LiLa Linking Latin Tutorial

Matteo Pellegrini[1], Eleonora Litta[1], Marco Passarotti[1], Rachele Sprugnoli[1], Francesco Mambrini[1] and Giovanni Moretti[1]

[1]Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milano, Italy

Abstract

By applying Linked Data and FAIR principles, the LiLa: Linking Latin project makes linguistic resources (e.g. textual corpora, lexica, dictionaries) for Latin interact on the web via a lexical basis made of a collection of lemmas known as the LiLa Lemma Bank. In this hands-on tutorial, participants learned how to link a Latin text to the LiLa Knowledge Base of linguistic resources. By the end of the tutorial participants should have a better understanding of the benefits of linking a Latin text to the LiLa Knowledge Base, and of the work required to help machines process linguistic data and produce quality resources.

Keywords

LLOD, linguistic resources, ontologies, Latin

1. Introduction

The LiLa Knowledge Base (KB)[1] [1] makes textual and lexical resources for Latin interact through the commonly used data model called *Resource Description Framework (RDF)* [2], and through the ontology models developed and shared by the Linguistic Linked Open Data (LLOD) community, which applies the principles of the Linked Data paradigm [3] to the (meta)data contained in linguistic resources. The use of a common vocabulary ensures syntactic and semantic interoperability between distributed linguistic resources described following the same structural and conceptual principles. The lexical resources for Latin linked so far to LiLa include a derivational lexicon [4],[2] a polarity lexicon [5],[3] an etymological dictionary [6],[4] an index of Greek loanwords into Latin [7],[5] a joint resource providing a manually checked subset of the Latin WordNet[6] and a valency lexicon[7] [8], and the bilingual Latin-English dictionary by Charlton Lewis and Charles Short [9].[8][9] As for textual resources, the ones currently linked to the Knowledge Base are Thomas Aquinas' *Summa Contra Gentiles* from the Index Thomisticus Treebank [10],[10] the Latin works by Dante Alighieri in the UDante corpus [11],[11] an anonymous Latin comedy[12] – *Querolus Sive Aulularia* [12] – and a 13th century treaty on arithmetic by Leonardo Fibonacci[13] – the *Liber Abbaci*. Based on a large collection of "canonical forms" (lemmas) – the so-called "Lemma Bank" –, LiLa achieves interoperability between resources by linking all those entries in lexical resources and tokens in corpora that point to the same lemma in the LiLa collection.

The LiLa: Linking Latin Tutorial at LDK 2021 was second in a series of tutorials that aim not only at presenting the LiLa project, but also at teaching participants how to connect a textual resource to its Knowledge Base. The first tutorial was given at the Linked Pasts 6 conference (University of London and British Library, December 2-16, 2020; held on-line due to the COVID-19 pandemic), targeting scholars interested in applying Linked Data principles to the study of ancient and historical worlds. Another tutorial took place shortly after the one given at LDK and presented in this paper, in the context of the 2nd International Conference of the European Association for Digital Humanities (EADH, Krasnoyarsk, September 21-25, 2021; held on-line), thus addressing a more general audience.

The main purpose of this series of tutorials is to present the LiLa project to different communities that can be interested, namely the one of Digital Humanities, both in general (as in the EADH conference) and with a more specific focus on the past and on Linked Data (as in the Linked Pasts conference), and the one of Computational Linguistics and Natural Language Processing, present at the LDK 2021 conference. This large-scale dissemination effort is especially important for the success of the project due to its open-ended and community-driven

Language, Data & Knowledge conference (LDK 2021), 1-3 September, Zaragoza, Spain

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

[2] http://lila-erc.eu/data/lexicalResources/WFL/Lexicon.

[3] http://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon.

[4] http://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon.

[5] http://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon.

[6] http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon.

[7] http://lila-erc.eu/data/lexicalResources/LatinVallex/Lexicon.

[8] https://lila-erc.eu/data/lexicalResources/LewisShort/Lexicon.

[9] At the time of the tutorial, this latter resource was not yet online, hence it was not presented.

[10] http://lila-erc.eu/data/corpora/ITTB/id/citationUnit/005.

[11] http://lila-erc.eu/data/corpora/DanteSearch/id/corpus.

