The Leibniz List as Linguistic Linked Data in the LiLa Knowledge Base

Lisa Sophie Albertelli[1], Giulia Calvi[1], Francesco Mambrini[1],

[1]Università Cattolica del Sacro Cuore, Milano

Correspondence: francesco.mambrini@unicatt.it

Abstract

This paper presents the integration of the Leibniz List, a concept list from the Concepticon project, into the LiLa Knowledge Base of Latin interoperable resources. The modeling experiment was conducted using W3C standards like Ontolex and SKOS. This work, which originated in a project for a university course, is limited to a short list of words, but it already enables interoperability between the Concepticon and the language resources in a LOD architecture like LiLa. The integration enriches the LiLa ecosystem, allowing users to explore Latin lexicon from an onomasiological perspective and links concepts to lexical entries from various dictionaries and corpus attestations. The work showcases how standard Semantic Web technologies can effectively model and connect historical concept lists within larger linguistic knowledge infrastructures and provides an example for further experiments with the Concepticon's data.

1 Introduction

The aim of the present study is to model one concept list from the Concepticon project (List et al., 2016)[1] as Linguistic Linked Open Data (LLOD) and to connect it to the Knowledge Base (KB) of linguistic resources for Latin made available by the Lila Linking Latin project.[2] Specifically, the study focuses on the concepts included in a list compiled by the philosopher G. W. Leibniz and now published in the Concepticon. The paper discusses how the Latin verbalizations of these concepts were linked to the lemmas of the Lila Lemma Bank with the help of two widely used ontologies such as SKOS and the Ontolex-Lemon model. Our work leverages the lemma-as-gateway approach promoted by LiLa to make Leibniz's concepts part of a network of interoperable linguistic resources; at the same time, it integrates the concept-based perspective of the Concepticon into the LiLa ecosystem for the first time. The introduction of a concept list from this project allows us to widen the range and type of lexical resources available in LiLa and enables researchers interested in an onomasiological approach to lexicon (from the concepts to the words used to express them) to make use of the network of data in the KB. While the concept list described here is quite small, the work is a first step in modeling and integrating a similar resource.

The paper is organized as follows. Sections 1.1 and 1.2 introduce the Concepticon and the Leibniz List respectively. Section 1.3 provides a short overview of LiLa. Section 2 describes the work undertaken to model the data and the final results. Section 3 summarizes the conclusions and future perspectives.

1.1 The Concepticon

In the history of linguistics, several researchers have created lists of basic concepts in various domains with the goal of recording how these concepts are verbalized in one or more languages. Those lists were motivated by different research agendas, such as addressing the problem of subgrouping in historical linguistics (Swadesh, 1950), detecting deep genetic relationships among languages (Dolgopolsky, 1964) or providing standardized naming tests in clinical studies (Ardila, 2007).

The Concepticon (List et al., 2016) is a resource that attempts to collect the available concept lists and to provide a mapping between their entries. The project maintains a unified database freely available online where all the diverse lists documenting the same concepts can be accessed and searched. In fact, while not using W3C standards like RDF or SPARQL for data dissemination, the Concepticon adopts the Cross-Linguistic Data For-

[1] https://concepticon.clld.org/.

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

mats (CLDF),[3] itself rooted in principles closely related to those of Linked Data.

In the Concepticon, a concept list is a collection of locally defined concepts, each associated with an identifier and a label that indicates how it is expressed in one or more target languages. To give an example, the concept identified as `Luniewska-2016-299-2` from the concept list compiled by Łuniewska et al. (2016) is glossed with labels in 25 languages, including e.g. English ('ant'), Afrikaans ('mier'), and Finnish ('muurahainen').[4]

Within the framework of the project, all the entries from the different lists are mapped onto concept sets; a concept set is defined as a group of labels referring to the same concept. Each concept set is provided with a unique global identifier, a unique label and a human-readable definition. These sets are also classified into semantic fields, based on those used in the World Loanword Database (Haspelmath and Tadmor, 2009), and into ontological categories, which roughly mirror the distribution of words into parts of speech (List et al., 2016, 2394).[5] Concept sets are also organized with a series of ad-hoc relations among them, such as "broader", "narrower", and "similar". Thus, the aforementioned concept `Luniewska-2016-299-2` is linked to a set labeled ANT, belonging to the semantic field 'animals' and to ontological category 'person/thing', and glossed with the definition: "[a]ny of the black, red, brown, or yellow insects of the family Formicidae characterized by a large head and by living in organized colonies."[6] This set groups entries from 151 lists.

