Linking the *Neulateinische Wortliste* to the LiLa Knowledge Base of Interoperable Resources for Latin

Federica Iurescia, Eleonora Litta, Marco Passarotti,
Matteo Pellegrini, Giovanni Moretti, Paolo Ruffolo
Università Cattolica del Sacro Cuore, Milan
{federica.iurescia}{eleonoramaria.litta}{marco.passarotti}
{matteo.pellegrini}{giovanni.moretti}{paolo.ruffolo}@unicatt.it

Abstract

This paper describes the process of interlinking a lexical resource consisting of a list of more than 20,000 Neo-Latin words with other resources for Latin. The resources are made interoperable thanks to their linking to the LiLa Knowledge Base, which applies Linguistic Linked Open Data practices and data categories to describe and publish on the Web both textual and lexical resources for the Latin language.

1 Introduction

The Latin language shows a diachronic span covering more than two millennia, from the first literary texts in the 3rd century BC until today, when, for instance, Latin is the official language of the Vatican State. Moreover, having been for centuries the *lingua franca* of what is now referred to as the European area, Latin has been used in several different places by people with different cultural backgrounds, who produced texts of different typologies, thus resulting in a substantial degree of diatopic, diastratic and diaphasic variation.

Such a variation concerns every level of metalinguistic analysis, including morphology (Korkiakangas and Passarotti, 2011), syntax (Ponti and Passarotti, 2016), semantics (Perrone et al., 2021) and the lexicon.

As for the latter, despite the closed-corpus status of the Latin language (with a few exceptions of newly coined terms), there is not one fully comprehensive lexical resource that features the entire Latin lexicon. Yet, throughout the centuries, the lexicographic work on Latin has produced several dictionaries, lexica and glossaries covering specific eras (and/or areas) of the Latin language. For instance, the Latin-English dictionary by Lewis & Short (Lewis and Short, 1879) includes lexical entries about words from the Classical era, while the glossary by du Cange (du Cange et al., 1883–1887) and the Frankfurt Latin Lexicon (Mehler et al., 2020) concern Medieval Latin.

Over the last two decades, the research area dealing with linguistic resources for Latin has grown substantially, leading to the current availability of a large number of (annotated) corpora, including five treebanks available in the *Universal Dependencies* collection (de Marneffe et al., 2021) and several retro-digitised and newly built lexical resources. Such a situation raised the issue of the interoperability between the resources for Latin (like for many other languages), which are stored in separate silos and cannot interact. Starting in 2018, the *LiLa: Linking Latin* project[1] addressed this issue, by building a Linked Data Knowledge Base of interoperable resources for Latin. In the LiLa Knowledge Base, interoperability between resources is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. As a consequence, the core of the LiLa Knowledge Base consists of a large collection of Latin lemmas (called Lemma Bank), published as Linked Data and following the vocabulary and categories of the OntoLex-Lemon model (McCrae et al., 2017; Passarotti et al., 2020).

Given the central role played by the Lemma Bank in the architecture of LiLa, its lexical coverage is of the utmost importance.[2] In order to enhance the Lemma Bank with lemmas belonging to the so-called Neo-Latin or Modern Latin variety, we have recently started the process of linking the lexical entries of the *Neulateinische Wortliste* (NLW) by J. Ramminger, a dictionary of Latin from Petrarch up to the 18th century (Ramminger, 2016).[3] The dictionary currently includes about 21k entries, promising to allow for a relevant

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

[2] Before the work described in this paper, the LiLa Lemma Bank included about 200k lemmas for approximately 130k words. One word can have more than one lemma, like in the case of graphical variants: see Section 3.

[3] http://nlw.renaessancestudier.org.

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023), pages 82–87
May 5, 2023 ©2023 Association for Computational Linguistics

widening of the lexical coverage of the Lemma Bank. This paper describes the stages of this ongoing linking process, detailing the ones that we have already accomplished and outlining the future work.