[12] http://lila-erc.eu/data/corpora/Querolus/id/citationUnit/QuerolussiveAulularia.

[13] http://lila-erc.eu/data/corpora/CorpusFibonacci/id/corpus/Liber%20Abbaci.

nature: the LiLa Knowledge Base will serve its purpose only if scholars on the one hand use it, on the other hand contribute to its growth by linking texts to it. Consistently with this purpose, the LDK tutorial was intended for those scholars who want to publish Latin texts on the web (e.g. computational linguists, theoretical linguists, classicists, philologists) and to make them interact with the linguistic resources already linked to LiLa. No prior experience of Natural Language Processing and Linked Data technologies was expected but it was advised that participants had basic understanding of the concepts of lemmatisation, Part-of-Speech (PoS) tagging and Linked Data. The hands-on section of the tutorial involved processing and linking a Latin text to the KB so knowledge of Latin was declared as preferable. However, participants who did not know Latin but were nevertheless interested in the project and its methods were also welcome to join. By inviting to its tutorials participants from a larger audience, the LiLa project intends to stress its characterisation as an open-ended and a community-driven project, thus ensuring long life and sustainability to the KB and its connections.

2. The Tutorial

The second LiLa tutorial was held as part of the introductory workshop and tutorial activity program of the Language, Data and Knowledge conference (LDK) in Zaragoza, Spain, on September 1st, 2021. Due to the COVID-19 pandemic emergency resulting in travel restrictions, the conference was held as a hybrid in-person/online event, with lectures and presentations being given whether from the conference venue and shared on Zoom, or remotely on Zoom and projected at the venue. The LiLa tutors could not travel to Zaragoza and presented via Zoom. This was the second time the LiLa tutorial was presented remotely. The first time, however, the tutorial activities were spread on a full day, while for time constraints imposed by the conference rich schedule organisation, this time the tutorial was held during the morning alone, 9:00 to 13:00 with an half-hour coffee break between sections. The number of registered participants was 70, but on the day only 24 of them (including the host and members of the LiLa team) actually took part in the Zoom session.

All the materials necessary to participate in the event were provided by the LiLa team before and during the tutorial on a GitHub repository[14] containing the text to be uploaded in the Text Linker, the SPARQL queries discussed and a list of other useful links, namely to the website of the tutorial,[15] to the LiLa Lemma Bank query interface,[16] to the LiLa SPARQL endpoints,[17] and to a selection of other potentially interesting SPARQL queries that can be used in such endpoints.[18]

The activity consisted of two main sections, dedicated respectively to Theory and Practice: the first one included a presentation of the structure of the LiLa KB, while the second part contained a step-by-step demonstration on how to prepare and link a Latin text to LiLa, and ultimately exploit the linked content for research purposes. At the end of the first section, participants were invited to answer to a quiz consisting of five questions, using the SLIDO platform. The number of participants who actually took the quiz was 7. The questions were intended to evaluate the participants' comprehension of what had just been explained and possibly consolidating some of the fundamental concepts underlying the structure of LiLa.

In the following subsections, we detail the content of the two sections of the workshop activities.

2.1. Part 1. Theory: Overview of LiLa's Architecture

In the first part of the tutorial, we have started by highlighting the motivation behind the LiLa enterprise, namely the observation that the many resources and tools already available for Latin are characterised by different conceptual models (for instance, they use different tagsets and annotation schemes), and thus they are scattered, unconnected and unable to interact with one another, as they cannot be queried and used in a unified fashion.

The LiLa project addresses this problem by following the principles of the Linked Data paradigm, aiming at building an open-ended Knowledge Base to which new and existing resources and tools can be connected. Hence, all entities are assigned Uniform Resource Identifiers (URIs) in the HTTP protocol, so as to make them univocally findable and accessible to users. Resources are then interlinked to one another using the RDF data model, where information is expressed in terms of triples that connect a subject to an object through a predicate; accordingly, data can be browsed using the SPARQL query language, specifically designed for RDF data. LiLa also makes use of other ontologies, the most relevant of which being OntoLex [13], that has by now reached the status of a *de facto* standard for publishing lexical data as LLOD.

