Journal of Open Humanities Data

The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms

COLLECTION:
REPRESENTING THE ANCIENT WORLD THROUGH DATA

DATA PAPER

FRANCESCO MAMBRINI
MARCO CARLO PASSAROTTI
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The dataset contains a list of 215,102 Latin dictionary forms (known as canonical forms or lemmas). The dataset is a set of 1,699,687 Resource Description Framework (RDF) triples that describe, using a series of Web Ontology Language (OWL) ontologies for Linguistic Linked Data, the morphological properties of these forms. The dataset is used to link together a series of corpora and dictionaries in the interoperable network of language resources published by the *LiLa: Linking Latin* project.

CORRESPONDING AUTHOR:
Marco Carlo Passarotti
CIRCSE Research Center, Università Cattolica del Sacro Cuore, Milan, IT
marco.passarotti@unicatt.it

KEYWORDS:
Latin; lemmatization; linked open data; interoperability

TO CITE THIS ARTICLE:
Mambrini, F., & Passarotti, M. C. (2023). The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms. *Journal of Open Humanities Data*, 9: 28, pp. 1–5. DOI: https://doi.org/10.5334/johd.145

1 OVERVIEW

REPOSITORY LOCATION

Zenodo: https://doi.org/10.5281/zenodo.8300851.

CONTEXT

The dataset was produced by the ERC project *LiLa: Linking Latin*; it has been used in all the project publications, including Pellegrini et al. (2022); Sprugnoli, Mambrini, Passarotti, and Moretti (2023); Sprugnoli, Moretti, and Passarotti (2020). A full list of publications is available at https://lila-erc.eu/output (last accessed: 26, October,2023).

2 METHOD

STEPS

Our goal was to generate a dataset with all dictionary forms (known as canonical forms or lemmas) that can be adopted by projects dealing with the lemmatization of Latin, independently from the particular strategies adopted by each project. We wanted to provide each lemma with a stable Uniform Resource Identifier (URI) and to model its linguistic properties with the help of OWL ontologies for Linguistic Linked Open Data (LOD) (Khan et al., 2022).

The starting point was the list of lemmas used by the morphological analyzer LEMLAT 3.0 (Passarotti, Budassi, Litta, & Ruffolo, 2017). The list of Latin lexical items used by the software was compiled from three sources: a set of dictionaries of Classical Latin (Georges & Georges, 1913-1918; Glare, 2012; Gradenwitz, 1904), the *Onomasticon* by Forcellini (Budassi & Passarotti, 2016), and the Medieval glossary of Du Cange et al. (Cecchini et al., 2018). For the Classical words, LEMLAT also includes information on the derivational history of words (Litta, Passarotti, & Culy, 2016).

LEMLAT's lemmas have undergone a twofold process of revision. Firstly, we manually identified and merged the duplicate entries from Classical and Medieval Latin. Secondly, we generated all possible inflected forms that may be chosen as lemmas (such as the present, perfect and future participles of verbs, or de-adjectival adverbs) that were not already in LEMLAT. Finally, we developed an OWL ontology,[1] based on widely adopted standards for Linguistic Linked Data like `Ontolex` (McCrae, Bosque-Gil, Gracia, Buitelaar, & Cimiano, 2017), to express the different classes and properties used in lemmatization (Passarotti et al., 2020), and we modeled the lemmas according to it.

A first version was published in 2020 and it included 196,853 lemmas; it has been revised and extended (see below under "Quality control"). Version 1.2 now includes 215,102 canonical forms. The information was originally stored in a relational database (MariaDB); the RDF triples were generated from this source. Both versions (RDF, and Structured Query Language (SQL) format) are provided, together with instructions on how to generate the RDF from the database.

QUALITY CONTROL

Quality control has been constantly performed during the linking process of lexical and textual resources within the LiLa network. The Lemma Bank has been used to interlink 10 lexical resources and 9 textual corpora of various size.[2] Linking those resources means to match the lemma string used to lemmatize the entries in the original resource to the entries in our Lemma Bank. This process helped us in identifying several missing lemmas (in particular for proper nouns) and several duplicated entries. After manual revision, the former were added to the collection, while the latter were merged with pre-existing entries.

For an example of this workflow involving the ca. 1.7 million lemmatized words of the *Opera Latina* by LASLA, see Fantoli, Passarotti, Mambrini, Moretti, and Ruffolo (2022).

1 https://github.com/CIRCSE/LiLaOntologies (last accessed: 26, October, 2023).

2 A list is available at: https://lila-erc.eu/data-page (last accessed: 26, October, 2023).

DOI: 10.5334/johd.145

3 DATASET DESCRIPTION

Object name The LiLa Lemma Bank (V1.2).