2 Data

The NLW is a lexical resource that collects entries from the so-called Neo-Latin lexicon. These were retrieved mainly from literary sources and partly from secondary literature, such as scientific publications on Neo-Latin (Schoeck et al., 1990). The diachronic range covered by the resource spans between 1300 and 1700, on the basis of a decision taken by field experts, as explained by the author in the documentation available on the website.[4]

In the NLW, Neo-Latin is considered as the diachronic development of a specific diastratic variety of the language, namely Latin written production influenced by the linguistic ideals of Renaissance Humanists. These ideals may be subsumed under two general purposes: recovery of the language of Classical Antiquity, and enriching the lexicon with new entries that mirror contemporary changes in the society, e.g. *typographus* 'typographer'. However, the NLW does not feature the entire Neo-Latin lexicon according to these criteria, but it reflects its author's interests, as stated in the documentation.

The word list, consisting of 21,352 entries, was provided by the author in .docx format. The content of each entry is organised into a set of fields. The first one contains the citation form(s) of the lemma and all its graphical variants, followed by morphological information about its inflectional category, e.g. the endings of other forms of the word and a shortcut for the gender: for instance, "-a, -um" (the feminine and neuter of the nominative singular) for first class adjectives like *bizarrus* 'moody'; "-i, m." (the genitive singular and the gender) for second declension masculine nouns like *almirarchus* 'admiral'; "-ire, -ivi, -itum" (the present active infinitive, first-person singular of the perfect and supine) for fourth conjugation regular verbs like *semiambio* 'to half-circle'. The other fields feature a translation into German of the lemma and examples of its usage in textual sources, a set of administrative metadata (i.e. date of the creation), a numeric unique identifier for the entry, and philological and etymological information. Information about the presence of the lemma in a set of Classical and Medieval Latin dictionaries and lexicographic databases is provided as well.[5]

3 The *Neulateinische Wortliste* in LiLa

The LiLa Knowledge Base follows the principles of the Linguistic Linked Open Data paradigm. It adopts the RDF data model (Lassila and Swick, 1998), where information is coded in terms of triples that connect a subject to an object through a property. Each instance of an item ("individual") belongs to a specific class. The structure of the data is expressed by means of subclass relations and/or restrictions on the domain and range of properties – i.e., on the kinds of elements that they can have as subject and object, respectively. Classes and properties of existing ontologies are reused when possible, new ones are introduced if necessary.

As was hinted above, the core class of the LiLa Knowledge Base is `lila:Lemma`.[6] The lemmas of the Lemma Bank, to which the entries of lexical resources and the tokens of textual resources are linked, belong to this class. The lemma is simply defined as the citation form of a word, as it is recorded in dictionaries. Therefore, it is treated as a subclass of the class of forms in the OntoLex vocabulary (`ontolex:Form`).[7] This is the vocabulary that is used for the inclusion of lexical resources into the LiLa Knowledge Base: their entries belong to the class `ontolex:LexicalEntry`,[8] and they are connected to the corresponding `lila:Lemma` in the Lemma Bank by means of the property `ontolex:canonicalForm`;[9] entries of different resources that refer to the same word are linked to the same lemma in the Lemma Bank, thus achieving the desired interoperability.

As a consequence, the very first step of our procedure consisted in going through the entries of the

[4] `http://nlw.renaessancestudier.org/varia/einleit.htm`.

[5] *Thesaurus Linguae Latinae* (`https://tll.degruyter.com/about`), the *Ausführliches lateinisch-deutsches Handwörterbuch* (Georges, 1998), the *Lexicon totius latinitatis* by Forcellini (Forcellini, 1965), the *Dictionary of Medieval Latin from British Sources* (Latham et al., 2018), and the *Dictionary of Medieval Latin from Celtic Sources* (Devine et al., 1998).

[6] `http://lila-erc.eu/lodview/ontologies/lila/Lemma`. In this notation, a shorthand of the ontology where the class or property is defined precedes the colon, that is followed by the name of the class or property (in camel style with or without capitalisation of the first letter, respectively).

[7] `http://www.w3.org/ns/lemon/ontolex#Form`.

[8] `http://www.w3.org/ns/lemon/ontolex#LexicalEntry`.

[9] `http://www.w3.org/ns/lemon/ontolex#canonicalForm`.