The way in which interoperability is concretely achieved is by providing a large collection of lemmas, called the Lemma Bank, to which the entries of lexical resources, the tokens of textual resources, and the output of NLP tools can be linked. As a consequence of the central role of lemmas in the architecture of LiLa, in order

[14] https://github.com/CIRCSE/Tutorials/tree/main/LDK21.

[15] https://lila-erc.eu/ldk-2021/.

[16] https://lila-erc.eu/query/.

[17] https://lila-erc.eu/sparql/.

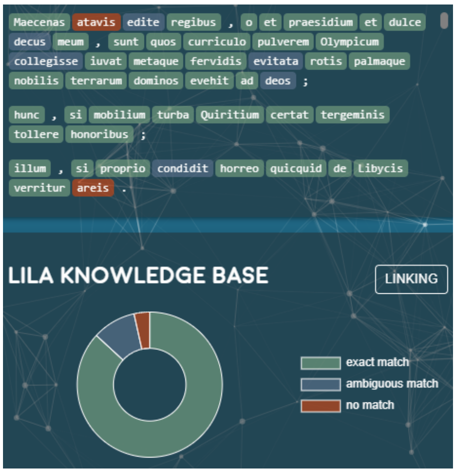
[18] https://github.com/CIRCSE/SPARQL-queries.

to be inserted into the Knowledge Base, a text needs to be not only online, but also lemmatised and PoS-tagged – ideally, with the Universal PoS tagset [14] used in the Universal Dependencies project [15].

2.2. Part 2. Practice: Preparing and Linking a Latin Text To LiLa

The second part of the tutorial was intended to show how a raw text can be prepared for inclusion into the Knowledge Base using (a beta version of) a tool called the Text Linker, that provides automatic lemmatisation and Part-of-Speech tagging, the results of which subsequently need to be manually checked and modified by the user. For this purpose, a Latin text was prepared in raw format and provided for the participants so that they could upload it into the Text Linker tool for a demonstration of the way it works. In any case, participants were let free to use a different Latin text of their choice, if they so wished. The provided text was from Book 1 of the *Odes* by Horace. The raw text was taken from the Musisque Deoque archive [16], and it was pre-processed so as to delete a series of metadata such as the Roman numbers identifying each poem and the numbering of lines.

The text was then given as input to the Text Linker. The output is an analysis where tokens that are unambiguously linked to one lemma of the Lemma Bank are displayed in green, ambiguous tokens for which several links are available are displayed in blue, and tokens that are left not linked are displayed in red, as is shown in Figure 1.



Maecenas atavis edite regibus , o et praesidium et dulce
decus meum , sunt quos curriculo pulverem Olympicum
collegisse iuvat metaque fervidis evitata rotis palmaque
nobilis terrarum dominos evehit ad deos ;

hunc , si mobilium turba Quiritium certat tergeminis
tollere honoribus ;

illum , si proprio condidit horreo quicquid de Libycis
verritur areis .

Figure 1: Output of the Text Linker on a portion of the sample text provided to participants

At this point, the task of the user consists in checking whether the lemma and PoS automatically assigned to unambiguous tokens are correct and selecting the correct alternative among the ones listed for ambiguous tokens. As for tokens whose analysis cannot be linked to the Lemma Bank by the Text Linker, this can be due to several reasons: if it is due to an error in the automatic tagging, the user has to correct the tag, while if it is due to the absence of the corresponding lemma in the Lemma Bank, the LiLa team has to add it.[19] Lastly, this might be caused by a discrepancy between the PoS assigned to the lemma by the text linker and the one of the corresponding lemma in the Lemma Bank; in such case, the user has to decide whether to link the lemma anyway or not.

These operations were performed by LiLa instructors on a selection of tokens, illustrating (some of) the problems that users can encounter in the process. Some time was then left to participants to go through other tokens by themselves.

