The LiLa Knowledge Base as a Reference Model for Interoperability between Linguistic Resources

Flavio Massimiliano Cecchini and **Marco Passarotti**
Università Cattolica del Sacro Cuore
Largo A. Gemelli 1
20155 Milan (MI), Italy
{flavio.cecchini,marco.passarotti}@unicatt.it

Relevant UniDive working groups: WG1,WG2

1 Introduction

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 still locked in data silos, failing to provide a comprehensive overview of the annotations available in these separate collections.

A current approach to interlinking linguistic resources builds in 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. What is still missing, however, is a fine-grained level of interaction between linguistic resources capable of stretching beyond descriptive metadata over to individual word occurrences in a text or entries in a lexicon.

To this end, the *LiLa: Linking Latin* project has built 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 (Passarotti et al., 2020). By using standard and language-independent data categories and ontologies, the architecture of the LiLa Knowledge Base is highly portable across languages and wants to become a reference model to address the issue of communication gaps between linguistic resources.

By presenting the lemma-based architecture of the LiLa Knowledge Base, this abstract discusses how LiLa harmonizes the different criteria for lemmatization that can be found in annotated corpora for Latin, to make distributed linguistic resources interact on the Web.

The final goal is to produce a framework capable of being expanded by any existing or future linguistic resource dealing with or related to Latin, and through which it will be possible to perform linguistically informed queries over all possible annotation levels and written documents, bridging the discrepancies between competing and conflicting annotation standards (intralinguistic harmonization). The more general structure of such a framework is by all means not limited to Latin, but represents a model for any other similar endeavor on other languages; in fact, it is possible to envision a connection between knowledge bases sharing the same principles of LiLa. Current examples in LiLa of expansions beyond a mere Latin horizon are the inclusion in the knowledge base of Proto-Italic and Proto-Indo-European roots (Mambrini and Passarotti, 2020), the treatment of Ancient Greek loanwords as part of the Latin lexicon (Franzini et al., 2020), and the addition of a selection of Latin loanwords used in Medieval Italian.[1]

2 The LiLa Knowledge Base

2.1 Harmonizing Criteria of Lemmatization

Since lemmatization is a layer of annotation and organization of linguistic data common to different kinds of resources, LiLa uses lemmas as the most productive interface between lexical resources, annotated corpora and NLP tools. As a consequence, the core of the LiLa Knowledge Base is a large collection of Latin lemmas (called Lemma Bank): interoperability is achieved by linking all those entries in lexical resources to tokens in corpora that point to the same lemma.

The Lemma Bank is a collection of individuals

[1] https://github.com/CIRCSE/DanteLatinLoanWords

that belong to the Lemma class as defined in the LiLa ontology: lemmatization is the task of indexing series of inflected forms under one form that is conventionally identified as the canonical form of citation. As such, the Lemma is safely subsumed under the general class of Form as defined in the Ontolex ontology (McCrae et al., 2017), a *de facto* standard in the Linked Data publication of lexical resources. Relying on the concepts of Ontolex, we define a Lemma as a Form that is linked to a Lexical Entry via the property "canonical form".

While the process of selecting the canonical forms to be used as lemmas tends to follow a standardized series of language-dependent conventions (e. g. for Latin nouns, usually the nominative singular form is chosen), building and structuring a repository of canonical forms that may serve as a hub in LiLa is complicated by the fact that different corpora, lexica and tools (for Latin as for many other languages) adopt different strategies to solve the conceptual and linguistic challenges posed by lemmatization, namely (1) the form of the lemma and (2) lemmatization criteria.

With respect to the former, citation forms for the same lexical item chosen to represent the lemma can differ in (a) graphical representation (***v***oluptas vs. ***u***oluptas 'satisfaction'), (b) spelling (sul***f***ur vs. sul***ph***ur 'brimstone'), (c) ending and possibly inflectional type (diamet***er*** vs. diamet***ros*** vs. diamet***rus*** 'diameter']), or (d) in the paradigmatic slot representing the lemma (*sequor* 'to follow', first person singular of the passive/deponent[2] present indicative vs. *sequo*, first person singular of the active present indicative). In such cases, if two citation forms for the same item belong to different inflectional categories (the case of diamet***er*** vs. diamet***ros*** vs. diamet***rus***), or to different paradigmatic slots (the case of *sequor* vs. *sequo*), in the Lemma Bank of LiLa they are considered as two separate lemmas connected via a property called "lemma variant".[3] If they differ solely by spelling, the two citation forms are stored in the lexical collection of LiLa as separate Ontolex "written representations"[4] of the same lemma (the case of sul***f***ur vs. sul***ph***ur).