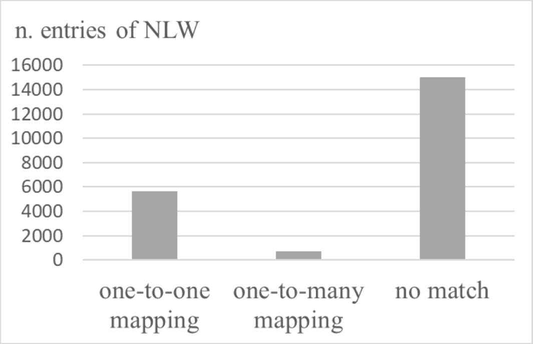

Figure 1: Mapping from NLW entries to LiLa lemmas

NLW and looking for the corresponding lemma(s) in the Lemma Bank. This was done by matching the string of the entry as it appears in the first field of the data that were delivered to us (see Section 2) with the different graphical variants of the lemmas in the Lemma Bank – coded as different `writtenRepresentations` using the OntoLex vocabulary.[10] This allowed us to unambiguously link 5,651 entries to their corresponding lemmas. In other cases (716 entries), however, more than one lemma matched the string of the NLW entry, so a disambiguation is needed to select which lemma is the correct one. Lastly, there are 14,985 entries of the NLW that are not found in the Lemma Bank: in these cases, we need to add a new lemma to be able to link those entries. Figure 1 shows this distribution visually.

In what follows, we describe our procedure i) to automatically generate new lemmas with all the relevant information (Section 3.1), and ii) to disambiguate between different homographic lemmas that match the string of a single NLW citation form (Section 3.2).

3.1 Automatic generation of new lemmas

In the Lemma Bank, several pieces of information are associated to each lemma by means of a set of dedicated properties. Among other things, each lemma is assigned a part of speech through the property `lila:hasPOS`;[11] information on the inflectional category – verbal conjugations, nominal declensions, adjectival classes – is provided through the property `lila:hasInflectionType`;[12] additionally, gender (masculine/feminine/neuter) is coded for nouns through the property `lila:hasGender`[13] and gradation (positive/comparative/superlative) for adjectives through the property `lila:hasDegree`.[14] When generating new lemmas for the entries of the NLW that have no match among the already existing lemmas of the Lemma Bank, to infer all these features we exploited the morphological information provided by the NLW entries.

Firstly, we isolated the information about the inflectional category (and, for nouns only, the gender) as a set of separate codes, e.g., the code "-i, m." identifying masculine 2nd declension nouns. This yielded a classification in almost a thousand (993) distinct codes. However, many of them (730) are attested in only one entry (*hapaxes*), and the overwhelming majority (920) are attested in less than 10 entries. At this stage, we focused on the 73 codes that are attested in more than 10 entries. Because of the frequency distribution of codes, this is sufficient to cover for most of the entries of the NLW (19,935 out of 21,352). Since the other codes often correspond to more marginal and not fully regular cases, they are best left for a successive stage of manual or semi-automatic insertion (when they are not already linked to existing lemmas of the Lemma Bank).

The 71 codes then underwent a process of normalisation, whereby some entries that are coded differently in the NLW data are attributed to the same class. In some cases, this is necessary because the coding of a single class is not uniform, due to inconsistencies in the way in which the original data have been compiled by hand. For instance, first class adjectives are coded sometimes as "-a, -um", sometimes as "-a -um", sometimes as "-a, .-um", sometimes with other minor variations, that are obviously not relevant to the morphological classification of the data. In other cases, the normalisation is motivated by the fact that different codes reflect a classification that is more fine-grained than the one of the Lemma Bank, so they can be conflated into a single class for our purposes. For instance, verbs of the first conjugation are coded in different ways in the NLW according to their strategy to form the perfect active indicative and the supine, e.g., by suffixation of *-avi* and *-atum* (see the verb *concentro* 'to concentrate', with code "-are, -avi, -atum") or by suffixation of *-ui* and *-tum* (see the

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

[11] http://lila-erc.eu/ontologies/lila/hasPOS.

[12] http://lila-erc.eu/ontologies/lila/hasInflectionType.

[13] http://lila-erc.eu/ontologies/lila/hasGender.

[14] http://lila-erc.eu/ontologies/lila/hasDegree.

NLW code	regex match	POS	Infl. Type	Gender
-ei, f.		NOUN	n5	f
-i, m.	^[a-z]+(us|(e|i)r)$	NOUN	n2	m
-i, m.	^[A-Z]+(us|(e|i)r)$	PROPN	n2	m
-i, m.	^[a-z]+os$	NOUN	n2e	m
-i, m.	^[A-Z]+os$	PROPN	n2e	m

Table 1: Mapping from the NLW morphological codes to the LiLa vocabulary

verb *triseco* 'to trisect', "-are, -ui, -ctum"), respectively. However, this difference is not reflected in the inflectional classification adopted in the Lemma Bank, and both these words would simply be assigned to the first conjugation class. Therefore, the two codes – together with all the other variants for the same conjugation – were normalised to a single one (namely, "-are") at this stage.