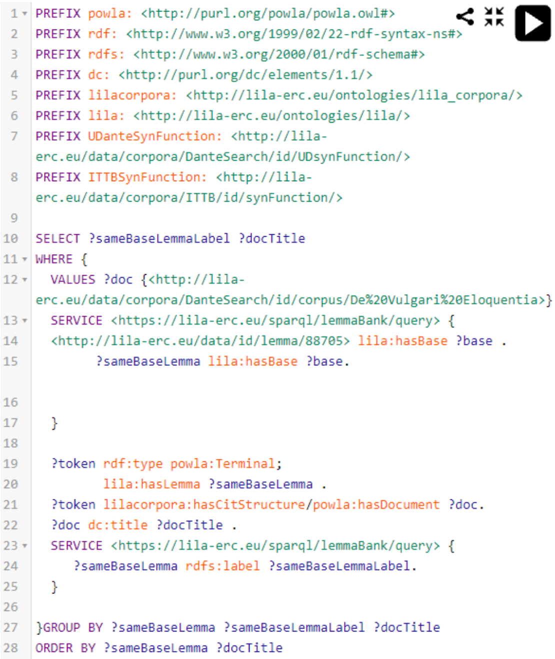
The next step was the automatic conversion of the lemmatised and PoS-tagged text to the RDF serialisation format used in LiLa, namely Turtle. After this conversion, a URI is assigned to the text itself as well as to each of its tokens, making it possible to perform the actual linking of the text to the LiLa Knowledge base.

The last part of the tutorial was dedicated to the illustration of the way in which both the newly added text and the entire Knowledge Base can be queried using the SPARQL language. For this purpose, some SPARQL queries were pre-prepared by LiLa instructors, who illustrated in detail to participants how the queries worked. Here, we will show one of them, reported in Figure 2 below.

The SPARQL language is structured in subject-predicate-object triples as the RDF data model it is conceived for. Strings preceded by a question mark are variables, while URIs are enclosed by angle brackets. The first part of the query (lines 1-8) consists in a definition of prefixes that can be used as shortcuts to refer to URIs in the body of the query, where the more reader-friendly string ending with a colon (e.g. "lila": in line 6) substitutes a part of the URI of the relevant resource (in this case, `http://lila-erc.eu/ontologies/lila/`). Therefore, the string "lila:hasBase" in ll. 14-15 is a shortcut for the full URI of the property `http://lila-erc.eu/ontologies/lila/hasBase`, defined in the LiLa ontology.

The query itself consists of two blocks. The SELECT block (l. 10) indicates the variables whose values should be displayed in the output of the query, while the WHERE block details the conditions that should be satisfied for an element to be returned by the query. The query of Figure 2 refers to the endpoint of the corpora linked to LiLa and, given the lexical base of a lemma, it finds all the

[19] We plan to make a form available to users to propose a new lemma for insertion in the Lemma Bank.

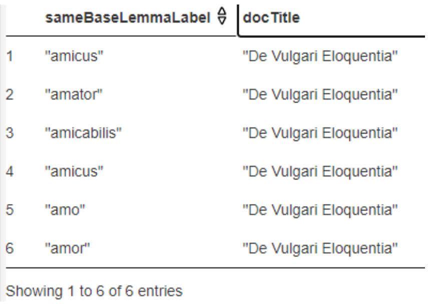


```
1  PREFIX powla: <http://purl.org/powla/powla.owl#>
2  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4  PREFIX dc: <http://purl.org/dc/elements/1.1/>
5  PREFIX lilacorpora: <http://lila-erc.eu/ontologies/lila_corpora/>
6  PREFIX lila: <http://lila-erc.eu/ontologies/lila/>
7  PREFIX UDanteSynFunction: <http://lila-
   erc.eu/data/corpora/DanteSearch/id/UDsynFunction/>
8  PREFIX ITTBSynFunction: <http://lila-
   erc.eu/data/corpora/ITTB/id/synFunction/>
9
10 SELECT ?sameBaseLemmaLabel ?docTitle
11 WHERE {
12   VALUES ?doc {<http://lila-
   erc.eu/data/corpora/DanteSearch/id/corpus/De%20Vulgari%20Eloquentia>}
13   SERVICE <https://lila-erc.eu/sparql/lemmaBank/query> {
14   <http://lila-erc.eu/data/id/lemma/88705> lila:hasBase ?base .
15         ?sameBaseLemma lila:hasBase ?base.

16

17   }
18
19   ?token rdf:type powla:Terminal;
20          lila:hasLemma ?sameBaseLemma .
21   ?token lilacorpora:hasCitStructure/powla:hasDocument ?doc.
22   ?doc dc:title ?docTitle .
23   SERVICE <https://lila-erc.eu/sparql/lemmaBank/query> {
24     ?sameBaseLemma rdfs:label ?sameBaseLemmaLabel.
25   }
26
27 }GROUP BY ?sameBaseLemma ?sameBaseLemmaLabel ?docTitle
28 ORDER BY ?sameBaseLemma ?docTitle
```

