Linking the Corpus CLaSSES to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin

Irene De Felice[1,*], Lucia Tamponi[2], Federica Iurescia[3] and Marco Passarotti[3]

[1]Università di Genova, Italy
[2]Università di Pisa, Italy
[3]Università Cattolica del Sacro Cuore, Milan, Italy

Abstract

In this paper, we describe the process of linking the corpus CLaSSES (which collects non-literary Latin texts of different periods and places) to the LiLa Knowledge Base of linguistic resources for Latin made interoperable through their publication as Linked Data. The paper details the RDF modeling of the (meta)data provided by CLaSSES and presents three queries on data from different resources that interact in LiLa.

Keywords

Latin, Textual resources, Linguistic Linked Open Data

1. Introduction

The Latin language shows a large diversity, in the light of its wide usage in terms both of diachrony (spanning across two millennia) and diatopy (all over Europe and beyond). Such diversity is mirrored in the set of linguistic resources currently available for Latin, ranging from collections of literary texts of the Classical era,[1] to corpora of documentary texts of the Medieval times,[2] dictionaries,[3] and glossaries.[4]

Like for many other languages, one limitation that affects the wealth of resources for Latin is their sparseness, which prevents the full exploitation of the data they provide. The LiLa Knowledge Base was built to overcome such limitation, making distributed resources for Latin interact through their publication as Linked Data, by using a set of commonly used vocabularies provided by ontologies for the representation of linguistic (meta)data.

Among the resources interlinked in LiLa is the CLaSSES corpus, which enhances the set of lexical and textual data made interoperable by the Knowledge Base with a peculiar kind of non-literary Latin texts (such as inscriptions, writing tablets, and letters) written in different periods and provinces of the Roman Empire, thus contributing to extend the coverage of LiLa with a typology of texts not present so far in the Knowledge Base.

This paper details the process of linking CLaSSES to LiLa, and is organized as follows. Section 2 presents the corpus CLaSSES. Section 3 describes the LiLa Knowledge Base. Section 4 discusses the modeling and the linking of CLaSSES into LiLa. Section 5 reports three examples of queries that exploit the interoperability of CLaSSES with other resources in LiLa. Finally, Section 6 provides some conclusions on the results of the linking, and outlines directions of future work.

2. CLaSSES

CLaSSES (*Corpus for Latin Sociolinguistic Studies on Epigraphic textS*) is a digital resource created by the Laboratory of Phonetics and Phonology at Pisa University. Freely accessible on the internet,[5] it consists of over 3,400 non-literary Latin texts such as inscriptions, private letters, ink tablets, ostraka and papyri from various periods (6th century BCE to 6th century CE) and regions of the Roman Empire. The goal of CLaSSES is to use non-literary texts that exhibit (ortho-)graphic variants as a source to study the sociolinguistic variation of Latin [3, 4]. The identification of these spelling variants is the most crucial aspect of the corpus: words like *dedet*

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
irene.defelice@edu.unige.it (I. De Felice); lucia.tamponi@fileli.unipi.it (L. Tamponi); federica.iurescia@unicatt.it (F. Iurescia); marco.passarotti@unicatt.it (M. Passarotti)

Although this paper was conceived and discussed jointly by the authors, solely for academic purposes scientific responsibility is to be divided up as follows: I. De Felice wrote Sections 2 and 4.2; L. Tamponi Sections 5, 5.1, 5.2, 5.3; F. Iurescia wrote Section 4.1; M. Passarotti wrote Sections 1 and 3. Section 6 is to be attributed to all authors.

[1]Such as the LASLA corpus: https://www.lasla.uliege.be/cms/c_8570411/fr/lasla-textes-latins.
[2]Such as the corpus of Computational Historical Semantics: https://www.comphistsem.org/home.html.
[3]Such as the bilingual Latin-English dictionary curated by Ch. T. Lewis and Ch. Short [1].
[4]Such as the Medieval Latin *Glossarium Mediae et Infimae Latinitatis* by du Cange [2].
[5]http://classes-latin-linguistics.fileli.unipi.it.