Currently, the Concepticon links 30,222 concepts from 160 concept lists to 2,495 concept sets. The project data are available on GitHub, where the lists and sets are distributed as tab-separated text files (`tsv`).[7]

1.2 The Leibniz list

In a letter to G.B Podestà, Gottfried Wilhelm Leibniz (1646-1716) advocated for the collection of language data to enhance the comparison of different languages and the study of their evolution (on the exchange see Rothman, 2021, 211-240). To this end, he emphasized the importance of words expressing "things of daily use" (*res usitatiores*). The letter was published as part of the complete edition of Leibniz's works curated by Dutens (Leibniz, 1768), and the list, which contains 128 entries, is included in the Concepticon.[8]

Leibniz himself categorized the concepts into six classes: numbers (*nomina numeralia*), age and kinship (*propinquitates et aetates*), body parts (*partes corporis*), things necessary for life (*necessitates*), natural being (*naturalia*), and actions (*actions*). The dataset distributed with the Concepticon reproduces Leibniz's list with a minimalist set of metadata. Each concept is assigned a Latin label, is accompanied by a brief English definition (gloss), and is uniquely identified by a composite string that (following the project schema) includes the name of the compiler (Leibniz), the year of the publication (1768), the total number of concepts (128) and a progressive number from 1 to 128. Furthermore, Leibniz's categorization in six classes is also reported with the Latin original labels. Finally, the dataset links each of Leibniz's concepts to the corresponding concept set, whose label (the Concepticon gloss) is also included in the table. Thus, for instance, the first item in the list is identified as `Leibniz-1768-128-1`, labeled *unum* in Latin and glossed as 'one'; the concept is linked to the set identified with the id 1493 and the Concepticon gloss 'ONE'.[9]

1.3 The LiLa Knowledge Base

The LiLa KB is a network of textual and lexical resources in Latin or documenting Latin words, all modeled as Linked Open Data (Passarotti et al., 2020). The core element that keeps the network connected is the LiLa Lemma Bank, a collection of more than 230,000 canonical forms that are used as lemmas to index lexical entries and to lemmatize texts (Mambrini and Passarotti, 2023). Currently, LiLa connects 17 lexicons, providing translations and definitions of Latin words into languages like Portuguese (Dezotti et al., 2024) or Czech (Gamba et al., 2024), and documenting aspects like Indo-European etymology (Mambrini and Passarotti, 2020), or borrowing from Greek (Franzini et al.,

[3] https://cldf.clld.org/.

[4] This concept from the list by Łuniewska et al. (2016) can be viewed online at: https://concepticon.clld.org/values/Luniewska-2016-299-2.

[5] The schema containing all the ontological categories, semantic fields and relations can be seen online at: https://github.com/concepticon/concepticon-data/blob/master/concepticondata/concepticon.json.

[6] https://concepticon.clld.org/parameters/587.

[7] https://github.com/concepticon/concepticon-data/.

[8] https://concepticon.clld.org/contributions/Leibniz-1768-128.

[9] https://concepticon.clld.org/parameters/1493.