Figure 2: Sample SPARQL query on the LiLa endpoints

lemmas sharing the same lexical base in a document.[20]

To go into some more detail, this query looks into the resource of Dante's *De Vulgari Eloquentia* (l. 12).[21] In l. 13, the keyword SERVICE is used for a federated query, referring to another endpoint, the one of the Lemma Bank. This is needed in order to retrieve the base of a specific lemma (see the URI of l. 14, corresponding to the lemma *amo*) and extract all the lemmas that share the same base (l. 15). The query then looks for all the tokens (i.e., elements of type `powla:terminal`, cf. l. 19) that are lemmatised under a lemma that has the same base (l. 20) in the selected document (l. 21), whose title is also extracted (l. 22). Then, we go back to the Lemma Bank (l. 23) to retrieve the human-readable label of the lemmas with the same base (l. 24). Lastly, tokens of the same lemma are grouped together (l. 27) and results are displayed in alphabetical order of lemma (l. 28). Of all the variables used in the query to extract the relevant information, only the ones listed in the SELECT block are returned in the results, that are displayed in a table-like fashion where the headers of the column correspond to those variables, as is shown in Figure 3 below.



	sameBaseLemmaLabel	docTitle
1	"amicus"	"De Vulgari Eloquentia"
2	"amator"	"De Vulgari Eloquentia"
3	"amicabilis"	"De Vulgari Eloquentia"
4	"amicus"	"De Vulgari Eloquentia"
5	"amo"	"De Vulgari Eloquentia"
6	"amor"	"De Vulgari Eloquentia"

Showing 1 to 6 of 6 entries

Figure 3: Output of the query of Figure 2

Four other queries were illustrated to participants, exploiting information from the various resources linked to LiLa. One of them lists and counts the occurrences of the forms of a lemma in a document.[22] Another one looks for all instances of *amor* in a document, retrieves all the possible senses of the word from the LatinWordNet and lists all the definitions of the connected synsets.[23] Yet another one lists lemmas with a positive polarity.[24] The last one retrieves all lemmas of a text having a same suffix.[25]

3. User feedback

Due to the hybrid nature of the tutorial, user feedback was difficult to achieve. A collaborative document was made available before the tutorial, but it was not modified by participants during or after the activity. However we believe that the description of the activity contained in the document can be a useful how-to reminder for those who attended.

Two questions were addressed during the morning, both related to possible inclusions of data into the KB. A participant asked if we had looked into the Latin content from the Wiktionary pages, and whether we were planning to link to it. Another one asked whether we were planning to include definitions from published dictionaries too. In both cases, including such additional information would be reasonable and welcome. Indeed, as was hinted in the Introduction, the inclusion of a bilingual Latin-English dictionary – namely the one by Lewis

[20] In LiLa, Classical Latin words are grouped by morphological family, through the property defined in the LiLa ontology `hasBase`. This connects to a node that groups together all words belonging to the same family, e.g. in our case all words belonging to the family of *amo* 'to love', such as *amicus* 'friend', *amator* 'lover', *amor* 'love' etc.

[21] During the tutorial, the participants were invited to use the URI of the document they had uploaded to the LemmaBank. Since the linking of these documents to LiLa was only temporary, it is not possible to use such documents for the query shown here.

[22] https://github.com/CIRCSE/Tutorials/blob/main/LDK21/queries/form.rq.

[23] https://github.com/CIRCSE/Tutorials/blob/main/LDK21/queries/latinwordnet.rq.

[24] https://github.com/CIRCSE/Tutorials/blob/main/LDK21/queries/sentiment.rq.

[25] https://github.com/CIRCSE/Tutorials/blob/main/LDK21/queries/suffix.rq.

and Short [9] – was among the projects of the LiLa group at the time of the tutorial, and it has now been completed.