(CIL-I^2-9-26) and *Vivia* (ILLRP-S-99-8) are categorized as “non-classical" forms in comparison to the standard spelling of Classical Latin, which would be *dedit* and *Vibia* respectively. CLaSSES is divided into four sections based on the place of provenance of the texts: Rome and Italy, Roman Britain, Sardinia, Egypt and Eastern Mediterranean. The database includes 3,415 texts, which were first automatically tokenized, resulting in 46,888 tokens. Then, expert annotators lemmatized the entire corpus manually, given the high number of incomplete and misspelt words that cannot be easily processed by automatic tools. They also provided a meta-linguistic and extra-linguistic annotation, including additional information about each document (place of provenance, dating, text type, author/addressee) and about each token of the corpus (graphic form, language). Finally, the linguistic annotation identifies non-classical variants and classifies them according to the variation phenomena [5, 6].

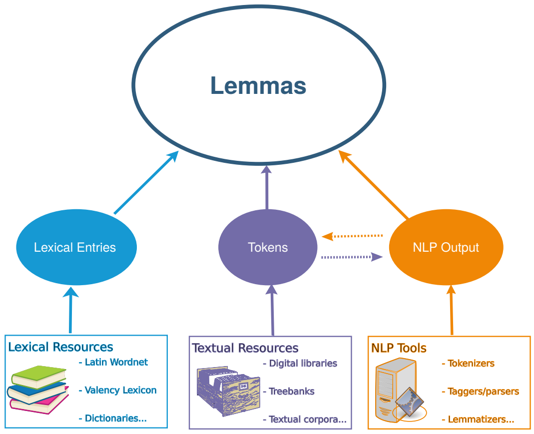

Figure 1: The architecture of LiLa.

3. The LiLa Knowledge Base

The aim of the “LiLa - Linking Latin” ERC project (2018-2023)[6] was to reach interoperability between the wealth of existing lexical and textual resources that have been developed in the last decades for Latin. One of the main problems that LiLa solved is the fact that such resources and tools are often characterized by different conceptual and structural models, which makes it difficult for them to interact with one another.

To this goal, LiLa undertook the creation of an open-ended Knowledge Base, following the principles of the Linked Data paradigm.[7] All content involved or referenced in the linguistic resources connected in LiLa is made unambiguously findable and accessible by assigning an HTTP Uniform Resource Identifier (URI) to each data point. Data reusability and interoperability between resources are achieved by establishing links between different URIs and by using web standards such as: [a] the RDF data model, which is based on triples: (i) a predicate-property connects (ii) a subject (a resource) with (iii) its object (another resource, or a literal) [7]; and [b] SPARQL, a query language specifically devised for RDF data.

Furthermore, the LiLa Knowledge Base makes reference to classes and properties of already existing ontologies to model the relevant information. The main ones are POWLA for corpus data [8], OLiA for linguistic annotation [9], and Ontolex-Lemon for lexical data [10, 11].

Within this framework, LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and NLP tools. Consequently, the architecture of the LiLa Knowledge Base is highly lexically based (Figure 1), grounding on a simple, but effective assumption that strikes a good balance between feasibility and granularity: textual resources are made of (occurrences of) words (“tokens”), lexical resources describe properties of words (in “lexical entries”), and NLP tools process words (producing “NLP outputs”).[8]

The core of the Knowledge Base is the so-called Lemma Bank,[9] a collection of about 200,000 Latin lemmas taken from the database of the morphological analyzer LEMLAT [12]. Interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma.

4. CLaSSES into LiLa

4.1. Modeling (Meta)data

The Lemma Bank of the LiLa Knowledge Base is modeled as a collection of Lexical Forms of Ontolex-Lemon. Lexical Forms are the inflected forms of Lexical Entries and are assigned one, or more graphical variants (`ontolex:writtenRep`).[10] One of the Lexical Forms of a Lexical Entry is linked to the latter by the property `ontolex:canonicalForm`, to model that it is the form that is canonically chosen to represent the entire lexical entry, i.e., the lemma. As a consequence, the Lemma Bank is not a lexical resource (as it does not contain Lexical Entries), rather it is a collection of Ontolex-Lemon Lexical Forms that can be used as Canonical Forms in the re-

[6] https://lila-erc.eu/.
[7] https://www.w3.org/DesignIssues/LinkedData.html.
[8] In Figure 1 the arrows going from and to the node for “NLP Output” represent the fact that tokens that are the output of a specific NLP tool (a tokenizer) become the input of further tools (like, for instance, a syntactic parser).
[9] http://lila-erc.eu/lodview/data/id/lemma/LemmaBank.
[10] http://www.w3.org/ns/lemon/ontolex#writtenRep.

sources for Latin to be interlinked in the LiLa Knowledge Base.