As for lemmatization criteria, differences are such that, on occasion, a word form might be traced back to multiple lemmas. For instance, this is the case of participles (i. e. verbal adjectives), which can be conceived either as parts of a verbal inflectional paradigm, or as independent lemmas. Accordingly, a participle can either be lemmatized under its corresponding verb, or under a dedicated participial lemma, either systematically or only when the participle has grown into an autonomous lexical item (e. g. *doctus* 'learned', morphologically the perfect participle of *doceo* 'to teach'). The same holds true for deadjectival adverbs (e. g. *aequaliter* 'evenly' from *aequalis* 'equal'), which are either lemmatized as forms of their base adjectives, or treated as independent lemmas. In such cases, we make use of a special sub-class of Lemma called Hypolemma. Typical hypolemmas in the LiLa Lemma Bank are participles for verbal lemmas and adverbs for adjectival lemmas. Lemmas and hypolemmas are linked to each other via the symmetric property "has hypolemma"/"is hypolemma".[5]

Thanks to this organization of citation forms, the Lemma Bank of LiLa makes it possible to harmonize the different lemmatization strategies that can be found in linguistic resources for Latin, by connecting all the occurrences of a same lexical item in various textual corpora, regardless of the citation forms chosen for their lemmatization. Similar issues and patterns are of course not limited to Latin, but can be found in any language: while their exact realization can vary, the framework described here remains a valid approach to treat them.

2.2 A Specific Example

As an example, we consider the lemma *claudeo*/*claudo* 'to limp'. In (Glare, 2012), the entry for this lemma includes both the second conjugation (*claudeo*) and the third conjugation variant (*claudo*), the latter also featuring the graphical variant *cludo*. In (Georges, 1998), alongside the citation forms *claudeo* and *claudo* we also find their respective morphologically passive ("deponent") counterparts *claudeor* and *claudor*.

In LiLa, these citation forms are represented by four lemmas, distinguished by inflectional category: *claudeo* and *claudo*, as well as their corresponding deponent forms *claudeor* and *claudor*,

[2] That is, morphologically passive while syntactically and semantically active.

[3] https://lila-erc.eu/lodview/ontologies/lila/lemmaVariant

[4] http://www.w3.org/ns/lemon/ontolex#writtenRep

[5] https://lila-erc.eu/lodview/ontologies/lila/hasHypolemma. https://lila-erc.eu/lodview/ontologies/lila/isHypolemma

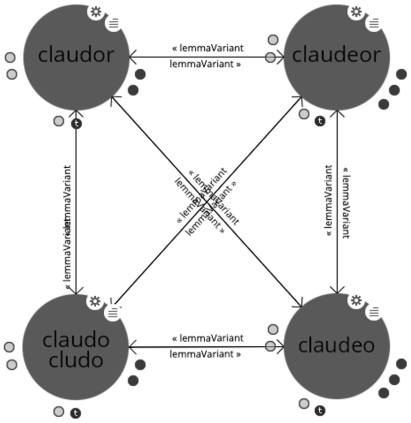

Figure 1: Citation forms for a lexical entry in LiLa.

are citation forms of different lemmas, as they follow different inflectional categories (active and passive second conjugation, respectively); *cludo*, on the other hand, is merged with *claudo*, sharing the same inflection. The LiLa Lemma Bank connects these four lemmas (see Figure 1) via the 'lemma variant' property, while *cludo* and *claudo* are represented as written representations, i. e. graphical variants of the same lemma. Finally, all four lemmas are connected to their respective hypolemmas for present, future and perfect participles.

In doing so, LiLa harmonizes different lemmatization strategies, granting interoperability. In the example of *claudeo*/*claudo*, all occurrences of this lexical item in the lemmatized corpora and lexica available in LiLa can be joined together by using a set of five connected citation forms, regardless of which citation form is used in a specific resource.

3 Annotation

Within the LiLa project, several textual resources have been annotated[6] and interlinked through the Lemma Bank. Annotation always includes lemmas and parts of speech, and in some cases also morphological features and syntactic relations, complying to the Universal Dependencies (UD) standard (de Marneffe et al., 2021). In particular, we performed the UD conversion of the *Index Thomisticus* treebank (Cecchini et al., 2018) and built the *UDante* treebank (Cecchini et al., 2020). This annotational endeavor and its interaction with the LiLa Knowledge Base has brought forth an ongoing effort of harmonization between Latin treebanks and the definition of more comprehensive and typologically grounded guidelines for Latin.

[6] https://lila-erc.eu/data-page/

References

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

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. The Association for Computational Linguistics (ACL).

Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020, Bologna, Italy, March 1–3 2021)*, pages 99–105, Turin, Italy. Associazione italiana di linguistica computazionale (AILC), Accademia University Press.

Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John Philip McCrae. 2013. Linguistic Linked Open Data (LLOD). Introduction and Overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (*LDL*-2013): Representing and linking lexicons, terminologies and other language data*, pages i – xi, Pisa, Italy. The Association for Computational Linguistics (ACL).

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

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

Karl Ernst Georges. 1998. *Ausführliches lateinisch-deutsches Handwörterbuch*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany. Reprint of first edition of 1913–1918, Hannover, Germany: Hahnsche Buchhandlung.

Peter Geoffrey William Glare. 2012. *Oxford Latin Dictionary*, second edition. Oxford Languages. Oxford University Press, Oxford, UK.

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 (ELRA).

John Philip McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The

OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, pages 587–597, Leiden, Netherlands. Lexical Computing CZ s.r.o.

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.