeled and described using ontologies such as Ontolex ([9]) and encoded in a graph-based data structure in RDF.

3. QUERYING THE KNOWLEDGE BASE

At the time of writing, there are two ways for querying the LiLa KB: through the Query Interface (https://lila-erc.eu/query/) or using the SPARQL endpoint. The Query Interface is a user-friendly graphical web application for searching the lemmas in the Lemma Bank, suitable for those unfamiliar with SPARQL. Users can search for a specific lemma or part of it or compose their own query by dragging and dropping any combination of query modules: each query module allows to filter the results with respect to a grammatical or morphological feature (such as gender, PoS, presence of a suffix) by choosing an option from a drop-down menu. Results can be saved as a CSV file. Alternatively, it is possible to copy the underlying SPARQL query and view the complete lemma description or the corresponding graph representation. Figure 1 shows a query retrieving all common nouns with masculine gender having the suffix -(t)or: this query has 1,528 results and the first three lemmas in alphabetical order are *abactor* "a cattle-stealer" and *abbreviator* "epitomist".

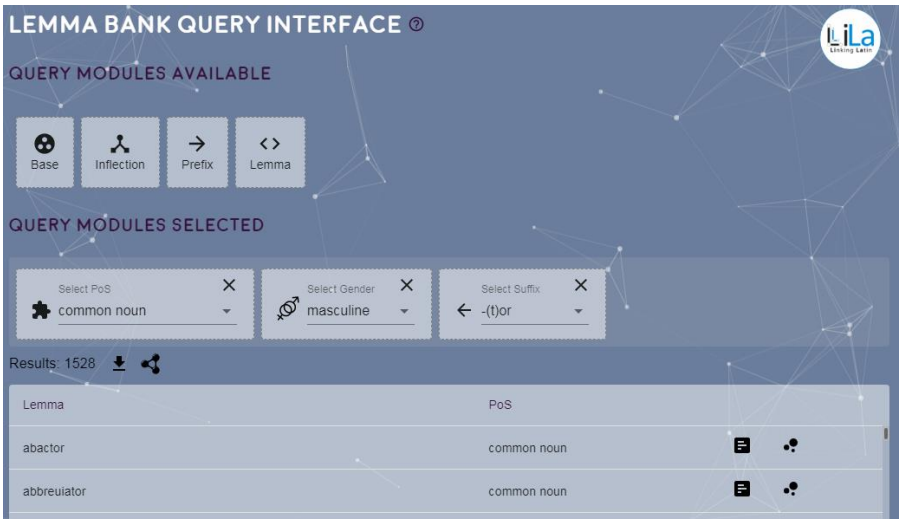

Figure 1. Screenshot of the Lemma Bank Query Interface.

Via the SPARQL endpoint (https://lila-erc.eu/sparql/) it is instead possible to access the ever-growing collection of connected resources beyond the Lemma Bank and perform more complex searches. We release and constantly update a set of queries in a dedicated GitHub repository to facilitate the use of the endpoint: https://github.com/CIRCSE/SPARQL-

[2] Both the collection of lemmas and the source data of the resources linked to LiLa (together with their TTL files, which provide the RDF triples) are freely available from the GitHub page of the host institution's CIRCSE research center: https://github.com/CIRCSE.

queries. For example, the query UDante-sentiment.rq in the repository works on 3 different interlinked resources, i.e., LatinAffectus, the Lemma Bank and UDante to retrieve all lemmas in UDante that appears in the sentiment lexicon with a negative polarity and count the total number of occurrences per lemma. This query results in the following top 5 lemmas with a negative sentiment: *peccatum* "sin" (17 occurrences), *litigium* "quarrel" (16), *mors* "death" (15), *malus* "bad" (12), *iniura* "injurious" (11).

4. UPCOMING RESOURCES

We are currently working on modelling and linking the two following resources:

1. the bilingual "Latin Dictionary" curated by Ch. T. Lewis and Ch. Short and published by Harper and Oxford University Press in 1879 ([8]).
2. the LASLA corpus developed by the homonymous laboratory in Liége, Belgium, which currently includes more than 150 texts from around 20 authors for a total of approximately 1,700,000 words ([6]).

REFERENCES

[1] Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American* 284 (5): 34–43.

[2] Chiarcos, Christian, Philipp Cimiano, Thierry Declerck, and John. P. McCrae. 2013. "Linguistic Linked Open Data (Llod). Introduction and Overview." In *Proceedings of the 2nd Workshop on Linked Data in Linguistics: Representing and Linking Lexicons, Terminologies and Other Language Data*.

[3] Chiarcos, Christian, Sebastian Nordhoff, and Sebastian Hellmann. 2012. *Linked Data in Linguistics*. Heidelberg: Springer.

[4] De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. "Universal Dependencies." *Computational Linguistics* 47 (2): 255–308.

[5] De Vaan, Michiel. 2008. *Etymological Dictionary of Latin and the Other Italic Languages*. Vol. 7. Boston: Brill, Leiden.

[6] Denooz, Joseph. 2007. "Opera Latina: Le Nouveau Site Internet Du Lasla." *Journal of Latin Linguistics* 9 (3): 21–34.

[7] Eger, Steffen, 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*.

[8] Lewis, Charlton Thomas. 1884. *Harpers' Latin Dictionary: A New Latin Dictionary Founded on the Translation of Freund's Latin-German Lexicon*. Harper & brothers.

[9] McCrae, John. P., Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon Model: Development and Applications." In *Proceedings of ELex 2017 Conference*.

[10] Passarotti, Marco, 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* 58 (1): 177–212.

[11] Saalfeld, Alexander. 1874. *Index Graecorum Vocabulorum in Linguam Latinam Translatorum Quaestiunculis Auctus.* F. Berggold.

[12] Sprugnoli, Rachele, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020. "Overview of the EvaLatin 2020 Evaluation Campaign." In *Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages*.