In particular, textual resources are connected to the Lemma Bank through the property `lila:hasLemma`,[11] which links a token in a corpus with its lemma in the Lemma Bank. In LiLa, textual resources are modeled as objects of the type Corpus from the POWLA ontology .[12] Each Corpus includes one, or more `powla:Document`,[13] which are the parts in which the corpus is divided, like for instance the different texts that it contains, or its sections. In the case of the Corpus entitled CLaSSES, there are 10 documents, corresponding to as many sections of the resource.[14] Every document of CLaSSES is assigned two layers, namely (1) a Document Layer, which collects all the tokens of a section, and (2) a Citation Layer, which records the full citation path of each token of a section.

For instance, Figure 2 shows the modeling of one token from CLaSSES. The token (*sacra*) is linked to its lemma in the Lemma Bank (*sacer*) by the `lila:hasLemma` property, and to the Document Layer by the `POWLA:hasLayer` property.[15] The properties `lila:isLayer`,[16] `lila:hasCitSubUnit`[17] and `POWLA:hasChild`[18] link the Citation Layer to the token. In the example of Figure 2, the token *sacra* occurs in the inscription number 27 of volume S of the Document entitled *Inscriptiones latinae liberae rei publicae*, to which both its Document and Citation Layers are linked through the property `POWLA:hasDocument`.[19]

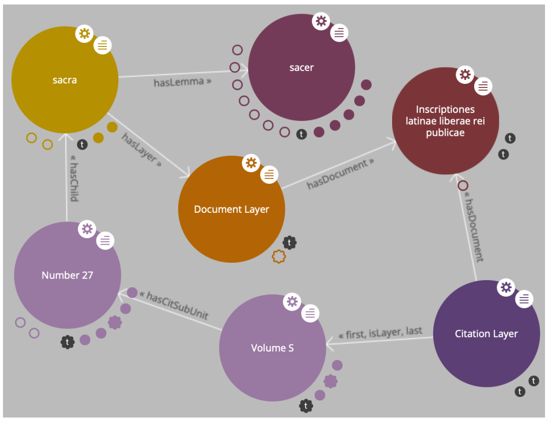

Figure 2: A token of CLaSSES in LiLa.

4.2. Linking Process

Out of the 46,888 tokens of CLaSSES, only those that are assigned a lemma are linked to the Lemma Bank of LiLa. Around 14k tokens of CLaSSES are not lemmatized due to the fragmentary nature of the texts contained therein. By exploiting the original lemmatization of the corpus, the automatic linking of the tokens of CLaSSES resulted in the following three output categories.

1. Perfect match (or one-to-one lemma; 25,279 items), i.e. whenever the lemma-PoS couple in CLaSSES was linked to one single lemma-PoS couple in the LiLa Lemma Bank. For such cases, we conducted an evaluation of the mapping on 10% of the couples. The data were randomly selected; to ensure that the sample was representative, the original PoS distribution was maintained.

   In CLaSSES, 3,490 different couples are recorded, thus the evaluation was conducted on 349 couples. Only 7 errors were found, all due to a wrong PoS tagging in the source data that caused a mapping error. Thus, the rate error is very low, i.e., 2%.
2. No match (or one-to-zero lemma; 5,366 items), i.e. when the lemma in CLaSSES was not associated with any lemma in LiLa. In this case, with the addition of the new lemma in LiLa we have enriched the Lemma Bank. Proper names are the category more affected, since inscriptions typically feature a wide range of anthroponyms which can identify the committee of the text (e.g., in public texts), the honorand (e.g., in sacred inscriptions) or the name of the dead on epitaphs [13]. In addition, given the wide geographical extension of our corpus, CLaSSES features local proper names typical of specific areas (e.g., Sardinia, or Roman Britain) that do not occur easily in Classical texts; an example from Sardinia [14, 15] is *Scribonissa* in ANRW-B61-6 [15, p. 45]. A few lemmas pertaining to other parts of speech were also added to the Lemma Bank, consisting mainly of hapax, like *ansata* in BTT-196-47 (lemma *ansatus* 'provided with handles'),[20] *infrascribo* in CEL-I-232-8 (lemma *infrascribo* 'to write lower down'),[21] *internumero* in BTT-645-48 (lemma *internumero* 'to reckon among other things').[22]
3. Ambiguous match (one-to-many lemmas; 1,503 items), i.e. when the lemma in CLaSSES was associated with several possible lemmas in LiLa. In most cases, the correct lemma between two or more possible ones was identified manually by a