Such normalised codes were then used to generate the morphological information according to the tagset adopted in the Lemma Bank, as illustrated in Table 1. In some cases, a direct mapping is possible. For instance, if a word is assigned the code "-ei, f." in the NLW, then it can be reliably inferred that it is a feminine noun of the 5th declension (n5) – e.g., *faceties* 'witticism'. In other cases, however, the code by itself does not allow for a direct mapping, and it needs to be complemented with information on the character string of the citation form. For such cases, we specified different regular expressions that the string of the NLW citation form needs to match for the corresponding lemma to be assigned a given part of speech, inflection type and gender in the LiLa Knowledge Base. For instance, the code "-i, m." is used for masculine nouns of the second declension in the NLW. However, such nouns are classified differently in the LiLa Knowledge Base according to their shape: as for their part of speech, they are considered to be proper nouns if they start with a capital letter, common nouns otherwise; as for their inflection type, they are grouped with regular second declension nouns (n2) if they end with "us", "er", or "ir" (e.g., *vicenuntius* 'deputy envoy', *cultrifer* 'knife man', *proseptemuir* 'deputy member of the consortium of The Seven Men'), with irregular ones (n2e) if they end with "os" (e.g., in Greek loanwords like *misanthropos* 'misanthropist'). By applying such mappings to the cases of entries of the NLW with no match in the Lemma Bank, we enhanced it with 13,477 new lemmas.[15]

3.2 Automatic disambiguation between homographic lemmas

In order to disambiguate automatically at least some of the cases where more than one lemma in the Lemma Bank matched the string of the entry of the NLW, we used the same mappings discussed in Section 3.1 and exemplified in Table 1. For instance, the string of the citation form of the NLW entry *formularius* 'compositor' matches two different lemmas of the Lemma Bank, one of them being a noun[16] and the other one an adjective.[17] However, since the NLW entry in question is assigned the code "-i, m.", we know that the entry is a second declension noun. Therefore, we can safely link it to the lemma with the corresponding part of speech and morphological features in the Lemma Bank.

This procedure was applied to all the cases of one-to-many mapping between the NLW and the Lemma Bank, again excluding the 214 cases with more than one citation form, that are left for manual disambiguation because they cannot be categorised automatically. Out of the 501 remaining ambiguous cases, 359 were automatically disambiguated, and each of them is consequently linked to a single lemma in the Lemma Bank at the end of the process.

4 Conclusion and Future Work

In this paper, we have described the ongoing process of linking a dictionary of Neo-Latin to the LiLa Knowledge Base.

Based on the lexical entries of the dictionary, the collection of lemmas that represents the core component of LiLa was enhanced with more than 13,000 new items.[18] Such an extension of the LiLa Lemma Bank promises to improve its lexical coverage of the Neo-Latin texts that we plan to link to the Knowledge Base in the near future. In particular, the texts will be taken from the CAMENA corpus, that counts about 50 million tokens.[19]

Besides the citation form (the lemma) and the translation(s) in German of the words (modelled as individuals belonging to the class

[15] We excluded 976 entries of the NLW providing more than one citation form, as these cannot always be treated automatically (see also the discussion in Section 4).

[16] http://lila-erc.eu/data/id/lemma/103663.

[17] http://lila-erc.eu/data/id/lemma/103662.

[18] The Lemma Bank can be queried at https://lila-erc.eu/query/.

[19] http://mateo.uni-mannheim.de/camenahtdocs/camena_e.html.

`ontolex:LexicalSense`),[20] the lexical entries of the NLW feature also a number of sample attestations of their use in Neo-Latin texts. We modelled and published this information as Linked Data, using the Frequency, Attestation and Corpus (FrAC) module of OntoLex-Lemon (Chiarcos et al., 2022a).