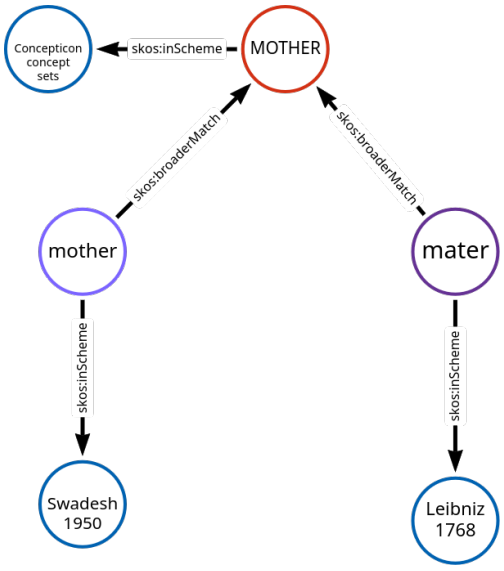

Figure 1: Relations between concepts, concept lists and concept sets

2020). Also, LiLa's lemmas are linked to about 12M tokens from more than 500 Latin texts, including well-known corpora like the LASLA's *Opera Latina* (Fantoli et al., 2022).

LiLa relies on a series of widely used ontologies for Linguistic Linked data to model language resources as RDF. In particular, for lexical information LiLa adopts the community standard Ontolex-Lemon (McCrae et al., 2017). Lemmas from the Lemma Bank are defined as instances of a subclass of `ontolex:Form` (Passarotti et al., 2020);[10] whenever a new lexicon modeled with Ontolex is linked to the KB, either its lexical entries are connected to the appropriate lemma via the property `ontolex:canonicalForm`, or its forms are mapped to LiLa's lemmas. This modeling choice provides great interoperability between LiLa and the network of resources from the Linguistic Linked Open Data Cloud (Cimiano et al., 2020, 29-41). It also makes the integration of new lexical and lexicalized Latin resources (such as the Leibniz list) very straightforward, as will be made clear in Section 2.

2 Modelling the Concepticon's Leibniz List

In this section, we explore how we translated Leibniz's Latin lexicalizations of his concepts by relying on the same model that is used by LiLa, and how we linked this information to the Lemma Bank. Moreover, we show that once the lexicalization of a concept is modeled as LOD, it becomes easy to integrate much of the information provided by the Concepticon using a popular W3C standard, namely the Simple Knowledge Organization System (SKOS).[11]

The lexical information provided in the Leibniz List is readily expressed with the Ontolex-Lemon model. Intuitively, the concepts collected by Leibniz (like all concepts mapped by the Concepticon, which point to notions and ideas not organized into formal ontologies) are perfect examples of instances of the class "Lexical Concept" in Ontolex.[12] While the Concepticon dataset only provides labels for them, a full lexicalization via Ontolex enables lexicographers to extend the range of possible linguistic metadata that can be attached to the words and, especially, to connect those words to a wealth of additional linguistic information. Note that, as the lists in the Concepticon start from concepts, generally (and effectively with the Leibniz list) ambiguity and polysemy do not pose a problem: each concept in the list is verbalized by a single lexical entry. If multiple lists use the same word to verbalize different concepts (e.g. "river bank" and "financial institution" with en. *bank*), curators will have to choose whether to create one single lexical entry with multiple senses, or multiple entries with a different form of mapping provided between them. Anyway, this case did not occur in our work.

To generate RDF representations of the lexical entries, lexical concepts and senses, we started from the TSV file downloaded from the Concepticon project and we modeled it using the software OpenRefine and a dedicated RDF plugin.[13] With such a limited list, the mapping to the LiLa lemmas was conducted manually, relying on the LiLa's Lemma Query Interface (Passarotti et al., 2024). For the lexical entries and senses (which in Ontolex reify the relation between words and concepts) we defined custom URIs within the LiLa namespace.[14] To collect all lexical entries connected to the list, we also created a lexicon using the Ontolex' `lime` model for lexicons and metadata.[15] For the concepts and concept lists, on the

[10] See `http://lila-erc.eu/ontologies/lila/Lemma`.

[11] `https://www.w3.org/2004/02/skos/`.

[12] See the documentation at `https://www.w3.org/2016/05/ontolex/#lexical-concept`.

[13] See `https://openrefine.org/` and `https://github.com/AtesComp/rdf-transform`.