[11] https://lila-erc.eu/ontologies/lila/hasLemma.
[12] http://purl.org/powla/powla.owl#Corpus.
[13] http://purl.org/powla/powla.owl#Document.
[14] http://lila-erc.eu/data/corpora/CLaSSES/id/corpus.
[15] http://purl.org/powla/powla.owl#hasLayer.
[16] https://lila-erc.eu/lodview/ontologies/lila_corpora/isLayer.
[17] https://lila-erc.eu/lodview/ontologies/lila_corpora/hasCitSubUnit.
[18] http://purl.org/powla/powla.owl#hasChild.
[19] http://purl.org/powla/powla.owl#hasDocument.
[20] http://lila-erc.eu/data/id/lemma/89148.
[21] http://lila-erc.eu/data/id/lemma/142756.
[22] http://lila-erc.eu/data/id/lemma/142757.

disambiguation based on the linguistic context of the document; this happens, for instance, in the case of homographs, as for the word *dico*, linked both to *dĭco*, 'to proclaim' or 'to dedicate'[23] and to *dīco*, 'to name', 'to utter'.[24] On rare occasions (29 tokens), it was however not possible to disambiguate between the lemmas available in LiLa: as a consequence, we linked the ambiguous tokens to all their corresponding lemmas. This was due to the fragmentary nature of some texts, where an analyzable context for disambiguation was not available. This is the case, for example, of BTT-609-16 *mallus* (context: [...] *mallus alu*[...] [...]*us*), for which two senses are equally possible, that of 'pole'[25] and 'appletree'.[26]

5. Querying CLaSSES in LiLa

Thanks to the interoperability of CLaSSES with the other resources for Latin linked to the LiLa Knowledge Base, research questions related to non-literary texts can be empirically investigated on the several different textual resources interlinked in the Knowledge Base by running queries on the SPARQL endpoint of LiLa[27]. By focusing on the question of spelling variants attested in the inscriptions, in what follows we shortly consider two case studies, i.e., consonant doubling (see 5.1), and the writing of long /i:/ through the diphthong <EI> (see 5.2). Moreover, we report and briefly discuss a query that exploits the information on derivational morphology recorded in the Lemma Bank (see 5.3).

5.1. Consonant doubling

As is known, the spelling of Latin long consonants through *geminatio consonantium* was introduced at the end of the third century BCE [16, 17, 18]. Consonant doubling, however, generalized slowly, so it was seldom omitted in the second century BCE in inscriptions. For example, in the 2nd-century inscriptions included in CLaSSES, 28 tokens (20 lemmas) display single for double consonants over 72 spellings with *geminatio consonantium*. These tokens can be easily retrieved through the function "Search for linguistic phenomena" available in the CLaSSES online search interface, by selecting the label "single pro double consonant". Thanks to the interoperability between distributed resources provided by LiLa, it is possible to search the occurrences of the lemmas for these tokens in the corpora interlinked through the Lemma Bank. This is particularly useful for both quantitative and qualitative linguistic analysis. For example, among the forms found in CLaSSES, it is possible to find occurrences of the same lemma either in the form with a single consonant or with a double consonant – such as the name *Mummius* in the *tituli mummiani*, which is displayed either in the forms *Mumius*, in CIL-I^2-628, or *Mummius*, in CIL-I^2-627, 629 [13]. The presence of the alternation between <C> and <CC> in these inscriptions can be interpreted as a sign of an incomplete generalization of consonant doubling at this stage. However, it is fundamental to exclude the possibility that the form *Mumius* occurring in our corpus represents a commonly attested variant of the proper name *Mummius*. This information is not readily retrievable in the available sources, since such spelling variants of proper names are generally not recorded in the dictionaries. However, by collecting the occurrences of the lemma *Mummius* in the textual resources interlinked through LiLa, it is possible to ascertain that the variant without consonant doubling is never attested in any of the texts provided by such resources (e.g., Cicero's *De Lege Agraria*, *In Verrem* and Tacitus' *Annales*, included in the LASLA corpus).[28] Thus, we may assume that the form *Mumius* found in CIL-I^2-628 is a hint of the incomplete generalization of *geminatio consonantium*, in line with the chronology proposed in the literature.