Furthermore, we have seen in Section 3.1 that the NLW provides a morphological classification of lemmas that is sometimes more fine-grained than the one adopted in the Lemma Bank, and was thus not exploited in our procedure to automatically generate new lemmas. However, this is a potentially useful piece of information, that we plan to model in Linked Data, using the Morphology module (morph) of OntoLex-Lemon (Chiarcos et al., 2022b).

Lastly, we have seen in Sections 3.1 and 3.2 that those entries of the NLW that have more than one citation form were left out from our automatic procedure. This is motivated by the fact that the nature of the different citation forms and the relation between them can be diverse, and consequently require a different modelling. In some cases (e.g., *typographicus/typograficus* 'typographic'), they are simply graphical variants, that should be treated as written representations of the same lemma. In other cases, they would be considered as different lemmas, connected to each other through the property `lila:lemmaVariant`,[21] according to the current practice of the LiLa Knowledge Base – e.g., because they have different genders, as in *cibulus*(M)/*cibulum*(N) 'morsel'. Since an `ontolex:LexicalEntry` cannot have more than one `ontolex:canonicalForm` relation, such cases require the introduction of different (sub-)entries, whose organisation can be modelled using classes and properties of the Lexicography module (lexicog)[22] of OntoLex-Lemon.

After converting the NLW into a RDF serialisation (Turtle), we published the resource as Linked Data in the LiLa Knowledge Base, so to make it interoperable with the other lexical and textual resources for Latin already included therein.[23]

[20] `https://www.w3.org/ns/lemon/ontolex#LexicalSense`.

[21] `http://lila-erc.eu/ontologies/lila/lemmaVariant`.

[22] `https://www.w3.org/2019/09/lexicog/`.

[23] The URI (Uniform Resource Identifier) of the NLW is `http://lila-erc.eu/data/lexicalResources/NLW/Lexicon`. The Turtle file is available at `https://github.com/CIRCSE/NeulateinischeWortliste`.

5 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

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4018–4027.

Christian Chiarcos, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. 2022b. Computational Morphology with OntoLex-Morph. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 78–86.

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

Kieran Devine, Francis J Smith, and Anthony Harvey. 1998. *Database of Medieval Latin from Celtic Sources*.

Charles du Fresne sieur du Cange, bénédictins de la congrégation de Saint-Maur, d. Pierre Carpentier, Johann Christoph Adelung, G. A. Louis Henschel, Lorenz Diefenbach, and Léopold Favre. 1883–1887. *Glossarium mediae et infimae latinitatis*. Favre, Niort, France.

Egidio Forcellini. 1965. *Lexicon totius latinitatis*. Arnaldo Forni, Bologna, Italy.

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

Timo Korkiakangas and Marco Passarotti. 2011. Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics*, 26(2):103–114.

Ora Lassila and Ralph R. Swick. 1998. Resource Description Framework (RDF) Model and Syntax Specification.

Ronald E. Latham, David R. Howlett, and Richard K. Ashdowne, editors. 2018. *Dictionary of Medieval Latin from British Sources*. British Academy (through Oxford University Press), Oxford, UK.

Charlton Thomas Lewis and Charles Short. 1879. *A Latin Dictionary*. Clarendon Press, Oxford, UK.

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.

Alexander Mehler, Bernhard Jussen, Tim Geelhaar, Alexander Henlein, Giuseppe Abrami, Daniel Baumartz, Tolga Uslu, and Wahed Hemati. 2020. The Frankfurt Latin Lexicon. From Morphological Expansion and Word Embeddings to SemioGraphs. *Studi e Saggi Linguistici*, LVIII(1):121–155.

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.

Valerio Perrone, Simon Hengchen, Marco Palma, Alessandro Vatri, Jim Q Smith, and Barbara McGillivray. 2021. Lexical semantic change for Ancient Greek and Latin. *Computational approaches to semantic change*, pages 287–310.

Edoardo Maria Ponti and Marco Passarotti. 2016. Differentia compositionem facit. A slower-paced and reliable parser for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 683–688, Portorož, Slovenia. European Language Resources Association (ELRA).

Johann Ramminger. 2016. Ein Wörterbuch des Lateinischen von Petrarca bis 1700.

Richard J Schoeck, Martina Rütt, and H-W Bartz. 1990. A Step Towards a Neo-latin Lexicon: A First Wordlist Drawn from "Humanistica Lovaniensia". *Humanistica Lovaniensia*, 39:340–365.