Format names and versions V1.2: Turtle serialization of RDF; SQL file. The first version of the Turtle RDF was included in the ILC-CNR for CLARIN-IT repository[3] under a more restrictive license (CC-BY-SA-NC 4.0).

Creation dates 2020-11-25 to 2023-08-30.

Dataset creators Marco Carlo Passarotti (supervisor), Flavio Massimiliano Cecchini (developer), Greta Franzini (annotator), Federica Iurescia (annotator), Eleonora Litta (annotator), Francesco Mambrini (annotator), Giovanni Moretti (developer), Giulia Pedonese (annotator), Matteo Pellegrini (annotator), Paolo Ruffolo (developer), Rachele Sprugnoli (annotator), Marinella Testori (annotator). Affiliation of all (at the time of data development): Università Cattolica del Sacro Cuore, Milan, Italy.

Language Latin; English for metadata.

License Creative common Attribution - ShareAlike 4.0 International (CC BY-SA 4.0).

Repository name Zenodo, GitHub.

Publication date 2023-08-30 (V1.2).

4 REUSE POTENTIAL

LOD PUBLICATION

Any project wishing to publish linguistic information about Latin words or texts may use the URIs from the Lemma Bank to link their data. Indeed, the dataset provides easily reusable unique identifiers for a wide set of Latin canonical forms, which are already linked to a wealth of textual and lexical information. The dataset relies on a W3C de-facto standard for lexical information (`Ontolex`): any project adopting this model may easily reuse our data. For instance, Wikidata Latin lexemes (Nielsen, 2020) provide links to the LiLa lemmas with an ad-hoc property `LiLa Linking Latin URI`.[4]

LINGUISTIC RESEARCH

The LOD paradigm used to connect resources via the URIs from the LiLa Lemma Bank allows researchers to run sophisticated queries across multiple layers of information. Users can, for instance, know how many derivative words that are etymologically linked to an Indo-European root exist in Latin, and where they are attested (Mambrini & Passarotti, 2020), or what the distribution is of negative and positive words in the lyrics of Horace (Sprugnoli et al., 2023).

The LiLa project provides a SPARQL endpoint, with pre-compiled queries that showcase some of these applications.[5] As other SPARQL services start to provide access to data linked to the LiLa Lemma Bank (such as the Wikidata query service, where it is now possible to run federated queries to the LiLa SPARQL endpoint),[6] this potential will only grow.

LANGUAGE LEARNING

The interoperability between resources can also be leveraged in the context of language learning. Latin is still widely studied in universities and secondary schools worldwide; however, the need for newer methods from current research on computational and corpus linguistics to facilitate the students' access to the language is strongly felt in the community, especially in the domain of word usages and meanings (Beyer, Schulz, Beyer, & Schulz, 2022). The capability of crossing multiple lexical resources (like sentiment, valency and word-formation lexicons)

3 http://hdl.handle.net/20.500.11752/OPEN-532 (last accessed: 26, October, 2023).

4 https://www.wikidata.org/wiki/Q117023407.

5 https://lila-erc.eu/sparql/ (last accessed: 26, October, 2023).

6 https://query.wikidata.org/ (last accessed: 26, October, 2023). A sample query is available at: https://w.wiki/7Si$ (last accessed: 26, October, 2023).

with textual attestations can be extremely helpful. Word lists can be easily generated to help teachers, including, for instance, nouns with positive polarity, grouped by derivational patterns (such as verb-to-noun derivations involving the suffix *-(t)io(n)*) in reversed frequency order, based on the number of occurrences in one or more reference corpora. The web-based query interface provided by LiLa can be used to that purpose.[7]

NATURAL LANGUAGE PROCESSING

The Lila Lemma Bank can support NLP tasks such as (1) lemmatization, (2) Part-of-Speech (PoS) tagging and (3) morphological analysis. As for (1), tools for automatic lemmatization can benefit from the connections of the canonical forms in the Lemma Bank to their occurrences in the interlinked corpora. These connections can be used to build a large lemmatized meta-corpus for Latin that may serve as a training set for a stochastic tool. As for (2), all forms in the Lemma Bank are assigned a PoS and that information can be exploited by PoS taggers in both training and testing phases. As for (3), the Lemma Bank enhances the canonical forms with morphological features such as gender and inflectional category, as well as derivational information on word formation.

FUNDING STATEMENT

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.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Francesco Mambrini: Conceptualization, Writing – original draft. Marco Carlo Passarotti: Conceptualization, Supervision, Funding acquisition.