5.2. <EI> for /i:/

The linking of the tokens of CLaSSES to the Lemma Bank of LiLa can also shed light on the writing of /i:/ through <EI> in Latin sources. It is known from the literature [19, 20, 21, 22] that, in the 'urban' Latin of the city of Rome, the monophthongization of the diphthong [ej] took place in two steps: (i) [ej] > [e:],[29] between the 3rd and mid-2nd century BCE; (ii) [e:] > [i:], between the 2nd and 1st century BCE. The data from CLaSSES, obtained through the function "Search for linguistic phenomena" (label "Diphthong - Classical <I> /ī/ = <EI>"), confirm the traditional picture, indicating that the spelling <EI> for /i:/ is either a conservative spelling retained in earlier documents, or an archaizing feature that characterizes the solemn register of later public and official inscriptions. More in detail, in CLaSSES the spelling <EI> for /i:/ is found in 225 occurrences (99 lemmas), mainly in older public inscriptions, before the 1st century BCE (212 occurrences over 225). A more comprehensive view of this phenomenon can be obtained thanks to the interoperability between different Latin corpora made possible in LiLa. By running a query on the corpora interlinked in the Knowledge Base, it is possible to collect all the tokens linked to the 99 lemmas concerned and select those

[23] http://lila-erc.eu/data/id/lemma/99301.
[24] http://lila-erc.eu/data/id/lemma/99302.
[25] http://lila-erc.eu/data/id/lemma/111421.
[26] http://lila-erc.eu/data/id/lemma/111423.
[27] https://lila-erc.eu/sparql/.
[28] https://lila-erc.eu/data/corpora/Lasla/id/corpus.
[29] Possibly a long lax [i:] [20, 23].

where the spelling <EI> for /i:/ takes place.

For instance, of particular interest is the form *sei* for *sī* 'if' that is found in Archaic Latin. By using LiLa, it is possible to find that out of the 22,161 occurrences of *si* in the corpora interlinked therein, 10 show the form *sei*. One relevant example is from Plautus' *Epidicus* (*Ep.* 567, twice). These 2 occurrences of *sei*, which in LiLa are recorded as 2 tokens from the LASLA corpus, testify to the above-mentioned first step of the monophthongization process ([ej] > [e:]), which takes place in the age of Plautus and which is attested elsewhere in his works.

5.3. Derivational Morphology

So far, we have discussed some very easy queries on specific lexical items that can be performed to compare information provided by CLaSSES to that provided by other resources. However, LiLa allows not only to explore and compare single corpora at the lexical level (via the Lemma Bank), but also to conduct in-depth linguistic analysis, concerning, for instance, morphology. For example, it is possible to compare the type and number of affixes found in CLaSSES, investigating how many (and which type of) derivational morphemes are represented in non-literary texts. The list of affixes that build up the lexicon of CLaSSES can be accessed with a SPARQL query that retrieves all the lemmas in the CLaSSES corpus that feature an affix (either prefix, or suffix) in their morphological form, and reports the number of their occurrences therein (see Listing 1[30]).

Morphological information was not annotated in CLaSSES. Thus, the link to LiLa allows to conduct more in-depth linguistic analyisis; most importantly, it also allows users to compare different corpora with relation to specific linguistic features. For instance, it is possible to investigate to what extent the derivational morphology found in non-literary texts deviates from that of Classical texts by performing the very same query on the LASLA corpus, by simply replacing the URI for CLaSSES in the SPARQL query (as subject of the `powla:hasSubDocument` property) with that for the LASLA corpus: http://lila-erc.eu/data/corpora/Lasla/id/corpus.