[14] An example for a lexical entry is: `http://lila-erc.eu/data/lexicalResources/Leibniz-1768-128/le_19`.

[15] See the documentation at: `https://www.w3.org/2016/`

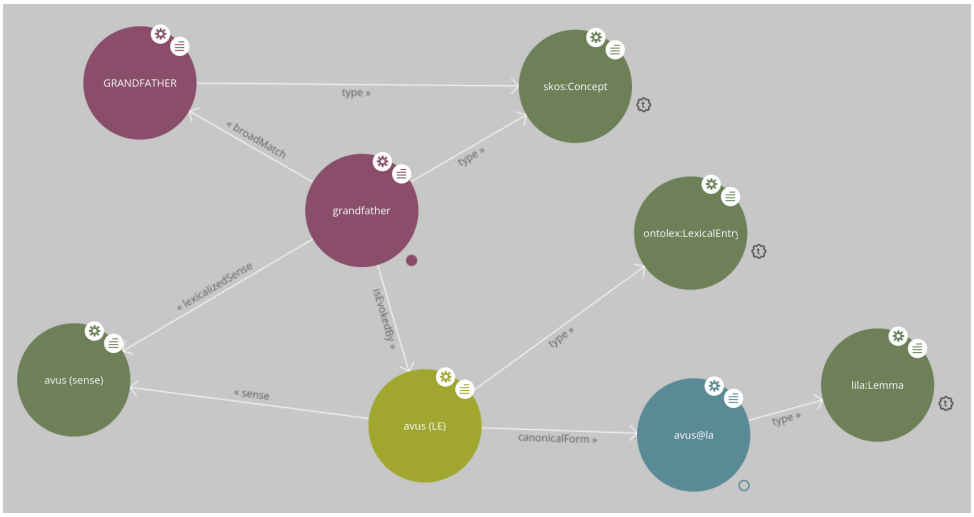

Figure 2: The concept and lexical entry "avus" (grandfather) in the Leibniz List, the Concepticon and LiLa (color code: crimson is used for the SKOS concepts; dark green for the OWL classes (right), and the lexical sense (left); yellow-green for the Lexical Entry; steel blue for the Lemma).

other hand, we reused the unique identifiers and web URLs of the Concepticon project.

As said, the Concepticon aligns all the different concept lists into concept sets. Once again, the nature of these notions is not difficult to capture using standard vocabularies of the Semantic Web. The properties and classes defined in SKOS can be leveraged to express the mapping and the simple organization (which includes broad/narrow, or "see also" relations) provided by the project. The class of `skos:Concept` is both intuitively and factually appropriate to represent the entries in the concept lists; glosses and definitions such as those found in the Concepticon are recorded via the `skos:definition` property. Each list represents an informal and historically independent collection of (SKOS) concepts, which is compatible with the definition of a `skos:ConceptScheme` (Allemang et al., 312).

The nature of concept sets is, on the other hand, less intuitive. While it would be possible to capture its specific essence by developing a dedicated Concepticon ontology, we preferred not to take this approach and rather rely on the available W3C standards only. From this perspective, the essential goal that concept sets are pursuing, i.e. the mapping of concepts from independent lists, can be readily captured in SKOS. In this perspective, concept sets are also instances of the `skos:Concept` class, not belonging to concept lists, but assigned to a dedicated Concepticon `skos:ConceptScheme`. The concepts from the different lists are then mapped onto the appropriate concept set using the standard SKOS mapping properties (Allemang et al., 310-2), and in particular `skos:broaderMatch` and `skos:narrowMatch`. Figure 1 schematizes this modeling approach with a fictitious example: the concepts for 'mother' (Lat. *mater*) from two different lists (Leibniz, 1768 and Swadesh, 1950) are linked to the respective dataset via the property `skos:inScheme`; the mapping between the two concepts is ensured via the `skos:broaderMatch` relation that connects the concepts to the Concepticon's concept set.