AUTHOR AFFILIATIONS

Francesco Mambrini orcid.org/0000-0003-0834-7562
CIRCSE Research Center, Università Cattolica del Sacro Cuore, Milan, IT
Marco Carlo Passarotti orcid.org/0000-0002-9806-7187
CIRCSE Research Center, Università Cattolica del Sacro Cuore, Milan, IT

REFERENCES

Beyer, A., Schulz, K., Beyer, A., & **Schulz, K.** (2022, December). New Insights and methods of vocabulary acquisition in Latin classes. *Forma y Función, 35*(2). DOI: https://doi.org/10.15446/fyf.v35n2.91129

Budassi, M., & **Passarotti, M.** (2016, August). Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 90–94). Berlin, Germany: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/W16-2110

Cecchini, F. M., Passarotti, M., Ruffolo, P., Testori, M., Draetta, L., Fieromonte, M., … Piantanida, G. (2018). Enhancing the Latin Morphological Analyser LEMLAT with a Medieval Latin Glossary. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)* (pp. 87–92). Torino: aAccademia University Press. DOI: https://doi.org/10.4000/books.aaccademia.3121

Fantoli, M., Passarotti, M., Mambrini, F., Moretti, G., & **Ruffolo, P.** (2022, June). Linking the LASLA Corpus in the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference* (pp. 26–34). Marseille, France: European Language Resources Association. Retrieved 2023-03-25, from https://aclanthology.org/2022.ldl-1.4

Georges, K. E., & **Georges, H.** (1913–1918). *Ausführliches lateinisch-deutsches Handwörterbuch.* Hannover: Hahnsche Buchhandlung.

7 https://lila-erc.eu/query/ (last accessed: 26, October, 2023).

Glare, P. G. W. (2012). *Oxford Latin Dictionary* (2nd ed.). Oxford, UK: Oxford University Press.

Gradenwitz, O. (1904). *Laterculi vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. Leipzig: Hirzel.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., García, E. G.-B., Gracia, J., … Truică, C.-O. (2022, September). When linguistics meets web technologies. Recent advances in modelling linguistic linked data. *Semantic Web, 13*(6), 987–1050. DOI: https://doi.org/10.3233/SW-222859

Litta, E., Passarotti, M., & **Culy, C.** (2016). Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the third italian conference on computational linguistics (clic-it 2016)* (pp. 185–189). Naples: Accademia University Press. DOI: https://doi.org/10.4000/books.aaccademia.1799

Mambrini, F., & **Passarotti, M.** (2020, May). Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 2020 globalex workshop on linked lexicography* (pp. 20–28). Marseille, France: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/2020.globalex-1.3

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & **Cimiano, P.** (2017). The OntoLex-Lemon Model: development and applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference* (pp. 587–597). Brno: Lexical Computing. Retrieved from https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

Nielsen, F. (2020, May). Lexemes in Wikidata: 2020 status. In *Proceedings of the 7th workshop on linked data in linguistics (ldl-2020)* (pp. 82–86). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.ldl-1.12

Passarotti, M., Budassi, M., Litta, E., & **Ruffolo, P.** (2017). The Lemlat 3.0 Package for Morphological Analysis of Latin. In G. Bouma & Y. Adesam (Eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language* (Vol. 133, pp. 24–31). Gothenburg: Linköping University Electronic Press.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., … Sprugnoli, R. (2020). Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici, 58*, 177–212. DOI: https://doi.org/10.3233/SSW210032

Pellegrini, M., Passarotti, M., Litta, E., Mambrini, F., Moretti, G., Corbetta, C., & **Verdelli, M.** (2022, October). Enhancing Derivational Information on Latin Lemmas in the LiLa Knowledge Base. A Structural and Diachronic Extension. *Prague Bulletin of Mathematical Linguistics, 119*(1), 67–92. DOI: https://doi.org/10.14712/00326585.023

Sprugnoli, R., Mambrini, F., Passarotti, M. C., & **Moretti, G.** (2023). The Sentiment of Latin Poetry. Annotation and Automatic Analysis of the Odes of Horace. *Italian Journal of Computational Linguistics (IJCOL), 9*(1), 53–71. DOI: https://doi.org/10.4000/ijcol.1125

Sprugnoli, R., Moretti, G., & **Passarotti, M.** (2020). Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas. *Italian Journal of Computational Linguistics (IJCOL), 6*(1), 29–45. DOI: https://doi.org/10.4000/ijcol.624

TO CITE THIS ARTICLE:
Mambrini, F., & Passarotti, M. C. (2023). The LiLa Lemma Bank: A Knowledge Base of Latin Canonical Forms. *Journal of Open Humanities Data*, 9: 28, pp. 1–5. DOI: https://doi.org/10.5334/johd.145

Submitted: 31 August 2023
Accepted: 26 October 2023
Published: 24 November 2023

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.