The affixes that most frequently occur in the CLaSSES corpus are three suffixes and a prefix:

- *-in*, 486 occurrences (7.3% of affixes extracted from the corpus);
- *-(t)or*, 456 occurrences (6.9%);
- *-t*, 442 occurrences (6.7%);
- *con-*, 440 occurrences (6.6%).

These affixes have a very different distribution in LASLA, in which only *con-* is among the most frequent affixes, with 32,763 occurrences (7,9%), whereas *-in* counts just for 1.2% of all affixes extracted from the corpus (5,137 occ.), *-(t)or* for 2.3% (9,593 occ.), and *-t* for 1% (4,024 occ.). Such differences are largely due to a number of lexemes that are highly frequent in epigraphic texts, in particular *dominus* 'master' (198 occ.) for the suffix *-in* and *imperator* 'general, emperor' (153 occ.) for the suffix *-(t)or*, which are most frequent in public inscriptions, or *libertus/liberta* 'freedman' (281 occ.) for the suffix *-t*, which is most frequent in funerary inscriptions, where the epitaph often refers to the civil status of freed slaves. Therefore, even if there is a major difference in dimension between the two corpora, a query such as the one here illustrated can bring to light specificities of the corpus CLaSSES that go beyond the lexical level and that could not be observed without comparison with other resources.

6. Conclusion and Future Work

The linking of CLaSSES into LiLa represents an added value for both the resources. As for CLaSSES, its (meta)data are now interoperable with the other resources interlinked in the Knowledge Base. As for LiLa, the non-literary texts of CLaSSES increased significantly its textual coverage, both in terms of size and in terms of register variation.

In the near future, we plan to model and interlink in LiLa other types of metadata provided by CLaSSES, such as information about the provenance and the dating of the texts. We plan to start from metadata on the time span of the texts, that we will model as Linked Data using data categories and properties from the CIDOC Conceptual Reference Model.[31]

Acknowledgments

The "LiLa - Linking Latin" 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. The work of Lucia Tamponi is partly funded by the PRIN Project "Ancient languages and writing systems in contact: a touchstone for language change", prot. 2017JBFP9H.

References

[1] C. T. Lewis, C. Short, A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary, Clarendon Press, Oxford, 1879.

[30] The query outputs a table with four columns: the label of the lemma (`?lemmaLabel`), the type of affix, either prefix or suffix (`?affixType`), the label of the affix (`?affixLabel`) and the total number of tokens for the lemma in CLaSSES (`(count(?tokenClasses) as ?count)`).

[31] https://www.cidoc-crm.org/.

Listing 1: A SPARQL query on the LiLa Knowledge Base

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX lila: <http://lila-erc.eu/ontologies/lila/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX powla: <http://purl.org/powla/powla.owl#>