Figure 2 visualizes the relations of concepts, words and forms in our final modeling of the Leibniz List. The crimson node at the center represents Leibniz's original concept *avus* 'grandfather'. The Latin lexicalization is expressed by the node below it, the lexical entry that evokes the concept; this lexical entry, in turn, is identified by the lemma *avus* from LiLa (`lila_lemma:90862`) on the bottom-right corner of the image. On the top-left corner, Leibniz's concept is linked to the Concepticon concept set GRANDFATHER, which serves as a potential gateway to concepts from 53 other lists.[16]

3 Conclusions

The present work originated from a final project for a university course on Linguistic Linked Open Data and Semantic Web.[17] The limited size of the

05/ontolex/#lexicon-and-lexicon-metadata.

[16] https://concepticon.clld.org/parameters/1383.

[17] The program of the class can be accessed at https://www8.unicatt.it/upl/proguc/MI/2024/ITA/LING/

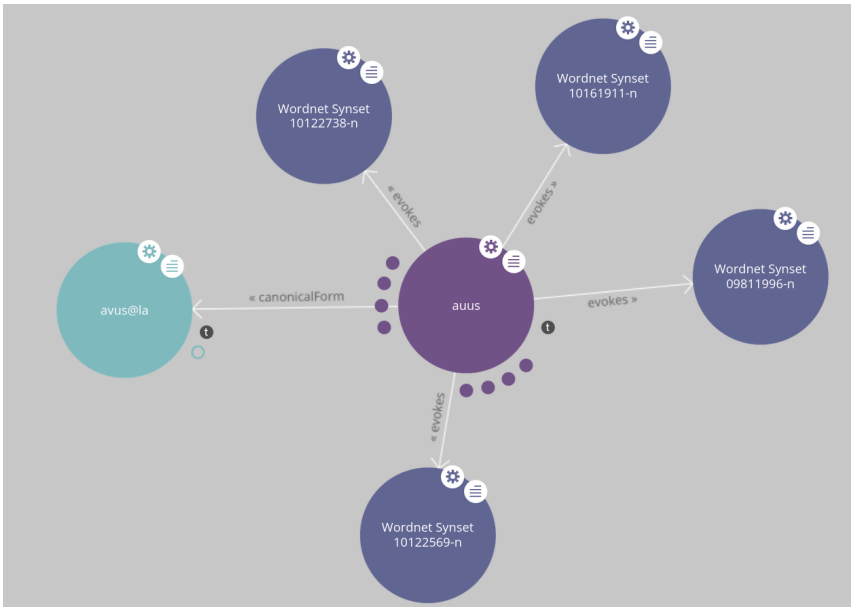

Figure 3: The LiLa lemma "avus" and the Latin WordNet

dataset allowed us to keep the effort proportionate to the class requirements, while at the same time enabling us to deliver a complete publishable result. In spite of its limited size, we believe that the results obtained go beyond the simple publication of a short word list, albeit of significant historical value.

The Concepticon project pursues the valuable goal of providing a single access point and a unified framework to concept lists. While the project's web interface and the underlying data are perfectly adequate to this aim, the integration into a LOD environment multiplies the usefulness of concept lists for linguistic studies. As shown in Figure 3, the same lemma "avus" (`lila_lemma:90862`) that is used as the canonical form of our example is also connected to an entry in the Latin WordNet in LiLa (Mambrini et al., 2021). The range of meanings of the Latin word that verbalizes Leibniz's concept included in the GRANDFATHER concept set is well captured by the image and the underlying data: the Latin word has four senses, which include, along with "the father of your father or mother" (`lwn:10161911-n`), also "someone from whom you are descended (but usually more remote than a grandparent)" (`lwn:09811996-n`), "the founder of a family" (`lwn:10122569-n`), and "person from an earlier time who contributed to the tradition shared by some group" (`lwn:10122738-n`). Researchers that, like Leibniz, are interested in collecting data to compare languages would find similar information about the polysemy of the words that verbalize the concepts invaluable. Interconnected knowledge bases like LiLa would provide the architecture to pursue this goal. A query to the LiLa's SPARQL endpoint would now allow to:[18] a) start from a Concepticon concept set like GRANDFATHER,[19] b) retrieve the Latin lexicalizations, c) access the wealth of information related to the Latin words, like the WordNet synsets associated with it, or all the corpus attestations of the word.