No further questions were addressed during the practical hands-on section.

4. Conclusions and Future Perspectives

In this paper, we have outlined the content of the LiLa: Linking Latin tutorial held at LDK 2021. In the future, we plan to continue this kind of dissemination of our research with further tutorials. These have the potential to increase the reach of our work among scholars from all the different communities potentially interested in the project, who can contribute to its success either simply as users, by exploiting the resources already linked to the LiLa Knowledge Base, or more actively as contributors, by linking new texts to it. To this aim, other LiLa events are already scheduled for December 2021 at a winter school in Padova and for March 2022 at a vacation school on Digital Humanities and Neo-Latin Studies.

The LiLa project is already part of the Nexus Linguarum intiative, that promotes the use of models based on the principles of Linked Data for linguistic resources.[26] In the future, we would like to make our resources interact also with the ones developed in the context of other LOD initiatives with a different focus than the strictly linguistic aspects related to the Ancient World, such as Pelagios [17] and Pleiades [18] for places.

Acknowledgments

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.

References

[1] M. Passarotti, F. Mambrini, G. Franzini, F. M. Cecchini, E. Litta, G. Moretti, P. Ruffolo, R. Sprugnoli, Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin, Studi e Saggi Linguistici 58 (2020) 177–212.

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

[3] C. Bizer, T. Heath, K. Idehen, T. Berners-Lee, Linked Data on the Web (LDOW2008), in: Proceedings of the 17th international conference on World Wide Web, 2008, pp. 1265–1266.

[4] M. Pellegrini, E. Litta, M. Passarotti, F. Mambrini, G. Moretti, The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources, in: Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021), 2021, pp. 101–109.

[5] R. Sprugnoli, F. Mambrini, G. Moretti, M. Passarotti, Towards the Modeling of Polarity in a Latin Knowledge Base, in: WHiSe@ ESWC, 2020, pp. 59–70.

[6] F. Mambrini, M. Passarotti, Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin, in: Proceedings of the 2020 Globalex Workshop on Linked Lexicography, 2020, pp. 20–28.

[7] G. Franzini, F. Zampedri, M. Passarotti, F. Mambrini, G. Moretti, Græcissare: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin., in: CLiC-it, 2020.

[8] F. Mambrini, M. Passarotti, E. Litta, G. Moretti, Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin, in: M. Alam, P. Groth, V. de Boer, T. Pellegrini, H. J. Pandit, E. Montiel, V. Rodríguez Doncel, B. McGillivray, A. Meroño-Peñuela (Eds.), Studies on the Semantic Web, IOS Press, Amsterdam, 2021.

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

[10] M. Passarotti, The Project of the Index Thomisticus Treebank, in: Digital Classical Philology, De Gruyter Saur, 2019, pp. 299–320.

[11] F. M. Cecchini, R. Sprugnoli, G. Moretti, M. Passarotti, UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works, in: CLiC-it, 2020.

[12] F. Gamba, Including a New Textual Resource into the LiLa Knowledge Base, Master's thesis, Università di Pavia, 2020.

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

[14] S. Petrov, D. Das, R. McDonald, A Universal Part-of-Speech Tagset, arXiv preprint arXiv:1104.2086 (2011).

[15] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics 47 (2021) 255–308. doi:10.1162/coli_a_00402.

[16] M. Manca, L. Spinazzè, P. Mastandrea, L. Tessarolo, F. Boschetti, Musisque Deoque: Text Retrieval on Critical Editions, Journal for Language Technology and Computational Linguistics 26 (2011) 127–138.

[17] L. Isaksen, R. Simon, E. T. Barker, P. de Soto Cañamares, Pelagios and the Emerging Graph of Ancient

[26] https://nexuslinguarum.eu.

World Data, in: Proceedings of the 2014 ACM conference on Web science, 2014, pp. 197–201.

[18] E. Barker, R. Simon, L. Isaksen, P. de Soto Canamares, The Pleiades Gazetteer and the Pelagios project, in: M. L. Berman, R. Mostern, H. Southall (Eds.), Enriching and Integrating Gazetteers, Indiana University Press, 2016, pp. 97–109.