SELECT ?lemmaLabel ?affixType ?affixLabel (count(?tokenClasses) as ?count)
WHERE {
      ?lemmaLiLa a lila:Lemma ;
            (lila:hasPrefix|lila:hasSuffix) ?affix ;
            rdfs:label ?lemmaLabel .
      ?affix rdfs:label ?affixLabel ;
            rdf:type ?affixType .
      ?tokenClasses lila:hasLemma ?lemmaLiLa ;
            powla:hasLayer ?docLayer ;
            rdfs:label ?tokenClassesLabel .
      ?docLayer powla:hasDocument ?subDocLayer .
      ?subDocLayer dc:title ?titlesubDocLayer .
      <http://lila-erc.eu/data/corpora/CLaSSES/id/corpus> powla:hasSubDocument ?subDocLayer .
}
GROUP BY ?lemmaLabel ?affixType ?affixLabel
ORDER BY DESC(?count)
```


[2] C. d. F. s. du Cange, bénédictins de la congrégation de Saint-Maur, d. P. Carpentier, J. C. Adelung, G. A. L. Henschel, L. Diefenbach, L. Favre, Glossarium mediae et infimae latinitatis, Favre, Niort, France, 1883–1887.

[3] G. Marotta, Talking stones. Phonology in Latin inscriptions?, Studi e Saggi Linguistici 53 (2015) 39–63.

[4] G. Marotta, Sociolinguistica storica ed epigrafia latina: il corpus CLaSSES I, Linguarum Varietas 5 (2016) 145–159.

[5] I. De Felice, G. Marotta, M. Donati, CLaSSES: A new digital resource for Latin epigraphy, Italian Journal of Computational Linguistics 1 (2015) 119–130.

[6] G. Marotta, F. Rovai, I. De Felice, L. Tamponi, CLaSSES: Orthographic variation in non-literary Latin, Studi e Saggi Linguistici 58 (2020) 39–65.

[7] O. Lassila, R. R. Swick, Resource Description Framework (RDF) Model and Syntax Specification, 1998.

[8] C. Chiarcos, POWLA: Modeling Linguistic Corpora in OWL/DL, in: E. Simperl, P. Cimiano, A. Polleres, O. Corcho, V. Presutti (Eds.), The Semantic Web: Research and Applications, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2012, pp. 225–239. doi:10.1007/978-3-642-30284-8_22.

[9] C. Chiarcos, M. Sukhareva, OLiA – Ontologies of Linguistic Annotation, Semantic Web 6 (2015) 379–386.

[10] P. Buitelaar, P. Cimiano, J. McCrae, E. Montiel-Ponsoda, T. Declerck, Ontology Lexicalization: The *lemon* Perspective, in: Proceedings of the Workshops-9th International Conference on Terminology and Artificial Intelligence (TIA 2011), 2011, pp. 33–36.

[11] J. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, P. Cimiano, The OntoLex-Lemon Model: Development and Applications, in: Proceedings of eLex, 2017, pp. 587–597.

[12] M. Passarotti, M. Budassi, E. Litta, P. Ruffolo, The Lemlat 3.0 Package for Morphological Analysis of Latin, in: Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language, 2017, pp. 24–31.

[13] L. Tamponi, Consonant gemination in Latin epigraphy between variation and standard, in: Latin Vulgaire-Latin Tardif XIV, Actes du XIVème colloque international sur le latin vulgaire et tardif, Brepols, Turnhout, Forthcoming.

[14] R. J. Rowland, Onomastic remarks on Roman Sardinia, Names 21 (1973) 82–102.

[15] G. Lupinu, Latino epigrafico della Sardegna: aspetti fonetici, Ilisso, Nuoro, 2000.

[16] M. Mancini, Lucilius and Nigidius Figulus on orthographic iconicity, Journal of Latin Linguistics 18 (2019) 1–34.

[17] M. Mancini, Repertori grafici e regole d'uso: il caso del latino <XS>, in: L. Agostiniani, M. P. Marchese (Eds.), Lingua, testi, storia. Atti della giornata di studi in ricordo di Aldo Luigi Prosdocimi, Giorgio

Bretschneider, Roma, 2019, pp. 13–53.
[18] L. Tamponi, La *geminatio consonantium*: studio su un corpus di epigrafi latine anteriori al I secolo d.C., Studi e Saggi Linguistici 60 (2022) 29–50.
[19] M. Niedermann, Phonétique historique du latin, Librairie C. Klincksieck, Paris, 1953.
[20] M. Leumann, Lateinische Laut- und Formenlehre, Beck, Munchen, 1977.
[21] W. S. Allen, Vox latina: A Guide to the Pronunciation of Classical Latin, Cambridge University Press, Cambridge, 1978.
[22] M. Mancini, *Dilatandis litteris*: uno studio su Cicerone e la pronunzia 'rustica', in: R. Bombi, G. Cifoletti, F. Fusco, L. Innocente, V. Orioles (Eds.), Studi linguistici in onore di Roberto Gusmani, Edizioni dell'Orso, Alessandria, 2006, pp. 1023–1046.
[23] M. Benedetti, G. Marotta, Monottongazione e geminazione in latino: nuovi elementi a favore dell'isocronismo sillabico, in: P. Molinelli, P. Cuzzolin, C. Fedriani (Eds.), Latin Vulgaire–Latin Tardif, Actes du Xe Colloque International sur le Latin Vulgaire et Tardif, Sestante, Bergamo, 2014, pp. 25–43.