In this work we have modeled a small subset of a larger resource. The Concepticon is different from other popular computational resources such as WordNet or BabelNet in that it adopts an onomasiological perspective and puts the notion of the concept at the center, instead of focusing on representing language-specific senses (List et al., 2016, 2393-4). The work presented here is (to our knowledge) the first attempt to model such a resource as Linguistic Linked Data. We hope that we succeeded in providing a valuable reference to extend the work to model other concept lists.

Our experiment has shown that simple and widely used W3C standards like SKOS and Ontolex are perfectly capable to capture the structure and the mapping of an ambitious project like the Concepticon and to easily integrate its data into a KB of linguistic resources.

`2J3A_Linguistic_linked_open_data_Mambrini.docx`.

[18] `https://lila-erc.eu/sparql/`.

[19] `https://concepticon.clld.org/parameters/1383`.

References

Dean Allemang, James A. Hendler, and Fabien Gandon. *Semantic Web for the working ontologist: effective modeling in RDFS and OWL*, 3rd edition. Morgan Kaufmann/Elsevier, Waltham, MA.

Alfredo Ardila. 2007. Toward the development of a cross-linguistic naming test. *Archives of Clinical Neuropsychology*, 22(3):297–307. Special Issue: Cultural Diversity.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham.

Lucas Consolin Dezotti, Marco Passarotti, and Francesco Mambrini. 2024. Modelling and linking an old Latin-Portuguese dictionary to the LiLa knowledge base. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11537–11547, Torino, Italia. ELRA and ICCL.

Aron B. Dolgopolsky. 1964. Gipoteza drevnejšego rodstva jazykovych semej severnoj evrazii s verojatnostej točky zrenija [a probabilistic hypothesis concerning the oldest relationships among the language families of northern eurasia]. *Voprosy Jazykoznanija*, 2:53–63.

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. 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*, pages 26–34, Marseille, France. European Language Resources Association.

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.

Federica Gamba, Passarotti Marco, and Paolo Ruffolo. 2024. Publishing the dictionary of medieval latin in the czech lands as linked data in the lila knowledge base. *Italian Journal of Computational Linguistics*, 10:95–116.

Martin Haspelmath and Uri Tadmor, editors. 2009. *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Gottfried Wilhem von Leibniz. 1768. Desiderata circa linguas populorum, ad Dn. Podesta. In Louis Dutens, editor, *Godefridi Guilielmi Leibnitii opera omnia, nunc primum collecta, in classes distributa, praefationibus et indicibus exornata*, volume 6, pages 228–231. Fratres des Tournes, Geneva.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A Resource for the Linking of Concept Lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2393–2400, Portorož, Slovenia. European Language Resources Association (ELRA).

Magdalena Łuniewska, Ewa Haman, Sharon Armon-Lotem, Bartłomiej Etenkowski, Frenette Southwood, Darinka Anđelković, Elma Blom, Tessel Boerma, Shula Chiat, Pascale Engel de Abreu, et al. 2016. Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*, 48(3):1154–1177.

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.

Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Further with Knowledge Graphs. Studies on the Semantic Web 53*, Amsterdam. IOS Press.

Francesco Mambrini and Marco Carlo Passarotti. 2023. The lila lemma bank: A knowledge base of latin canonical forms. *Journal of Open Humanities Data*.

John P 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, Brno, Czech Republic. 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*, 58:177–212.

Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2024. The services of the LiLa knowledge base of interoperable linguistic resources for Latin. In *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, pages 75–83, Torino, Italia. ELRA and ICCL.

E. Natalie Rothman. 2021. *The Dragoman Renaissance: Diplomatic Interpreters and the Routes of Orientalism*. Cornell University Press, Ithaca, NY.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16(4):157–167.